Differences between v5.3 and v5.4 of the
No Bullshit Guide to Math and Physics

Ivan Savov

2020-11-08

No bullshit guide to math and physics
by Ivan Savov

Published by Minireference Co.
Montréal, Québec, Canada
`minireference.com` | `@minireference` | `fb.me/noBSguide`

For inquiries, contact the author at `ivan@minireference.com`

Mathematics Subject Classifications (2010): `00A09`, `70-01`, `97I40`, `97I50`.

Library and Archives Canada Cataloguing in Publication

Savov, Ivan, 1982-, author
No bullshit guide to math & physics / Ivan Savov. — Fifth edition.

ISBN 978-0-9920010-0-1 (pbk.)

1. Mathematics–Textbooks. 2. Calculus–Textbooks.
3. Mechanics–Textbooks. I. Title. II. Title: No bullshit guide to math and physics.

QA39.3.S28 2014 511′.07 C2014-905298-7

Fifth edition
v5~~.3~~.4 git commit 1179:91e8d55
Previous editions: v1.0 2010, v2.0 2011, v3.0 2012, v4.0 2013, v5.0 2014.

`ISBN 978-0-9920010-0-1`

Contents

mostly new

new section

Concept map

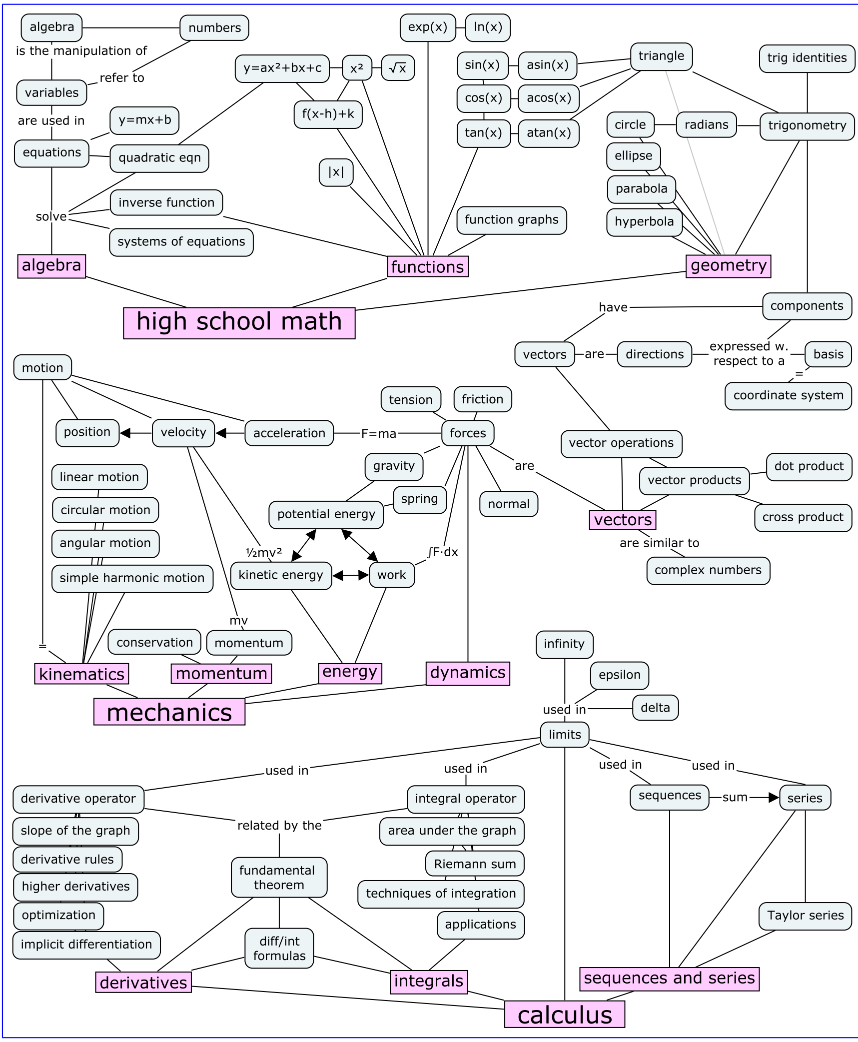


Figure 1: This diagram shows the connections between the concepts, topics, and subjects covered in the book. Seeing the connections between concepts is key to understanding math and physics. Consult the index on page 551 to find the exact location in the book where each concept is defined.

You can annotate the concept map with your current knowledge of each concept to keep track of your progress through the book.

- Add a single dot (•) next to all concepts you've heard of.
- Add two dots (••) next to concepts you think you know.
- Add three dots (•••) next to concepts you've used in exercises and problems.

By collecting some dots every week, you'll be able to move through the material in no time at all.

If you don't want to mark up your book, you can download a printable version of the concept map here: `bit.ly/mathphyscmap`.

This is a new idea to "gamify" the reading.

The idea is for readers to actively mark up concepts as unknown, known, and used and observe the progress: from a giant map with scary unknown terminology ...

... to a bunch of "completed levels" made up of concepts the reader knows well.

Looking for feedback about this idea from anyone who tries it out.
Did it feel motivational or just a chore?

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 $200 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 concepts 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 of students.

How?

The sections in this book are **self-contained tutorials**. 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

Non-students, don't worry: you don't 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 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~~ tutoring math and physics for more than ~~15 yearsas a private tutor. My tutoring experiencehas taught me how to explain concepts that people find difficult~~ 17 years. Through this experience, I learned to break complicated ideas into smaller, interconnected chunks that are easy 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~~think the best way to teach math and physics is to clearly define concepts and show the paths that connect them. 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.`~~ Minireference Co. to revolutionize the textbook industry. We make textbooks that don't suck.

Ivan Savov
Montreal, ~~2014~~ 2020

Introduction

The last ~~two centuries have~~ century has 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~~ use computers to automate repetitive tasks and increase your productivity. However, when it comes to building *new* things, understanding how technology works becomes important. One particularly useful skill is the ability to ~~create mathematical models of real-world situations~~**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 and engineering.

Calculus and mechanics can be difficult subjects. Understanding the material isn't hard *per se*, but it takes patience and practice to become comfortable with the new ideas. 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~~ The concept map in Figure 1 (page v) shows an overview of all the concepts and topics we'll discuss in the book. There are a lot of new things to learn, but don't worry—we'll navigate the material step by step and it will all make sense in the end.

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~~equations, functions, geometry, and trigonometry. 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~~ Adult learners can use this review chapter as therapy to recover from any ~~damaging educational~~ traumatizing math learning 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 understand the concepts of velocity and acceleration, and be able to predict the flight time of a ball thrown in the air.

In Chapter 3, we'll 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'll see physics is actually quite simple.

Chapter 5 covers topics from differential calculus and integral calculus. We'll study limits, derivatives, integrals, sequences, and series. You'll find that ~~120~~ 130 pages are enough to cover all the concepts in calculus, ~~as well as illustrate them with examples and practice exercises~~including practical applications.

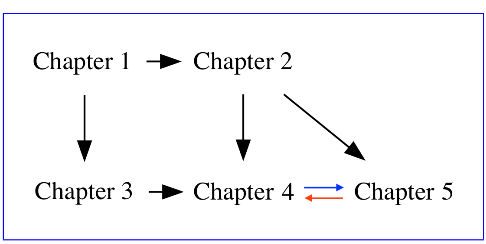


Figure 2: The prerequisite structure for the chapters in this book.

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~~ highlights the connections between them.

Are you ready for this? Let's dig in~~.~~!

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 group 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 for which items belong in that collection. Note also that each of the sets in the list contains all the sets above it, as illustrated in Figure 1.2. 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.

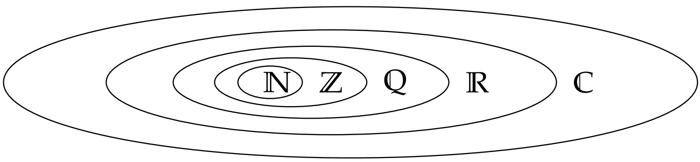


Figure 1.2: An illustration of the nested containment structure of the different number sets. The set of natural numbers is contained in the set of integers, which in turn is contained in the set of rational numbers. The set of rational numbers is contained in the set of real numbers, which is contained in the set of complex 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're 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~~ $\cancel{\frac{1}{4}}$ limiting—a rotisserie menu that offers $\frac{1}{2}$ of a chicken ~~, would be completely~~ would be totally confusing.

If you want to use division in your mathematical calculations, you'll need the rationals $\mathbb{Q}$. The ~~rationals are the set of~~ set of rational numbers corresponds to all numbers that can be expressed as *frac-*

tions of ~~integers:~~

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

the form $\frac{m}{n}$ where m and n are integers, and $n \neq 0$. 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~~ , therefore we say that $\sqrt{2}$ is *irrational* (not in the set $\mathbb{Q}$). An irrational number has an infinitely long decimal expansion that doesn't repeat. For example, $\pi = 3.141592653589793\ldots$ where the dots indicate that the decimal expansion of π continues all the way to infinity.

Combining the irrational numbers with 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~~ rational numbers $\mathbb{Q}$, 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 to 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 youare~~ 'll assume you're familiar with this ~~kind of stuff:~~ stuff:

$$2+3=5, \quad 45+56=101, \quad 65-66=-1, \quad 9\,999+1=10\,000.$$

You can visualize numbers as sticks of different length. Adding numbers is like adding sticks together: the resulting stick has a length equal to the sum of the lengths of the constituent sticks, as illustrated in Figure 1.3.

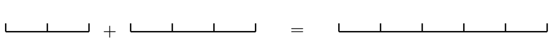


Figure 1.3: The addition of numbers corresponds to adding lengths.

Addition is *commutative*, which means that $a + b = b + a$. In other words, the order of the numbers in a summation doesn't matter. 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

Subtraction is the inverse operation of addition:

$$2 - 3 = -1, \quad 45 - 56 = -11, \quad 999 - 1 = 998.$$

Unlike addition, subtraction is not a commutative operation. The expression $a - b$ is not equal to the expression $b - a$, or written mathematically:

$$a - b \neq b - a.$$

Instead we have $b - a = -(a - b)$, which shows that changing the order of a and b in the expression changes its sign.

Subtraction is not associative either:

$$(a - b) - c \neq a - (b - c).$$

For example $(7 - 2) - 3 = 2$ while $7 - (2 - 3) = 8$.

Multiplication

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

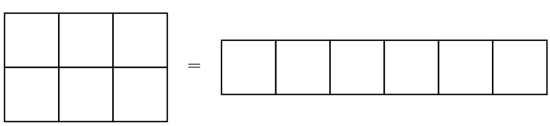


Figure 1.4: The area of a rectangle with width 3 m and height 2 m is equal to 6 m^2, which is equivalent to six squares with area 1 m^2 each.

Multiplication of numbers is also commutative, $ab = ba$~~;~~, and associative, $abc = (ab)c = a(bc)$. In modern math notation, no special symbol is ~~used~~ required 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.

Division is not a commutative operation since a/b is not equal to b/a. Division is not associative either: $(a \div b) \div c \neq a \div (b \div c)$. For example, when $a = 6$, $b = 3$, and $c = 2$, we get $(6/3)/2 = 1$ while $6/(3/2) = 4$.

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

The act of multiplying a number by itself many times is called *exponentiation*.

To visualize how exponents work, we can draw a connection between the value of exponents and the dimensions of geometric objects. Figure 1.5 illustrates how the same length 2 corresponds to different geometric objects when raised to different exponents. The number 2 corresponds to a line segment of length two, which is a geometric object in a one-dimensional space. If we add a line segment of length two in a second dimension, we obtain a square with area 2^2 in a two-dimensional space. Adding a third dimension, we obtain a cube with volume 2^3 in a three-dimensional space. Indeed, raising a base a to the exponent 2 is commonly called "a squared," and raising a to the power of 3 is called "a cubed."

The geometrical analogy about one-dimensional quantities as lengths, ~~two dimensional~~ two-dimensional quantities as areas, and ~~three dimensional~~ three-dimensional quantities as volumes is good to keep in mind.

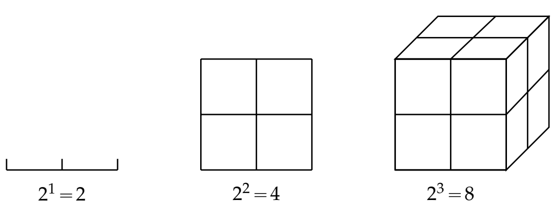


Figure 1.5: Geometric interpretation for exponents 1, 2, and 3. A length raised to exponent 2 corresponds to the area of a square. The same length raised to exponent 3 corresponds to the volume of a cube.

Our visual intuition works very well up to three dimensions, but and the associated mnemonic "Please Excuse My Dear Aunt Sally," might help you remember the order of operations.

For instance, the expression $5 \cdot 3^2 + 13$ is interpreted as "First find the square of 3, then multiply it by 5, and then add 13." Parentheses 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 $(5 \cdot 3)^2 + 13$, where parentheses indicate that the square acts on $(5 \cdot 3)$ as a whole and not on 3 alone.

Exercises

E1.1 Solve for the unknown x in the following equations:

a) $3x + 2 - 5 = 4 + 2$ **b)** $\frac{1}{2}x - 3 = \sqrt{3} + 12 - \sqrt{3}$

c) $\frac{7x-4}{2} + 1 = 8 - 2$ **d)** $5x - 2 + 3 = 3x - 5$

E1.2 Indicate all the number sets the following numbers belong to.

a) -2 **c)** $\cancel{8^{\frac{1}{3}}}$ $8 \div 4$ **d)** $\frac{5}{3}$

b) $\sqrt{-3}$ **e)** $\frac{\pi}{2}$

E1.3 Calculate the values of the following expressions:

a) $2^3 3 - 3$ **b)** $2^3(3 - 3)$ **c)** $\frac{4-2}{3^3}(6 \cdot 7 - 41)$

too complicated!

1.3 Number representations

We use the letters "a, b, c, ..." to write words. In a similar fashion, we use the digits $0, 1, 2, 3, 4, 5, 6, 7, 8$, and 9 to write numbers in the language of math. You can think of the digits 0 through 9 as the "letters" used to write numbers. For example, the number 334 consists of the digits 3, 3, and 4. Note that the same digit 3 denotes two different quantities depending on its position within the number. The first digit 3 corresponds to the value three hundred, while the second digit 3 corresponds to the value thirty.

Concepts

In this section, we'll review three important number representations:

- The *decimal notation* for integers, rationals, and real numbers consists of an integer part and a fractional part separated by a *decimal point*. For example, the decimal 32.17 consists of the integer 32 and the fractional part 0.17.
- The *fraction notation* for integers and rational numbers consists of a numerator divided by a denominator. Here are some sample math expressions with fractions: $\frac{1}{2}$, $\frac{3}{4}$, $\frac{3}{2} = 1\frac{1}{2}$, and $\frac{17}{100}$.

- The *number line* is a graphical representation for numbers that allows us to visualize numbers as geometric points on a line.

The same number a can be represented in multiple equivalent ways. It is often convenient to convert from one representation to another depending on the calculations we need to perform. For example, the number three can be expressed as the numeral 3, the decimal 3.0, the fraction $\frac{3}{1}$, or as the point that lies three units to the right of the origin on the number line. All these representations refer to the same quantity, but each representation is useful in different contexts.

The goal of this section is to get you comfortable working with all the number representations. ~~The decimal representation for numbers is very common in everyday life; you're likely already familiar with decimals, so we won't spend too much time on them. Instead we'll focus on reviewing fractions, as well as fraction operations like addition and multiplication. It's important to understand fractions because many math concepts like ratios, percents, and proportionality are best described in the language of fractions.~~

Positional notation for numbers

The Hindu–Arabic numeral system is the most widely used system for writing numbers today. It is a *decimal positional* system. The term *decimal* refers to the fact that it uses 10 unique symbols (the digits 0 through 9) to represent numbers. The system is *positional* because the value of each digit depends on its position within the number. Positional number systems are also called *place-value* systems.

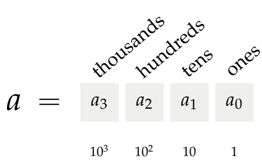


Figure 1.7: The place-value representation of the number $a = a_3a_2a_1a_0$.

Note the terminology used to refer to the individual digits of the numeral: we call a_3 the thousands, a_2 the hundreds, a_1 the tens, and a_0 the units.

Any natural number $a \in \mathbb{N}$, no matter how large, can be written as a sequence of digits:

$$\begin{aligned} a \equiv = \, & a_n \cdots a_2 a_1 a_0 \\ = \, & a_n \cdot 10^n + \cdots + a_2 \cdot 10^2 + a_1 \cdot 10 + a_0 \cdot 1, \end{aligned}$$

where the digits $a_0, a_1, \ldots$ come from the set $\{0,1,2,3,4,5,6,7,8,9\}$.

For example, the numeral 4235 corresponds to this calculation:

$$\begin{aligned} 4235 &= 4 \cdot 10^3 \quad + 2 \cdot 10^2 + 3 \cdot 10 + 5 \cdot 1 \\ &= 4 \cdot 1000 + 2 \cdot 100 + 3 \cdot 10 + 5 \cdot 1 \\ &= 4000 + 200 + 30 + 5. \end{aligned}$$

Note how the English pronunciation of the number, "four thousand, two hundred and thirty-five," literally walks you through the calculation.

~~When reading the digits of a number from left to right, each "step" we take to the right brings us to a digit that has a place value 10 times smaller than the previous digit. In the next section we'll learn how to extend this pattern one step further to the right in order to describe numbers smaller than one.~~

Decimal representation

~~We can use~~ Any number a less than one can be written as a *decimal* ~~*notation*~~*point* ~~to represent integers, rationals, and approximations to real numbers. The *decimal point*~~ followed by a sequence of digits, as illustrated in Figure 1.8.

$$\begin{aligned} a &= 0\,.\,a_{-1}a_{-2}a_{-3}\cdots \\ &= 0 + \frac{a_{-1}}{10^1} + \frac{a_{-2}}{10^2} + \frac{a_{-3}}{10^3} + \cdots\,. \end{aligned}$$

The decimal point indicates the beginning of the fractional part of a number. The place values of the digits to the right of the decimal point correspond to different decimal fractions. For example~~$0.7 = \frac{7}{10^1} = 7 \times 10^{-}$~~ ~~$0.07 = \frac{7}{100} = 7 \times 10^{-2}$, and $0.007 = \frac{7}{1000} = 7 \times 10^{-3}$. Note the positional logic used for decimals is the same as the positional logic used for integers: the place value of each digit decreases by a factor of 10 each time we take a "step" to~~ the ~~right.~~ digit 7 corresponds to three different decimal fractions depending on its position within the number:

$$0.7 = \frac{7}{10}\,, \qquad 0.07 = \frac{7}{100}\,, \qquad \text{and} \qquad 0.007 = \frac{7}{1000}\,.$$

~~Any number a less than one can be written using a decimal point~~

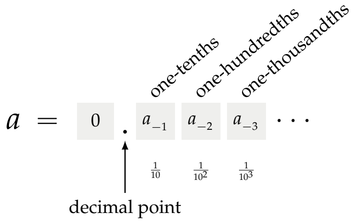


Figure 1.8: The decimal representation of a number smaller than one.

~~followed by a sequence of digits, as illustrated in Figure 1.8.~~

$$
\begin{aligned}
a &= 0\,.\,a_{-1}a_{-2}a_{-3}\cdots \\
&= 0 + \frac{a_{-1}}{10^1} + \frac{a_{-2}}{10^2} + \frac{a_{-3}}{10^3} + \cdots .
\end{aligned}
$$

The first digit to the right of the decimal point a_{-1} represents the *tenths*, the second digit a_{-2} represents the *hundredths*, the third the *thousandths*, and so on. ~~We can use decimal notation to describe fractions like one-half (0.5), one-quarter (0.25), and three-quarters (0.75).~~

In general, a number written in decimal notation has both an integer part and a fractional part:

$$
\begin{aligned}
a &= a_n \cdots a_2 a_1 a_0\,.\,a_{-1}a_{-2}a_{-3}\cdots \\
&= a_n \cdot 10^n + \cdots + a_2 \cdot 10^2 + a_1 \cdot 10 + a_0 + \frac{a_{-1}}{10^1} + \frac{a_{-2}}{10^2} + \frac{a_{-3}}{10^3} + \cdots
\end{aligned}
$$

The decimal point appears in the middle of the digits and acts as a separator. The digits to the left of the decimal point, $a_n \cdots a_2 a_1 a_0$, correspond to the integer part of the number, while the digits to the right of the decimal, $0.a_{-1}a_{-2}a_{-3}\cdots$, correspond to the fractional part of the number.

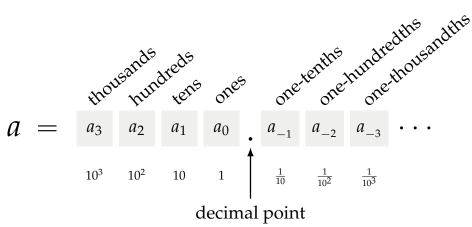


Figure 1.9: The decimal number a consists of an integer part $a_3a_2a_1a_0$ and a fractional part $0.a_{-1}a_{-2}\cdots$ separated by the decimal point.

Note the names for the different digits in the fractional part of the decimal in Figure 1.9. These names are used when we describe the fractional part of a decimal in words:

- "1.4" is read "one and four tenths," or you could informally describe the decimal as you see it written: "one point four."
- "45.37" is read "forty-five and thirty-seven hundredths," or sometimes "forty-five point three seven."
- A length measurement like "0.345 ~~in~~in" is read "three-hundred forty-five thousandths of an inch."

We can ~~write approximations for irrational numbersusing decimal notation~~use decimal notation to represent rational numbers like one-half (0.5), one-quarter (0.25), and three-quarters (0.75). We can also use decimal notation to write approximations to irrational numbers. For example, the irrational number $\sqrt{2}$ (the diagonal of a square with length one) is ~~approximatively~~ approximately equal to 1.41421. We say the approximation 1.41421 is "accurate to five decimals," because this is how many digits there are in its fractional part.

* * *

So far we ~~'ve discussed number representations that you are familiar with from your~~ reviewed the decimal representation for numbers, which is very familiar to us from everyday life. Perhaps you're starting to think that math isn't so bad after all? Some of you must be saying, "Wonderful, I'm becoming friends with numbers while avoiding uncomfortable topics like fractions." Sorry, but you're not getting off so easily because this is exactly what's coming up next. That's right, we're about to make friends with fractions, too.

Fractions

~~First let's review the definition of the set of rational numbers $\mathbb{Q}$. Every rational number 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. For example, the fraction $\frac{3}{8}$ describes having three parts out of a whole cut into eight parts, hence the name "three-eighths."

It's important to understand fractions because many math concepts like rational numbers ($\mathbb{Q}$), ratios, percents, and probabilities are best described in the language of fractions.

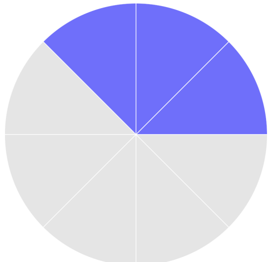


Figure 1.10: The fraction $\frac{3}{8}$ can be visualized as three slices from a pizza that has been cut into eight equal slices.

Definitions

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

$$a/b \cancel{=} = a \div b \cancel{=} = \frac{a}{b}\,.$$

The top and bottom parts of ~~a fraction~~ the fraction $\frac{a}{b}$ have special names:

- b is called the *denominator* of the fraction. It tells us how many parts make up the whole.
- a is called the *numerator*. It tells us the number of parts we have.

Fractions are the most natural way to represent rational numbers. Why natural? Check out these simple fractions:

$$\begin{aligned}
\tfrac{1}{1} &= 1.0\\
\tfrac{1}{2} &= 0.5\\
\tfrac{1}{3} &= 0.33333\ldots = 0.\overline{3}\\
\tfrac{1}{4} &= 0.25\\
\tfrac{1}{5} &= 0.2\\
\tfrac{1}{6} &= 0.166666\ldots = 0.1\overline{6}\\
\tfrac{1}{7} &= 0.14285714285714285\ldots = 0.\overline{142857}
\end{aligned}$$

~~Note that a line above some numbers means the digits underneath the line are repeated infinitely many times.~~ 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~~complicated 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 infinitely in the decimal, as in the case of $0.\overline{3}$, $0.1\overline{6}$, and $0.\overline{142857}$ shown above.

Fractions allow us to carry out 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 calculation as Wow that was complicated! This calculation is much simpler if we use fractions: Want to know how we did that? We multiplied the first term by $\frac{3}{3} = 1$ and the second term by $\frac{7}{7} = 1$ in order to obtain two equivalent fractions with the same denominator. This is one of the standard strategies when performing fraction ~~addition~~operations: rewriting them as equivalent fractions that have the same denominator. ~~Let's look at the procedure for adding fractionsin more detail~~

Equivalent fractions

The fractions $\frac{3}{8}$, $\frac{6}{16}$, and $\frac{12}{32}$ all correspond to the same number. Think about it—if you cut a pizza in 8 pieces and take 3 of them (see Figure 1.10), or you cut a pizza in 16 equal pieces and take 6 of them, you'll get the same amount of pizza in the end. All fractions of the form $\frac{3k}{8k}$ are *equivalent* to the fraction $\frac{3}{8}$, meaning they correspond to the same number.

Reciprocals

The mathematical term *reciprocal* is used to describe the notion of "flipping" a number. The reciprocal of y is $\frac{1}{y}$, which is read "one over y." Multiplication by the reciprocal $\frac{1}{y}$ is the same as division by y. The product of any number and its reciprocal equals one: $y \times \frac{1}{y} = \frac{y}{y} = 1$. The reciprocal of the fraction $\frac{m}{n}$ is the "flipped" fraction $\frac{n}{m}$. The product of $\frac{m}{n}$ and its reciprocal equals one: $\frac{m}{n} \times \frac{n}{m} = \frac{mn}{nm} = 1$.

Another way to denote the notion of "flipping" a number is to use the exponent negative one. The reciprocal of the number y is denoted y^{-1} and equals $\frac{1}{y}$. The reciprocal of the fraction $\frac{m}{n}$ is denoted $(\frac{m}{n})^{-1}$ and equals $\frac{n}{m}$. Using the negative exponent notation for reciprocals, we can write the "flip and multiply" rule for dividing fractions as

$\frac{a}{b} \div \frac{c}{d} = \frac{a}{b} \times \left(\frac{c}{d}\right)^{-1} = \frac{a}{b} \times \frac{d}{c} = \frac{ad}{bc}$. We'll discuss negative exponents more generally in Section 1.8.

Adding 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 can 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 of different wholes. Before we can add the numerators, we must rewrite the fractions so they have the same denominator, called a *common denominator*. We can obtain a common denominator by multiplying the first fraction by $\frac{d}{d} = 1$ and the second fraction by $\frac{b}{b} = 1$ in order to make the denominator of both fractions the same:

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

Now that we have fractions with the same denominator, we can add their numerators. Note it's okay to change the denominator of a fraction as long as we also change the numerator in the same way. Multiplying the ~~tops and bottoms of the fractions~~ top and the bottom of a fraction by the same number ~~(in this case $\frac{d}{d}$ or $\frac{b}{b}$)~~ is the same as multiplying by 1. So while the numbers of the fractions change, their equivalency is preserved:

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

~~**Finding the least common denominator** To add fractions they must share a common denominator. If you play around with the math, you'll quickly realize that two fractions can share many possible common denominators. Although any common denominator will do, generally you can make your life simpler by using the *least common denominator*—the smallest possible common denominator. The least common denominator of two fractions is the *least common multiple* of the two denominators LCM(b,d). The LCM of two numbers can be obtained using this formula:~~

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

~~where GCD(b,d) is the *greatest common divisor* of b and d—the largest number that divides both b and d.~~

Example To add $\frac{1}{6}$ and $\frac{1}{15}$, we ~~could~~ can use the product of the two denominators as the common denominator~~:~~, $6 \times 15 = 90$~~.~~ ~~Or, we could find the *least* common denominator by breaking each denominator into its smallest factors $6 = 3 \times 2$ and $15 = 3 \times 5$; then recognizing that 3 is the greatest common divisor of 6~~, and ~~15. We find the least common multiple is LCM$(6,15) = \frac{6\times15}{3} = 30$, then use common denominator 30 when performing the addition:~~ perform the fraction addition as follows:

$$\frac{1}{6} + \frac{1}{15} = \cancel{\frac{5\times1}{5\times6}}\,\frac{1}{6}\times\frac{15}{15} + \cancel{\frac{1\times2}{15\times2}}\,\frac{1}{15}\times\frac{6}{6} = \cancel{\frac{5}{30}}\,\frac{15}{90} + \cancel{\frac{2}{30}}\,\frac{6}{90} = \frac{21}{90} = \frac{7\times}{30}$$

Note ~~that using the least common denominator is not *required*—but it is the most efficient way to add fractions without having to deal with excessively large numbers. If you skip all this GCD and LCM business and use the larger common denominator $6 \times 15 = 90$, you'll arrive at the same answer after simplifying the result:~~

$$\cancel{\frac{1}{6} + \frac{1}{15} = \frac{15\times1}{15\times6} + \frac{1\times6}{15\times6} = \frac{15}{90} + \frac{6}{90} = \frac{21}{90} = \frac{7}{30}.}$$

I removed this section because it's not essential material, and readers found it confusing and stopped reading here...

Whole-and-fraction notation

A fraction greater than 1 like $\frac{5}{3}$ can also be denoted $1\frac{2}{3}$, which is read as "one and two-thirds." Similarly, $\frac{22}{7} = 3\frac{1}{7}$. We write the integer part of the number first, followed by the fractional part.

There is nothing wrong with writing fractions like $\frac{5}{3}$ and $\frac{22}{7}$. However, some teachers call these fractions *improper* and demand that all 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 infinitely in the decimal:~~

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

~~**Exercises**~~

~~Compute the value of the following expressions:~~

a) $\frac{1}{2} + \frac{1}{3}$
b) $\frac{1}{2} + \frac{1}{3} + \frac{1}{4}$
c) $3\frac{1}{2} + 2 - \frac{1}{3}$

a) $\frac{5}{6}$; **b)** $\frac{13}{12} = 1\frac{1}{12}$; **c)** $\frac{31}{6} = 5\frac{1}{6}$.

a) ~~To compute~~ $\frac{1}{2} + \frac{1}{3}$, ~~we rewrite both fractions using the common denominator 6, then compute the sum:~~ $\frac{1}{2} + \frac{1}{3} = \frac{3}{6} + \frac{2}{6} = \frac{5}{6}$. **b)** ~~You can use the answer from part (a), or compute the triple sum directly by setting all three fractions to a common denominator:~~ $\frac{1}{2} + \frac{1}{3} + \frac{1}{4} = \frac{6}{12} + \frac{4}{12} + \frac{3}{12} = \frac{13}{12}$. **c)** ~~Here we first rewrite~~ $3\frac{1}{2}$ ~~as~~ $\frac{7}{2}$, ~~then use the common denominator 6 for the computation:~~ $\frac{7}{2} + 2 - \frac{1}{3} = \frac{21}{6} + \frac{12}{6} - \frac{2}{6} = \frac{31}{6}$.

Number line

The *number line* is a very useful visual representation for numbers. Every number from the sets $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$ corresponds to some point on the number line. Developing a visual representation for numbers allows us to instantly compare the numbers' sizes based on their positions on the number line.

Figure 1.11 shows the natural numbers represented as equally spaced ~~notches~~ points on the number line. We can construct the entire set of natural numbers by starting from 0 and taking steps of length one to the right on the number line. That's what counting is—we just keep adding one.

Note that natural numbers never end. We can always keep adding one to every number and obtain a larger number. The number line therefore extends to the right to infinity.

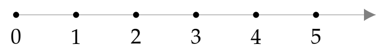


Figure 1.11: The natural numbers $\mathbb{N}$.

Numbers to the left of zero are negative, while numbers to the right of zero are positive. The number line extends indefinitely on both sides, going to negative infinity on the left ~~side~~ and positive infinity on the right~~side~~.

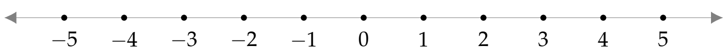


Figure 1.12: The integers $\mathbb{Z}$.

The set of integers corresponds to a ~~discrete set of~~ set of equally-spaced points on the number line ~~. Observe there are~~ with gaps of empty space between each integer. ~~The rational numbers $\mathbb{Q}$ and~~ We need the real numbers $\mathbb{R}$ to fill these gaps.

~~Recall that the set of rational numbers $\mathbb{Q}$ consists of all numbers that can be written as a fraction of two integers. Rational numbers allow us to refer to points between the integers. To find the location of the rational number $a\frac{m}{n}$~~ The set of real numbers $\mathbb{R}$ is the complete representation of all possible points on the number line~~, go to the integer a and then continue $\left(\frac{m}{n}\right)^{\text{th}}$ of the way to the next integer. For example, the rational number $\frac{-3}{2} = -1.5 = -1\frac{1}{2}$ corresponds to going to the point -1, then continuing halfway of the distance to the integer -2.~~

~~It is instructive to study the rational numbers in the interval between 0 and 1, which correspond to the fractions $\frac{m}{n}$ where $m \leqslant n$. There are infinitely many rational numbers between 0 and 1. The number lineis packed with them! For example, the infinite sequence of fractions $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \ldots$ consists of distinct rational numbers that all live in the interval between 0~~: every real number corresponds to some point on the number line, and ~~1. The rational numbers $\frac{1}{n}$ get closer and closer to 0 as n becomes larger and larger. The numbers of this sequence are densely packed next to each other, filling all the space near 0.~~

~~The rational numbers have the same structure everywhere~~ every point on the number line ~~. No matter which interval of the numberline you look at, you'll find it densely packed with rational numbers. In order to represent this density of numbers visually, we use a thick line to fill in~~ corresponds to some real number. Visually, the set of real numbers fills the entire number line ~~as illustrated~~ in **bold**, as shown in Figure 1.13. ~~Basically, the rationals are represented by points~~ In other words, there are real numbers everywhere!

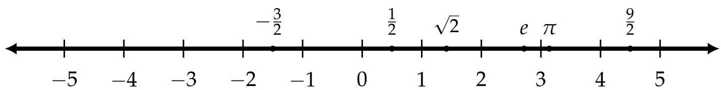


Figure 1.13: The ~~rationals $\mathbb{Q}$ and the reals~~ real numbers $\mathbb{R}$ ~~densely fill~~ cover the entire number line.

~~The~~ Recall that the set of real numbers ~~$\mathbb{R}$~~ includes all the ~~rationals $\mathbb{Q}$ and also~~ rational numbers like $-\frac{3}{2}$, $\frac{1}{2}$, and $\frac{9}{2}$, as well as irrational numbers like ~~$\sqrt{2} = 1.414213562\ldots$, $e = 2.7182818\ldots$, and $\pi = 3.14159265\ldots$. The visual representation for the reals is identical to the rationals: they also fill the entire~~ $\sqrt{2}$, e, and π. This means any number you are likely to run into when solving math problems can be visualized as a point on the number line. The number line can also be used to represent subsets of the real numbers. We'll talk about that in Section 1.23. For example, the subset of real numbers that are greater than two and smaller than four is shown in Figure 1.81 (page 159).

Discussion

Since we're still on the topic of number representations, I want to add some footnotes with "bonus material" related to the ideas we've covered in this section. Feel free to skip to the next section if you're in a hurry, because this is definitely not going to be on the exam!

~~Integers and divisibility~~

~~Recall the concepts of *greatest common divisor* (GCD) and *least common multiple* (LCM) we used to add fractions. The GCD and LCM are related to the notion of *divisibility* for natural numbers. For example, 3 divides 12 since $12/3 = 4$ and 4 is an integer, but 3 does not divide 7 since $7/3$ is not an integer. We say "b divides a" whenever a/b is an integer. In other words, if b divides a, then $a = kb$ for some integer k. If a/b is not an integer, we say "b does not divide a."~~

~~The *divisors* of the number x is the set of numbers that divide x. Every number can be written as a product of its divisors. For example, $12 = 3 \times 4$ since 3 and 4 are divisors of 12. The number 4 can be subdivided further as 2×2, so another expression for 12 in terms of its divisors is $12 = 3 \times 2 \times 2$. This procedure of splitting a number into smaller and smaller divisors terminates when we write the number in terms of its prime divisors. The set of *prime numbers* is the set of numbers that cannot be subdivided any further:~~

$$\{2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, \ldots\}$$

~~A number p is *prime* if it has no divisors other than 1 and itself. All other numbers are called *composite numbers*, meaning they can be written as a product of prime numbers.~~

~~The *greatest common divisor* of a and b, denoted $\text{GCD}(a, b)$, is the largest number that divides both a and b. For example, let's calculate $\text{GCD}(12, 20)$. Writing 12 as a product of its divisors, we find $12 = 3 \times 2 \times 2$. Similarly, $20 = 5 \times 2 \times 2$. By comparing the two expressions, we see that 2×2 is common to both expressions, so $\text{GCD}(12, 20) = 4$.~~

~~The *least common multiple* of two numbers, denoted $\text{LCM}(a, b)$, represents the smallest integer that has both a and b as divisors. We can calculate the $\text{LCM}(a, b)$ using the following formula:~~

$$\text{LCM}(a, b) = \frac{ab}{\text{GCD}(a, b)}.$$

~~Continuing the above example, we find $\text{LCM}(12, 20) = \frac{12 \times 20}{\text{GCD}(12,20)} = \frac{240}{4} = 60$. Observe that 12 divides 60 and also 20 divides 60.~~

The prime numbers are super cool, but not essential to anything else we'll cover in the book, so decided to cut.

I'm working on a separate handout on number theory and "math object kinds" where this text will be reused.

Elementary arithmetic procedures

The four basic arithmetic operations are addition, subtraction, multiplication, and division. We can perform these operations for numerals a and b of any size using only pen and paper. It is sufficient to follow one of the well-defined procedures (called algorithms) for manipulating the individual digits that make up the numbers. The Wikipedia articles on elementary arithmetic and long division offer an excellent discussion of these procedures.

[Algorithms for performing elementary arithmetic]
`https://en.wikipedia.org/wiki/Elementary_arithmetic`
`https://en.wikipedia.org/wiki/Long_division`

Computer representations

Whenever you want to store a number on a computer, you must choose an appropriate computer representation for this number. The two most commonly used types of numbers in the computer world are integers (`int`) and floating point numbers (`float`). Computer integers can accurately describe the set of mathematical integers $\mathbb{Z}$, but there are limitations on the maximum size of numbers that computers can store. We can use floating point numbers to store decimals with up to 15 digits of precision. The `int` and `float` numbers that computers provide are sufficient for most practical computations, and you probably shouldn't worry about the limited precision of computer number representations. Still, I want you to be aware of the distinction between the abstract mathematical concept of a number and its computer representation. The real number $\sqrt{2}$ is irrational and has an infinite number of digits in its decimal expansion. On a computer, $\sqrt{2}$ is represented as the approximation 1.41421356237310 (a `float`). For most purposes the approximation is okay, but sometimes the limited precision can show up in calculations. For example, `float(sqrt(2))*float(sqrt(2)) = 2.0000000000000004` $\neq 2$ and `float(0.1)+float(0.2) = 0.30000000000000004` $\neq 0.3$. The result of the computer's calculation is only accurate up to the 15$^{\text{th}}$ digit. That's pretty good if you ask me.

Scientific notation

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

Greek letters and complicated units
were confusing for nothing

Dealing with such numbers is much easier if we use *scientific notation*. ~~For example, the~~ The speed of light can be written as ~~$c = 2.99792458 \times 10^{8}$ m/s~~ 2.99792458×10^{8}, and the permeability of free space is denoted as ~~$\mu_0 = 1.256637 \times 10^{-6}$ N/A^2~~ 1.256637×10^{-6}. In both cases, we express the number as a decimal number between 1.0 and 9.9999 . . . followed by the number 10 raised to some ~~power~~exponent. 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 point 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 need to count the zeros!

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 won't go into a full discussion about the topic of *significant figures* here. Feel free to read the Wikipedia article on the subject to learn more.

Computer systems represent numbers using scientific notation, too. When entering a floating point number into the computer, separate the decimal part from the exponent by the character `e`, which stands for "exponent." For example, the speed of light is written as `2.99792458e8` and the permeability of free space is `1.256637e-6`.

~~Links~~Exercises

E1.4 Compute the value of the following expressions:

a) $\frac{1}{2} + \frac{1}{3}$ **b)** $\frac{1}{2} + \frac{1}{3} + \frac{1}{4}$ **c)** $3\frac{1}{2} + 2 - \frac{1}{3}$

Links

~~Numbers and number representations are fascinating topics connected to hundreds of other topics in math.~~ I encourage you to check the ~~Wikipedia~~ links provided below ~~for interesting historical context~~to learn more about numbers and number representations.

[History of the Hindu–Arabic system for representing numbers]
`https://en.wikipedia.org/wiki/Hindu-Arabic_numeral_system`

[Positional number representation systems]
`https://en.wikipedia.org/wiki/Positional_notation`

[Decimal representation]
https://en.wikipedia.org/wiki/Decimal_representation

[More ~~general number representation systems~~ details on scientific notation]
https://en.wikipedia.org/wiki/~~Numeral~~Scientific_~~system~~notation

[Info about significant figures calculations]
https://en.wikipedia.org/wiki/Significant_figures

1.4 Variables

In math we use a lot of *variables* and *constants*, which are placeholder names for *any* number or unknown. Variables allow us to perform calculations without knowing all the details.

Example You're having tacos for lunch today and wondering how many you can eat without going over your caloric budget. Your goal is to eat 800 calories for lunch and you want to do the calculation before getting to the restaurant because you fear your math abilities might be affected in the presence of tacos. You're not sure how many calories each taco contains, so you invent the variable c to denote this unknown. You also define the variable x to represent the number of tacos you will eat, and come up with the equation $800 = cx$ to represent the total number of calories of your lunch. Solving for x, you find the total number of tacos you should order is $x = \frac{800}{c}$. If the restaurant serves tacos that contain $c = 200$ calories each, then you should order $x = \frac{800}{200} = 4$ of them. If the restaurant serves only giant tacos worth $c = 400$ calories each, then you can only eat $x = \frac{800}{400} = 2$ of them. Observe we were able to solve for x even before knowing the value of c.

Variable names

There are common naming patterns for variables:

- x: ~~general name~~ name used for the unknown in equations~~(also used to denote a function 's input, as well as an object's position in physics)~~
- ~~v: velocity in physics problems~~
- ~~x_i, x_f: denote an object's initial and final positions in physics~~. We also use x to denote function inputs and the position of objects in physics.
- i, j, k, m, n: common names for integer variables

- a, b, c, d: letters near the beginning of the alphabet are often used to denote constants (fixed quantities that do not change).
- θ, ϕ: the Greek letters *theta* and *phi* are used to denote angles
- C: costs in business, along with P for profit, and R for revenue
- X: ~~a random variable~~ capital letters are used to denote random variables in probability theory

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 can 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 the equation $u^2 - 5u + 6 = 0$, which in ~~tern~~ turn can be rewritten as $(u-2)(u-3) = 0$ ~~.~~ using the techniques we'll learn in Section 1.7.

We now see that the solutions are $u_1 = 2$ 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 symbol m? It's much easier ~~in the long run~~ to use m throughout your calculation, and wait until the last step to substitute the value 140 kg when computing the final numerical answer.

1.5 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. We're looking for the unknown x such that $f(x)$ equals c. 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~~ *inverse function* (denoted f^{-1}) we "undo" the effects of f. We apply the inverse function f^{-1} to both sides of the equation to obtain

$$f^{-1}(f(x)) = \underline{x =} f^{-1}(c)\,.$$

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 ~~$x = f^{-1}(c)$~~ $x = f^{-1}(c)$.

The above example introduces the notation f^{-1} for denoting the ~~function's *inverse*~~ inverse function. This notation is ~~borrowed from the notion of inverse numbers:~~ inspired by the notation for reciprocals. Recall that multiplication by the reciprocal 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~~ inverse function. 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~~ an equation can have 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 ~~$f^{-1}(x) = \sqrt{x}$~~ $f^{-1}(y) = \sqrt{y}$, but 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:

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

many readers are confused by this one

The function-inverse relationship is *symmetric*—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.

FAQ ANS

Don't be surprised to see $-1x \Leftrightarrow -1x$ in the list of function inverses. Indeed, the opposite operation of multiplying by -1 is to multiply by -1 once more: $(-(-x) = x)$.

Example 1

If you want to solve the equation $x - 4 = 5$, you can apply the inverse function of $x - 4$, which is $x + 4$. After adding four to both sides of the equation, $x - 4 + 4 = 5 + 4$, we obtain the answer $x = 9$.

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 the 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~~ We'll show a formula for solving quadratic equations in Section 1.7. Solving cubic equations like $ax^3 + bx^2 + cx + d = 0$ ~~with pen and paper~~ using a formula is also possible, but at this point you might as well start using a computer to solve for the unknowns. See page 524 in Appendix D.

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 inverse functions is the key for solving all these types of equations, so be sure to practice using it.

Exercises

E1.5 Solve for x in the following equations:

a) $3x = 6$ **b)** $\log_5(x) = 2$ **c)** $\log_{10}(\sqrt{x}) = 1$

E1.6 Find the function inverse and use it to solve the problems.

1.6 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.

Let's define some terminology for referring to different parts of math expressions. 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.

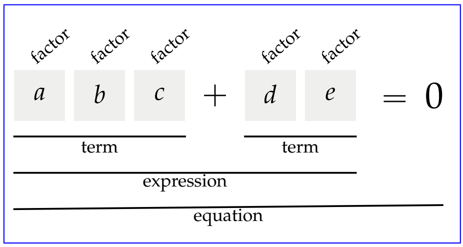


Figure 1.14: Diagram showing the names used to describe the different parts of the equation $abc + de = 0$.

Given any ~~four numbers $a, b, c,$ and d~~three numbers a, b, and c, 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; as a factor consisting 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~~ all algebra operations and illustrate what they're capable of.

~~Expanding brackets~~ moved to after Example

~~The distributive property is useful when dealing with polynomials. For instance,~~

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

We see t is contained in both terms on the right-hand side, so we can ~~rewrite~~ "factor it out" by rewriting the equation as

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

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

Expanding brackets

To *expand* a bracket is to multiply each term inside the bracket by the factor outside the bracket. The key thing to remember when expanding brackets is to apply the *distributive* property: $a(x+y) = ax + ay$. For longer expressions, we may need to apply the distributive property several times, until there are no more brackets left:

$$\begin{aligned}(a+b)(x+y+z) &= a(x+y+z) + b(x+y+z) \\ &= ax + ay + az + bx + by + bz.\end{aligned}$$

After expanding the brackets in this expression, we end up with six terms—one term for each of the six possible combinations of products between the terms in $(a+b)$ and the terms in $(x+y+z)$.

The distributive property is often used to manipulate expressions containing different powers of the variable x. For instance,

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

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

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

The bracket-expanding and simplification techniques demonstrated above are very common in math, and I recommend you solve some algebra practice problems to get the hang of them. Most math textbooks skip simplification steps and jump straight to the answer, since they assume readers are capable of doing simplifications on their own. It would be too long (and annoying) to show the simplifications for each expression. For example, the sentence "We can rewrite $(x+3)(x+2)$ as $x^2 + 5x + 6$," is the short version of the longer sentence, "We can apply the distributive property twice on $(x+3)(x+2)$ then combine the terms with the same power of x to get $x^2 + 5x + 6$."

It's not unusual for people to make math mistakes when expanding brackets and manipulating long algebra expressions. To avoid mistakes, use a step-by-step approach and apply only one operation in each step. Write legibly and keep the equations "organized" so it's easy to check the calculations performed in each step. 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 we sorted the terms in the final expression according to the different powers of x, with the terms containing x^2 grouped together, and the terms containing x grouped together. This approach helps keep things organized when dealing with expressions containing many terms.

Factoring

Factoring involves "taking out" the common parts of a complicated expression in order to make the expression more compact. Suppose we're given the expression $6x^2y + 15x$. We can simplify this expression by taking out the common factors and ~~writing~~ moving them in front of a bracket. Let's see how ~~this is done~~ to do this, step by step. ~~The expression~~

The expression $6x^2y + 15x$ has two terms~~and each term can be split~~. Let's split each term into its constituent factors:

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

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~~of factoring—notice the common factors are taken out and moved in front of the bracket:

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

~~Quadratic factoring~~Factoring quadratic expressions

A *quadratic expression* is an expression of the form $ax^2 + bx + c$. ~~When dealing with a quadratic function, it is often useful to rewrite the function as a product of two factors~~The expression contains a *quadratic term* ax^2, a *linear term* bx, and a constant term c. The numbers a, b, and c are called *coefficients*: the quadratic coefficient is a, the linear coefficient is b, and the constant coefficient is c. ~~Suppose you're given the quadratic function~~ $\cancel{f(x) = x^2 - 5x + 6}$

To *factor* the quadratic expression $ax^2 + bx + c$ is to rewrite it as the product of a constant and two factors like $(x + p)$ and $(x + q)$:

$$ax^2 + bx + c = a(x + p)(x + q).$$

Rewriting quadratic expressions in factored form helps us better understand and describe their properties.

Example Suppose we're 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 functionpositive, and for which~~ the properties of the function $f(x) = x^2 - 5x + 6$. Specifically, we're asked to find the function's *roots*, which are the values of x ~~is the function negative?~~ for which the function equals zero.

Factoring the expression $x^2 - 5x + 6$ ~~will help us see the properties of the function more clearly. To *factor* a~~ helps us see its properties more clearly, and makes our job of finding the roots of $f(x)$ easier. The factored form of this quadratic expression is ~~to express it as the product of two factors:~~

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

~~We now~~ Now we can see at a glance that the values of x for which $f(x) = 0$ are $x = 2$ and $x = 3$. When $x = 2$, the factor $(x - 2)$ is zero and hence $f(x) = 0$. Similarly, when $x = 3$, the ~~solutions (the 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~~factor $(x - 3)$ is zero so $f(x) = 0$.

~~For certain~~ How did we know that the factors of $x^2 - 5x + 6$ are $(x - 2)$ and $(x - 3)$ in the above example? For simple quadratics like the one above, ~~you~~ we can simply *guess* ~~what the factors will be~~the values of p and q in the equation $x^2 - 5x + 6 = (x+p)(x+q)$. Before we start guessing, let's look at the expanded version of the product between $(x+p)$ and $(x+q)$:

$$(x+p)(x+q) = x^2 + (p+q)x + pq.$$

Note the linear term on the right-hand side contains the sum of the unknowns $(p+q)$, while the constant term contains their product pq. If we want the equation $x^2 - 5x + 6 = x^2 + (p+q)x + pq$ to hold, we must find two numbers p and q whose sum equals -5 and whose product equals 6. After a couple of attempts we find $p = -2$ and $q = -3$. This guessing approach is an effective strategy for many of the factoring problems we will likely be asked to solve, since math teachers often choose simple numbers like ± 1, ± 2, ± 3, or ± 4 for the constants p and q. For more complicated quadratic expressions, ~~you~~we'll need to use the ~~, which will be the subject of the next section. For now let us continue with more algebra tricks~~quadratic formula, which we'll talk about in Section 1.7.

Common quadratic forms

make terminology explicit

Let's look at some common variations of quadratic expressions you might encounter when doing algebra calculations.

The quadratic expression $x^2 - p^2$ is called a *difference of squares*, and it can be obtained by multiplying the factors $(x+p)$ and $(x-p)$:

$$(x+p)(x-p) = x^2 \cancel{-xp} \cancel{+px} - p^2 = x^2 - p^2.$$

There's no linear term because the $-xp$ term cancels the $+px$ term. Any time you see an expression like $x^2 - p^2$, you can know it comes from a product of the form $(x+p)(x-p)$.

A *perfect square* is a quadratic expression that can be written as the product of repeated factors $(x+p)$:

$$x^2 + 2px + p^2 = (x+p)(x+p) = (x+p)^2.$$

Note $x^2 - 2qx + q^2 = (x-q)^2$ is also a perfect square.

Completing the square

NEW SECTION

~~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 due to the reasoning we follow to find the value of k. The constants~~

~~h~~ In this section we'll learn about an ancient algebra technique called *completing the square*, which allows us to rewrite *any* quadratic expression of the form $x^2 + Bx + C$ as a perfect square plus some constant correction factor $(x + p)^2 + k$. This algebra technique was described in one of the first books on *al-jabr* (algebra), written by Al-Khwarizmi around the year 800 CE. The name “completing the square” comes from the ingenious geometric construction used by this procedure. Yes, we can use geometry to solve algebra problems!

We assume the starting point for the procedure is a quadratic expression whose quadratic coefficient is one, $1x^2 + Bx + C$, and use capital letters B 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 geometric meaning of h~~ C to denote the linear and constant coefficients. The capital letters are to avoid any confusion with the quadratic expression $ax^2 + bx + c$, for which $a \neq 1$. Note we can always write $ax^2 + bx + c$ as $a(x^2 + \frac{b}{a}x + \frac{c}{a})$ and apply the procedure to the expression inside the brackets, identifying $\frac{b}{a}$ with B and ~~k in more detail in Section 1.13 (page 95). For now, let's focus on the algebra steps.~~ $\frac{c}{a}$ with C.

~~Let's try to find the values of k and h in the expression $(x - h)^2 + k$ needed to complete the square in the expression $x^2 + 5x + 6$. Assume the two expressions are equal, and then expand the bracket~~ First let's rewrite the quadratic expression $x^2 + Bx + C$ by splitting the linear term into two equal parts:

$$x^2 + \frac{B}{2}x + \frac{B}{2}x + C.$$

We can interpret the first three terms geometrically as follows: the x^2 term corresponds to a square with side length x, while the two $\frac{B}{2}x$ terms correspond to rectangles with sides $\frac{B}{2}$ and x. See the left side of Figure 1.15 for an illustration.

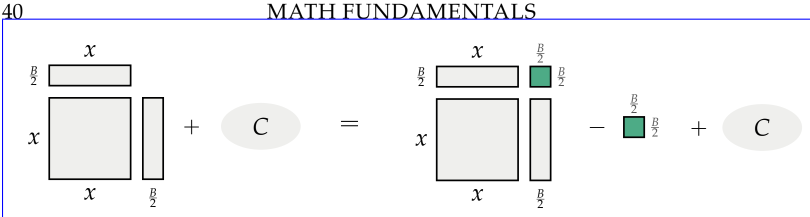


Figure 1.15: To complete the square in the expression $x^2 + Bx + C$, we need to add the quantity $(\frac{B}{2})^2$, which corresponds to a square (shown in darker colour) with sides equal to half the coefficient of the linear term. We also subtract $(\frac{B}{2})^2$ so the overall value of the expression remains unchanged.

The square with area x^2 and the two rectangles can be positioned to form a larger square with side length $(x + \frac{B}{2})$. Note there's a small piece of sides $\frac{B}{2}$ by $\frac{B}{2}$ missing from the corner. To *complete the square*, we can add a term $\left(\frac{B}{2}\right)^2$ to this expression. To preserve the equality, we also subtract $\left(\frac{B}{2}\right)^2$ from the expression to ~~obtain~~

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

~~Observe the structure in the above equation . On both sides of the equality there is one term which contains x^2 (~~obtain:

$$\begin{aligned} x^2 + \tfrac{B}{2}x + \tfrac{B}{2}x + C &= \underbrace{x^2 + \tfrac{B}{2}x + \tfrac{B}{2}x + \left(\tfrac{B}{2}\right)^2} - \left(\tfrac{B}{2}\right)^2 + C \\ &= \left(x + \tfrac{B}{2}\right)^2 - \left(\tfrac{B}{2}\right)^2 + C. \end{aligned}$$

The right-hand side of this equation describes the area of the square with side length $(x + \frac{B}{2})$, minus the area of the small square $\left(\frac{B}{2}\right)^2$, plus the constant C, as illustrated on the right side of Figure 1.15.

We can summarize the entire procedure in one equation:

$$x^2 + Bx + C = (x \underbrace{+ \tfrac{B}{2}}_{(1)})^2 + C \underbrace{- \left(\tfrac{B}{2}\right)^2}_{(2)}.$$

There are two things to remember when you want to apply the complete-the-square trick: (1) choose the constant inside the bracket to be $\frac{B}{2}$ (half of the linear coefficient), and (2) subtract $\left(\frac{B}{2}\right)^2$ outside the bracket in order to keep the equation balanced.

Solving quadratic equations

Suppose we want to solve the quadratic equation $x^2 + Bx + C = 0$. It's not possible to solve this equation with the digging-toward-the-x approach from Section 1.1 (since x appears in both the quadratic term ~~), one term that contains x^1~~ (x^2 and the linear term ~~), and constant terms. If the expressions are equal, then the coefficient of all the terms must be equal. By focusing on the quadratic terms in the equation (they are underlined) we see $A = 1$~~Bx). Enter the completing-the-square trick!

Example Let's find the solutions of the equation $x^2 + 5x + 6 = 0$. The coefficient of the linear term is $B = 5$, so we ~~rewrite the equation as~~ choose $\frac{B}{2} = \frac{5}{2}$ for the constant inside the bracket, and subtract $\left(\frac{B}{2}\right)^2 = \left(\frac{5}{2}\right)^2$ outside the bracket to keep the equation balanced. Completing the square gives

$$x^2 + \underline{5x}\,5x + 6 = \left(x\,\underline{-2hx} + h\,\frac{5}{2}\right)^2 + k\,6 - \left(\frac{5}{2}\right)^2 = 0.$$

Next we ~~look at the linear terms (underlined) and infer that $h = -2.5$. After rewriting, we obtain an equation in which k is the only unknown:~~

$$x^2 + 5x + 6 = x^2 - 2(-2.5)x + (-2.5)^2 + k.$$

~~Wemust pick a value of k that makes the constant terms equal:~~

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

~~After completing the square we obtain~~

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

~~The~~ use fraction arithmetic to simplify the constant terms in the expression: $6 - \left(\frac{5}{2}\right)^2 = 6 \cdot \frac{4}{4} - \frac{25}{4} = \frac{24-25}{4} = \frac{-1}{4} = -0.25$.

We're left with the equation

$$(x + 2.5)^2 - 0.25 = 0,$$

which we can now solve by digging toward x. First move 0.25 to 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.~~ to get $(x+2.5)^2 = 0.25$. Then take the square root on both sides to obtain $(x+2.5) = \pm 0.5$, which simplifies to $x = -2.5 \pm 0.5$. The two solutions are $x = -2.5 + 0.5 = -2$ and $x = -2.5 - 0.5 = -3$. You can verify these solutions by substituting the values in the original equation $(-2)^2 + 5(-2) + 6 = 0$ and similarly $(-3)^2 + 5(-3) + 6 = 0$. Congratulations, you just solved a quadratic equation using a 1200-year-old algebra technique!

~~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 $Ax^2 + Bx + C = A(x-h)^2 + k$, wehave $h = \frac{-B}{2A}$. Using this shortcut will save you some time, but you will still have to go through the algebra steps to find k~~In the next section, we'll learn how to leverage the complete-the-square trick to obtain a general-purpose formula for quickly solving quadratic equations.

~~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 $x^2 + 4x + 1 = (x+2)^2 - 3$.~~

Exercises

E1.7 Factor the following quadratic expressions:

a) $x^2 - 8x + 7$ **b)** $x^2 + 4x + 4$ **c)** $x^2 - 9$

Hint: Guess the values p and q in the expression $(x+p)(x+q)$.

E1.8 ~~Expand the following expressions:~~ Solve the equations by completing the square.

a) ~~$(a+b)^2$~~ $x^2 + 2x - 15 = 0$

b) ~~$(a+b)^3$~~

~~**c)** $(a+b)^4$~~

~~**d)** $(a+b)^5$~~

~~Can you spot a pattern in the coefficients of the different expressions? Do you think there is a general formula for $(a+b)^n$?~~ $x^2 + 4x + 1 = 0$

~~The coefficients of the expression of $(a+b)^n$ for different values of n correspond to the rows in *Pascal's triangle*. Check out the to learn the general formula and see an interesting animation of how it can be constructed.~~

1.7 Solving quadratic equations

What would you do if asked to solve for x in the quadratic equation $\cancel{x^2 = 45x + 23}$ $2x^2 = 4x + 6$? 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 is called the *standard form* of the quadratic equation. We obtain this form by moving all the numbers and xs to one side and leaving only 0 on the other side. ~~This is called the *standard form* of the quadratic equation.~~ For example, to transform the quadratic equation $\cancel{x^2 = 45x + 23}$ $2x^2 = 4x + 6$ into standard form, ~~subtract $45x + 23$~~ we subtract $4x + 6$ from both sides of the equation to obtain $\cancel{x^2 - 45x - 23 = 0}$ $2x^2 - 4x - 6 = 0$. What are the values of x that satisfy this equation?

Quadratic formula

~~**Claim**~~

The solutions to the equation $ax^2 + bx + c = 0$ ~~are~~ for $a \neq 0$ are

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

~~This result is called the *quadratic formula*, and~~ The quadratic formula is usually abbreviated $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$, where the sign "$\pm$" stands for both "$+$" and "$-$." The notation "$\pm$" allows us to express both solutions x_1 and x_2 in one equation, but you should keep in mind there are really two solutions.

Let's see how the quadratic formula is used to solve the equation $\cancel{x^2 - 45x - 23 = 0}$ $2x^2 - 4x - 6 = 0$. Finding the two solutions requires the simple mechanical task of identifying ~~$a = 1$, $b = -45$, and $c = -23$~~ $a = 2$, $b = -4$, and $c = -6$, then plugging these values into the two parts of the formula:

changed numbers so prob. can be solved by hand

$$x_1 = \cancel{\frac{45+\sqrt{45^2-4(1)(-23)}}{2}}\;\frac{4+\sqrt{4^2-4(2)(-6)}}{4} = \cancel{45.5054\ldots}\;\frac{4+\sqrt{16+48}}{4} = \ldots$$

$$x_2 = \cancel{\frac{45-\sqrt{45^2-4(1)(-23)}}{2}}\;\frac{4-\sqrt{4^2-4(2)(-6)}}{4} = \cancel{-0.5054\ldots.}\;\frac{4-\sqrt{16+48}}{4} = \ldots$$

~~Verify using your calculator that both of the values above~~ We can easily verify that value $x_1 = 3$ and $x_2 = -1$ both satisfy the original equation $\cancel{x^2 = 45x+23}$ $2x^2 = 4x + 6$.

Proof of ~~claim~~the quadratic formula

~~Understanding proofs is an important aspect of learning mathematics.~~ Every claim made by a mathematician comes with a ~~proof~~*proof*, which is a step-by-step argument that shows why the claim is true. It's ~~not necessary to know the proofs of *all* math statements, but the more proofs you know the more solid your understanding of math will become. It's easy to spot~~ easy to see where a proof starts and where a proof ends in mathematical texts. Each proof begins with the heading *Proof* (usually in italics) and has the symbol "~~□~~□" at its end. The purpose of these demarcations is to give readers the option to skip the proof. It's ~~usually okay to skip proofs when reading a math book, but if you really want to hang with the mathematicians, you have to read the proofs~~not necessary to read and understand the proofs of all math statements, but reading proofs can often lead you to a more solid understanding of the material.

I want you to see the proof of the quadratic formula because it's an important result that you'll use very often to solve math problems. Reading the proof will help you understand where the quadratic formula comes from. ~~This proof is an example of an argument fromfirst principles, which means it uses only basic rules of math and doesn't depend on any advanced math knowledge. You can totally handle this! The proof uses~~ The proof relies on the completing-the-square technique from the previous section~~.~~, and some basic algebra operations. You can totally handle this!

Proof. ~~Starting~~ We're starting from the quadratic equation $ax^2 + bx + c = 0$, ~~we~~ and we're making the additional assumption that $a \neq 0$. We want to find the value or values of x that satisfy this equation.

The first thing we want to do is divide by a to obtain the equivalent equation

$$x^2 + \frac{b}{a}x + \frac{c}{a} = 0.$$

~~Next we 'll *complete the square* by asking, "What are the values of h and k that satisfy the equation~~

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

~~To find the values for h and k, let's expand the bracket on the left-hand side to obtain~~

$$x^2 - 2hx + h^2 + k \;=\; x^2 + \frac{b}{a}x + \frac{c}{a}.$$

~~We can 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 $h = -\frac{b}{2a}$~~We are allowed to divide by a since we assumed that $a \neq 0$.

~~Let's now substitute the value $h = -\frac{b}{2a}$ into the above equation and see what we have so far:~~

$$x^2 + \frac{b}{a}x + \frac{b^2}{4a^2} + k \;=\; x^2 + \frac{b}{a}x + \frac{c}{a}.$$

~~To determine the value of k, we need to ensure the constant terms on both sides of the equation are equal, and then isolate k~~Next we apply the *complete the square* trick to the quadratic expression, to obtain an equivalent expression of the form $(x+?)^2+?$. Recall that the trick for completing the square is to choose the number inside the bracket to be half the coefficient of the linear term of the quadratic expression, which is $\frac{b}{2a}$ in this case. We must also subtract the square of this term outside the bracket in order to maintain the equality. After completing the square, we're left with the following equation:

$$\frac{b^2}{4a^2} + k \;=\; \frac{c}{a} \qquad \Rightarrow \qquad k = \frac{c}{a} - \frac{b^2}{4a^2}.$$

~~Having found the values of both h and k, wecan write the equation $ax^2 + bx + c = 0$ in the form $(x-h)^2 + k = 0$ as follows:~~

$$\left(x + \frac{b}{2a}\right)^{2} + \frac{c}{a} - \frac{b^2}{4a^2} = 0.$$

From here~~on, we can~~, we use the standard ~~procedure for "digging" toward the~~ digging-toward-the-x ~~, which we saw in Section 1.1~~procedure. Move all constants to the right-hand side~~:~~

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

~~Next,~~ ,

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

and take the square root of both sides to undo the square function~~.~~ ~~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~~

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

Since any number and its opposite have the same square, taking the square root gives us two possible solutions, which we denote using the "$\pm$" symbol.

Next we subtract $\frac{b}{2a}$ from both sides of the equation to isolate x and obtain $x=-\frac{b}{2a}\pm\sqrt{\frac{b^2}{4a^2}-\frac{c}{a}}$. We 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,~~ , $\sqrt{\frac{b^2}{4a^2}-\frac{c}{a}} = \sqrt{\frac{b^2}{4a^2}-\frac{4a\cdot c}{4a\cdot a}} = \sqrt{\frac{b^2-4ac}{4a^2}}$ and add the fractions on the right-hand side to obtain $x=\frac{-b\pm\sqrt{b^2-4ac}}{2a}$. The solutions to the quadratic equation $ax^2+bx+c=0$ are

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

This completes the proof of the quadratic formula. □

The expression b^2-4ac is called the *discriminant* of the equation. The discriminant tells us important information about the solutions of the equation $ax^2 + bx + c = 0$. The solutions x_1 and x_2 correspond to real numbers if the discriminant is positive or zero: $b^2 - 4ac \geqslant 0$. When the discriminant is zero ($b^2 - 4ac = 0$), the equation has only one solution since $x_1 = x_2 = \frac{-b}{2a}$. If the discriminant is negative, $b^2 - 4ac < 0$, the quadratic formula requires computing the square root of a negative number, which is not allowed for real numbers.

Alternative proof~~of claim~~

To ~~have a proof~~prove the quadratic formula, we don't necessarily need to show the ~~derivation of~~ algebra steps we followed to obtain the formula as outlined above. The ~~claim~~ quadratic formula states that x_1 and x_2 are solutions. To prove the ~~claim~~ formula is correct we can simply plug x_1 and x_2 into the ~~quadratic equation and verify that the answers are zero~~equation $ax^2 + bx + c = 0$ to verify that x_1 and x_2 are solutions. 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 447 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 $f(x) = ax^2 + bx + c = (x - x_1)(x - x_2)$.~~ a constant and two factors:

$$f(x) = ax^2 + bx + c = a(x - x_1)(x - x_2).$$

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~~: use the quadratic formula to find the two solutions, x_1 and x_2, then rewrite the expression as ~~$(x - x_1)(x - x_2)$~~$a(x - x_1)(x - x_2)$.

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 real 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?~~ (for now). We'll come back to the idea of taking square roots of negative numbers in Section 3.5 (see page 232).

Links

[~~Intuitive visual derivation~~ Algebra explanation of the quadratic formula]
https://www.youtube.com/watch?v=r3SEkdtpobo

[Visual explanation of the quadratic formula derivation]
https://www.youtube.com/watch?v=EBbtoFMJvFc

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:

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

Any number times its reciprocal gives one: $bb^{-1} = \frac{b}{b} = 1$. 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 1

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

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

~~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~~

Proof. 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~~very recently, so we don't know each other too well yet. In fact, the only thing we know about logs 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

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

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

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

We have thus shown that $\log(x) + \log(y) = \log(xy)$. $\square$

Using this property, we can derive two other useful formulas:

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

and

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

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 is helpful when you need to compute a logarithm in a base that is not available on your calculator. Suppose you're 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 $\log_{10}(7)$.

Exercises

E1.16 Use the properties of logarithms to simplify the expressions

a) $\log(x) + \log(2y)$ **b)** $\log(z) - \log(z^2)$ **c)** $\log(x) + \log(y/x)$

d) $\log_2(8)$ **e)** $\log_3(\frac{1}{27})$ **f)** $\log_{10}(10000)$

1.10 The Cartesian plane

The Cartesian plane, named after famous philosopher and mathematician René Descartes, is used to visualize pairs of numbers (x, y).

Recall the number line representation for numbers that we introduced in Section 1.3.

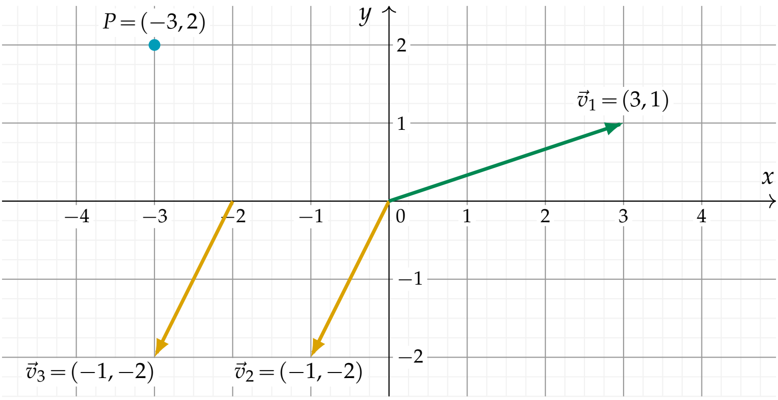


Figure 1.18: A Cartesian plane which shows the point $P = (-3, 2)$ and the vectors $\vec{v}_1 = (3, 1)$ and $\vec{v}_2 = \vec{v}_3 = (-1, -2)$.

mark where the vector starts and where it ends. Note that vectors $\vec{v}_2$ and $\vec{v}_3$ illustrated in Figure 1.18 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.

Graphs of functions

The Cartesian plane is great for visualizing functions. You can think of a function as a set of input-output pairs $(x, f(x))$. You can draw the *graph* of 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 1.19.

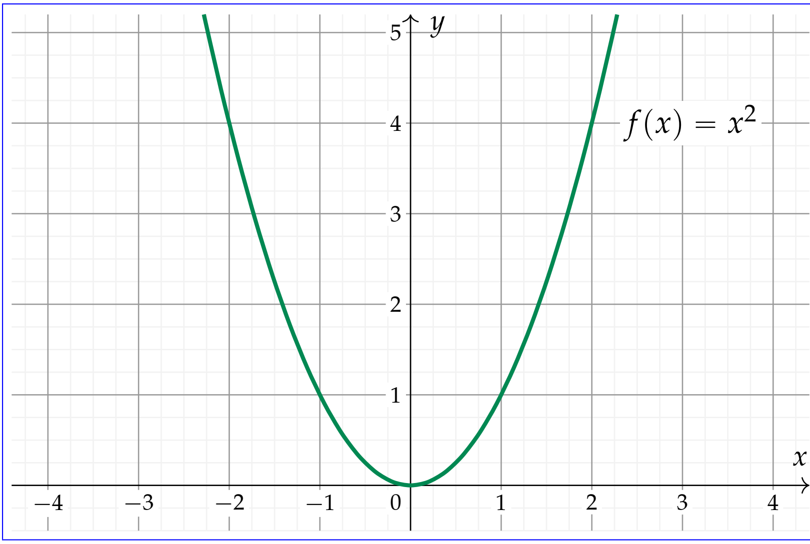


Figure 1.19: The graph of the function $f(x) = x^2$ consists of all pairs of points (x, y) in the Cartesian plane that satisfy $y = x^2$.

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.

~~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 $f(x)$.

Dimensions

The number line is ~~one dimensional~~one-dimensional. Every number x can be visualized as a point on the number line. The Cartesian plane has two dimensions: the x dimension and the y dimension. If we need to visualize math concepts in 3D, we can use a three-dimensional coordinate system with x, y, and z axes (see Figure 3.10 on page 223).

1.11 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 as you learn 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 image, and its inverse function~~**properties**. 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 produces 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: $f : \mathbb{R} \to \mathbb{R}$.

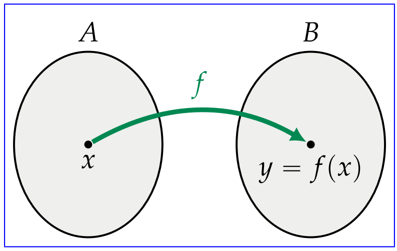


Figure 1.20: An abstract representation of a function f from the set A to the set B. The function f is the arrow which *maps* each input x in A to an output $f(x)$ in B. The output of the function $f(x)$ is also denoted y.

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

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

- ~~The~~ A: the *source set* of the function describes the types of numbers that the function takes as inputs.
- $\text{Dom}(f)$: the *domain* of a function is the set of allowed input values for the function.
- B: the *target set* of a function describes the type of outputs the function has. The target set is sometimes called the *codomain*.
- ~~The~~ $\text{Im}(f)$: the *image* ~~*or range*~~ of the function ~~f~~ is the set of all possible output values of the function.
- ~~The~~ The image is sometimes called the *range*.

See Figure 1.21 for an illustration of these concepts. The purpose of introducing all this math terminology is so we'll have words to distinguish the general types of inputs and outputs of the function (real numbers, complex numbers, vectors) from the specific properties of the function like its domain and image.

Let's look at an example to illustrate the difference between the source set and the domain of a function~~describes the type of outputs the function has.~~ . Consider the square root function $f : \mathbb{R} \to \mathbb{R}$ defined as $f(x) = \sqrt{x}$, which is shown in Figure 1.22. The source set of f is the set of real numbers—yet only nonnegative real numbers are allowed as inputs, since $\sqrt{x}$ is not defined for negative

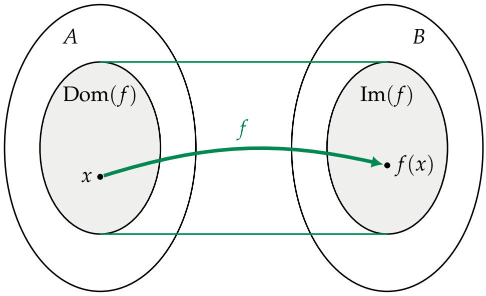


Figure 1.21: Illustration of the input and output sets of a function $f\colon A \to B$. The *source set* is denoted A and the *domain* is denoted $\mathrm{Dom}(f)$. Note that the function's domain is a subset of its source set. The *target set* is denoted B and the *image* is denoted $\mathrm{Im}(f)$. The image is a subset of the target set.

numbers. Therefore, the domain of the square root function is only the nonnegative real numbers: $\mathrm{Dom}(f) = \mathbb{R}_+ = \{x \in \mathbb{R} \mid x \geqslant 0\}$. Knowing the domain of a function is essential to using the function correctly. In this case, whenever you use the square root function, you need to make sure that the inputs to the function are nonnegative numbers.

The complicated-looking expression between the curly brackets uses *set notation* to define the set of nonnegative numbers $\mathbb{R}_+$. In words, the expression $\mathbb{R}_+ = \{x \in \mathbb{R} \mid x \geqslant 0\}$ states that "$\mathbb{R}_+$ is defined as the set of all real numbers x such that x is greater than or equal to zero." We'll discuss set notation in more detail in Section 1.23. For now, you can just remember that $\mathbb{R}_+$ represents the set of nonnegative real numbers.

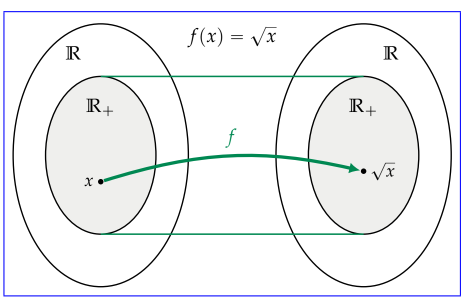


Figure 1.22: The input and output sets of the function $f(x) = \sqrt{x}$. The domain of f is the set of nonnegative real numbers $\mathbb{R}_+$ and its image is $\mathbb{R}_+$.

To illustrate the ~~subtle~~ difference between the image of a function and its ~~codomain, consider~~ target set, let's look at the function $f(x) = x^2$ shown in Figure 1.23. The quadratic function is of the form

$f\colon \mathbb{R} \to \mathbb{R}$. The function's ~~domain~~ source set is $\mathbb{R}$ (it takes real numbers as inputs) and its ~~codomain~~ target set is $\mathbb{R}$ (the outputs are real numbers too); however, not all ~~outputs are possible~~ real numbers are possible outputs. The *image* of the function $f(x) = x^2$ consists only of the nonnegative real numbers ~~$[0, \infty) \equiv \{y \in \mathbb{R} \mid y \geqslant 0\}$.~~

~~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 $f(x)$~~ $\mathbb{R}_+ = \{y \in \mathbb{R} \mid y \geqslant 0\}$, since $f(x) \geqslant 0$ for all x.

~~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.~~

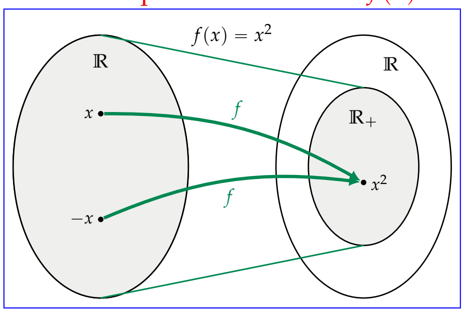


Figure 1.23: The function $f(x) = x^2$ is defined for all reals: $\mathrm{Dom}(f) = \mathbb{R}$. The image of the function is the set of nonnegative real numbers: $\mathrm{Im}(f) = \mathbb{R}_+$.

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

Function properties

We'll now introduce some additional terminology for describing three important function properties. Every function is a mapping from a source set to a target set, but what kind of mapping is it?

- ~~A function is *one-to-one* or~~ A function is *injective* ~~if it maps different inputs to different outputs~~ if it maps two different inputs to two different outputs. If x_1 and x_2 are two input values that are not equal $x_1 \neq x_2$, then the output values of an injective function will also not be equal $f(x_1) \neq f(x_2)$.
- A function is *surjective* if its image is equal to its target set. For every output y in the target set of a surjective function, there is at least one input x in its domain such that $f(x) = y$.
- A function is *bijective* if it is both injective and surjective.

I know this seems like a lot of terminology to get acquainted with, but it's important to have names for these function properties. We'll need this terminology to give a precise definition of the *inverse function* in the next section.

Injective property We can think of *injective* functions as pipes that transport fluids between containers. Since fluids cannot be compressed, the "output container" must be at least as large as the "input container." If there are two distinct points x_1 and x_2 in the input container of an injective function, then there will be two distinct points $f(x_1)$ and $f(x_2)$ in the output container of the function as well. In other words, injective functions don't smoosh things together.

In contrast, a function that doesn't have the injective property can map several different inputs to the same output value. The function $f(x) = x^2$ is not injective since it sends inputs x and $-x$ to the same output value $f(x) = f(-x) = x^2$, as illustrated in Figure 1.23.

The maps-distinct-inputs-to-distinct-outputs property of injective functions has an important consequence: given the output of an injective function y, there is only one input x such that $f(x) = y$. If a second input x' existed that also leads to the same output $f(x) = f(x') = y$, then the function f wouldn't be injective. For each of the outputs y of an injective function f, there is a *unique* input x such that $f(x) = y$. In other words, injective functions have a unique-input-for-each-output property.

Surjective property A function is ~~*ontoor if it covers the entire output set(in other words, if the image*~~ *surjective* if its outputs cover the entire target set: every number in the target set is a possible output of the function for some input. For example, the function $f\colon \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^3$ is surjective: for every number y in the target set $\mathbb{R}$, there is an input x, namely $x = \sqrt[3]{y}$, such that $f(x) = y$. The function $f(x) = x^3$ is surjective since its image is equal to ~~the function's codomain).~~ its target set, $\text{Im}(f) = \mathbb{R}$, as shown in Figure 1.24.

On the other hand, the function $f\colon \mathbb{R} \to \mathbb{R}$ defined by the equation $f(x) = x^2$ is not surjective since its image is only the nonnegative numbers $\mathbb{R}_+$ and not the whole set of real numbers (see Figure 1.23). The outputs of this function do not include the negative numbers of the target set, because there is no real number x that can be used as an input to obtain a negative output value.

Bijective property A function is bijective if it is both injective and surjective. ~~In this case, f is~~ When a function $f : A \to B$ has both

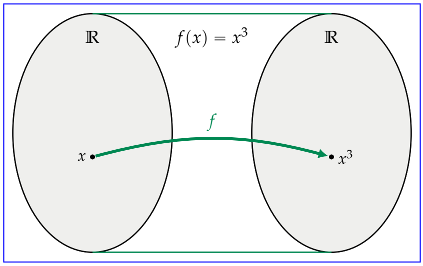


Figure 1.24: For the function $f(x) = x^3$ the image is equal to the target set of the function, $\text{Im}(f) = \mathbb{R}$, therefore the function f is surjective. The function f maps two different inputs $x_1 \neq x_2$ to two different outputs $f(x_1) \neq f(x_2)$, so f is injective. Since f is both injective and surjective, it is a *bijective* function.

the injective and surjective properties, it defines 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 $x \in X$,~~ numbers of the source set A and the numbers of the target set B. This means for every input value x, there is exactly one corresponding output value y, and for every output value y, there is exactly one input value x such that $f(x) = y$~~(injective part).~~ . An example of a bijective function is the function $f\colon \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^3$ (see Figure 1.24). For every input x in the source set $\mathbb{R}$, the corresponding output y is given by $y = f(x) = x^3$. For every output value y in the target set $\mathbb{R}$, the corresponding input value x is given by $x = \sqrt[3]{y}$.

~~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 . 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~~ A function is not bijective if it lacks one of the required properties. Examples of non-bijective functions are $f(x) = \sqrt{x}$, which is not surjective and $f(x) = x^2$, which is neither injective nor surjective.

Counting solutions Another way to understand the injective, surjective, and bijective properties of functions is to think about the solutions to the equation $f(x) = b$, where b is a number in the target set B. The function f is injective if the equation $f(x) = b$ has *at most* *one* solution for every number b. The function f is surjective if the equation $f(x) = b$ has *at least one* solution for every number b. If the function f is bijective then it is both injective and surjective, which means the equation $f(x) = b$ has *exactly one* solution.

Inverse function

We used inverse functions repeatedly in previous chapters, each time describing the inverse function informally as an "undo" operation. Now that we have learned about bijective functions, we can give a the precise definition of the inverse function and explain some of the details we glossed over previously.

Recall that a *bijective* function $f : A \to B$ is a *one-to-one correspondence* between the numbers in the source set A and numbers in the target set B: for every output y, there is exactly one corresponding input value x such that $f(x) = y$. The *inverse function*, denoted f^{-1}, is the function that takes any output value y in the set B and finds the corresponding input value x that produced it $f^{-1}(y) = x$.

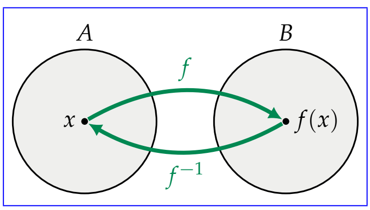


Figure 1.25: The inverse f^{-1} undoes the operation of the function f.

For every bijective function $f : A \to B$, there exists an inverse function $f^{-1} : B \to A$ that performs the *inverse mapping* of f. If we start from some x, apply f, and then apply f^{-1}, we'll arrive—full circle—back to the original input x:

$$f^{-1}(\,f(x)\,) = x.$$

In Figure 1.25 the function f is represented as a forward arrow, and the inverse function f^{-1} is represented as a backward arrow that puts the value $f(x)$ back to the ~~same output~~ x it came from.

Similarly, we can start from any y in the set B and apply f^{-1} followed by f to get back to the original y we started from:

$$f(\,f^{-1}(y)\,) = y.$$

In words, this equation tells us that f is the "undo" operation for the function f^{-1}, the same way f^{-1} is the "undo" operation for f. ~~For example $f(x) = x^2$ is not injective since the~~

If a function is missing the injective property or the surjective property then it isn't bijective and it doesn't have an inverse. Without the injective property, there could be two inputs ~~2 and -2 are both mapped to the output value 4~~ x and x' that both produce the same output $f(x) = f(x') = y$. In this case, computing $f^{-1}(y)$ would be impossible since we don't know which of the two possible inputs x or x' was used to produce the output y. Without the surjective property, there could be some output y' in B for which the inverse function f^{-1} is not defined, so the equation $f(f^{-1}(y)) = y$ would not hold for all y in B. The inverse function f^{-1} exists only when the function f is bijective.

Wait a minute! We know the function $f(x) = x^2$ is not bijective and therefore doesn't have an inverse, but we've repeatedly used the square root function as an inverse function for $f(x) = x^2$. What's going on here? Are we using a double standard like a politician that espouses one set of rules publicly, but follows a different set of rules in their private dealings? Is mathematics corrupt?

Don't worry, mathematics is not corrupt—it's all legit. We can use inverses for non-bijective functions by imposing *restrictions* on the source and target sets. The function $f(x) = x^2$ is not bijective when defined as a function $f : \mathbb{R} \to \mathbb{R}$, but it *is* bijective if we define it as a function from the set of nonnegative numbers to the set of nonnegative numbers, $f : \mathbb{R}_+ \to \mathbb{R}_+$. Restricting the source set to $\mathbb{R}_+ = \{x \in \mathbb{R} \mid x \geqslant 0\}$ makes the function injective, and restricting the target set to $\mathbb{R}_+$ also makes the function surjective. The function $f : \mathbb{R}_+ \to \mathbb{R}_+$ defined by the equation $f(x) = x^2$ is bijective and its inverse is $f^{-1}(y) = \sqrt{y}$.

It's important to keep track the of restrictions on the source set we applied when solving equations. For example, solving the equation $x^2 = c$ by restricting the solution space to nonnegative numbers will give us only the positive solution $x = \sqrt{c}$. We have to manually add the negative solution $x = -\sqrt{c}$ in order to obtain the complete solutions: $x = \sqrt{c}$ or $x = -\sqrt{c}$, which is usually written $x = \pm\sqrt{c}$. The possibility of multiple solutions is present whenever we solve equations involving non-injective functions.

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:

$$f\circ g\,(x) \cancel{\equiv} = f(\,g(x)\,) = z.$$

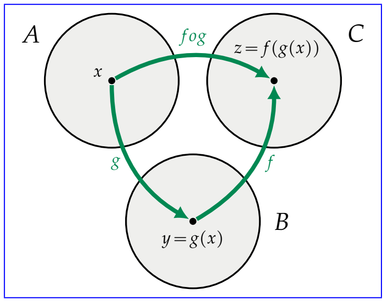


Figure 1.26: The function composition $f \circ g$ describes the combination of first applying the function g, followed by the function f: ~~$f \circ g\,(x) \equiv f(g(x))$~~ $f \circ g\,(x) = f(g(x))$.

Figure 1.26 illustrates ~~this concept~~ the concept of function composition. 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 ~~$f \circ g\,(x) \equiv f(g(x))$~~ $f \circ g\,(x) = f(g(x))$.

Don't worry too much about the "$\circ$" symbol—it's just a convenient math notation I wanted you to know about. Writing $f \circ g$ is ~~just as good~~ the same as writing $f(g(x))$. The important takeaway from Figure 1.26 is that functions can be combined by using the outputs of one function as the inputs to the next. This is a very useful idea for building math models. You can understand many complicated input-output transformations by describing them as compositions of simple functions.

~~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. For every input value x, there is exactly one corresponding output value y. This means that we can start from any output value y and find the corresponding input value x that produced it.~~

~~The inverse f^{-1} undoes the operation of the function f.~~

Example 1 Consider the function $g\colon \mathbb{R}_+ \to \mathbb{R}_+$ given by $g(x) = \sqrt{x}$, and the function $f\colon \mathbb{R} \to \mathbb{R}_+$ defined by $f(x) = x^2$. The composite function $f \circ g\,(x) = (\sqrt{x})^2 = x$ is defined for all nonnegative reals. The composite function $g \circ f$ is defined for all real numbers, and we have $g \circ f\,(x) = \sqrt{x^2} = |x|$.

~~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,~~

Example 2 The composite functions $f \circ g$ and ~~then apply f^{-1}, you'll arrive—full circle—back to the original input x~~ $g \circ f$ describe different operations. If $g(x) = \ln(x)$ and $f(x) = x^2$, the functions $g \circ f\,(x) = \ln(x^2)$ and $f \circ g\,(x) = (\ln x)^2$ have different domains and produce different outputs, as you can verify using a calculator.

Using the notation "$\circ$" for function composition, we can give a concise description of the properties of a bijective function $f : A \to B$ and its inverse function $f^{-1} : B \to A$:

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

~~In Figure 1.25 the function f is represented as a forward arrow, and the inverse function f^{-1} is represented as a backward arrow that puts the value $f(x)$ back to the~~

$$(f^{-1} \circ f)(x) = x \qquad \text{and} \qquad (f \circ f^{-1})(y) = y,$$

for all x ~~it came from~~in A and all y in B.

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 $\sqrt[n]{\ }$~~ squiggle notation $\sqrt{\ }$ for square roots and superscripts 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 $\cos(60^\circ)$, 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: ~~$\cos(60^\circ) \equiv \frac{1}{2}$~~$\cos(60^\circ) = \frac{1}{2}$. Therefore, seeing $\cos(60^\circ)$ 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. All scientific calculators have 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: $\{\{\text{in} = x_1, \text{out} = f(x_1)\}, \{\text{in} = x_2, \text{out} = f(x_2)\}, \{\text{in} = x_3, \text{out} = f(x_3)\}, \ldots\}$. 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~~ , where the first number of each pair represents an input value and the second represents the output value given by the function.

We can also build a **table of values** by ~~picking some random inputs~~ writing the input values in one column and recording the ~~output of the function in the second column:~~

input $= x$	$\to$	$f(x) =$ output
0	$\to$	$f(0)$
1	$\to$	$f(1)$
55	$\to$	$f(55)$
x_4	$\to$	$f(x_4)$.

~~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~~ corresponding output values in a second column. You can choose inputs at random or focus on the important-looking ~~x value~~ x values in the function's domain.

You can create a table of values for any function you want to study. Follow the example shown in Table 1.1. Use the input values that interest you and fill out the right side of the table by calculating the value of $f(x)$ for each input x.

input $= x$	$\rightarrow$	$f(x) =$ output
0	$\rightarrow$	$f(0)$
1	$\rightarrow$	$f(1)$
55	$\rightarrow$	$f(55)$
x_4	$\rightarrow$	$f(x_4)$

Table 1.1: Table of input-output values of the function $f(x)$. The input values $x = 0$, $x = 1$ and $x = 55$ are chosen to "test" what the function does.

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 1.27.

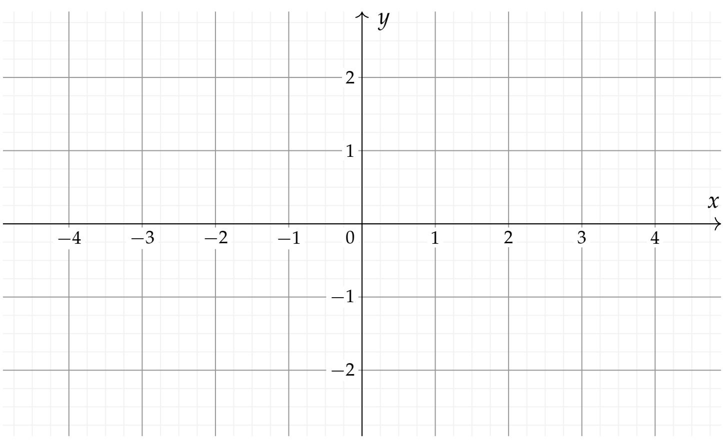


Figure 1.27: An empty (x, y)-coordinate system that you can use to ~~plot the graph of *any*~~ draw function ~~$f(x)$~~graphs. The graph of $f(x)$ consists of all the points for which $(x, y) = (x, f(x))$. See Figure 1.19 on page 60 for the graph of $f(x) = x^2$.

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:

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

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~~ connections to other functions. An example of a mathematical ~~fact is $\sin(30°) = \frac{1}{2}$. An example of a mathematical relation~~ connection is the equation ~~$\sin^2 x + \cos^2 x = 1$, which indicates~~ $\log_B(x) = \frac{\log_b(x)}{\log_b(B)}$, which describes a link between the ~~sin function and the cos function~~ logarithmic function base B and the logarithmic function base b.

The more you know about a function, the more "paths" your brain builds to connect to that function. Real math knowledge is not about memorization; it ~~requires establishing a graph~~ is about establishing a network 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*~~See the concept maps on page v for an illustration of the paths that link math concepts. Mathematical thought is the usage of ~~this graph to produce calculations and mathematical arguments called proofs~~these associations to carry out calculations and produce mathematical arguments. 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 $(\sin(x))^2$~~knowing about the connection between logarithmic functions will allow you compute the value of $\log_7(e^3)$, even though calculators don't have a button for logarithms base 7. We find $\log_7(e^3) = \frac{\ln e^3}{\ln 7} = \frac{3}{\ln 7}$, which can be computed using the $\boxed{\text{ln}}$ button.

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 some of the many paths linking math concepts~~;~~, but it's ~~up to~~ on you to reinforce these paths ~~by using what you've learned to practicesolving problems~~through practice.

~~**Example**~~

Example 3 Consider the function f from the real numbers to the real numbers ($f\colon \mathbb{R} \to \mathbb{R}$) defined ~~by the quadratic expression,~~

$$\cancel{f(x) = x^2 + 2x + 3.}$$

as $f(x) = x^2 + 2x - 3$. The value of f when $x = 1$ is $\cancel{f(1) = 1^2 + 2(1) + 3 = 1 + 2}$
When $x = 2$, the output is $\cancel{f(2) = 2^2 + 2(2) + 3 = 4 + 4 + 3 = 11}$ $f(2) = 2^2 + 2(2)$
What is the value of f when $x = 0$? You can use algebra to rewrite this function as $f(x) = (x+3)(x-1)$, which tells you the graph of this function crosses the x-axis at $x = -3$ and at $x = 1$. The values above will help you plot the graph of $f(x)$.

~~**Example 2**~~

Example 4 Consider the exponential function with base 2 ~~:~~

$$\cancel{f(x) = 2^x.}$$

defined by $f(x) = 2^x$. 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$\cancel{:\ (-1, \tfrac{1}{2^1} = \tfrac{1}{2}), (-2, \tfrac{1}{2^2} = \tfrac{1}{4}), (-3, \tfrac{1}{2^3} = \tfrac{1}{8})}$, so we have the points $(-1, \frac{1}{2})$, $(-2, \frac{1}{4})$, $(-3, \frac{1}{8})$, etc. You can plot these $(x, f(x))$ coordinates in the Cartesian plane to obtain the graph of the function.

Discussion

To describe a function we specify its source and target sets $f\colon A \to B$, then give an equation of the form $f(x) =$ "expression involving x" that defines the function. Since functions are defined using equations, does this mean that functions and equations are the same thing? Let's take a closer look.

In general, any equation containing two variables describes a *relation* between these variables. For example, the equation $x - 3 = y - 4$ describes a relation between the variables x and y. We can isolate the variable y in this equation to obtain $y = x + 1$ and thus find the value of y when the value of x is given. We can also isolate x to obtain $x = y - 1$ and use this equation to find x when the value of y is given. In the context of an equation, the relationship between the variables x and y is symmetrical and no special significance is attached to either of the two variables.

We also can describe the same relationship between x and y as a function $f : \mathbb{R} \to \mathbb{R}$. We choose to identify x as the input variable and y as the output variable of the function f. Having identified y with the output variable, we can interpret the equation $y = x + 1$ as the definition of the function $f(x) = x + 1$.

Note that the equation $x - 3 = y - 4$ and the function $f(x) = x + 1$ describe the same relationship between the variables x and y. For example, if we set the value $x = 5$ we can find the value of y by solving the equation $5 - 3 = y - 4$ to obtain $y = 6$, or by computing the output of the function $f(x)$ for the input $x = 5$, which gives us the same answer $f(5) = 6$. In both cases we arrive at the same answer, but modelling the relationship between x and y as a function allows us to use the whole functions toolbox, like function composition and function inverses.

* * *

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.

~~To build mathematical intuition, it is essential you understand functions' graphs. Trying to memorize the definitions and the properties of functions is a difficult task. Remembering what the function "looks like" is comparatively easier.~~

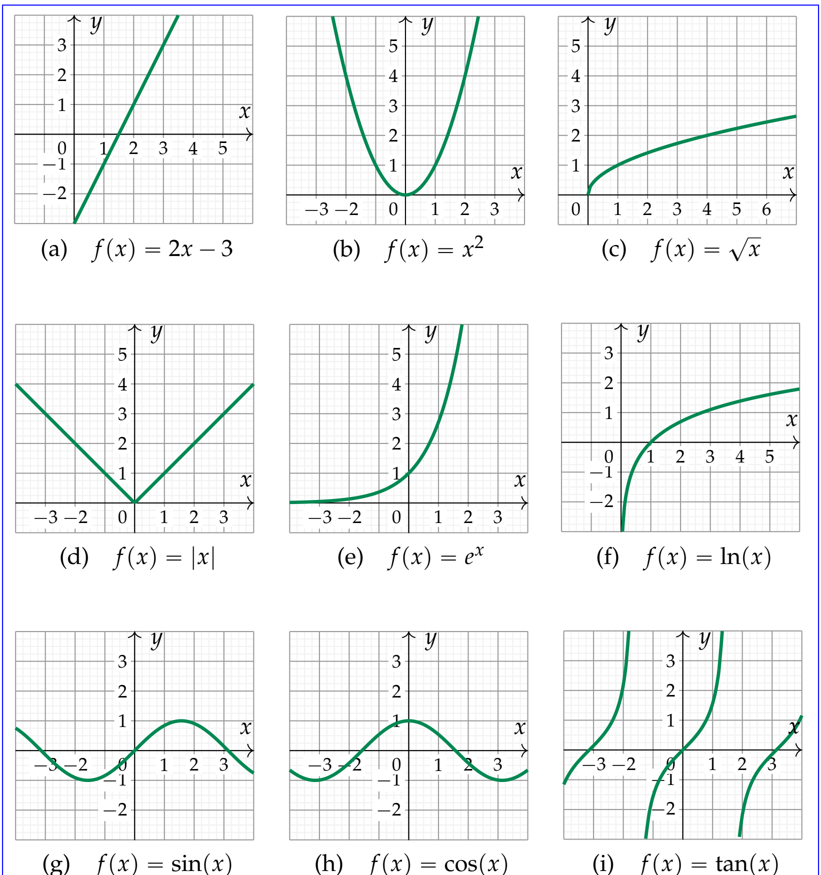
1.12 Functions reference

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 you~~, so you'll know how to handle the situation when a teacher tries to pull some trick on you at the final.

If you~~are~~'re seeing these functions for the first time, don't worry about remembering all the facts and properties on the first reading.

We~~will~~'ll use these functions throughout the rest of the book~~so you will~~, so you'll have plenty of time to become familiar with them. ~~Just remember~~ Remember to return to this section if you ever get stuck on a function.

To build mathematical intuition, it's essential you understand functions' graphs. Memorizing the definitions and properties of functions gets a lot easier with visual accompaniment. Indeed, remembering what the function "looks like" is a great way to train yourself to recognize various types of functions. Figure 1.28 shows the graphs of some of the most important functions we'll use in this book.



(a) $f(x) = 2x - 3$ (b) $f(x) = x^2$ (c) $f(x) = \sqrt{x}$

(d) $f(x) = |x|$ (e) $f(x) = e^x$ (f) $f(x) = \ln(x)$

(g) $f(x) = \sin(x)$ (h) $f(x) = \cos(x)$ (i) $f(x) = \tan(x)$

Figure 1.28: We'll see many types of function graphs in the next pages.

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~~ is the value of the function when $x = 0$.

~~The equation of the line $f(x) = mx + b$ is so important that it's worth taking the time to contemplate it for a few seconds.~~ Consider what relationship the equation of $f(x)$ describes for different values of m and b. What happens when m is positive? What happens when m is negative?

~~I'll leave some blank space here to give you "pages-turned" credit for taking the time.~~

Graph

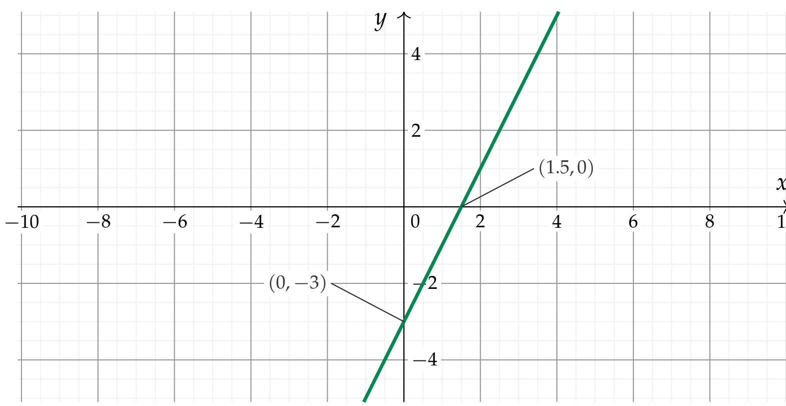


Figure 1.29: The graph of the function $f(x) = 2x - 3$. The slope is $m = 2$. The y-intercept of this line is ~~at $y = -3$~~$b = -3$. The x-intercept is at $x = \frac{3}{2}$.

Properties

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

- ~~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 1.29 is $2x - 1y = 3$.

Given the general equation of a line $Ax + By = C$ with $B \neq 0$, you can convert to the function form $y = f(x) = mx + b$ ~~using $b = \frac{C}{B}$ and~~ by computing the slope $m = \frac{-A}{B}$ and the y-intercept $b = \frac{C}{B}$.

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.

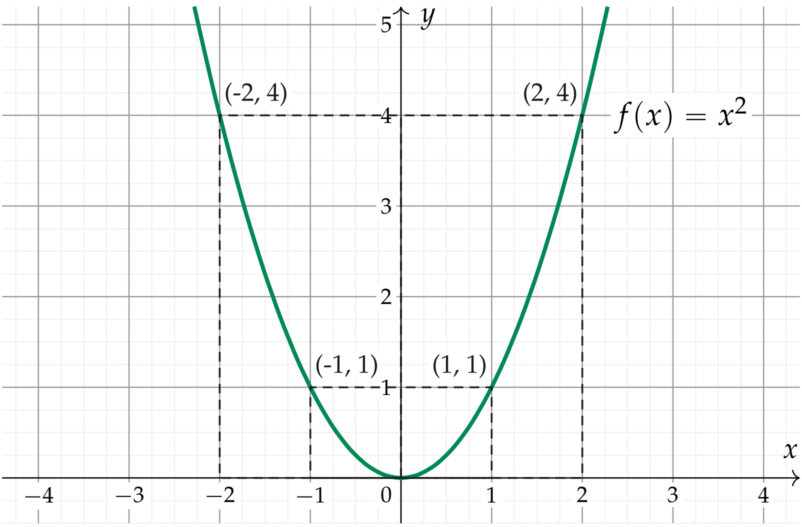


Figure 1.30: Plot of the quadratic function $f(x) = x^2$. The graph of the function passes through the following (x, y) coordinates: $(-2, 4)$, $(-1, 1)$, $(0, 0)$, $(1, 1)$, $(2, 4)$, $(3, 9)$, etc.

Properties

- Domain: ~~$x \in \mathbb{R}$.~~ $\mathbb{R}$. The function $f(x) = x^2$ is defined for all ~~input values $x \in \mathbb{R}$~~numbers.
- Image: ~~$f(x) \in [0, \infty)$.~~ $\mathbb{R}_+ = \{y \in \mathbb{R} \mid y \geqslant 0\}$. The outputs are ~~never negative:~~ nonnegative numbers since $x^2 \geqslant 0$, for all ~~$x \in \mathbb{R}$~~real numbers x.
- 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.

The set expression $\{y \in \mathbb{R} \mid y \geqslant 0\}$ that we use to define the nonnegative real numbers ($\mathbb{R}_+$) is read "the set of real numbers that are greater than or equal to zero."

Square root

The square root function is denoted

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

The square root $\sqrt{x}$ is the inverse function of the square function x^2 ~~for $x \geqslant 0$~~when the two functions are defined as $f : \mathbb{R}_+ \to \mathbb{R}_+$. The symbol $\sqrt{c}$ refers to the *positive* solution of $x^2 = c$. Note that $-\sqrt{c}$ is also a solution of $x^2 = c$.

Graph

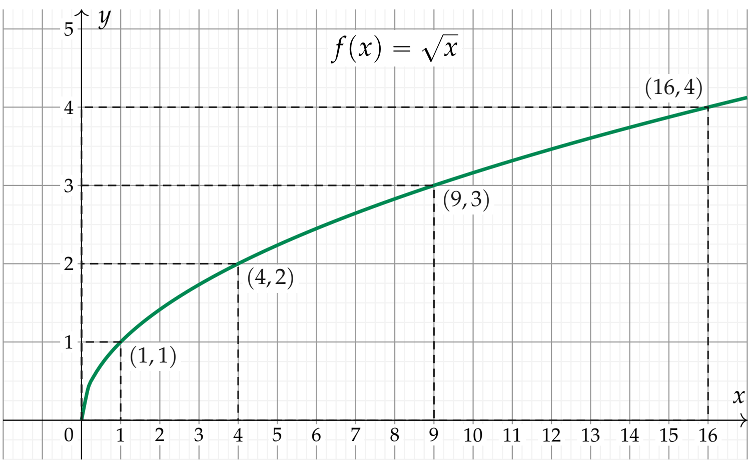


Figure 1.31: The graph of the function $f(x) = \sqrt{x}$. The domain of the function is ~~$x \in [0, \infty)$. You~~ $\mathbb{R}_+$ because we can't take the square root of a negative number.

Properties

- Domain: ~~$x \in [0, \infty)$.~~ $\mathbb{R}_+ = \{x \in \mathbb{R} \mid x \geqslant 0\}$. The function $f(x) = \sqrt{x}$ is only defined for nonnegative inputs~~$x \geqslant 0$.~~. 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: ~~$f(x) \in [0, \infty)$.~~ $\mathbb{R}_+ = \{y \in \mathbb{R} \mid y \geqslant 0\}$. The outputs of the function $f(x) = \sqrt{x}$ are ~~never negative:~~ nonnegative numbers since $\sqrt{x} \geqslant 0$~~, for all $x \in [0, \infty)$~~.

In addition to *square* root, there is also *cube* root ~~$f(x) = \sqrt[3]{x} \equiv x^{\frac{1}{3}}$~~ $f(x) = \sqrt[3]{x} = x^{\frac{1}{3}}$, 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 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 number. 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:

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

Graph

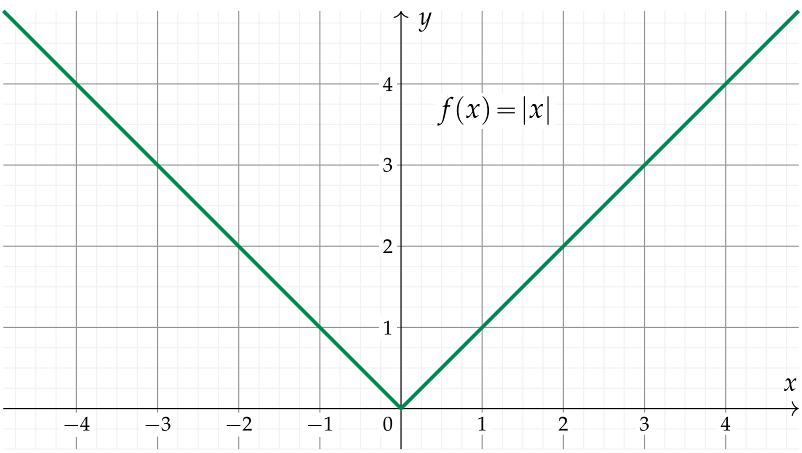


Figure 1.32: The graph of the absolute value function $f(x) = |x|$.

Properties

- ~~Always returns a nonnegative number~~ Domain: $\mathbb{R}$. The function $f(x) = |x|$ is defined for all inputs.
- Image: $\mathbb{R}_+ = \{y \in \mathbb{R} \mid y \geqslant 0\}$
- The combination of squaring followed by square-root is equivalent to the absolute value function:

$$\sqrt{x^2} \cancel{\equiv} = |x|,$$

You can add two polynomials by adding together their coefficients:

$$\begin{aligned} f(x) + g(x) &= (a_n x^n + \cdots + a_1 x + a_0) + (b_n x^n + \cdots + b_1 x + b_0) \\ &= (a_n + b_n)x^n + \cdots + (a_1 + b_1)x + (a_0 + b_0). \end{aligned}$$

The subtraction of two polynomials works similarly. We can also multiply polynomials together using the general algebra rules ~~of~~ for expanding brackets. ~~The notion of polynomial division also exists, but that's a more advanced topic that we won't discuss for now. Instead let's focus on the basics.~~

~~Solving polynomial equations~~

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~~ values of x that ~~makes~~ make 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:
Remember, if we perform the same operations on both sides of the equation, the resulting equation has the same solutions. Therefore, the values of x that satisfy

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

$x^2 - 3x - 10 = 0,$ namely $x = -2$ and $x = 5$, also satisfy

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

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

For a first-degree polynomial equation, $P_1(x) = mx + b = 0$, the solution is $x = \frac{-b}{m}$: just move b to the other side and divide by m.

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.

There is also a formula for polynomials of degree 3 and 4, but they are complicated. For polynomials with order $\geqslant 5$, 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 SymPy solve the standard-form equation $C(x) = 0$, call the function `solve(expr,var)`, where the expression `expr` corresponds to $C(x)$, and `var` is the variable you want to solve for. For example, to solve $x^2 - 3x + 2 = 0$, 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~~ solutions to any equation of the form `expr = 0`. ~~Indeed, we can verify that $x^2 - 3x + 2 = (x-1)(x-2)$, so~~ In this case, we see the solutions are $x = 1$ and $x = 2$~~are~~.

Another way to solve the equation is to factor the polynomial $C(x)$ using the function `factor` like this:

```
>>> factor(x**2 - 3*x + 2)             # usage: factor(expr)
 (x - 1)*(x - 2)
```

We see that $x^2 - 3x + 2 = (x-1)(x-2)$, which confirms the two roots are indeed $x = 1$ and $x = 2$.

To learn more about SymPy, check out Appendix D on page 519, which talks about all the SymPy functions that are available to you.

Substitution trick

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

$$g(x) = x^4 - 7x^2 + 10 = 0.$$

Imagine this problem is on your exam, where you are not allowed to use 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

$$g(y) = y^2 - 7y + 10 = 0,$$

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 $x = \pm\sqrt{\alpha}$ and $x = \pm\sqrt{\beta}$, 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 - 7*y + 10, y)
[2, 5]
>>> solve(x**4 - 7*x**2 + 10, x)
[sqrt(2), -sqrt(2), sqrt(5), -sqrt(5)]
```

Note how the second-degree polynomial has two roots, while the fourth-degree polynomial has four roots.

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, ..., a_n, a polynomial function can take on many different shapes. Consider the following observations about the symmetries of polynomials:

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 1.33.

In the remainder of this book, we'll meet the function $\sin(x)$ many times. We'll define the function $\sin(x)$ more formally as a trigonometric ratio in Section 1.15. In Chapter 3 we'll use $\sin(x)$ and $\cos(x)$ (another trigonometric ratio) to work out the *components* of vectors. Later in Chapter 4, we'll 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 on 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

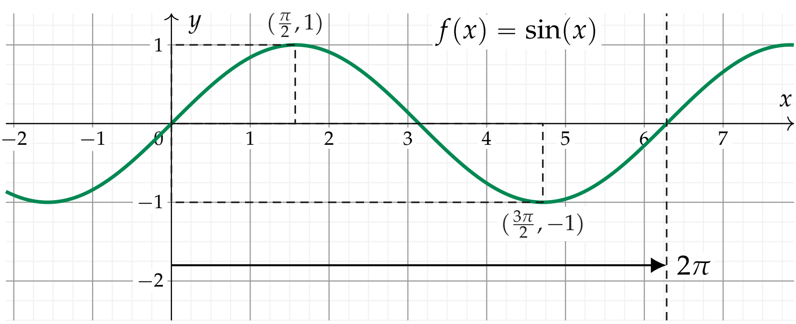


Figure 1.33: The graph of the function $y = \sin(x)$ passes through the following (x, y) coordinates: $(0,0)$, $(\frac{\pi}{6}, \frac{1}{2})$, $(\frac{\pi}{4}, \frac{\sqrt{2}}{2})$, $(\frac{\pi}{3}, \frac{\sqrt{3}}{2})$, $(\frac{\pi}{2}, 1)$, $(\frac{2\pi}{3}, \frac{\sqrt{3}}{2})$, $(\frac{3\pi}{4}, \frac{\sqrt{2}}{2})$, $(\frac{5\pi}{6}, \frac{1}{2})$, and $(\pi, 0)$. For ~~$x \in [\pi, 2\pi]$~~ x between π and 2π, the function's graph has the same shape ~~as~~ it has for ~~$x \in [0, \pi]$~~ x between 0 and π, but with negative values.

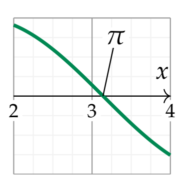


Figure 1.34: The function $f(x) = \sin(x)$ crosses the x-axis at $x = \pi$.

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.

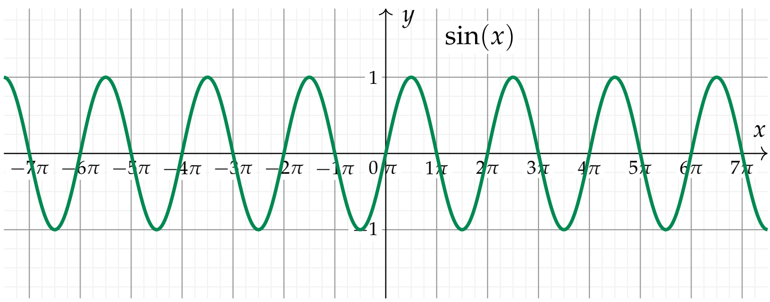


Figure 1.35: The graph of $\sin(x)$ from $x = 0$ to $x = 2\pi$ repeats periodically everywhere else on the number line.

Properties

- Domain: ~~$x \in \mathbb{R}$.~~ $\mathbb{R}$. The function $f(x) = \sin(x)$ is defined for all input values~~$x \in \mathbb{R}$~~.
- Image: ~~$\sin(x) \in [-1, 1]$.~~ $\{y \in \mathbb{R} \mid -1 \leqslant y \leqslant 1\}$. The outputs of the sine function are always between -1 and 1.
- Roots: ~~$[\ldots, -3\pi, -2\pi, -\pi, 0, \pi, 2\pi, 3\pi, \ldots]$~~$\{\ldots, -3\pi, -2\pi, -\pi, 0, \pi, 2\pi, 3$ 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: ~~$\csc(x) \equiv \frac{1}{\sin x}$~~ $\csc(x) = \frac{1}{\sin x}$ (csc is read *cosecant*)
- The inverse function of $\sin(x)$ is denoted as $\sin^{-1}(x)$ or $\arcsin(x)$, not to be confused with ~~$(\sin(x))^{-1} = \frac{1}{\sin(x)} \equiv \csc(x)$. Sometimes the function $\sin^{-1}(x)$ is denoted "$\arcsin(x)$."~~ $(\sin(x))^{-1} = \frac{1}{\sin(x)} = \csc(x)$.
- The number $\sin(\theta)$ is the length-ratio of the vertical side and the hypotenuse in a right-angle triangle with angle θ at the base.

Cosine

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

Graph

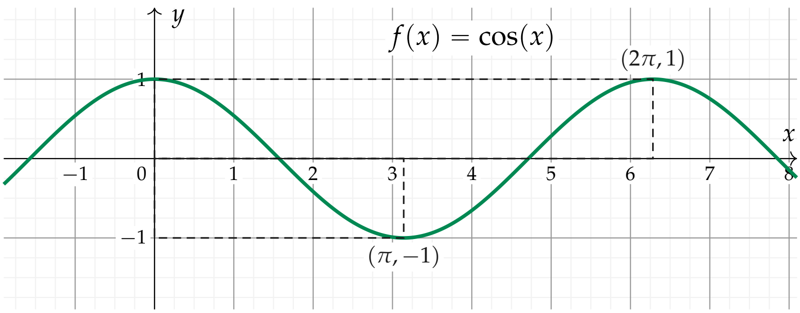


Figure 1.36: The graph of the function $y = \cos(x)$ passes through the following (x, y) coordinates: $(0, 1)$, $(\frac{\pi}{6}, \frac{\sqrt{3}}{2})$, $(\frac{\pi}{4}, \frac{\sqrt{2}}{2})$, $(\frac{\pi}{3}, \frac{1}{2})$, $(\frac{\pi}{2}, 0)$, $(\frac{2\pi}{3}, -\frac{1}{2})$, $(\frac{3\pi}{4}, -\frac{\sqrt{2}}{2})$, $(\frac{5\pi}{6}, -\frac{\sqrt{3}}{2})$, 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 again at $x = 2\pi$.

Properties

- Domain: $\cancel{x \in \mathbb{R}}$ $\mathbb{R}$
- Image: $\cancel{\cos(x) \in [-1, 1]}$ $\{y \in \mathbb{R} \mid -1 \leqslant y \leqslant 1\}$
- Roots: $\cancel{[\ldots, -\frac{3\pi}{2}, -\frac{\pi}{2}, \frac{\pi}{2}, \frac{3\pi}{2}, \frac{5\pi}{2}, \ldots]}$ $\{\ldots, -\frac{3\pi}{2}, -\frac{\pi}{2}, \frac{\pi}{2}, \frac{3\pi}{2}, \frac{5\pi}{2}, \ldots\}$
- Relation to sin: $\sin^2 x + \cos^2 x = 1$
- Relation to sec: $\cancel{\sec(x) \equiv \frac{1}{\cos x}}$ $\sec(x) = \frac{1}{\cos x}$ (sec is read *secant*)
- The inverse function of $\cos(x)$ is denoted $\cos^{-1}(x)$ or $\arccos(x)$.
- The cos function is *even*: $\cos(x) = \cos(-x)$
- The number $\cos(\theta)$ is the length-ratio of the horizontal side and the hypotenuse in a right-angle triangle with angle θ at the base

Tangent

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

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

Graph

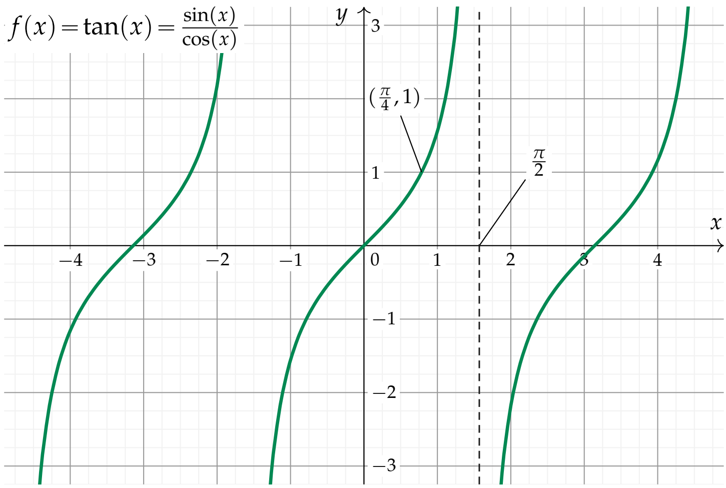


Figure 1.37: The graph of the function $f(x) = \tan(x)$.

Properties

- Domain: $\{x \in \mathbb{R} \mid x \neq \frac{(2n+1)\pi}{2}$ for any $n \in \mathbb{Z}\}$ ~~.~~
- Image: ~~$x \in \mathbb{R}$.~~ $\mathbb{R}$
- The function tan is periodic with period π.
- The tan function "blows up" at values of x where $\cos x = 0$. These are called *asymptotes* of the function and their locations are $x = \ldots, \frac{-3\pi}{2}, \frac{-\pi}{2}, \frac{\pi}{2}, \frac{3\pi}{2}, \ldots$.
- Value at $x = 0$: $\tan(0) = \frac{0}{1} = 0$, because $\sin(0) = 0$.
- Value at $x = \frac{\pi}{4}$: $\tan\left(\frac{\pi}{4}\right) = \frac{\sin(\frac{\pi}{4})}{\cos(\frac{\pi}{4})} = \frac{\frac{\sqrt{2}}{2}}{\frac{\sqrt{2}}{2}} = 1$.
- The number $\tan(\theta)$ is the length-ratio of the vertical and the horizontal sides in a right-angle triangle with angle θ.
- The inverse function of $\tan(x)$ is denoted $\tan^{-1}(x)$ or $\arctan(x)$.
- The inverse tangent function is used to compute the angle at the base in a right-angle triangle with horizontal side length ℓ_h and vertical side length ℓ_v: $\theta = \tan^{-1}\left(\frac{\ell_v}{\ell_h}\right)$.

Exponential

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

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

Graph

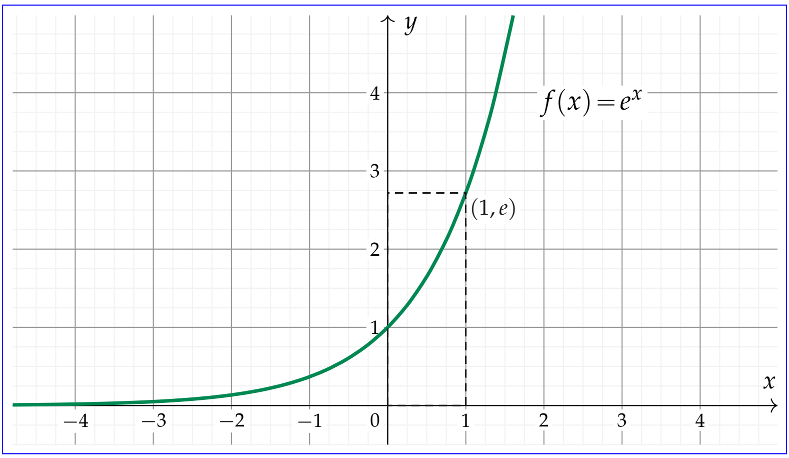


Figure 1.38: The graph of the exponential function $f(x) = e^x$ passes through the following ~~(x, y) coordinates~~points: $(-2, \frac{1}{e^2})$, $(-1, \frac{1}{e})$, $(0, 1)$, $(1, e)$, $(2, e^2)$, ~~$(3, e^3 = 20.08\ldots)$~~$(3, e^3)$, ~~$(5, 148.41\ldots)$~~$(4, e^4)$, ~~and $(10, 22026.46\ldots)$~~etc.

Properties

- Domain: $\cancel{x \in \mathbb{R}}$ $\mathbb{R}$
- Image: $\cancel{e^x \in (0, \infty)}$ $\{y \in \mathbb{R} \mid y > 0\}$
- $f(a)f(b) = f(a + b)$ since $e^a e^b = e^{a+b}$
- ~~The derivative (the slope of the graph) of the exponential function is the exponential function:~~ $\cancel{f(x) = e^x \quad \Rightarrow \quad f'(x) = e^x}$

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 1.38. 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 $f(x) = A1^x = A$.

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

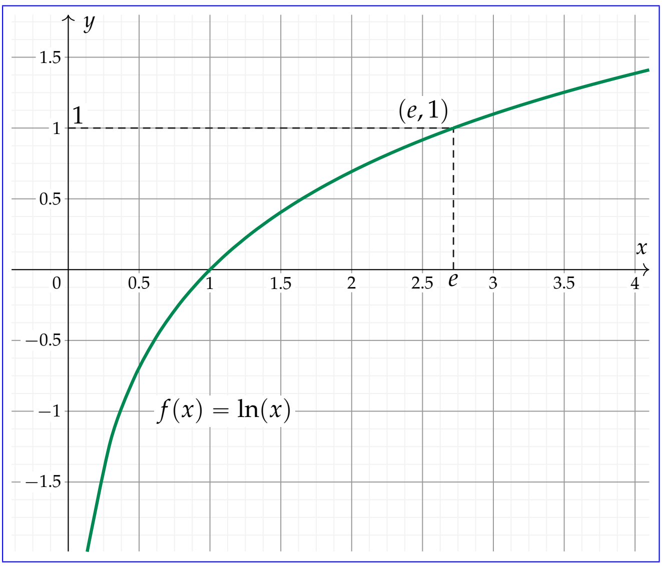


Figure 1.39: The graph of the function $\ln(x)$ passes through the following ~~(x, y)~~ coordinates: $(\frac{1}{e^2}, -2)$, $(\frac{1}{e}, -1)$, $(1, 0)$, ~~$(e, 1)$~~$(e, 1)$, ~~$(e^2, 2)$~~$(e^2, 2)$, ~~$(e^3, 3)$~~$(e^3, 3)$, ~~$(148.41\ldots, 5)$~~$(e^4, 4)$, ~~and $(22026.46\ldots, 10)$~~etc.

Properties

- Domain: $\{x \in \mathbb{R} \mid x > 0\}$
- Image: $\mathbb{R}$

Exercises

E1.17 Find the domain, the image, and the roots of $f(x) = 2\cos(x)$.

E1.18 What are the degrees of the following polynomials? Are they even, odd, or neither?

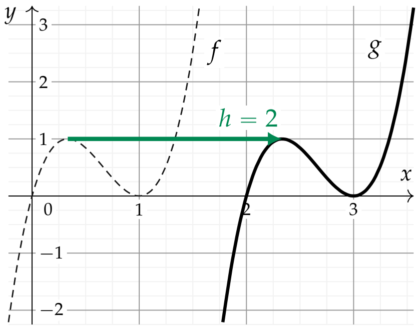


Figure 1.42: The graph of the function $g(x) = f(x-2)$ has the same shape as the graph of $f(x)$ translated to the right by two units.

Figure 1.42 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

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

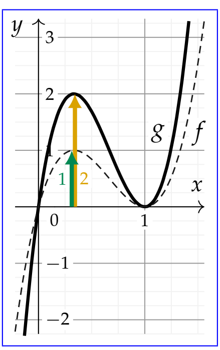


Figure 1.43: The graph of the function $g(x) = 2f(x)$ looks like $f(x)$ vertically stretched by a factor of two.

Figure 1.44 shows the function ~~$f(x) = 6.75(x^3 - 2x^2 + x)$, as well as the function~~ $g(x)$, which is $f(x)$ compressed horizontally by a factor of $a = 2$:

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

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

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 points of $f(x)$ are compressed toward the y-axis by a factor of 2.

General quadratic function

~~The general quadratic function takes the form~~Any quadratic function can be written in the form:

$$f(x) = \underline{A}a(x-h)^2 + k,$$

where x is the input, and ~~A, h~~a, h, and k are ~~the~~ parameters. This is called the ~~*parameters*~~*vertex form* of the quadratic function, and the coordinate pair (h,k) is called the *vertex* of the parabola. This equation can be obtained by starting from the basic quadratic function x^2 (see Figure 1.30) and applying three transformations: a horizontal translation by h units, a vertical scaling by a, and finally a vertical translation by k units.

Parameters

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

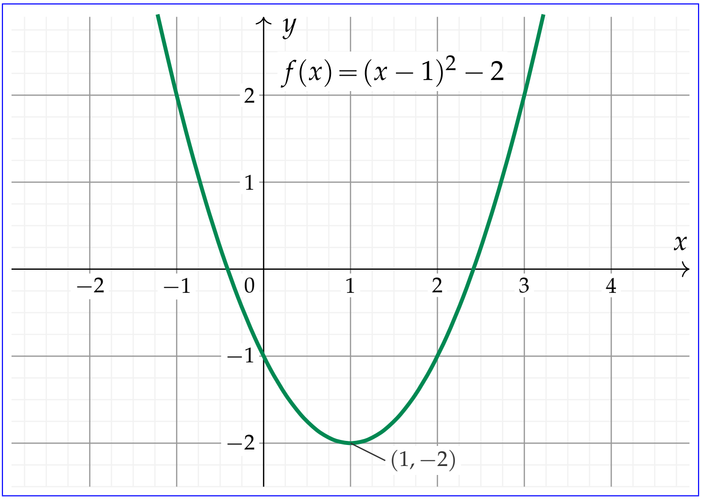


Figure 1.45: The graph of the function $f(x) = (x-1)^2 - 2$ is the same as the ~~basic~~ function $f(x) = x^2$, but shifted one unit to the right and two units down.

~~Graph~~

The graph in Figure 1.45 illustrates a quadratic function with parameters ~~$A = 1$~~ $a = 1$, $h = 1$ (one unit shifted to the right), and $k = -2$ (two units shifted down).

We can also write a quadratic function as a second-degree polynomial $f(x) = ax^2 + bx + c$. This is called the *standard form* of the quadratic function. Given a quadratic expression in standard form $ax^2 + bx + c$, we can find its equivalent expression in vertex form $a(x-h)^2 + k$ using the complete-the-square trick we learned in Section 1.6.

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

$$f(x) = \cancel{A}a(\cancel{x-a}x - x_1)(\cancel{x-b}x - x_2),$$

where ~~a and b~~ x_1 and x_2 are the two roots ~~. Another common way of writing a quadraticfunction is $f(x) = Ax^2 + Bx + C$.~~

~~Properties~~

of the quadratic. Given a quadratic function $f(x) = ax^2 + bx + c$, we can find its roots using the quadratic formula: $x_1 = \frac{-b+\sqrt{b^2-4ac}}{2a}$ and $x_2 = \frac{-b-\sqrt{b^2-4ac}}{2a}$ (see Section 1.7).

- ~~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 $x_1 \neq x_3$.~~

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.

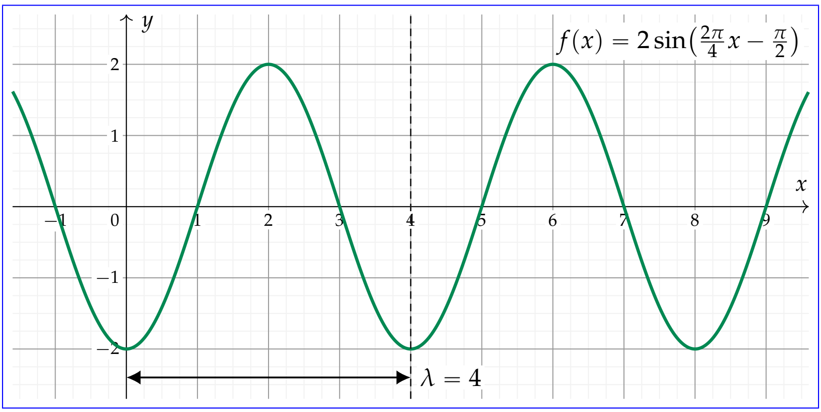


Figure 1.46: The graph of the function $f(x) = 2\sin\left(\frac{2\pi}{4}x - \frac{\pi}{2}\right)$, which has amplitude $A = 2$, wavelength $\lambda = 4$, and phase shift $\phi = \frac{\pi}{2}$.

Parameters

- A: the amplitude describes the distance above and below the x-axis that the function reaches as it oscillates.
- 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" sine 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.

Exercises

E1.20 Given the functions $f(x) = x + 5$, $g(x) = x - 6$, $h(x) = 7x$, and $q(x) = x^2$, find the formulas for the following composite functions:

a) $q \circ f$ **b)** $f \circ q$ **c)** $q \circ g$ **d)** $q \circ h$

In each case, describe how the graph of the composite function is related to the graph of $q(x)$.

E1.21 Find the amplitude A, the wavelength λ, and the phase shift ϕ for the function $\cancel{f(x) = 5\sin(62.83t - \frac{\pi}{8})}$ $f(x) = 5\sin(62.83x - \frac{\pi}{8})$.

E1.22 Choose the coefficients a, b, and c for the quadratic function $f(x) = ax^2 + bx + c$ so that it passes through the points $(0, 5)$, $(1, 4)$, and $(2, 5)$.
Hint: Find the equation $f(x) = A(x - h)^2 + k$ first.

E1.23 Find the values α and β that will make the function $g(x) = 2\sqrt{x - \alpha} + \beta$ pass through the points $(3, -2)$, $(4, 0)$, and $(7, 2)$.

1.14 Geometry

The word "geometry" comes from the Greek roots *geo*, which means "earth," and *metron*, which means "measurement." This name is linked to one of the early applications of geometry, which was to measure the total amount of land contained within a certain boundary region. Over the years, the study of geometry evolved to be more abstract. Instead of developing formulas for calculating the area of specific regions of land, mathematicians developed general area formulas that apply to *all* regions that have a particular shape.

In this section we'll present ~~a number of~~ formulas for calculating the perimeters, areas, and volumes for various shapes (also called "figures") commonly encountered in the real world. For two-dimensional figures, the main quantities of interest are the figures' areas and the figures' perimeters (the length of the walk around the figure). For three-dimensional figures, the quantities of interest are the surface area (how much paint it would take to cover all sides of the figure), and volume (how much water it would take to fill a container of this shape). The formulas presented are by no means an exhaustive list of everything there is to know about geometry, but they represent a core set of facts that you want to add to your toolbox.

Cosine rule The cosine rules states the following equations are true:

$$a^2 = b^2 + c^2 - 2bc\cos(\alpha),$$
$$b^2 = a^2 + c^2 - 2ac\cos(\beta),$$
$$c^2 = a^2 + b^2 - 2ab\cos(\gamma).$$

These equations are useful when you know two sides of a triangle and the angle between them, and you want to find the third side.

Circle

The circle is a beautiful shape. If we take the centre of the circle at the origin $(0,0)$, the circle of radius r corresponds to the equation

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

This formula describes the set of points (x, y) with a distance from the centre equal to r.

Area

The area enclosed by a circle of radius r is given by $A = \pi r^2$. A circle of radius $r = 1$ has area π.

Circumference and arc length

The circumference of a circle of radius r is ~~given by the formula~~

$$C = 2\pi r.$$

~~The~~

$$C = 2\pi r.$$

A circle of radius $r = 1$ has circumference 2π. This is the total length you can measure by following the curve all the way around to trace the outline of the entire circle. For example, the circumference of a circle of radius 3 m is $C = 2\pi(3) = 18.85$ m. This is how far you'll need to walk to complete a full turn around a circle of radius $r = 3$ m.

~~The area enclosed by a circle of radius~~ 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 $\theta = 57°$ as shown in Figure 1.49. What is the arc's length ℓ? 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

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

The arc length ℓ depends on r, the angle θ, and a factor of $\frac{2\pi}{360}$.

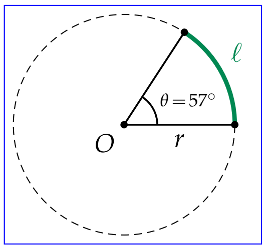


Figure 1.49: The arc length ℓ equals $\frac{57}{360}$ of the circle's circumference $2\pi r$.

Radians

While scientists and engineers commonly use degrees as a measurement unit for angles, mathematicians prefer to measure angles in *radians*, denoted rad.

Measuring an angle in radians is equivalent to measuring the arc length ℓ on a circle with radius $r = 1$, as illustrated in Figure 1.50.

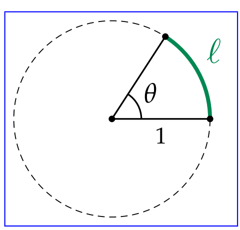


Figure 1.50: The angle θ measured in radians corresponds to the arc length ℓ on a circle with radius 1. The full circle corresponds to the angle 2π rad.

The conversion ratio between degrees and radians is

$$2\pi \text{ rad} = 360°.$$

When the angle θ is measured in radians, the arc length is given by:

$$A\ell = \pi r^2 \theta.$$

~~Circles are~~ To find the arc length ℓ, we simply multiply the circle radius r times the angle θ measured in radians.

Note the arc-length formula with θ measured in radians is simpler than the arc-length formula with θ measured in degrees, since we don't need the conversion factor of $360°$.

The geometry of circles is so important that we dedicated a whole section (Section 1.17) to ~~them~~pursuing this topic in more detail. For now, let's continue discussing some other important geometric shapes.

Sphere

A sphere of radius r is described by the equation $x^2 + y^2 + z^2 = r^2$. The surface area of the sphere is $A = 4\pi r^2$, and its volume is given by $V = \frac{4}{3}\pi r^3$.

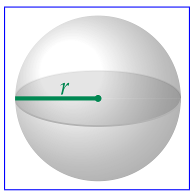


Figure 1.51: A sphere of radius r has surface area $4\pi r^2$ and volume $\frac{4}{3}\pi r^3$.

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 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 combined volume of the four pistons, which are cylindrical in shape. The owner's manual tells you the radius of each piston is 43.75 mm, and the height of each piston is 83.1 mm. Verify the total engine volume is 1998789 $\text{mm}^3 \approx 2$ L.

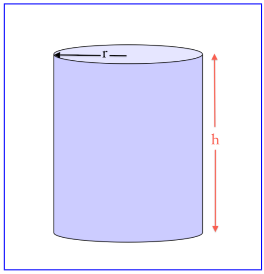


Figure 1.52: A cylinder with radius r and height h has volume $\pi r^2 h$.

Cones and pyramids

The volume of a square pyramid with side length a and height h is given by the formula $V = \frac{1}{3}a^2h$. The volume of a cone of radius r and height h is given by the formula $V = \frac{1}{3}\pi r^2 h$. Note the factor $\frac{1}{3}$ appears in both formulas. These two formulas are particular cases of the general volume formula that applies to all pyramids:

$$V = \tfrac{1}{3}Ah,$$

where A is the area of the pyramid's base and h is its height. This formula applies for pyramids with a base that is a triangle (triangular pyramids), a square (square pyramids), a rectangle (rectangular pyramids), a circle (cones), or any other shape.

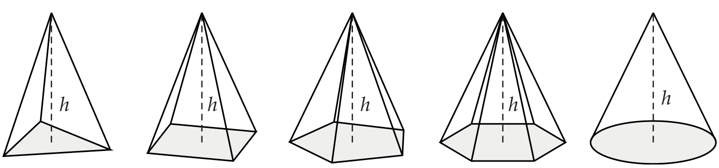
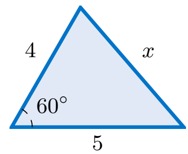


Figure 1.53: The volumes of pyramids and cones are described by the formula $V = \frac{1}{3}Ah$, where A is the area of the base and h is the height.

~~The System is obsessed with the pyramid shape. Many large organizations are structured like pyramids: the top boss tells vice-presidents what to do, vice-presidents tell directors what to do, directors tell upper management what to do, and so on until the commands reach regular employees. This pyramid-like structure allows for tight control of information and budgets within the organization. Pyramid structures are not *necessarily* bad; yet we often find some of the worst aspects of human nature concentrated at the tops of society's pyramids. It's wise to keep an eye on pyramid-shaped power structures, and watch out for any shenanigans the big bosses may try to pull.~~

Exercises

E1.24 Find the length of side x in the triangle below.



Hint: Use the cosine rule.

E1.25 Find the volume and the surface area of a sphere with radius 2.

E1.26 On a rainy day, Laura brings her bike indoors, and the wet bicycle tires leave a track of water on the floor. What is the length of the water track left by the bike's rear tire (diameter 73 cm) if the wheel makes five full turns along the floor?

1.15 Trigonometry

~~We can put any three lines together to make a triangle. What 's more, if~~ If one of the ~~triangle's angles~~ angles in a triangle 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~~ concepts like vectors and complex numbers~~, as well as physics subjects like oscillations and waves~~.

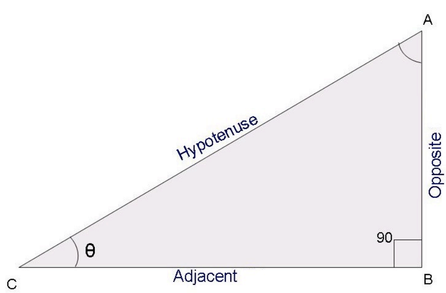


Figure 1.54: A right-angle triangle. The angle at the base is denoted θ and the names of the sides of the triangle are indicated.

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.

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 ~~$|\text{hyp}|^2$~~ hyp^2, we obtain

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

~~which~~ Since $\frac{\text{adj}}{\text{hyp}} = \cos\theta$ and $\frac{\text{opp}}{\text{hyp}} = \sin\theta$, this equation can be rewritten as

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

This is a powerful *trigonometric identity* that describes an important ~~relationship~~ relation between sine and cosine functions. In case you've never seen this notation before, the expression $\cos^2\theta$ is used to denote $(\cos(\theta))^2$.

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,~~ at Figure 1.55 and you'll see a circle. This is the unit circle because it has a radius of 1. You see it, yes?"

"Yes."

"Now imagine a point P that moves along the circle of radius 1, starting from the point $P(0) = (1, 0)$. The x and y coordinates of the point ~~$P(\theta) = (P_x(\theta), P_y(\theta))$~~ $P(\theta) = (P_x(\theta),\ P_y(\theta))$ 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 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.

The unit circle

The *unit circle* is a circle of radius one centred at the origin. The unit circle consists of all points (x, y) that satisfy the equation $x^2 + y^2 = 1$. A point ~~$P = (P_x, P_y)$~~ P on the unit circle has coordinates $(P_x, P_y) = (\cos\theta, \sin\theta)$, where θ is the angle P makes with the x-axis.

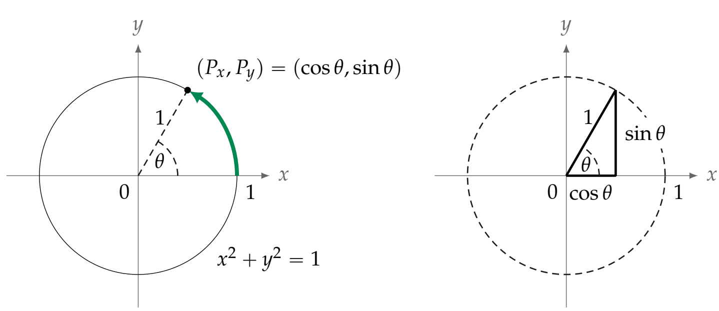


Figure 1.55: The unit circle corresponds to the equation $x^2 + y^2 = 1$. The coordinates of the point P on the unit circle are $P_x = \cos\theta$ and $P_y = \sin\theta$.

~~You should be familiar with the values of sin~~ Figure 1.56 shows the

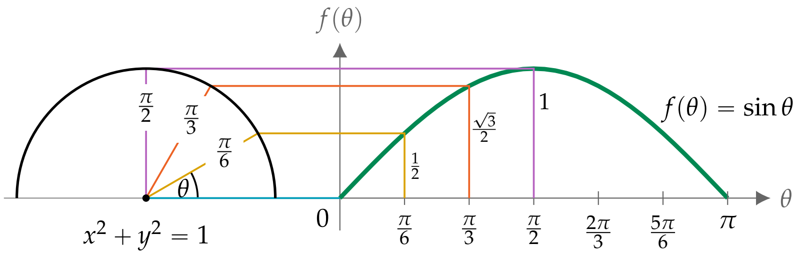


Figure 1.56: The function $f(\theta) = \sin\theta$ describes the vertical position of a point P that travels along the unit circle. The ~~first half~~ graph shows the values of ~~a cycle is shown~~the function $f(\theta) = \sin\theta$ for angles between $\theta = 0$ and $\theta = \pi$.

graph of the function $f(\theta) = \sin\theta$. The values $\sin\theta$ for the angles 0, $\frac{\pi}{6}$ (30°), $\frac{\pi}{3}$ (60°), and $\frac{\pi}{2}$ (90°) are marked. There are three values to remember: $\sin\theta = 0$ when $\theta = 0$, $\sin\theta = \frac{1}{2}$ when $\theta = \frac{\pi}{6}$ (30°), and $\sin\theta = 1$ when $\theta = \frac{\pi}{2}$ (90°). See Figure 1.33 (page 88) for a graph of $\sin\theta$ that shows a complete cycle around the circle. Also see Figure 1.36 (page 91) for the graph of $\cos\theta$.

Instead of trying to memorize the values of the functions $\cos\theta$ and ~~cos for all angles that are multiples of~~ $\sin\theta$ separately, it's easier to remember them as a combined "package" $(\cos\theta, \sin\theta)$, which describes the x- and y-coordinates of the point P for the angle θ. Figure 1.57 shows the values of $\cos\theta$ and $\sin\theta$ for the angles 0, $\frac{\pi}{6}$ (30°)~~or~~, $\frac{\pi}{4}$ (45°)~~. All of them are shown in Figure 1.58~~, $\frac{\pi}{3}$ (60°), and $\frac{\pi}{2}$ (90°). These are the most common angles that often show up on homework and exam questions. For each angle, the x-coordinate (the first number in the bracket) is $\cos\theta$, and the y-coordinate (the second number in the bracket) is $\sin\theta$.

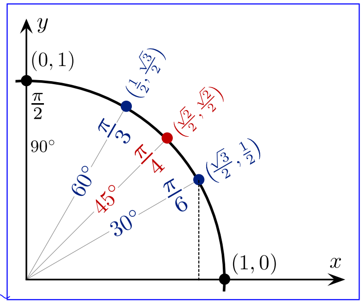


Figure 1.57: The ~~unit circle. The~~ combined $(\cos\theta, \sin\theta)$ coordinates ~~of~~ for the ~~point~~ points on the unit circle ~~$(\cos\theta, \sin\theta)$ are indicated for several important values of~~ at the ~~angle θ~~most common angles: 0, $\frac{\pi}{6}$ (30°), $\frac{\pi}{4}$ (45°), $\frac{\pi}{3}$ (60°), and $\frac{\pi}{2}$ (90°).

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

$$\sin(30°) = \sin(\tfrac{\pi}{6}) = \tfrac{1}{2}.$$

~~Knowing this, you can determine all the other angles . Let's start with $\cos(30°)$. We know that at 30°, point P~~ Note the values of $\cos\theta$ and $\sin\theta$ for the angles shown in Figure 1.57 are all combinations of the fractions $\frac{1}{2}$, $\frac{\sqrt{2}}{2}$, and $\frac{\sqrt{3}}{2}$. The square roots appear as a consequence of the trigonometric identity $\cos^2\theta + \sin^2\theta = 1$. This identity tells us that the sum of the squared coordinates of each point 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,~~

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

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

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

~~To find the~~ is equal to one. Let's look at what this equation tells us for the angle $\theta = \frac{\pi}{6}$ (30°). Remember that $\sin(30°) = \frac{1}{2}$

(the length of the dashed line in Figure 1.57). We can plug this value into the equation $\cos^2(30°) + \sin^2(30°) = 1$ to find the value: $\cos(30°) = \sqrt{1 - \sin^2(30°)} = \sqrt{1 - \frac{1}{4}} = \sqrt{\frac{3}{4}} = \frac{\sqrt{3}}{2}$.

The coordinates $\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$ for the angle $\theta = \frac{\pi}{4}$ (45°) are obtained from a similar calculation. We know the values of $\sin\theta$ and $\cos\theta$ must be equal for that angle, so we're looking for the number a that satisfies the equation $a^2 + a^2 = 1$, which is $a = \frac{1}{\sqrt{2}} = \frac{\sqrt{2}}{2}$. The values of $\cos(60°)$ and $\sin(60°)$ ~~, observe the symmetry of the circle. 60 degrees measured~~ can be obtained from a symmetry argument. Measuring 60° from the x-axis is the same as ~~30 degrees measured~~ measuring 30° from the y-axis~~. From this, we know~~, so $\cos(60°) = \sin(30°) = \frac{1}{2}$ ~~. Therefore, $\sin(60°) = \frac{\sqrt{3}}{2}$~~ and $\sin(60°) = \cos(30°) = \frac{\sqrt{3}}{2}$.

~~To find the values of sin and cos for~~ We can extend the calculations described above for all other angles that are multiples of $\frac{\pi}{6}$ (30°) and $\frac{\pi}{4}$ (45°~~, we need to find the value a such that~~

$$\cancel{a^2 + a^2 = 1,}$$

~~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~~

$$\cancel{\frac{\sqrt{2}}{2} = \cos(45°) = \sin(45°).}$$

~~All other angles in the circlebehave like the three angles above, with one difference:~~) to obtain the $\cos\theta$ and $\sin\theta$ values for the whole unit circle, as shown in Figure 1.58.

PLZ STAY CALM AND UNIT CIRCLE Don't be intimidated by all the information shown in Figure 1.58! You're not expected to memorize all these values. The primary reason for including this figure is so you can appreciate the symmetries of the sine and cosine values that we find as we go around the circle. The values of $\sin\theta$ and $\cos\theta$ for all angles are the same as the values for the angles between 0° and 90°, but one or more of their ~~components~~ coordinates 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~~-coordinate is negative since the point lies to the left of the y-axis. Another use for Figure 1.58 is to convert between angles measured in degrees and radians, since both units for angles are indicated.

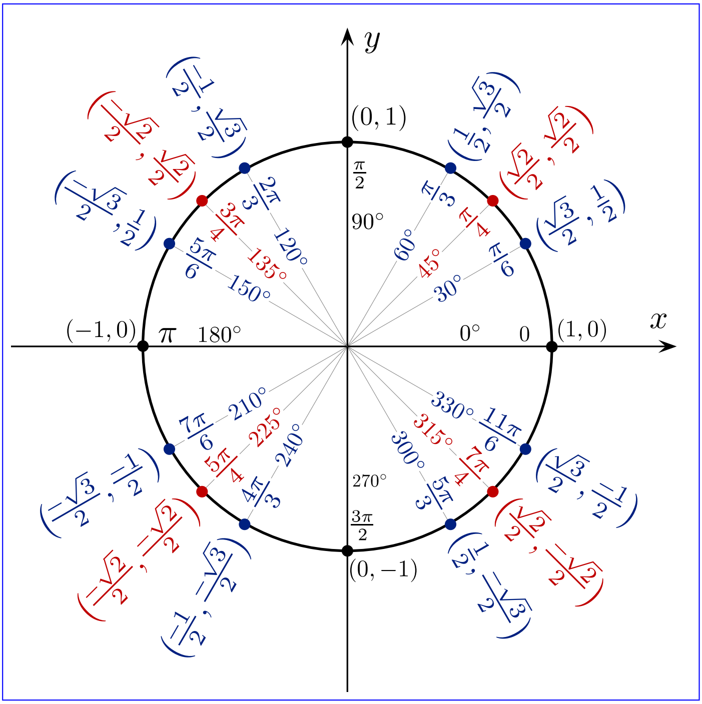


Figure 1.58: The coordinates of the point on the unit circle $(\cos\theta, \sin\theta)$ are indicated for all multiples of $\frac{\pi}{6}$ (30°) and $\frac{\pi}{4}$ (45°). Note the symmetries.

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 $\cancel{r_x = r\cos\theta}\,x = r\cos\theta$, and the length of the *opposite* side as $\cancel{r_y = r\sin\theta}\,y = r\sin\theta$. It is extremely important you get comfortable with this notation.

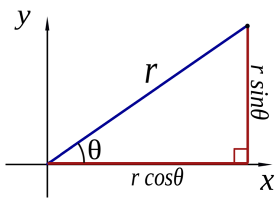


Figure 1.59: The x- and y-coordinates of a point at the angle θ and distance of r from the origin are given by $x = r\cos\theta$ and $y = r\sin\theta$.

The reasoning behind the above calculations is as follows:

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

and

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

Calculators

~~Make sure to set your calculator to the correct units for working with angles. What should you type into your~~ Watch out for the units of angle measures when using calculators and computers. Make sure you know what kind of angle units the functions sin, cos, and tan expect as inputs, and what kind of outputs the functions $\sin^{-1}$, $\cos^{-1}$, and $\tan^{-1}$ return.

For example, let's see what we should type into the calculator to compute the sine of 30 degrees~~? If your~~. If the calculator is set to degrees, we simply type: ~~, ,~~ $\boxed{3}$, $\boxed{0}$, $\boxed{\texttt{sin}}$, $\boxed{=}$, and obtain the answer 0.5.

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

1. Change the `mode` of the calculator so it works in degrees.
2. Convert 30° to radians

$$30\,\cancel{[^\circ]}^\circ \times \frac{\cancel{2\pi\,[\text{rad}]}\ 2\pi\text{ rad}}{\cancel{360\,[^\circ]}\ 360^\circ} = \frac{\pi}{6}\,\cancel{[\text{rad}]}\ \text{rad},$$

and type: ~~, , , , on your~~ $\boxed{\pi}$, $\boxed{/}$, $\boxed{6}$, $\boxed{\texttt{sin}}$, $\boxed{=}$ on the calculator.

Try computing $\cos(60^\circ)$, $\cos(\frac{\pi}{3}\text{ rad})$, and $\cos^{-1}(\frac{1}{2})$ using your calculator to make sure you know how it works.

Exercises

E1.27 Given a circle with radius $r = 5$, find the x- and y-coordinates of the point at $\theta = 45°$. What is the circumference of the circle?

E1.28 Convert the following angles from degrees to radians.

a) $30°$ **b)** $45°$ **c)** $60°$ **d)** $270°$

Links

[Unit-circle walkthrough and tricks by patrickJMT on YouTube]
`http://bit.ly/1mQg9Cj` ~~and~~ and `http://bit.ly/1hvA702`

1.16 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 sum of the squares of the sides of a triangle is equal to the square of the hypotenuse.

2. ~~sico + sico~~Sine angle sum

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

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

3. ~~coco − sisi~~Cosine angle sum

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

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 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 obtain

$$\begin{aligned}\cos(2x) &= \cos^2(x) - \sin^2(x) \\ &= 2\cos^2(x) - 1 = 2\left(1 - \sin^2(x)\right) - 1 = 1 - 2\sin^2(x).\end{aligned}$$

The formulas for expressing $\sin(2x)$ and $\cos(2x)$ in terms of $\sin(x)$ and $\cos(x)$ are called *double angle formulas*.

If we rewrite the double-angle formula for $\cos(2x)$ to isolate the $\sin^2$ or the $\cos^2$ term, we obtain the *power-reduction formulas*:

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

~~Self similarity~~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:

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

This follows because adding a multiple of 2π brings us back to the same point on the unit circle.

Furthermore, sin and cos ~~are self similar within each~~ have symmetries with respect to zero,

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

within each π half-cycle,

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

and within each full 2π cycle~~:~~,

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

Take the time to revisit Figure 1.33 (page 88), Figure 1.36 (page 91), and Figure 1.58 (page 114) to visually confirm that all the equations shown above are true. Knowing the points where the functions take on the same values (symmetries) or take on opposite values (anti-symmetries) is very useful in calculations.

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:

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

~~Sum formulas~~

$$\sin(a) + \sin(b) = 2\sin\left(\frac{1}{2}(a+b)\right)\cos\left(\frac{1}{2}(a-b)\right),$$

$$\sin(a) - \sin(b) = 2\sin\left(\frac{1}{2}(a-b)\right)\cos\left(\frac{1}{2}(a+b)\right),$$

$$\cos(a) + \cos(b) = 2\cos\left(\frac{1}{2}(a+b)\right)\cos\left(\frac{1}{2}(a-b)\right),$$

$$\cos(a) - \cos(b) = -2\sin\left(\frac{1}{2}(a+b)\right)\sin\left(\frac{1}{2}(a-b)\right).$$

Formulas for sums and products

~~Product formulas~~

Here are some formulas for transforming sums into products:

$$\sin(a) + \sin(b) = 2\sin\left(\tfrac{1}{2}(a+b)\right)\cos\left(\tfrac{1}{2}(a-b)\right),$$

$$\sin(a) - \sin(b) = 2\sin\left(\tfrac{1}{2}(a-b)\right)\cos\left(\tfrac{1}{2}(a+b)\right),$$

$$\cos(a) + \cos(b) = 2\cos\left(\tfrac{1}{2}(a+b)\right)\cos\left(\tfrac{1}{2}(a-b)\right),$$

$$\cos(a) - \cos(b) = -2\sin\left(\tfrac{1}{2}(a+b)\right)\sin\left(\tfrac{1}{2}(a-b)\right).$$

And here are some formulas for transforming products into sums:

$$\sin(a)\cos(b) = \frac{1}{2}\tfrac{1}{2}\Big(\sin(\sin(a+b)a + \sin(a-b)b) + \sin(a-b)\Big),$$

$$\sin(a)\sin(b) = \frac{1}{2}\tfrac{1}{2}\Big(\cos(\cos(a-b)a-b) - \cos(a+b)\cos(a+b)\Big),$$

$$\cos(a)\cos(b) = \frac{1}{2}\tfrac{1}{2}\Big(\cos(\cos(a-b)a-b) + \cos(a+b)\cos(a+b)\Big).$$

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.~~

Exercises

E1.29 Given $a = \pi$ and $b = \frac{\pi}{2}$, find

a) $\sin(a+b)$ **b)** $\cos(2a)$ **c)** $\cos(a+b)$

E1.30 Simplify the following expressions and compute their value without using a calculator.

a) $\cos(x) + \cos(\pi - x)$ **b)** $2\sin^2(x) + \cos(2x)$

c) $\sin(\frac{5\pi}{4})\sin(-\frac{\pi}{4})$ **d)** $2\cos(\frac{5\pi}{4})\cos(-\frac{\pi}{4})\cos(\pi)$

1.17 ~~Circle~~Circles and polar coordinates

~~The~~ In this section, we'll review what we know about circles and define the *polar coordinate system*, a specialized coordinate system for describing circles and other circular shapes.

Formulas

A *circle* is a set of points located at a constant distance from a centre point. ~~This geometric 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 a 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~~More generally, the circle's centre can be located at any point ~~(p,q) on the plane~~(h,k) in the plane, as illustrated in Figure ~~??~~1.60.

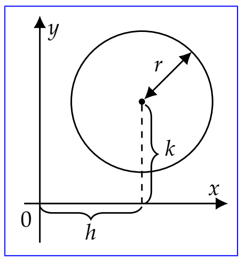


Figure 1.60: A circle of radius r centred at the point ~~(p,q)~~ (h,k) is described by the ~~formula $(x-p)^2 + (y-q)^2 = r^2$~~equation $(x-h)^2 + (y-k)^2 = r^2$.

~~Explicit function~~Describing circles using functions

The ~~equation of a circle~~ circle equation $x^2 + y^2 = r^2$ is a *relation* ~~or an *implicit function*~~ ~~involving~~ between the variables x and y. ~~To obtain an *explicit function*~~ If we want to describe the circle using a function $y = f(x)$~~for the circle~~, we can solve for y ~~to obtain~~ in the equation $x^2 + y^2 = r^2$ to obtain

$$y = f_t(x) = \sqrt{r^2 - x^2}, \qquad -r \leqslant x \leqslant r,$$

and

$$y = f_b(x) = -\sqrt{r^2 - x^2}, \qquad -r \leqslant x \leqslant r.$$

~~The explicit expression is really~~ Describing a circle requires two functions, ~~because a vertical line crosses the circle in two places. The first function corresponds to~~ f_t and f_b, because there are two values of y that satisfy the equation $x^2 + y^2 = r^2$ for each value of x. The function f_t describes the top half of the circle, ~~and the second function corresponds to~~ while the function f_b describes 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~~ You might be wondering why a simple geometric shape like a circle requires such complicated-looking formulas like $f_t(x)$ and $f_b(x)$ to describe it. Surely there's a better way to describe circles that doesn't involve quadratic expressions and square roots? There is! If instead of using the Cartesian coordinates (x, y) ~~of any point on the circle in terms of~~ we use the polar coordinates $r\angle\theta$, then the equation of a circle becomes very simple. We'll learn about that next.

The polar coordinate system

NEW STUFF

Figure 1.61 shows the *polar coordinate system*, which consists of concentric circles at different distances from the origin (also called the *pole*), and radial lines extending from the origin in all directions. We can specify the location of any point in the plane using the *polar coordinates* $r\angle\theta$, where r measures the ~~distance of the point~~point's distance from the origin, and θ ~~is~~ describes the angle measured ~~from the x~~in the counterclockwise direction starting from the r-axis. For example, the point $Q = 2\angle 60^\circ$ is located at the distance of $r = 2$ units form the origin, in the direction $\theta = 60^\circ$.

~~To convert from the polar coordinates $r\angle\theta$ to~~ Compare the polar coordinate system shown in Figure 1.61 with the Cartesian coordinate system in Figure 1.18. In the Cartesian coordinate system, we interpret the coordinate pair (x, y) as the instructions "Walk a distance of x units in the direction of the x-axis, and a distance of y units in the direction of the y-axis." In a polar coordinate system, we interpret the coordinates $r\angle\theta$ as the instructions "Turn toward the direction θ and walk a distance of r units in that direction." Both types of coordinates give instructions for getting to a particular point in the plane, with Cartesian coordinates giving the instructions in the form of two distances, while polar coordinates give the instructions in the form of a distance and a direction.

A Cartesian coordinate pair (x, y) is made of x and y coordinates, while a polar coordinate pair $r\angle\theta$ is made of r and θ coordinates. In this book, we use the angle symbol $\angle$ (read "at an angle of") to separate the polar coordinates r and θ, in order to emphasize the difference between Cartesian and polar coordinates. However, some other books use the notation (r, θ) for polar coordinates, so you have

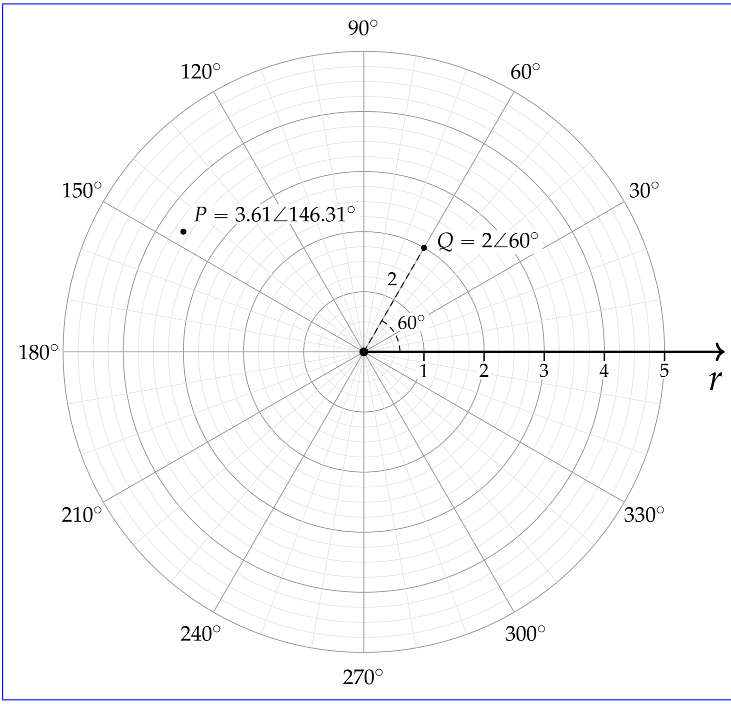


Figure 1.61: ~~Polar coordinates $r\angle\theta$~~ We can ~~be used~~ use the polar coordinate system to describe ~~any~~ points in the two-dimensional plane. The polar coordinates $r\angle\theta$ describe the point ~~(x,y)~~located at the distance r from the origin in the direction θ.

to watch out—the coordinate pair $(20, 30)$ could be either a (x, y) coordinate pair or a (r, θ) coordinate pair, depending on the context.

Note the polar coordinates that describe a given point are not unique, meaning the same point can be described in multiple ways. The point $Q = 2\angle 60°$ is equally described by the polar coordinates $2\angle -300°$, since a clockwise turn of $300°$ is the same as a counterclockwise turn of $60°$. We can also describe the same point Q using the polar coordinates $-2\angle 240°$ and $-2\angle -120°$, which tell us to turn in the direction opposite to $60°$ and measure a negative distance $r = -2$. While all of these polar coordinates for Q are equivalent, the preferred way to specify polar coordinates is with positive r values and angles $|\theta| \leqslant 180°$.

Converting between Cartesian and polar coordinates

Figure 1.62 shows a point whose location is described both in terms of Cartesian coordinates (x, y) and polar coordinates $r\angle\theta$. The triangle formed by the coordinates $(0,0)$, ~~use the~~ $(x,0)$, and (x,y) is a right-angle triangle. This means we can apply our knowledge of the trigonometric functions sin, cos, and tan to obtain formulas for converting between Cartesian coordinates (x,y) and polar coordinates $r\angle\theta$.

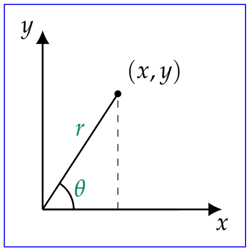


Figure 1.62: Polar coordinates $r\angle\theta$ can describe any point (x,y).

To convert from polar coordinates $r\angle\theta$ to (x,y) coordinates, we use the definitions of the trigonometric functions ~~cos and sin:~~ $\cos\theta = \frac{\text{adj}}{\text{hyp}} = \frac{x}{r}$ and $\sin\theta = \frac{\text{opp}}{\text{hyp}} = \frac{y}{r}$ to obtain the formulas:

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

For example, the Cartesian coordinates of the point $Q = 2\angle 60^\circ$ are given by $Q = (x,y) = (2\cos 60^\circ, 2\sin 60^\circ) = (1, \sqrt{3})$.

~~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: $\theta \in [0, 360^\circ)$. 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~~

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

~~In words, this expression describes the set of points~~ To convert from (x,y) ~~in the Cartesian plane with~~ coordinates to $r\angle\theta$ coordinates, we can use the circle equation $x^2 + y^2 = r^2$ and the definition of the tangent function $\tan\theta = \frac{\text{opp}}{\text{adj}} = \frac{y}{x}$, then solve for r and θ to obtain the formulas:

$$r = \sqrt{x^2 + y^2} \quad \text{and} \quad \theta = \begin{cases} \tan^{-1}\left(\frac{y}{x}\right) & \text{if } x > 0, \\ 180^\circ + \tan^{-1}\left(\frac{y}{x}\right) & \text{if } x < 0, \\ 90^\circ & \text{if } x = 0 \text{ and } y > 0, \\ -90^\circ & \text{if } x = 0 \text{ and } y < 0. \end{cases}$$

Finding the angle θ is a little tricky. We must use a different formula for computing θ depending on where the point is located, and there are four different cases to consider. The basic idea is to use the inverse tangent function $\tan^{-1}$, which is also called arctan, or `atan` on computer systems. By convention, the function $\tan^{-1}$ returns values between -90° ($-\frac{\pi}{2}$ rad) and 90° ($\frac{\pi}{2}$ rad), which correspond to points with positive x-coordinates~~that are described by $r\cos\theta$ and with y-coordinates that are described by $r\sin\theta$, where~~. If the x-coordinate of the point is negative, we must add 180° (π rad) to the output of the inverse-tangent calculation to obtain the correct angle. When $x = 0$ we can't compute the fraction $\frac{y}{x}$ because we cannot divide by zero, so we must handle the cases with $x = 0$ separately as described in the above equation.

If you have access to a computer algebra system, the easiest way to calculate the angle θ ~~varies from 0° to 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~~ for the point (x, y) is to use the two-input inverse tangent function `atan2(y,x)`. The function `atan2` is the best way to compute the angle since it handles all four cases of converting Cartesian coordinates to polar coordinates automatically and always gives the correct angle. You can try some calculations with `atan2` using the computer algebra system at `https://live.sympy.org`.

For example, consider the point P with Cartesian coordinates $(-3, 2)$ shown in Figure 1.18 (page 59). To find the polar coordinates of this point we first calculate the distance from the centre, $r = \sqrt{(-3)^2 + 2^2} = \sqrt{}$
To find the angle 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~~

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

~~The top half of the circle is also described by $\{(x, y) \in \mathbb{R}^2 \mid y = \sqrt{r^2 - x^2},\ x \in [-r,$ where the parameter used is the~~ we note that the x-coordinate of P is negative, so the angle θ we're looking for is given by the formula $\theta = 180^\circ + \tan^{-1}\!\left(\frac{2}{-3}\right) = 146.31^\circ$. The angle of the point $P = (-3, 2)$ can also be obtained from `atan2(2,-3)`. The polar coordinates of the point P are $3.61\angle 146.31^\circ$ (see Figure 1.61).

Equations in polar coordinates

Equations in polar coordinates serve to describe relations between the variables r and θ. For example, the equation of a circle with radius 2 in polar coordinates is simply $r = 2$. If we substitute

$r = \sqrt{x^2 + y^2}$ and square both sides of the equation, we obtain the equation $x^2 + y^2 = 2^2$ that we saw in the beginning of this section.

We can use the substitutions $x = r\cos\theta$ and $y = r\sin\theta$ to convert equations from Cartesian coordinates x and y to polar coordinates r and θ. Consider the equation $2x - y = 3$, which describes the line shown in Figure 1.29 on page 77. We can rewrite this equation in polar coordinates as $2r\cos\theta - r\sin\theta = 3$, which is a relation between the polar coordinates r and θ.

As you can tell from these examples, polar coordinates are very convenient when dealing with circles, and less so when working with lines. Indeed, describing a circle in polar coordinates is as simple as $r = 2$, while in Cartesian coordinates we had to use the complicated-looking functions f_t and f_b (see page 120). The situation is the opposite for lines: the equation of a line in Cartesian coordinates is simple, $2x - y = 3$, while in polar coordinates the same line is described by a tangled mess involving sin and cos functions.

~~Area~~Functions in polar coordinates

~~The area of a circle of radius r is $A = \pi r^2$.~~ A function in polar coordinates is denoted $r(\theta)$ and describes how the distance r varies as a function of the angle θ.

~~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~~For example, a circle with radius 2 is described by the function $r(\theta) = 2$ in polar coordinates, as illustrated in Figure 1.63 (a). The circle is described by a constant function in polar coordinates, since the points in all directions have the same distance from the centre.

~~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 $\theta = 57°$. What is the arc's length ℓ?~~ As another example, we can transform the equation of the line $2x - y = 3$ to polar coordinates to obtain $2r\cos\theta - r\sin\theta = 3$, then isolate r to obtain the function

$$r(\theta) = \frac{3}{2\cos\theta - \sin\theta},$$

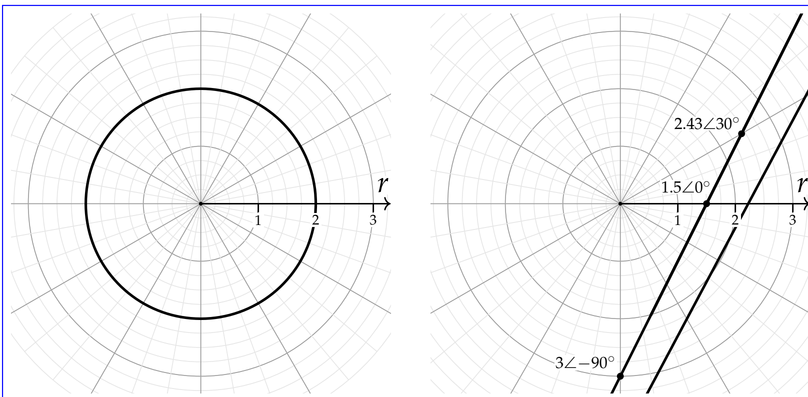
which describes the distance from the origin for different angles θ. For example, when $\theta = 0$, we find $r(0) = \frac{3}{2\cos 0 - \sin 0} = 1.5$, so we can plot the polar coordinates $1.5\angle 0^\circ$ on the function's graph.

~~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~~ The polar coordinates graph of the function $r(\theta)$ corresponds to all points with polar coordinates $r(\theta)\angle\theta$, for all possible values of θ~~is~~

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

~~The arc length ℓ depends on r, the angle θ,~~. This is analogous to how we obtain the graph of the function $f(x)$ in Cartesian coordinates by plotting the points $(x, f(x))$, for all possible values of the input variable x. Figure 1.63 shows the graphs of the two functions discussed here.

If you ever need to graph a function $r(\theta)$ by hand, you can compute the value of the function for several angles like $\theta = -90^\circ$, $\theta = 0^\circ$, $\theta = 90^\circ$, then plot these points in the polar coordinate system. For example, to graph the function $r(\theta) = \frac{3}{2\cos\theta - \sin\theta}$, we can compute $r(-90^\circ) = \frac{3}{2\cos(-90^\circ) - \sin(-90^\circ)} = 3$, which gives us the point $3\angle -90^\circ$ on the graph. We can similarly compute $r(0^\circ) = 1.5$ and ~~a factor of $\frac{2\pi}{360}$~~$r(30^\circ) = 2.43$, which gives us the points $1.5\angle 0^\circ$ and $2.43\angle 30^\circ$.



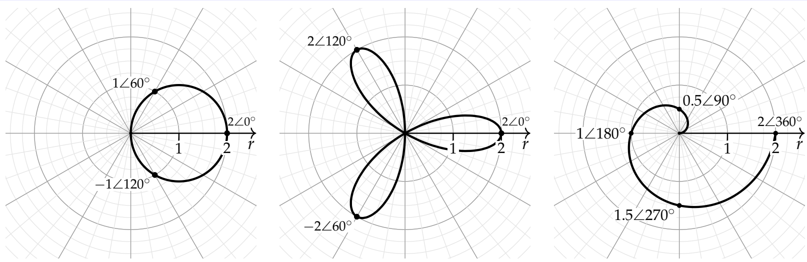
(a) $r(\theta) = 2$ (b) $r(\theta) = \frac{3}{2\cos\theta - \sin\theta}$

Figure 1.63: The ~~arc length ℓ equals $\frac{57}{360}$~~ graphs of functions in polar coordinates are obtained by computing the ~~circle's circumference $2\pi r$~~distance $r(\theta)$ in all directions θ varying from 0° to 360°.

~~**Radians**~~

Figure 1.64 shows the polar coordinates graphs of three other interesting functions. Look at the points $r\angle\theta$ indicated in each graph and check that they satisfy the corresponding function $r(\theta)$.

~~Though degrees are commonly used as a measurement unit for angles, it's much better to measure angles in *radians*~~



(a) $r(\theta) = 2\cos\theta$ (b) $r(\theta) = 2\cos(3\theta)$ (c) $r(\theta) = \frac{2}{360^\circ}\theta$

Figure 1.64: The graphs of three functions in polar coordinates: (a) a circle, (b) a three-petalled rose, and (c) an Archimedean spiral.

Discussion

The polar coordinate system is an alternative way of describing points in space using polar coordinates $r\angle\theta$ instead of the usual Cartesian coordinates (x, y). See the concept map in Figure 1.65. Your knowledge and experience with the trigonometric functions sin, ~~since radians are the *natural* units for measuring angles. The conversion ratio between degrees and radians is~~

$$2\pi[\text{rad}] = 360^\circ.$$

~~When measuring angles in radians, the arc length is given by:~~

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

~~Measuring angles in radians is equivalent to measuring arc length on a circle with radius $r = 1$~~cos, and tan is what allows you to convert between Cartesian and polar coordinates.

The formulas for converting between Cartesian coordinates (x, y) and polar coordinates $r\angle\theta$ covered in this section are important, and you should consider them "required material." I expect you to become totally fluent with these formulas now, because we'll need

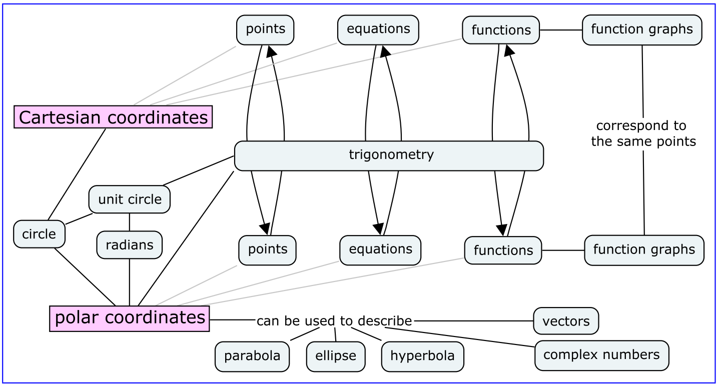


Figure 1.65: Cartesian coordinates (x, y) and polar coordinates $r\angle\theta$ are two equivalent systems for representing points, equations, and functions.

them later in the book when we learn about vectors (Section 3.2) and complex numbers (Section 3.5).

In contrast, the three next sections are not "required material." We'll now switch gears to "entertainment mode" and learn about three bonus geometry topics: ellipses, parabolas, and hyperbolas. I want you to know about these shapes, but I don't expect you to be fluent with all the definitions and equations. You can take it easy for the next three sections because none of the material will be "on the exam." You deserve a break after all the polar coordinates formulas!

Exercises

E1.31 ~~On a rainy day, Laura brings her bike indoors, and the wet bicycle tires leave a track of water on the floor. What is the length of the water track left by the bike's rear tire (diameter 73cm) if the wheel makes five full turns along the floor?~~ Convert the given points from Cartesian to polar coordinates:

a) $(3, 1)$ **b)** $(-1, -2)$ **c)** $(0, -6)$

Convert the points from polar to Cartesian coordinates:

d) $10\angle 30°$ **e)** $10\angle -345°$ **f)** $10\angle 120°$

E1.32 ~~Describe the circle of radius 3 centred at (1,4) in terms of Cartesian coordinates and in terms of a parametric equation.~~ Draw the graph of the function $r(\theta) = \frac{2}{\sin\theta}$ in polar coordinates for θ varying from 0 to 180°. What is the equivalent description of this function in Cartesian coordinates?

Links

[Visual introduction to polar coordinates]
https://www.youtube.com/watch?v=stU63ST6ung

[Professor Dave explains equations in polar coordinates]
https://www.youtube.com/watch?v=jwLUapqnwkk

1.18 Ellipse

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

Parameters

Figure 1.66 shows an ellipse with all its parameters annotated:

- ~~a: the half-length~~ F_1, F_2: the two *focal points* of the ellipse ~~along the x-axis, also known as the semi-major axis~~
- ~~b: the~~ 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
- a: the semi-major axis of the ellipse is the half-length of the ellipse along the ~~y~~x-axis. The distance between V_1 and V_2 is $2a$.
- ~~ε: the *eccentricity*~~ b: the semi-minor axis of the ellipse ~~, $\varepsilon = \sqrt{1-\frac{b^2}{a^2}}$~~
- ~~F_1, F_2: the two *focal points*~~ is the half-width of the ellipse along the y-axis. The distance between V_3 and V_4 is $2b$.
- ~~r_1~~c: the distance ~~from a point on the ellipseto~~ of the focal points from the centre of the ellipse. The distance between F_1
- ~~r_2: the distance from a point on the ellipseto~~ and F_2 is $2c$.
- ε: the *eccentricity* of the ellipse, $\varepsilon = \sqrt{1-\frac{b^2}{a^2}} = \frac{c}{a}$.

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

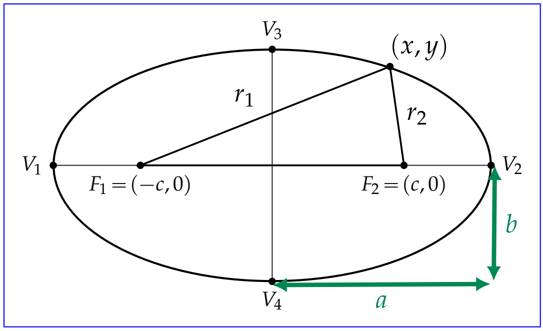


Figure 1.66: An ellipse with semi-major axis a and semi-minor axis b. The locations of the focal points F_1 and F_2 are indicated.

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.

~~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~~ The parameters a and b determine the shape 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~~ focal points F_1 and F_2 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~~The coordinates of the two focal points are

$$F_1 = (-c, 0) \qquad \text{and} \qquad F_2 = (c, 0),$$

where $c = \sqrt{a^2 - b^2}$ is the *focal distance*.

The *eccentricity* of an ellipse is given by the equation

$$\varepsilon = \sqrt{1 - \frac{b^2}{a^2}} = \frac{c}{a}.$$

The parameter ε (the Greek letter *epsilon*) varies between 0 and 1 and describes how much the shape of the ellipse differs from the shape of a circle. When $\varepsilon = 0$ the ellipse is a circle with radius a, and both focal points are located at the centre. As the eccentricity ε increases, the ellipse becomes more elongated and the focal points spread farther apart.

Polar coordinates

~~In polar coordinates, the ellipse can be described by a~~ Consider a polar coordination system whose centre is located at the focus F_2. We can describe the ellipse by specifying the function $r_2(\theta)$~~as illustrated in Figure ??. This function gives the distance of a point E from~~, which describes the distance from the focus F_2 to the point E on the ellipse as a function of the angle θ ~~. Recall~~ (see Figure 1.67). Recall that for functions in polar coordinates, the angle θ is the independent variable that varies from 0 to 2π (360°), and the dependent variable is the distance $r_2(\theta)$.

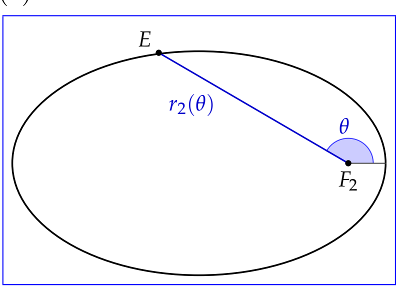


Figure 1.67: The ~~shape of~~ function $r_2(\theta)$ in polar coordinates specifies the distance between the point E on the ellipse ~~is described by~~ and the ~~function $r_2(\theta)$~~focal point F_2 for all angles.

The ~~equation of the ellipse in polar coordinates depends on the length of the semi-major axis a and the eccentricity ε. The equation~~ function 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~~semi-major axis. The distance is smallest when $\theta = 0$ with ~~$r_2(0) = a(1-\varepsilon) = q$~~ $r_2(0) = a - c = a(1-\varepsilon)$ and largest when $\theta = \pi$ with ~~$r_2(\pi) = a + a\varepsilon = a(1+\varepsilon)$~~$r_2($

Calculating the orbit of the Earth

~~To a close approximation, the~~ 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~~km. We substitute these values into the general formula for $r_2(\theta)$ and obtain the following equation:

$$r_2(\theta) = \cancel{\frac{149\,556\,484.56}{1+0.01671123\cos(\theta)}\ [\text{km}]}\frac{149\,556\,484}{1+0.01671123\cos(\theta)}\ \text{km}.$$

The point where the Earth is closest to the Sun is called the *perihelion*. It occurs when $\theta = 0$, which happens around the 3$^{\text{rd}}$ 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 3$^{\text{rd}}$ of July.

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

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

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

Google "perihelion" and "aphelion" to verify that the above predictions are accurate. It's kind of cool that a mathematical formula can describe the motion of our planet, don't you think?

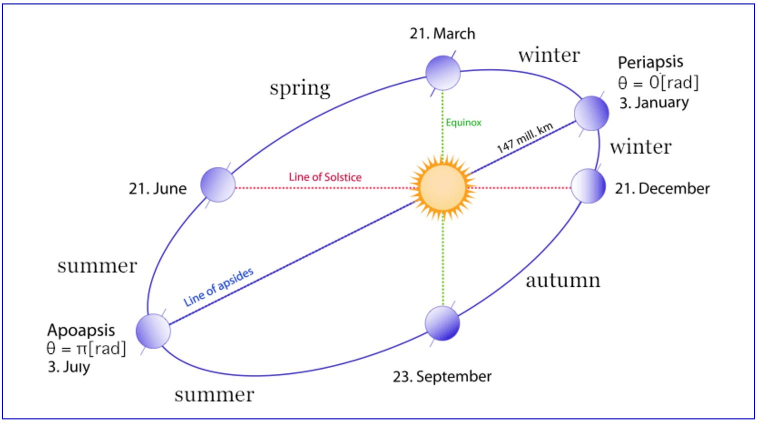


Figure 1.68: The orbit of the Earth around the Sun. Key points of the orbit are labelled. The seasons in the Northern hemisphere are also indicated.

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~~ the values of $\theta(t)$ with t measured in days shown in Table **??**. 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 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~~

Exercises

E1.33 The focal points of the ellipse with equation $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ are $F_1 = (-c, 0)$ and $F_2 = (c, 0)$, as illustrated in Figure 1.66. Use the definition of the ellipse ~~had nothing to do with seasons! Seasons are predominantly caused by the~~ $r_1 + r_2 = \text{const.}$ to compute the value of the parameter c in terms of the parameters a and b.

Links

[Interactive graph of an ellipse]
https://www.desmos.com/calculator/kgmh67lroj

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

1.19 Parabola

NEW STUFF

The *parabola* is another important geometric shape. In this section, we'll see how we can describe parabolas using their geometric properties, as well as in terms of algebraic equations.

Parameters

Figure 1.69 shows a parabola with all its parameters annotated:

- f: the *focal length* of the parabola
- $F = (0, f)$: the *focal point* of the parabola
- $\{(x, y) \in \mathbb{R}^2 \mid y = -f\}$: the *directrix* line to the parabola
- r: the distance from point P on the parabola to the focal point F
- ℓ: the closest distance from a point P on the parabola to the parabola's directrix line

Geometric definition

The shape of a parabola is determined by a single parameter f, called the ~~*axial tiltof 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* 21^{st} *of June. On that day, the Earth's spin axis is tilted toward the Sun so the Northern hemisphere receives*~~ *focal length*. For a parabola with focal

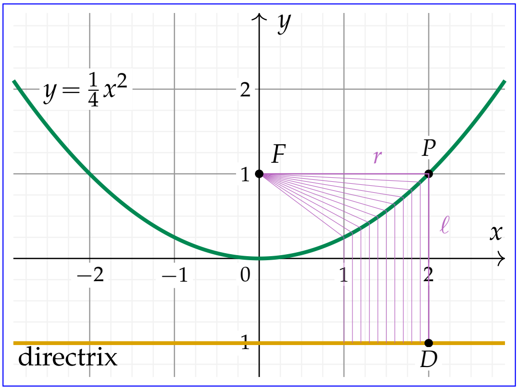


Figure 1.69: The parabola is defined geometrically as the set of points whose distance form the focal point r is equal to their distance from the directrix ℓ. The figure shows the point P on the parabola that has distance $r = 2$ from F, and distance $\ell = 2$ from the point D on the directrix. This parabola can be described algebraically using the equation $y = \frac{1}{4}x^2$.

length f, the focal point is at $F = (0, f)$ and the *directrix* line has the equation $y = -f$. The parabola is defined as the set of points P for which the distance from the focal point and the directrix are equal:

$$r = \ell,$$

where $r = d(P, F)$ is the distance from the point P to the focal point F, and $\ell = d(P, D)$ is the distance from P to the point D on the directrix that is closest to the point P.

Figure 1.69 shows a parabola opening upward with focal length $f = 1$ centred at the origin. The parabola is the set of points that are equidistant from the focal point $F = (0, 1)$ and the directrix line located at $y = -1$.

Algebraic description

The shape of a parabola with focal length f opening upward corresponds to the graph of the quadratic functions $f(x) = \frac{1}{4f}x^2$. This is a special case of the general formula for quadratic functions $f(x) = a(x-h)^2 + k$, which you're already familiar with from Section 1.13 (see page 99). The parabola shown in Figure 1.69 is centred at the origin, so the displacement parameters h and k are both zero. The coefficient a in the general formula is related to the focal length f through the relation $a = \frac{1}{4f}$, so in the case of focal length $f = 1$ the coefficient is $a = \frac{1}{4}$. See Figure 1.69.

The formula $y = \frac{1}{4f}x^2$ is specific to the case of a parabola opening upward, but similar algebraic expressions exist for parabolas opening downward and sideways. The parabola with focal length f opening downward is described by the equation $y = -\frac{1}{4f}x^2$. The parabola opening to the left and to the right are described by relations $x = -\frac{1}{4f}y^2$ and $x = \frac{1}{4f}y^2$. With your knowledge of the ~~most sunlight.~~ displacement parameters h and k used for general quadratic equations (see page 99), you can also obtain algebraic expressions for parabolas that are not centred at the origin.

Polar coordinates

In the previous section we connected the geometric definition of parabolas with quadratic algebraic expressions. When learning math, it's important to note connections of this sort because they are the bridges between different mathematical domains. If one day you have to solve a geometry problem involving parabolas, you could use algebraic equations to describe the parabolas and solve the problem using algebra. If on another day you encounter an algebra problem involving a quadratic equation, you could visualize the quadratic equation as a parabolic shape and solve the problem using geometric reasoning. Being able to travel between math domains like this is a mark of true math fluency.

In the spirit of further bridge-building, I want to show you the equation of a parabola in polar coordinates. We choose a coordinate system centred at the focal point F. The polar-coordinates equation for the parabola with focal length f opening to the left is

$$r(\theta) = \frac{2f}{1+\cos\theta}.$$

Figure 1.70 shows a particular instance of this formula when the parabola has focal length $f = 1$. Try substituting the values $\theta = 0$ and $\theta = 90^\circ$ ($\frac{\pi}{2}$ radians) in the polar equation to verify that it correctly describes the points on the parabola.

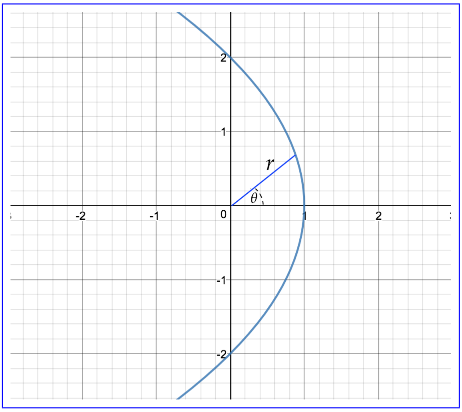


Figure 1.70: The parabola described by $r(\theta) = \frac{2}{1+\cos\theta}$ in polar coordinates.

The key point I want you to take away from this section is that algebraic formulas can be very useful for describing geometric shapes. The parabola illustrated in Figure 1.70 can be described in three equivalent ways: geometrically through its focal length $f = 1$ and directrix line $x = 2$; algebraically as the relation $x = 1 - \frac{1}{4}y^2$ in Cartesian coordinates; or as the function $r(\theta) = \frac{2}{1+\cos\theta}$ in polar coordinates.

Parabola applications

Parabolic shapes are of special importance in optics and communications. Using parabolic lenses, mirrors, and antennas, it's possible to focus the energy emitted from a distant object into a single point. This is due to the *reflective property* of parabolas, which states that all light rays coming from far away are redirected toward the focal point of the parabolic shape. The reflective property makes parabolas useful for many practical communication applications.

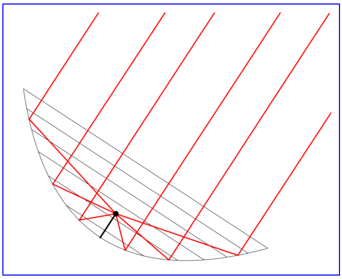


Figure 1.71: The reflective property of parabolas tells us all radio waves coming from infinity are reflected toward the focal point of the parabola.

Figure 1.71 illustrates the setup for a radio communication scenario in which a ground station is trying to detect a signal coming from a satellite in orbit. The satellite is very far away so the signal received on Earth is very weak. A parabolic satellite dish antenna collects the signal from a large surface area and focuses all of it on the focal point of the parabola. A radio receiver placed at the focal point of the parabola receives a much stronger signal, since the focal point is where the power from the whole dish surface is concentrated. This is thanks to the reflective property of the parabolic shape: all radio waves coming from the far-away satellite get reflected toward the focal point of the parabola.

Exercises

E1.34 Consider some arbitrary point $P = (x, y)$ that lies on the parabola with focal length f centred at the origin as illustrated in Figure 1.69. Use the geometric definition of the parabola $r = \ell$ to obtain a relation between the x- and y-coordinates of the point P.

Hint: The distance between points $A = (A_x, A_y)$ and $B = (B_x, B_y)$ is given by $d(A, B) = \sqrt{(A_x - B_x)^2 + (A_y - B_y)^2}$.

Hint: Recall the definitions of $r = d(P, F)$ and $\ell = d(P, D)$.

Links

[Interactive graph of a parabola]
https://www.desmos.com/calculator/4ddfrv7wvx

[Further reading about ~~Earth-Sun geometry~~ parabolas on Wikipedia]
~~http~~https://~~www~~en.~~physicalgeography~~wikipedia.~~net~~org/~~fundamentals~~wik

1.20 Hyperbola

The *hyperbola* is another fundamental shape of nature. ~~A horizontal hyperbola is~~

Parameters

- F_1, F_2: the *focal points* of the hyperbola
- r_1: the distance from a point of the hyperbola to F_1
- r_2: the distance from a point of the hyperbola to F_2

- a: the semi-major axis of the hyperbola is the distance from the origin to the vertices V_1 and V_2
- b: the semi-minor axis of the hyperbola is the distance from a focus to the nearest asymptote
- c: the distance of the focal points from the centre. The distance between F_1 and F_2 is $2c$.
- ε: *eccentricity* of the hyperbola, $\varepsilon = \sqrt{1 + \frac{b^2}{a^2}} = \frac{c}{a}$

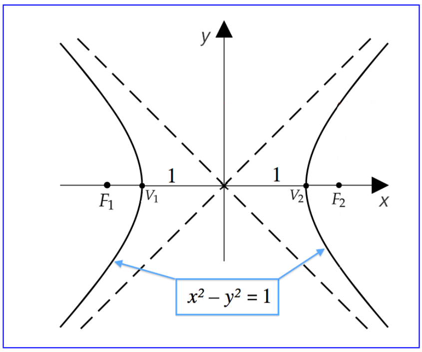


Figure 1.72: The graph of the unit hyperbola $x^2 - y^2 = 1$. The graph has two branches opening to the sides, and its eccentricity is $\varepsilon = \sqrt{1 + \frac{1}{1}} = \sqrt{2}$.

The graph of a hyperbola consists of two separate *branches*, as illustrated in Figure 1.72. The dashed lines are called the *asymptotes* of the hyperbola. The graph of the hyperbola approaches these lines but never touches them. The equations that describe these asymptotes are $y = \frac{b}{a}x$ and $y = -\frac{b}{a}x$.

Definition

A hyperbola is defined as the set of points such that the absolute value of the difference of the distances to the two focal points is constant:

$$|r_1 - r_2| = \text{const.}$$

Another way to define a hyperbola is as the set of points (x, y) ~~which~~ that satisfy the equation

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

The ~~numbers a~~ coordinates of the two focal points of this hyperbola are

$$F_1 = (-c, 0) \qquad \text{and} \qquad F_2 = (c, 0),$$

where the *focal distance* is $c = \sqrt{a^2 + b^2}$. The coordinates of the *vertices* V_1 and ~~b are arbitrary constants. This hyperbola passes through the points~~ V_2 are $(-a, 0)$ and $(a, 0)$. ~~The eccentricity of this hyperbolais defined as~~

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

~~Eccentricity is an important parameter of the hyperbola, as it~~

The hyperbola's *eccentricity* is defined by the equation

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

The eccentricity is a number greater than 1 that determines the hyperbola's shape. Recall ~~the~~ that an 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~~Read on to find out.

~~Graph~~

~~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 separate *branches*, as illustrated in Figure 1.72. We'll focus our discussion mostly on the right branch of the hyperbola.~~

Hyperbolic trigonometry

~~The trigonometric functions sin and cos describe the~~

The study of the geometry of the ~~unit circle~~points on the unit circle is called *circular trigonometry*. The geometry of the unit circle is described by the trigonometric functions $\sin\theta$ and $\cos\theta$. The function $\cos\theta$ defines the x-coordinates of the points on the unit circle, and $\sin\theta$ defines their y-coordinates. 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~~

Similarly, the study of the geometry of the ~~points on the unit circle is called~~ unit hyperbola is called ~~*circular*~~ *hyperbolic trigonometry*.

~~Instead of looking at a~~ Doesn't that sound awesome? Next time your friends ask what you have been up to, tell them you are learning about hyperbolic trigonometry. Whereas we trace the path of the point P on the unit circle $x^2 + y^2 = 1$, ~~let's trace out~~ we'll instead trace the path of a point Q on the unit hyperbola $x^2 - y^2 = 1$. We~~will~~'ll 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 right branch of 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 ~~$\mu \in [0, \infty)$~~ μ actually measures an area. The area of the highlighted region in Figure 1.73 corresponds to $\frac{1}{2}\mu$.

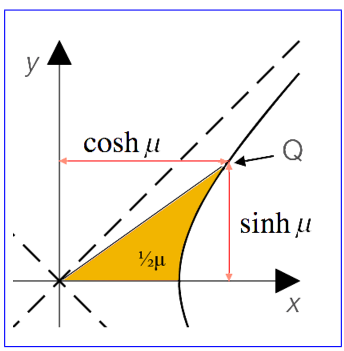


Figure 1.73: The functions $\cosh\mu$ and $\sinh\mu$ are defined as the x- and y-coordinates of a point moving on the unit hyperbola $x^2 - y^2 = 1$.

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

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.~~ , as illustrated in Figure 1.75.

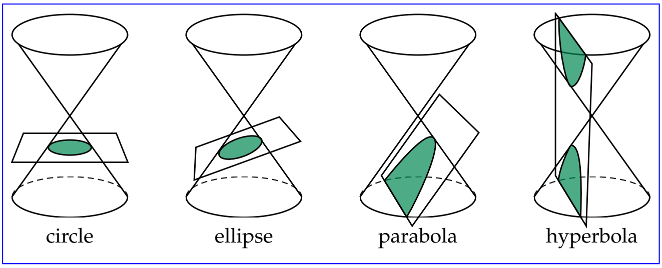


Figure 1.75: Taking slices through a cone at different angles produces different geometric shapes: a circle, an ellipse, a parabola, or a hyperbola.

~~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~~ All four conic sections can be described by the same ~~equation,~~ function in polar coordinates:

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

where q is the curve's closest distance to a focal point ~~.~~ and ε is the curve's eccentricity. For a circle~~$q = a$, for an ellipse$q = a(1-\varepsilon)$,~~ , $q = R$ (the radius) and the eccentricity parameter is $\varepsilon = 0$. For an ellipse, $q = a(1-\varepsilon)$ and the eccentricity parameter varies between 0 and ~~for a hyperbola $q = a(\varepsilon - 1)$. In the context of~~ 1 $(0 \leqslant \varepsilon < 1)$. Note we include the case $\varepsilon = 0$ since a circle is a special case of an ellipse. For a parabola, ~~the length q is sometimes referred to as~~ $q = f$ (the focal length~~and denoted f~~) and the eccentricity is $\varepsilon = 1$. For a hyperbola, $q = a(\varepsilon - 1)$ and the eccentricity is $\varepsilon > 1$.

We can use the eccentricity parameter ε to classify all four curves. Depending on the ~~parameter~~ value of ε, the equation $r(\theta)$ defines either a circle, an ellipse, a parabola, or a hyperbola. Table 1.3 summarizes all our observations regarding conic sections.

The motion of the planets is explained by Newton's law of gravitation. The gravitational interaction between two bodies ~~is always~~

Conic section	Equation	Polar ~~equation~~ function	Eccentr
Circle	$\cancel{x^2 + y^2 = a^2}$ $x^2 + y^2 = R^2$	$\cancel{r(\theta) = a}$ $r(\theta) = R$	$\varepsilon = 0$
Ellipse	$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$	$r(\theta) = \frac{a(1-\varepsilon^2)}{1+\varepsilon\cos(\theta)}$	$\cancel{\varepsilon = \sqrt{1}}$
Parabola	$\cancel{y^2 = 4qx}$ $y^2 = 4fx$	$\cancel{r(\theta) = \frac{2q}{1+\cos(\theta)}}$ $r(\theta) = \frac{2f}{\cancel{1+\cos(\theta)}}$	$\varepsilon = 1$
Hyperbola	$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$	$r(\theta) = \frac{a(\varepsilon^2-1)}{1+\varepsilon\cos(\theta)}$	$\cancel{\varepsilon = \sqrt{1}}$

Table 1.3: The four conic sections and their eccentricity parameters.

always leads one of the two bodies to follow a trajectory described by one of the ~~four conic sections~~ conic sections for which the other body is the focal point. Figure 1.76 illustrates four different trajectories for a satellite near planet F. The circle ($\varepsilon = 0$) and the ellipse ($\cancel{\varepsilon \in [0, 1)}$ $0 \leqslant \varepsilon < 1$) 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.

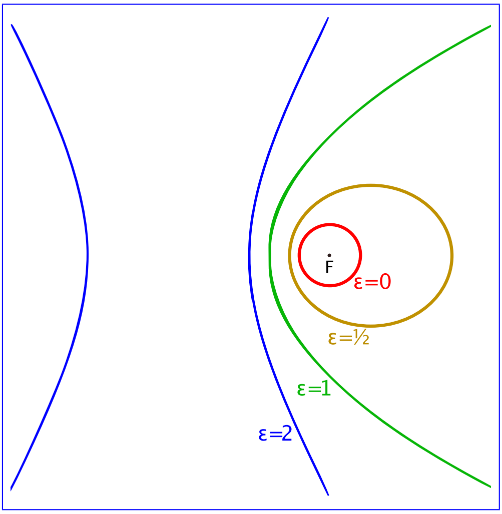


Figure 1.76: Four different trajectories for a satellite moving near a planet.

Links

[Interactive graph of a hyperbola]
https://www.desmos.com/calculator/2mnsk5o8vn

[Lots of information about ~~ellipses and orbits~~ conic sections on Wikipedia]
https://en.wikipedia.org/wiki/Conic_section
http://en.wikipedia.org/wiki/Eccentricity_(mathematics)

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

* * *

I'd love to continue this geometric digression and tell you more about the properties and applications of conic sections, but there are more pressing math topics to discuss! In the next section we'll learn how to solve systems of equations with multiple unknown variables. After that, we'll learn how interest calculations work, and finally we'll conclude the chapter by introducing terminology and notation for describing mathematical sets.

1.21 Solving systems of linear equations

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~~ *two* equations and must solve for ~~two~~ *two* unknowns? For example,

$$x + 2y = 5,$$
$$3x + 9y = 21.$$

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

  $$a_1x + b_1y = c_1,$$
  $$a_2x + b_2y = c_2,$$

  where as, bs, and cs are given constants.

Principles

If you have n equations and n unknowns, you can solve the equations ~~simultaneously~~ *simultaneously* and find the values of the unknowns. There are several different approaches for solving equations simultaneously. We'll ~~learn about~~ show three of these approaches ~~in this section~~for the case $n = 2$.

Solution techniques

When solving for two unknowns in two equations, the best approach is to *eliminate* one of the variables from the equations. By combining the two equations appropriately, we can simplify the problem to the problem of finding one unknown in one equation.

Solving by substitution

We want to solve the following system of equations:

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

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

$$3y = 6.$$

We find $y = 2$, but what is x? Easy. To solve for x, plug the value $y = 2$ into any of the equations we started from. Using the equation $x = 5 - 2y$, we find $x = 5 - 2(2) = 1$.

Solving by subtraction

Let's ~~return to our set of equationsto see another approach for solving~~now look at another way to solve the same system of equations:

$$x + 2y = 5,$$
$$3x + 9y = 21.$$

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:

$$3x + 6y = 15,$$
$$3x + 9y = 21.$$

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:

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

The $3x$ terms cancel. This subtraction eliminates the variable x because we multiplied the first equation by 3. We find $y = 2$. To find x, substitute $y = 2$ into one of the original equations:

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

from which we deduce that $x = 1$.

Solving by equating

There is a third way to solve the system of equations

$$x + 2y = 5,$$
$$3x + 9y = 21.$$

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

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

Though the variable x is unknown to us, we know two facts about it: x is equal to $5 - 2y$ and x is equal to $7 - 3y$. Therefore, we can eliminate x by equating the right-hand sides of the equations:

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

We solve for y by adding $3y$ to both sides and subtracting 5 from both sides. We find $y = 2$ then plug this value into the equation $x = 5 - 2y$ to find x. The solutions are $x = 1$ and $y = 2$.

Discussion

The ~~three elimination techniques presented here~~ repeated use of the three algebraic techniques presented in this section will allow you to solve any system of n linear equations in n unknowns. Each time you ~~perform~~ eliminate one variable using a substitution, a subtraction, or an elimination by equating, you're simplifying the problem to a problem of finding $(n-1)$ unknowns in a system of $(n-1)$ equations. There is ~~actually an entire~~ an entire math course called linear algebra, in which you'll develop a ~~more advanced,~~ systematic approach for solving systems of linear equations.

Geometric solution

Solving a system of two linear equations in two unknowns can be understood geometrically as finding the point of intersection between two lines in the Cartesian plane. In this section we'll explore this correspondence between algebra and geometry to develop yet another way of solving systems of linear equations.

The algebraic equation $ax + by = c$ containing the unknowns x and y can be interpreted as a *constraint* equation on the set of possible values for the variables x and y. We can visualize this constraint geometrically by considering the coordinate pairs (x, y) that lie in the Cartesian plane. Recall that every point in the Cartesian plane can be represented as a coordinate pair (x, y), where x and y are the coordinates of the point.

Figure 1.77 shows the geometrical representation of three equations. The line ℓ_a corresponds to the set of points (x, y) that satisfy the equation $x = 1$, the line ℓ_b is the set of points (x, y) that satisfy the equation $y = 2$, and the line ℓ_c corresponds to the set of points that satisfy $x + 2y = 2$.

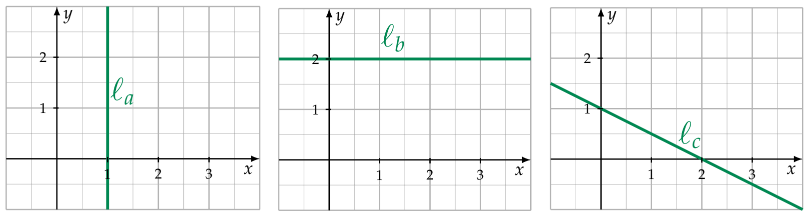


Figure 1.77: Graphical representations of three linear equations.

You can convince yourself that the geometric lines shown in Figure 1.77 are equivalent to the algebraic equations by considering individual points (x, y) in the plane. For example, the points $(1, 0)$, $(1, 1)$, and $(1, 2)$ are all part of the line ℓ_a since they satisfy the equation $x = 1$. For the line ℓ_c, you can verify that the line's x-intercept $(2, 0)$ and its y-intercept $(0, 1)$ both satisfy the equation $x + 2y = 2$.

The Cartesian plane as a whole corresponds to the set $\mathbb{R}^2$, which describes all possible pairs of coordinates. To understand the equivalence between the algebraic equation $ax + by = c$ and the line ℓ in the Cartesian plane, we can use the following precise math notation:

$$\ell : \{(x, y) \in \mathbb{R}^2 \mid ax + by = c\}.$$

In words, this means that the line ℓ is defined as the subset of the pairs of real numbers (x, y) that satisfy the equation $ax + by = c$. Figure 1.78 shows the graphical representation of the line ℓ.

You don't have to take my word for it, though! Think about it and convince yourself that all points on the line ℓ shown in Figure 1.78 satisfy the equation $ax + by = c$. For example, you can check that the x-intercept $(\frac{c}{a}, 0)$ and the y-intercept $(0, \frac{c}{b})$ satisfy the equation $ax + by = c$.

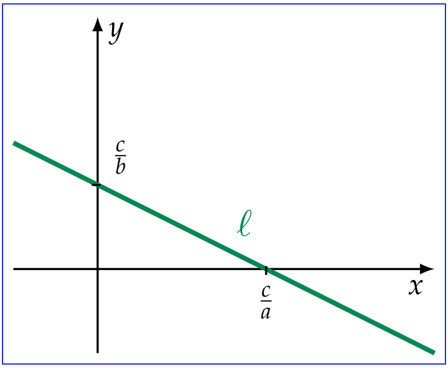


Figure 1.78: Graphical representation of the equation $ax + by = c$.

Solving the system of two equations

$$a_1x + b_1y = c_1,$$

$$a_2x + b_2y = c_2,$$

corresponds to finding the intersection of the lines ℓ_1 and ℓ_2 that represent each equation. The pair (x, y) that satisfies both algebraic equations simultaneously is equivalent to the point (x, y) that is the intersection of lines ℓ_1 and ℓ_2, as illustrated in Figure 1.79.

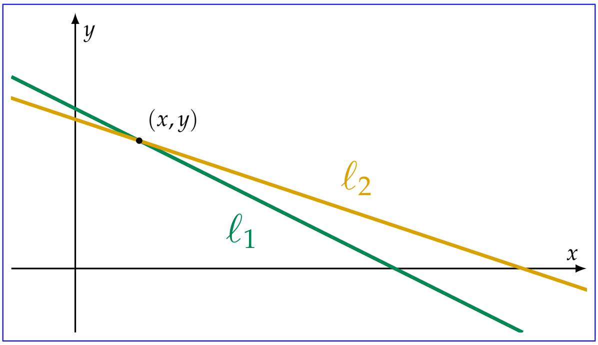


Figure 1.79: The point (x, y) that lies at the intersection of lines ℓ_1 and ℓ_2.

Example Let's see how we can use the geometric interpretation to solve the system of equations

$$x + 2y = 5,$$

$$3x + 9y = 21.$$

We've already seen three different *algebraic* techniques for finding the solution to this system of equations; now let's see a *geometric* approach for finding the solution. I'm not kidding you, we're going to solve the exact same system of equations a fourth time!

The first step is to draw the lines that correspond to each of the equations using pen and paper or a graphing calculator. The second step is to find the coordinates of the point where the two lines intersect as shown in Figure 1.80. The point $(1,2)$ that lies on both lines ℓ_1 and ℓ_2 corresponds to the x and y values that satisfy both equations simultaneously.

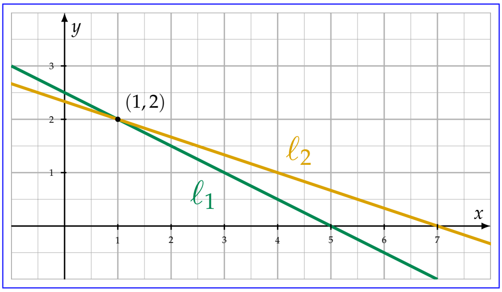


Figure 1.80: The line ℓ_1 with equations $x + 2y = 5$ intersects the line ℓ_2 with equation $3x + 9y = 21$ at the point $(1,2)$.

Visit the webpage at `www.desmos.com/calculator/exikik615f` to play with an interactive version of the graphs shown in Figure 1.80. Try changing the equations and see how the graphs change.

Exercises

E1.35 Plot the lines ℓ_a, ℓ_b, and ℓ_c shown in Figure 1.77 (page 149) using the Desmos graphing calculator. Use the graphical representation of these lines to find: **a)** the intersection of lines ℓ_c and ℓ_a, **b)** the intersection of ℓ_a and ℓ_b, and **c)** the intersection of lines ℓ_b and ℓ_c.

E1.36 Solve the system of equations simultaneously for x and y:

$$\begin{aligned} 2x + 4y &= 16, \\ 5x \; - \; y &= 7. \end{aligned}$$

E1.37 Solve the system of equations for the unknowns x, y, and z:

$$\begin{aligned} 2x + y - 4z &= 28, \\ x \; + y \; + \; z &= 8, \\ 2x - y - 6z &= 22. \end{aligned}$$

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~~ scenario 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~~ is 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 effective APR of 6.183%. After six years you will owe

$$L_6 = \exp\!\left(\frac{6}{100}\right)^6 \times 1000 = \$1433.33.$$

The nominal APR is 6% in each case, yet, the more frequent the compounding schedule, the more money you'll owe after six years.

Exercises

E1.39 Studious Jack borrowed \$40 000 to complete his university studies and made no payments since graduation. Calculate how much money he owes after 10 years in each of the scenarios.

a) Nominal annual interest rate of 3% compounded monthly

b) Effective annual interest rate of 4%

c) Nominal annual interest rate of 5% with infinite compounding

E1.40 Entrepreneurial Kate borrowed \$20 000 to start a business. Initially her loan had an effective annual percentage rate of 6%, but after five years she negotiated with the bank to obtain a lower rate of 4%. How much money does she owe after 10 years?

1.23 Set notation

A *set* is the mathematically precise notion for describing a group of objects. You don't need to know about sets to perform simple math; but more advanced topics require an understanding of what sets are and how to denote set membership, set operations, and set containment relations. This section introduces all the relevant concepts.

Definitions

- *set*: a collection of mathematical objects
- S, T: the usual variable names for sets
- $s \in S$: this statement is read "s is an element of S" or "s is in S"
- $\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}$: some important number sets: the naturals, the integers, the rationals, and the real numbers, respectively.
- $\varnothing$: the *empty set* is a set that contains no elements
- ~~{ definition }~~{ ..... }: the curly brackets ~~surround the definition of a set~~are used to define sets, and the expression inside the curly brackets describes ~~what the set contains~~the set contents.

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 that are in both S and T.
- $S \setminus T$: *set difference* or *set minus*. The set difference $S \setminus T$ corresponds to the elements of S that are not in T.

Set relations:

- $\subset$: is a strict subset of
- $\subseteq$: is a subset of or equal to

Here is a list of special mathematical shorthand symbols and their corresponding meanings:

- $\in$: element of
- $\notin$: not an element of
- $\forall$: for all
- $\exists$: there exists
- $\nexists$: there doesn't exist
- $|$: such that

These symbols are used in math proofs because they allow us to express complex mathematical arguments succinctly and precisely.

An *interval* is a subset of the real line. We denote an interval by specifying its endpoints and surrounding them with either square brackets "[" or round brackets "(" to indicate whether or not the corresponding endpoint is included in the interval.
Sometimes we encounter intervals that consist of two disjointed parts. We use the notation $[a, b] \cup [c, d]$ to denote the union of the two intervals, which is the set of numbers *either* between a and b (inclusive) *or* between c and d (inclusive).

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.

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 "$\mathbb{R}_+$ is a subset of $\mathbb{R}$") to be the set of non-negative real numbers: or expressed more compactly, 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."

Note we used the ~~symbol "≡~~"is defined as" symbol "$\stackrel{\text{def}}{=}$" instead of the basic "=" to give ~~you~~ an extra hint that we're defining a new variable $\mathbb{R}_+$ that is equal to the set expression on the right. In this book, we'll sometimes use the symbol "~~≡" whenever we define~~ $\stackrel{\text{def}}{=}$" when defining new variables and math quantities. Some other books use the notation ":=" or "≡" for this purpose. The meaning of "~~≡~~$\stackrel{\text{def}}{=}$" is identical to "=" but it tells ~~you~~ us the variable on the left of the equality is new.

Example 2: Even and odd integers

In both of the above examples, we use the ~~mathematical~~ *set-builder* notation $\{\ldots \mid \ldots\}$ to define the sets. Inside the curly braces we first describe the general kind of mathematical 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 the fundamental number sets we defined in Section 1.2 in the beginning of the book. 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:

$$\mathbb{N}\cancel{\equiv} \stackrel{\text{def}}{=} \{0, 1, 2, 3, 4, 5, 6, \ldots\}.$$

We use the notation $\mathbb{N}^*$ to denote the set of *positive natural numbers*. The set $\mathbb{N}^*$ is the same as $\mathbb{N}$ but excludes zero.

The integers are the numbers derived by adding or subtracting 1 some number of times:

$$\mathbb{Z}\cancel{\equiv} \stackrel{\text{def}}{=} \{x \mid x = \pm n, n \in \mathbb{N}\}.$$

~~When~~ If we allow for divisions between integers, we ~~get the rational numbers:~~ require the set of rational numbers to represent the results:

$$\mathbb{Q}\cancel{\equiv} \stackrel{\text{def}}{=} \left\{ \cancel{z}\ \cancel{z = \frac{x}{y} \text{ where } x \text{ and } y \text{ are in}}\ \frac{m}{n} \,\middle|\, m \in \mathbb{Z},\ \cancel{\text{and } y \neq 0}\ n \in \mathbb{N}^* \right\}\cancel{.},$$

In words, this expression is telling us that every rational number can be written as a fraction $\frac{m}{n}$, where m is an integer ($m \in \mathbb{Z}$), and n is a positive natural number ($n \in \mathbb{N}^*$).

The broader class of real numbers also includes all rationals as well as irrational numbers like $\sqrt{2}$ and π:

$$\mathbb{R}\cancel{\equiv} \stackrel{\text{def}}{=} \{\pi, e, -1.53929411\ldots, 4.99401940129401\ldots, \ldots\}.$$

Finally, we have the set of complex numbers:

$$\mathbb{C}\cancel{\equiv} \stackrel{\text{def}}{=} \{1, i, 1+i, 2+3i, \ldots\},$$

where $\cancel{i \equiv \sqrt{-1}}$ $i \stackrel{\text{def}}{=} \sqrt{-1}$ is the unit imaginary number.

Note that the definitions of $\mathbb{R}$ and $\mathbb{C}$ are not very precise. Rather than give a precise definition of each set inside the curly braces 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. And every real number is also a complex number. See Figure 1.2 (page 8) for an illustration of the subset relationship between the number sets.

CLARIFICATIONS about fractions and rationals nums. imported from French transl.

Rational numbers and fractions

So far in this book, we've used the notions of "fraction" and "rational number" somewhat interchangeably. Now that we've learned about sets, we can clarify the differences and equivalencies between these related concepts.

The same rational number $\frac{2}{3}$ can be written as a fraction in multiple, equivalent ways. The fractions $\frac{2}{3}$, $\frac{4}{6}$, $\frac{6}{9}$, $\frac{8}{12}$, and $\frac{2k}{3k}$ all correspond to the same rational number. Keep in mind the existence of these *equivalent fractions* whenever checking whether two rational numbers are equal. For example, one person could obtain the answer $\frac{2}{3}$ to a given problem, while another person obtains the answer $\frac{4}{6}$. Since the two fractions look different, we might think these are different answers, when in fact both answers correspond to the same rational number.

A *reduced fraction* is a fraction of the form $\frac{m}{n}$ such that the numbers m and n are the smallest possible. We can obtain the reduced fraction by getting rid of any common factors that appear both in the numerator and denominator. For example,

$$\frac{4}{6} = \frac{2\cdot 2}{3\cdot 2} = \frac{2\cdot \cancel{2}}{3\cdot \cancel{2}} = \frac{2}{3},$$

where we cancelled the common factor 2 to obtain the equivalent reduced fraction. Reduced fractions are a useful representation for the set of rational numbers, because each rational number corresponds to a unique reduced fraction. Two rational numbers are equal if and only if they correspond to the same reduced fraction.

Subsets of the real line

Recall that the real numbers $\mathbb{R}$ have a graphical representation as points on the number line. See Figure 1.13 on page 24 for a reminder. The number line is also useful for representing various subsets of the real numbers, which we call *intervals*. We can graphically represent an interval 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

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

or graphically as in Figure ~~??~~1.81.

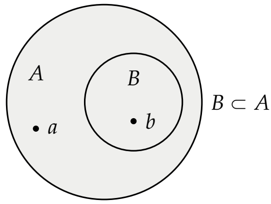


Figure 1.83: Venn diagram showing an example of the set relation $B \subset A$. The set B is strictly contained in the set A.

~~is an abstract mathematical notion, but the~~ The picture helps us visualize ~~the situation~~this abstract mathematical notion.

Mathematicians use two different symbols to describe set containment, in order to specify either a *strict* containment relation or a *subset-or-equal* relation. The two types of containment relations between sets are similar to the *less-than* ($<$) and *less-than-or-equal* ($\leqslant$) relations between numbers. A strict containment relation is denoted by the symbol $\subset$. We write $B \subset A$ if and only if every element of B is also an element of A, and there exists at least one element of A that is not an element of B. Using set notation, the previous sentence is expressed as

$$B \subset A \qquad \Leftrightarrow \qquad \forall b \in B, b \in A \text{ and } \exists a \in A \text{ such that } a \notin B.$$

For example, the expression $E \subset \mathbb{Z}$ shows that the even numbers are a strict subset of the integers. Every even number is an integer, but there exist integers that are not even (the odd numbers). Some mathematicians prefer the more descriptive symbol $\subsetneq$ to describe strict containment relations.

A subset-or-equal relation is denoted $B \subseteq A$. In writing $B \subseteq A$, a mathematician claims, "Every element of B is also an element of A," but makes no claim about the existence of elements that are contained in A but not in B. The statement $B \subset A$ implies $B \subseteq A$; however, $B \subseteq A$ does not imply $B \subset A$. This is analogous to how $b < a$ implies $b \leqslant a$, but $b \leqslant a$ doesn't imply $b < a$, since a and b could be equal.

Set operations

Venn diagrams also help us visualize the subsets obtained from set operations. Figure 1.84 illustrates the set union $A \cup B$, the set intersection $A \cap B$, and the set difference $A \backslash B$, for two sets A and B.

The union $A \cup B$ describes all elements that are in either set A or set B, or both. If $e \in A \cup B$, then $e \in A$ or $e \in B$.

which is the set of elements that are in A or B but not in C.

Another example of a complicated set expression is

$$(A \cap B) \cup (B \cap C) = \{b, c, d\},$$

which describes the set of elements in both A and B or in both B and C. As you can see, set notation is a compact, precise language for writing complicated set expressions.

Example 4: Word problem

A startup is looking to hire student interns for the summer. ~~Let S denote the whole set of students looking for a summer internship.~~ Define C to be the subset of students who are good with computers, M the subset of students who know math, D the students with design skills, and L the students with good language skills.

Using set notation, we can specify different subsets of the students the startup might hire. Let's say the startup is a math textbook publisher; they want to hire students from the set $M \cap L$—the students who are good at math and who also have good language skills. A startup that builds websites needs both designers and coders, and therefore would choose students from the set $D \cup C$.

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'll look at ~~some mathematical arguments that make~~ a simple mathematical proof that makes use of the new symbols.

Simple proof example

Claim: Given $J(n) = 3n + 2 - n$, $J(n) \in E$ for all $n \in \mathbb{Z}$.

The claim is that the function $J(n)$ ~~is always~~ outputs an even number, whenever the input n is an integer. ~~This means no matter which integer number n we choose, the function $J(n) = 3n + 2 - n$ will always output an even number~~To prove this claim, we have to show that the expression $3n + 2 - n$ is even for all numbers $n \in \mathbb{Z}$.

~~**Proof:**~~

Proof. We want to show $J(n) \in E$ for all $n \in \mathbb{Z}$. Let's first review the definition of the set of even numbers ~~$E = \{m \in \mathbb{Z} \mid m = 2n, n \in \mathbb{Z}\}$~~$E \stackrel{\text{def}}{=} \{m \in \mathbb{Z} \mid$
A number is even if it is equal to $2n$ for some integer n. Next let's simplify the expression for $J(n)$ as follows:

$$J(n) = 3n + 2 - n = 2n + 2 = 2(n + 1).$$

Observe that the number $(n+1)$ is always an integer whenever n is an integer. Since the output of $J(n) = 2(n+1)$ is equal to $2m$ for some integer m, we've proven that $J(n) \in E$, for all $n \in \mathbb{Z}$. ~~□~~

~~Less simple proof example: Square root of 2 is irrational~~

~~The following is an ancient mathematical proof expressed in terms of modern math symbols.~~

~~**Claim:** $\sqrt{2} \notin \mathbb{Q}$. The claim is that $\sqrt{2}$ is not part of the set of rational numbers. Recall the definition of the set of rational numbers: $\mathbb{Q} = \{\frac{m}{n} \mid m, n \in \mathbb{Z}, n \neq 0\}$. If $\sqrt{2} \notin \mathbb{Q}$, this means no numbers $m \in \mathbb{Z}$ and $n \in \mathbb{Z}$ exist such that $m/n = \sqrt{2}$. Using mathematical notation, the previous sentence is expressed as~~

$$\nexists\, m \in \mathbb{Z}, n \in \mathbb{Z} \;\Big|\; m/n = \sqrt{2}.$$

```
This proof is not
neither intuitive
nor educational so
  > /dev/null
```

~~To prove the claim, we'll use a technique called *proof by contradiction*. We 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 (page 156) 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~~

$$\frac{m^2}{n^2} = 2 \qquad \Rightarrow \qquad m^2 = 2n^2.$$

~~If we analyze the last equation in more detail, we can conclude that m cannot be an odd number, or written "$m \notin O$" in math. 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 the factor 2 and every number containing a factor 2 is even, not odd. If m is an integer ($m \in \mathbb{Z}$) and m is not odd ($m \notin O$) then it must be that m is even ($m \in E$).~~

~~If m is even, then it contains 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 the equation $m^2 = 2n^2$ once more, this time substituting $m = 2q$ into the equation:~~

$$\underline{(2q)^2 = 2n^2 \quad \Rightarrow \quad 2q^2 = n^2.}$$

~~By a similar reasoning as before, we can conclude n cannot be odd ($n \notin O$) so n must be even ($n \in E$). We've shown that both m and n must be even numbers, which means they both contain a factor 2. However, this statement contradicts our initial assumption that m and n do not have any common factors!~~

~~The fact that we arrived at a contradiction means there must be a mistake in our reasoning. Since each step we carried out was 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. □~~ □

Sets as solutions to equations

Another context where sets come up is when describing solutions to equations and inequalities. In Section 1.1 we learned how to solve for the unknown x in equations. To solve the equation $f(x) = c$ is to find all the values of x that satisfy this equation. For simple equations like $x - 3 = 6$, the solution is a single number $x = 9$, but more complex equations can have multiple solutions. For example, the solution to the equation $x^2 = 4$ is the set $\{-2, 2\}$, since both $x = -2$ and $x = 2$ satisfy the equation.

Please update your definition of the math verb "to solve" (an equation) to include the new notion of a ~~*solution set*~~*solution set*—the set of values that satisfy the equation. A solution set is the mathematically precise way to describe an equation's solutions:

- The solution set to the equation $x - 3 = 6$ is the set $\{9\}$.
- The solution set for the equation $x^2 = 4$ is the set $\{-2, 2\}$.

- The solution set of $\sin(x) = 0$ is the set ~~$\{x \mid x = \pi n, \forall n \in \mathbb{Z}\}$~~$\{x \mid x = \pi n, \forall n$
- The solution set for the equation $\sin(x) = 2$ is $\emptyset$ (the empty set), since there is no number x that satisfies the equation.

The SymPy function solve returns the solutions of equations as a list. To solve the equation $f(x) = c$ using SymPy, we first rewrite it as expression that equals zero $f(x) - c = 0$, then call the function solve:

```
>>> solve(x-3 -6, x)       # usage: solve(expr, var)
[9]

>>> solve(x**2 -4, x)
[-2, 2]

>>> solve(sin(x), x)
[0, pi]                    # found only solutions in [0,2*pi)

>>> solve(sin(x) -2, x)
[]                         # empty list = empty set
```

~~Solution sets~~

In the next section we'll learn how the notion of a solution set is used for describing the solutions to systems of equations.

Solution sets to systems of equations

Let's revisit what we learned in Section 1.21 about the solutions to systems of linear equations, and define their solution sets more precisely. The solution set for the system of equations

$$a_1x + b_1y = c_1,$$

$$a_2x + b_2y = c_2,$$

corresponds to the intersection of two sets:

$$\underbrace{\{(x,y) \in \mathbb{R}^2 \mid a_1x + b_1y = c_1\}}_{\ell_1} \cap \underbrace{\{(x,y) \in \mathbb{R}^2 \mid a_2x + b_2y = c_2\}}_{\ell_2}.$$

Recall that the lines ℓ_1 and ℓ_2 are the geometric interpretation of these sets. Each line corresponds to a set of coordinate pairs (x, y) that satisfy the equation of the line. The solution to the system of equations is the set of points at the intersection of the two lines $\ell_1 \cap \ell_2$. Note the word *intersection* is used in two different mathematical contexts: the solution is the *set intersection* of two sets, and also the *geometric intersection* of two lines.

Let's take advantage of this correspondence between set intersections and geometric line intersections to understand the solutions to systems of equations in a little more detail. In the next three sections, we'll look at three possible cases that can occur when trying to solve a system of two linear equations in two unknowns. So far we've only discussed Case A, which occurs when the two lines intersect at a point, as in the example shown in Figure 1.85. To fully understand the possible solutions to a system of equations, we need to think about all other cases; like Case B when $\ell_1 \cap \ell_2 = \varnothing$ as in Figure 1.86, and Case C when $\ell_1 \cap \ell_2 = \ell_1 = \ell_2$ as in Figure 1.87.

Case A: One solution. When the lines ℓ_1 and ℓ_2 are non-parallel, they will intersect at a point as shown in Figure 1.85. In this case, the solution set to the system of equations contains a single point:

$$\{(x,y) \in \mathbb{R}^2 \mid x+2y=2\} \cap \{(x,y) \in \mathbb{R}^2 \mid x=1\} = \{(1,\tfrac{1}{2})\}.$$

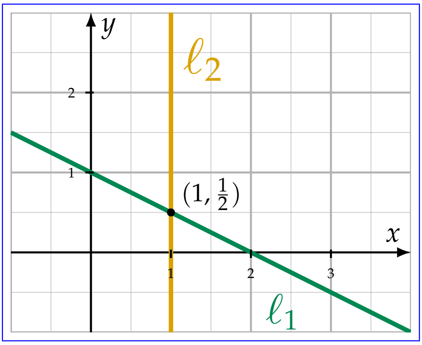


Figure 1.85: Case A: The intersection of the lines with equations $x+2y=2$ and $x=1$ is the point $(1,\frac{1}{2}) \in \mathbb{R}^2$.

Case B: No solution. If the lines ℓ_1 and ℓ_2 are parallel then they will never intersect. The intersection of these lines is the empty set:

$$\{(x,y) \in \mathbb{R}^2 \mid x+2y=2\} \cap \{(x,y) \in \mathbb{R}^2 \mid x+2y=4\} = \varnothing.$$

Think about it—there is no point (x,y) that lies on both ℓ_1 and ℓ_2. Using algebra terminology, we say this system of equations has no solution, since there are no numbers x and y that satisfy both equations.

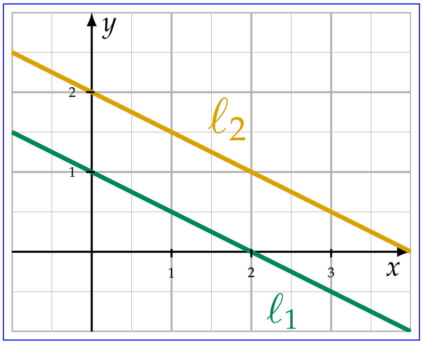


Figure 1.86: Case B: The lines with equations $x + 2y = 2$ and $x + 2y = 4$ are parallel and do not intersect. Using set notation, we can describe the solution set as $\varnothing$ (the empty set).

Case C: Infinitely many solutions. If the lines ℓ_1 and ℓ_2 are parallel and overlapping then they intersect everywhere. This case occurs when one of the equations in a system of equations is a multiple of the other equation, as in the case of equations $x + 2y = 2$ and $3x + 6y = 6$. The lines ℓ_1 and ℓ_2 that correspond to these equations are shown in Figure 1.87. Any point (x, y) that satisfies $x + 2y = 2$ also satisfies $3x + 6y = 6$. Since both equations describe the same geometric line, the intersection of the two lines is equal to the lines: $\ell_1 \cap \ell_2 = \ell_1 = \ell_2$. In this case, the solution to the system of equations is described by the set $\{(x, y) \in \mathbb{R}^2 \mid x + 2y = 2\}$.

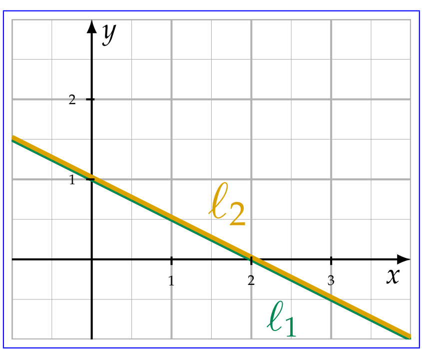


Figure 1.87: Case C: the line ℓ_1 described by equation $x + 2y = 2$ and the line ℓ_2 described by equation $3x + 6y = 6$ correspond to the same line in the Cartesian plane. The intersection of these lines is the set $\{(x, y) \in \mathbb{R}^2 \mid x + 2y = 2\} = \ell_1 = \ell_2$.

We need to consider all three cases when thinking about the solutions to systems of linear equations: the solution set can be a point (Case A), the empty set (Case B), or a line (Case C). Observe that the same mathematical notion (a set) is able to describe the solutions in all three cases even though the solutions correspond to very different geometric objects. In Case A the solution is a set that contains a single point $\{(x,y)\}$. In Case B the solution is the empty set $\emptyset$. And in Case C the solution set is described by the infinite set $\{(x,y) \in \mathbb{R}^2 \mid ax + by = c\}$, which corresponds to a line ℓ in the Cartesian plane. I hope you'll agree with me that set notation is useful for describing mathematical concepts precisely and for handling solutions to linear equations.

Sets are also useful for describing the solutions to inequalities, which is what we'll learn about next.

Inequalities

In this section, we'll learn how to solve inequalities. The solution set to an inequality is an *interval*—a subset of the number line. Consider the inequality $x^2 \leqslant 4$, which is equivalent to asking the question, "For which values of x is x^2 less than or equal to 4?" The answer to this question is the interval ~~$[-2,2] = \{x \in \mathbb{R} \mid -2 \leqslant x \leqslant 2\}$~~ $[-2,2] = \{x \in \mathbb{R} \mid -$

Working with inequalities is essentially the same as working with their endpoints. To solve the inequality $x^2 \leqslant 4$, we first solve $x^2 = 4$ to find the endpoints and then use trial and error to figure out which part of the space to the left and right of the endpoints satisfies the inequality.

It's important to distinguish the different types of inequality conditions. The four different types of inequalities are

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

Depending on the type of inequality, the answer will be either a *open* or *closed* interval.

To solve inequalities we use the techniques we learned for solving equations: we perform simplifying steps **on both sides of the inequality** until we obtain the answer. The only new aspect when dealing with inequalities is the following. When multiplying an inequality by a negative number on both sides, we must flip the direction of the inequality:

$$f(x) \leqslant g(x) \qquad \Rightarrow \qquad -f(x) \geqslant -g(x).$$

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: 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* of a function is the set of all possible outputs of the function: For example, the function $f(x) = x^2$ has the image set $\text{Im}(f) = \mathbb{R}_+$ since the outputs it produces are always nonnegative.

Discussion

Knowledge of the precise mathematical jargon introduced in this section is not crucial to understanding basic mathematics. That said, I wanted to expose you to some technical math notation here because this is the language in which mathematicians think and communicate. Most advanced math textbooks will assume you understand technical math notation, so it's good to be prepared.

Exercises

E1.41 Given the three sets $A = \{1, 2, 3, 4, 5, 6, 7\}$, $B = \{1, 3, 5\}$, and $C = \{2, 4, 6\}$, compute the following set expressions.

a) $A \backslash B$ **b)** $B \cup C$ **c)** $A \cap B$ **d)** $B \cap C$

e) $A \cup B \cup C$ **f)** $A \backslash (B \cup C)$ **g)** $(A \backslash B) \cup C$

E1.42 Find the values of x that satisfy the following inequalities.

a) $2x < 3$ **b)** $-4x \geqslant 20$ **c)** $|2x - 3| < 5$

d) $3x + 3 < 5x - 5$ **e)** $\frac{1}{2}x - 2 \geqslant \frac{1}{3}$ **f)** $(x + 1)^2 \geqslant 9$

Express your answer as an interval with appropriate endpoints.

1.24 Math problems

We've now reached the first section of problems in this book. The purpose of these problems is to give you a way to comprehensively practice your math fundamentals. ~~In the real world, you'll rarely have to solve equations by hand; however, knowing how to manipulate math expressions and solve math equations~~ Knowing how to solve math problems is a very useful skill to develop. At times, honing your math chops might seem like tough mental work, but at the end of each problem, you'll gain a stronger foothold on all the ~~subjects~~ topics you've been learning about. You'll also experience a small *achievement buzz* after each problem you vanquish.

~~I have a special message to readers who are learning math just for fun: you can either try the problems in this section or skip them. Since you have no upcoming exam on this material, you could skip ahead to Chapter 2 without any immediate consequences. However (and it's a big however), those readers who don't~~ Sit down and take a crack at these ~~problems will be missing a significant opportunity.~~

~~Sit down to do them later~~ practice problems today, or another time when you're properly caffeinated. If you ~~take the initiative to~~ make time for ~~math~~some math practice, you'll ~~find yourself developing lasting~~ develop long-lasting comprehension and true math fluency. ~~Without the practice of solving problems, however,~~

Without solving any problems, you're ~~extremely~~ likely to forget most of what you've learned in the next ~~month or two, simple as that. You 'll~~ few months. You might still remember the big ideas, but the details will be fuzzy and faded. By solving some of the practice problems, you'll remember a lot more stuff. Don't break the pace now: with math, it's very much *use it or lose it*!

~~By solving some of the problems in this section, you'll remember a lot more stuff.~~ Make sure you ~~step away from the pixels~~ put your phone away while you're ~~solving~~ working on the problems. You don't need fancy technology to do math; grab a pen and some paper from the printer and you'll be fine. ~~Do yourself a favour: put your phone in airplane-mode, close the lid of your laptop, and move away from desktop computers. Give yourself some time to think. Yes, I know you can look up the answer to any question in five seconds on the internet, and you can use to solve any math problem, but that is like outsourcing the thinking. Mathematicians~~ The great mathematicians like Descartes, Hilbert, Leibniz, and Noether did most of their work with pen and paper and they did well. Spend some time with math the way they did.

P1.1 Solve for x in the equation $x^2 - 9 = 7$.

P1.2 Solve for x in the equation $\cos^{-1}\left(\frac{x}{A}\right) - \phi = \omega t$.

P1.3 Solve for x in the equation $\frac{1}{x} = \frac{1}{a} + \frac{1}{b}$.

P1.4 Use a calculator to find the values of the following expressions:

P1.5 Compute the following expressions involving fractions:

a) $\dfrac{1}{2} + \dfrac{1}{4}$ **b)** $\dfrac{4}{7} - \dfrac{23}{5}$ **c)** $1\frac{3}{4} + 1\frac{31}{32}$

P1.6 Use the basic rules of algebra to simplify the following expressions:

a) $ab\,\dfrac{1}{a}\,b^2cb^{-3}$ **b)** $\dfrac{abc}{bca}$

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 terms 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

function $\text{tr}(t)$. Using mathematical symbols, we can represent this relationship as

$$\text{tr}(t) = \frac{d}{dt}\{\text{ba}(t)\}.$$

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 $\text{tr}(t)$, and you want to compute the final account balance at the end of the month. Since $\text{tr}(t)$ is the derivative of $\text{ba}(t)$, you can use an integral (the inverse operation of the derivative) to obtain $\text{ba}(t)$. Knowing the balance of your account at the beginning of the month, you can predict the balance at the end of the month by using the following integral calculation:

$$\text{ba}(30) = \text{ba}(0) + \int_0^{30} \text{tr}(t)\, dt.$$

This calculation makes sense since $\text{tr}(t)$ represents the instantaneous changes in $\text{ba}(t)$. If you want to find the overall change in the account balance from day 0 until day 30, you can compute the total of all the transactions on the account.

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 the equations of motion in the first place? Now that you know Newton's mathematical techniques (calculus), you can learn to derive the equations of motion by yourself.

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
- $\cancel{x_i \equiv x(0), v_i \equiv v(0)}$$x_i \stackrel{\text{def}}{=} x(0), v_i \stackrel{\text{def}}{=} v(0)$: the initial conditions

first use of the new notation \eqdef

Hint: Differentiate the function with respect to t.

P2.7 You're on a mission to Jupiter where you design an experiment to measure the planet's gravitational acceleration. In the experiment, you let go of a ball from a height of 4[m] and watch it fall to the ground. When the ball hits the ground, its speed is 14[m/s].

1. What is the gravitational acceleration on Jupiter?
2. Find the position of the ball as a function of time.

Hint: Use the fourth equation of motion.

P2.8 You're pulling a 5[kg] cart in a straight path. The position of the cart as a function of time is $x(t) = 6t^2 + 2t + 1$[m].

1. Find the velocity and acceleration of the cart as functions of time.
2. Calculate the force you're using to pull the cart.

Hint: Take the derivative of the position with respect to time. Use Newton's 2nd law $F = ma$.

P2.9 A remote controlled car has a mass of 0.5[kg]. The electric engine pushes the car with a force of 1.0[N] starting from rest at point A.

1. Find the acceleration, velocity, and position of the car as functions of time, assuming $x = 0$ at point A.
2. Calculate the velocity of the car at $t = 4$[s].
3. What is the car's velocity when it is 9[m] away from point A?

Hint: Use Newton's 2nd law and integration.

P2.10 Below is an acceleration-vs-time graph of a particle. At $t = 0$[s], the particle starts moving from rest at $x = 0$[m]. The particle's acceleration from $t = 0$[s] to ~~$t = 3$~~$t = 2$[s] is given by $a(t) = 3t$[m/s^2]. After $t = 2$[s], the acceleration is constant $a = 6$[m/s^2].

1. Find the velocity $v(2)$ and position $x(2)$ of the particle at $t = 2$[s].
2. Construct the functions of time that describe the acceleration, the velocity, and the position of the particle after $t = 2$[s].
3. How much time is needed for the particle to reach $x = 49$[m]?
4. At what distance from the origin will the particle's velocity reach 12[m/s]?

FIX

Chapter 3

Vectors

In this chapter we'll 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~~forces, velocities, and accelerations.

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 geometric objects, represented by arrows in space. For instance, 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).

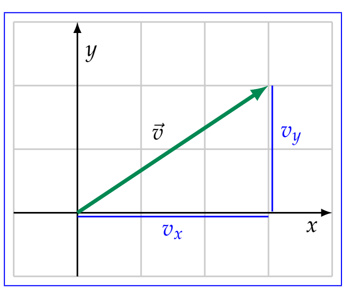


Figure 3.1: The vector $\vec{v} = (3,2)$ ~~can be represented as~~ is an arrow in the Cartesian plane. The horizontal component of $\vec{v}$ is $v_x = 3$ ~~. The~~ and the vertical component ~~of $\vec{v}$~~ is $v_y = 2$.

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 current problems in computer graphics, probability theory, machine learning, and other fields of science and mathematics. It's all about vectors these days, so you'd best get to know them.

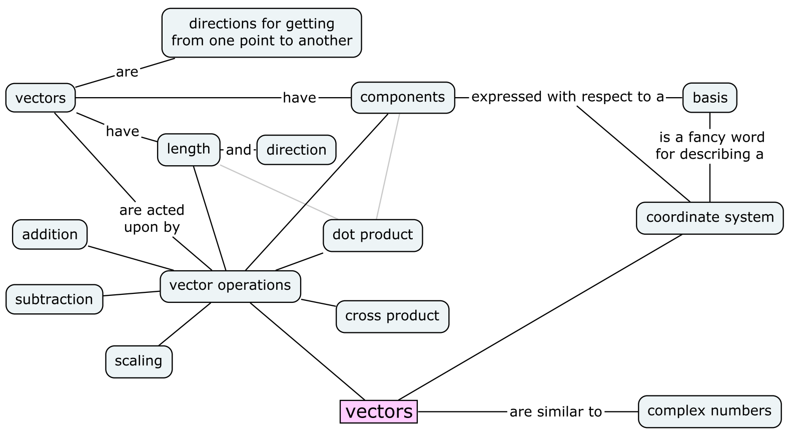


Figure 3.2: This figure illustrates the new concepts related to vectors. As you can see, there is quite a bit of new vocabulary to learn, but don't be fazed—all these terms are just fancy ways of talking about arrows.

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,~~ stop, which is 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)~~ 90° to the right so you're facing east, and walk another 3 km in that direction. You arrive at X as promised by your friend.

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 because you like the vegetation that grows there. 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 number addition: $a + b = b + a$. Maybe an analogous property holds for vectors? Indeed, it does:

$$\vec{d} = 2\text{km}\,\hat{N} + 3\text{km}\,\hat{E} = 3\text{km}\,\hat{E} + 2\text{km}\,\hat{N}.$$

The displacements in the $\hat{N}$ ~~directions and the~~ and $\hat{E}$ 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 just 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 $\hat{N} \equiv \hat{y}$~~$\hat{E} = \hat{x}$ and $\hat{N} = \hat{y}$—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*,

$$(v_x, v_y)\cancel{\equiv}= v_x\,\hat{x} + v_y\,\hat{y}.$$

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 ~~20m/s~~200 kilometres per hour: we must also specify the direction in which the ball is moving. Both of the two velocities

$$\vec{v}_1 = (\cancel{20}200, 0) \qquad \text{and} \qquad \vec{v}_2 = (0, \cancel{20}200)$$

bcs km/h more intuitive than m/s

describe motion at the speed of ~~20m/s~~200 kilometres per hour; 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* its direction. The direction makes a big difference. If it turns out the tennis ball is hurtling toward you, you'd better get out of the way!

~~This section's main idea~~ The main idea in this chapter 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 , weneed to think about the basic mathematical operations that~~ We'll start by defining what vectors are. Then we'll describe all the mathematical operations we can perform ~~on vectors. We will define~~ with vectors, which include 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 geometric properties~~In Section 3.4 we'll also talk about two different kinds of vector products.

Definitions

~~The~~ A two-dimensional vector ~~$\vec{v} \in \mathbb{R}^2$ is equivalent~~ $\vec{v}$ corresponds to a *pair of numbers*~~$\vec{v} \equiv (v_x, v_y)$. We call~~ :

$$\vec{v} = (v_x, v_y),$$

where v_x ~~the x-component of $\vec{v}$,~~ is the *x-component* of the vector and v_y is ~~the y-component of $\vec{v}$.~~

~~Vector representations~~

its *y-component*. We denote the set of two-dimensional vectors as $\mathbb{R}^2$, since the components of a two-dimensional vector are specified by two real numbers. 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. 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.~~

mathematical shorthand $\vec{v} \in \mathbb{R}^2$ to define a two-dimensional vector $\vec{v}$. Vectors in $\mathbb{R}^2$ can be represented as arrows in the Cartesian plane. See the vector $\vec{v} = (3, 2)$ illustrated in Figure 3.1.

~~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)~~We can also define three-dimensional vectors like the vector $\vec{v} = (v_x, v_y, v_z) \in \mathbb{R}^3$, which has three components. Three-dimensional vectors can be represented as arrows in a coordinate system that has three axes, like the one shown in Figure 3.10 on page 223. A three-dimensional coordinate system is similar to the Cartesian coordinate system you're familiar with, and includes the additional z-axis that measures the height above the plane. In fact, there's no limit to the number of dimensions for vectors. We can define vectors in an n-dimensional space: $\vec{v} = (v_1, v_2, \ldots, v_n) \in \mathbb{R}^n$. For the sake of simplicity, we'll define all the vector operation formulas using two-dimensional vectors. Unless otherwise indicated in the text, all the formulas we give for two-dimensional vectors $\vec{v} \in \mathbb{R}^2$ also apply to n-dimensional vectors $\vec{v} \in \mathbb{R}^n$.

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: $\vec{u} + \vec{v} = (u_x + v_x, u_y + v_y)$~~ **Addition:** $\vec{u} + \vec{v} = (u_x + v_x, u_y + v_y)$
- ~~Subtraction: $\vec{u} - \vec{v} = (u_x - v_x, u_y - v_y)$~~ **Subtraction:** $\vec{u} - \vec{v} = (u_x - v_x, u_y -$
- ~~Scaling: $\alpha\vec{u} = (\alpha u_x, \alpha u_y)$~~ **Scaling:** $\alpha\vec{u} = (\alpha u_x, \alpha u_y)$
- ~~Dot product:~~ **Dot product:** $\vec{u} \cdot \vec{v} = u_x v_x + u_y v_y$
- ~~Length:~~ **Length:** $\|\vec{u}\| = \sqrt{\vec{u} \cdot \vec{u}} = \sqrt{u_x^2 + u_y^2}$. ~~We will also sometimes simply~~ The vector's length is also called the *norm* of the vector. We sometimes use the letter u to denote the length of the vector $\vec{u}$. ~~Cross product: $\vec{u} \times \vec{v} = (u_y v_z - u_z v_y, u_z v_x - u_x v_z, u_x v_y - u_y$~~

Note there is no vector division operation.

For vectors in a three-dimensional space $\vec{u} = (u_x, u_y, u_z) \in \mathbb{R}^3$ and $\vec{v} = (v_x, v_y, v_z) \in \mathbb{R}^3$, we can also define the **cross product** operation ~~The cross product is only defined for three-dimensional vectors like~~

~~$\vec{u}=(u_x,u_y,u_z)$ and $\vec{v}=(v_x,v_y,v_z)$~~$\vec{u}\times\vec{v}=(u_yv_z-u_zv_y,\ u_zv_x-u_xv_z,\ u_xv_y-u_y$
The dot product and the cross product are new operations that you probably haven't seen before. We'll talk more about dot products and the cross products in Section 3.4. For now let's start with the basics.

Vector representations

We'll use three equivalent ways to denote vectors in two dimensions:

- $\vec{v}=(v_x,v_y)$: component notation. The vector is written as a pair of numbers called the *components* or *coordinates* of the vector.
- $\vec{v}=v_x\hat{\imath}+v_y\hat{\jmath}$: unit vector notation. The vector is expressed as a combination of the unit vectors $\hat{\imath}=(1,0)$ and $\hat{\jmath}=(0,1)$.
- $\vec{v}=\|\vec{v}\|\angle\theta$: length-and-direction notation (polar coordinates). The vector is expressed in terms of its *length* $\|\vec{v}\|$ and the angle θ that the vector makes with the x-axis.

new fig; showcasing three of the colours in the noBS figs styles!

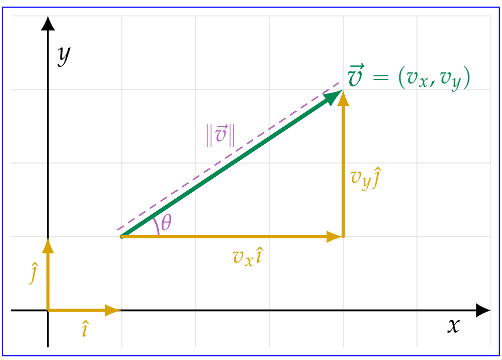


Figure 3.3: The vector $\vec{v}=(v_x,v_y)=v_x\hat{\imath}+v_y\hat{\jmath}=\|\vec{v}\|\angle\theta$.

~~Pay careful attention to the dot product and the cross product. Although they're called products, these operations behave much differently from taking the product of two numbers. Also note, there is no notion of vector division~~We use the component notation for doing vector algebra calculations since it is most compact. The unit vector notation shows explicitly that the vector $\vec{v}$ corresponds to the sum of $v_x\hat{\imath}$ (a displacement of v_x steps in the direction of the x-axis) and $v_y\hat{\jmath}$ (a displacement of v_y steps in the direction of the y-axis). The length-and-direction notation describes the vector $\vec{v}$ as a displacement of $\|\vec{v}\|$ steps in the direction of the angle θ. We'll use all three ways of denoting vectors throughout the rest of the book, and we'll learn how to convert between them.

Conversely, we can think of the vector $\vec{v}$ as being twice as long as the vector $\vec{w}$.

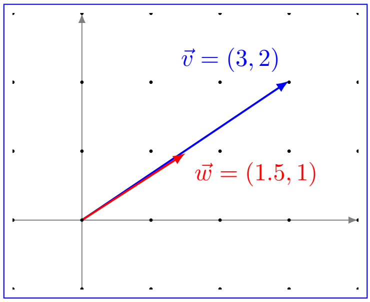


Figure 3.6: Vectors $\vec{v}$ and $\vec{w}$ are related by the equation $\vec{v} = 2\vec{w}$.

Multiplying a vector by a negative number reverses its direction.

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 two-dimensional vectors: we can ~~specify a vector~~ describe the vector $\vec{v} \in \mathbb{R}^2$ in terms of its length ~~$\|\vec{v}\|$ and its direction—the~~ $\|\vec{v}\|$ and its direction θ—the angle it makes with the x-axis. For example, the vector $(1,1)$ can also be written as $\sqrt{2}\angle 45^\circ$ ~~. This magnitude-and-direction~~ in polar coordinates. This length-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~~ It's therefore good to know the formulas for converting between the two ~~notations~~vector representations.

To convert the length-and-direction vector ~~$\|\vec{r}\|\angle\theta$~~ $\vec{v} = \|\vec{v}\|\angle\theta$ into an x-component and a y-component ~~(r_x, r_y)~~ (v_x, v_y), use the formulas

$$r v_x = \|\vec{v}\| \cos\theta \quad \text{and} \quad r v_y = \|\vec{v}\| \sin\theta.$$

To convert from component notation ~~(r_x, r_y)~~ $\vec{v} = (v_x, v_y)$ to length-and-direction ~~$\|\vec{r}\|\angle\theta$, use~~ $\|\vec{v}\|\angle\theta$, use

$$r = \|\vec{v}\| = \sqrt{r_x^2 + r_y^2}\,\cancel{\text{and}}\sqrt{v_x^2 + v_y^2}, \quad \theta = \tan^{-1}\frac{r_y}{r_x}. \begin{cases} \tan^{-1}\left(\frac{v_y}{v_x}\right) & \text{if } v_x > 0, \\ 180^\circ + \tan^{-1}\left(\frac{v_y}{v_x}\right) & \text{if } v_x < 0, \\ 90^\circ & \text{if } v_x = 0 \text{ a} \\ -90^\circ & \text{if } v_x = 0 \text{ a} \end{cases}$$

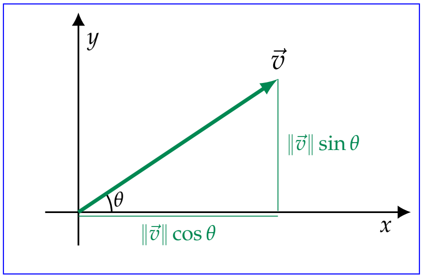


Figure 3.7: The x- and y~~-coordinates~~ -components of a vector with length ~~r and~~ $\|\vec{v}\|$ in the direction θ are given by ~~$r_x = r\cos\theta$~~ $v_x = \|\vec{v}\|\cos\theta$ and ~~$r_y = r\sin\theta$~~$v_y = \|\vec{v}\|\sin\theta$.

~~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 should 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~~ We've already seen these formulas in Section 1.17 (page 122), when we learned about the transformations between Cartesian and polar coordinates for points. The conversion procedure for vectors is exactly the same, including the trickiness around calculating θ when v_x is negative or zero. I invite you to revisit exercise E1.31 on page 128 to review the conversion operations between Cartesian coordinates and polar coordinates.

Unit vector notation

In ~~three~~ two dimensions, we can think of a vector ~~$\vec{v} = (v_x, v_y, v_z)$~~ $\vec{v} = (v_x, v_y)$ as a command to "Go a distance v_x in the x-direction ~~,~~ and 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 $\hat{k}$~~$\hat{\imath}$ and $\hat{\jmath}$. These are the unit vectors pointing in the x ~~, y, and z directions, respectively:~~ and y directions:

$$\hat{\imath} = (1,0,0), \qquad \text{and} \qquad \hat{\jmath} = (0,1,0),\cancel{\text{and}} = (0,0,1).$$

Any number multiplied by $\hat{\imath}$ corresponds to a vector with that number in the first coordinate. For example, ~~$3\hat{\imath} = (3,0,0)$. Similarly, $4\hat{\jmath} = (0,4,0)$ and $5\hat{k} = (0,0,5)$~~$3\hat{\imath} = (3,0)$ and $4\hat{\jmath} = (0,4)$.

In physics, we tend to perform a lot of numerical calculations with vectors; to make things easier, we often use unit vector notation:

$$v_x\hat{\imath} + v_y\hat{\jmath} + v_z \Leftrightarrow \qquad \Leftrightarrow \qquad (v_x, v_y, v_z).$$

Now find the y-component of the net force using the sin of the angles:

$$\begin{aligned} F_{\text{net},y} &= W_y + N_y + F_{f,y} \\ &= 300\sin(-90^\circ) + 260\sin(120^\circ) + 50\sin(30^\circ) \\ &= -49.8. \end{aligned}$$

Combining the two components of the vector, ~~we~~ you get the final answer:

$$\begin{aligned} \vec{F}_{\text{net}} &= (F_{\text{net},x}, F_{\text{net},y}) \\ &= (-86.7, -49.8) = -86.7\hat{\imath} - 49.8\hat{\jmath} \\ &= 100\angle 209.9^\circ, \end{aligned}$$

where you found the angle 209.9° by computing $\tan^{-1}(49.8/86.7)$ and adding 180° since the x-component is negative. 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 $12\hat{\imath}$, then the speed of the boat relative to the shore will be

$$12\hat{\imath} - 5\hat{\imath} = 7\hat{\imath},$$

since we must "deduct" the speed of the current from the speed of the boat relative to the water. See the vector diagram in Figure 3.8.

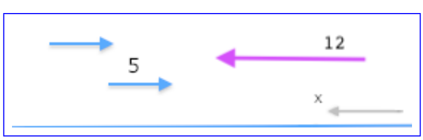


Figure 3.8: A boat travels with speed 12 knots against a current of 5 knots.

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. The situation is illustrated in Figure 3.9. 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 (in the $\hat{\jmath}$ direction).

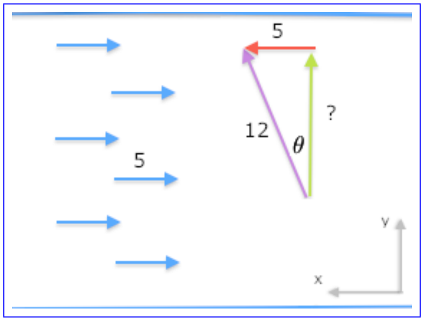


Figure 3.9: Part of the boat's thrust cancels the current.

Let's analyze the vector diagram. The opposite side of the triangle is parallel to the current flow and has length 5. We take the up-the-river component of the ~~speed~~ velocity $\vec{v}$ to be equal to $5\hat{\imath}$, so that it cancels exactly the $-5\hat{\imath}$ flow of the river. The hypotenuse has 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 $\frac{\text{opp}}{\text{hyp}} = \frac{5}{12} = \sin(\theta)$, 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:

$$\theta = \sin^{-1}\left(\tfrac{5}{12}\right) = 24.62^\circ.$$

The across-the-river component of the velocity can be calculated using $v_y = 12\cos(\theta) = 10.91$, or from Pythagoras' theorem if you prefer $v_y = \sqrt{\|\vec{v}\|^2 - v_x^2} = \sqrt{12^2 - 5^2} = 10.91$.

~~Vector dimensions~~Discussion

~~The most common types of vectors are two-dimensional vectors (like the ones~~ We did a lot of hands-on activities with vectors in this section and skipped over some of the theoretical details. Now that you've been exposed to the practical side of vector calculations, it's worth clarifying certain points that we glossed over.

Vectors vs. points

We used the notation $\mathbb{R}^2$ to describe two kinds of math objects: the set of points 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~~ the set of vectors in a ~~n-dimensional vectoris~~ two-dimensional space. The point $P = (P_x, P_y)$ and the vector $\vec{v} = (v_x, v_y)$ are both represented by pairs of real numbers, so we use the notation $P \in \mathbb{R}^2$ and $\vec{v} \in \mathbb{R}^2$ to describe them. This means that a pair of numbers $(3, 2) \in \mathbb{R}^2$ could represent the *coordinates* of a point, or the *components* of a vector, depending on the context.

Let's take a moment to review the definitions of points and vectors and clarify the types of operations we can perform on them:

- **Space of points** $\mathbb{R}^2$: the set of points $P = (P_x, P_y)$ corresponds to locations in the Cartesian plane. The point $P = (P_x, P_y)$ corresponds to the geometric instructions: "Starting at the origin $(0, 0)$, move P_x units along the x-axis and P_y units along the y-axis." The distance between points P and Q is denoted $d(P, Q)$.
- **Vector space** $\mathbb{R}^2$: the set of vectors $\vec{v} = (v_x, v_y)$ describes displacements in the Cartesian plane. The vector $\vec{v} = (v_x, v_y)$ corresponds to the instructions: "Starting anywhere, move v_x units along the x-axis and v_y units along the y-axis." Vectors can be combined and manipulated using the vector algebra operations $\vec{u} + \vec{v}$, $\vec{u} - \vec{v}$, $\alpha\vec{u}$, $\vec{u} \cdot \vec{v}$, and $\|\vec{v}\|$.

Note the geometric instructions for points and vectors are very similar; the only difference is the starting point. The coordinates of a point (P_x, P_y) specify a *fixed position* relative to the origin $(0, 0)$, while the components of a vector (v_x, v_y) describe a *relative displacement* that can have any starting point.

Let's look at some examples of calculations that combine points and vectors. Consider the points P and Q in the Cartesian plane, and the displacement vector $\vec{v}_{PQ}$ between them. The displacement vector $\vec{v}_{PQ}$ gives the "move instructions" for getting from point P to point Q and is defined by the equation:

$$\vec{v}_{PQ} = Q - P.$$

This equation says that subtracting two points produces a vector, which make sense if you think about it—the "difference" between two points is a displacement vector.

We can use the displacement vector $\vec{v}_{PQ}$ in calculations like this:

$$P + \vec{v}_{PQ} = P + (v_1, v_2, \ldots, v_n Q - P) \in \mathbb{R}^n = Q.$$

~~The rules of vector algebra apply in higher dimensions, but our ability to visualize stops at three dimensions~~In words, this calculation shows that "Starting at the point P and moving by $\vec{v}_{PQ}$ brings us to the point Q."

The above equations use addition and subtraction operations between a mix of points and vectors. This is rather unusual: normally we only use operations like "+" and "-" between math objects of the same kind. In this case, we're allowed to mix points and vectors because they both describe "move instructions" of the same kind.

Let's keep going. What other useful calculations can we do by combining points and vectors? Suppose we wanted to find the midpoint M that lies exactly in the middle between points P and Q. We can find the midpoint M using the displacement vector $\vec{v}_{PQ}$ and some basic vector algebra. If starting from P and moving by $\vec{v}_{PQ}$ brings us all the way to the point Q, then starting from P and moving by $\frac{1}{2}\vec{v}_{PQ}$ will bring us to the midpoint: $M = P + \frac{1}{2}\vec{v}_{PQ}$.

~~Coordinate system~~

The mathematical bridge between points and vectors allows us to use vector techniques to solve geometry problems. By learning to describe geometric objects like points, lines, and circles using vectors, we can do complicated geometry calculations using simple algebraic manipulations like vector operations. This exemplifies a general pattern in mathematics: applying techniques developed in one domain to solve problems in another domain.

~~Vector components depend on the coordinate system in which the vectors are represented. Throughout this section we used the x, y,~~

Example You come to class one day and there's a surprise quiz that asks you to write the formula for the distance $d(P, Q)$ between two points $P = (P_x, P_y)$ and ~~z axes as the coordinate system, and we described vectorsas components along each of these axes. This is a very convenient coordinate system ; we have a set of three *perpendicular* axes,~~ $Q = (Q_x, Q_y)$. You don't remember ever learning about such a formula and feel caught off guard. How can the teacher ask for a formula they haven't covered in class yet? This seems totally unfair!

After a minute of stressing out, you take a deep breath, come back to your senses, and resolve to give this problem a shot. You start by sketching a coordinate system, placing points P and ~~a set of~~ Q in it, and drawing the line that connects the two points. What is the formula that describes the length of this line?

The line from P to Q looks like the hypotenuse of a triangle, which makes you think that trigonometry could somehow be used to find the answer. Unfortunately, trying to remember the trigonometry formulas has only the effect of increasing your math anxiety. You take this as a sign that you should look for other options. In math, it's important to trust your gut instincts.

By a fortunate coincidence, you were recently reading about the connection between points and vectors, and specifically about the displacement vector $\vec{v}_{PQ} = Q - P$. The line in your sketch represents the vector $\vec{v}_{PQ}$. You realize that the distance between the points P and Q is the same as the length of the vector $\vec{v}_{PQ}$. You remember the formula for the length of a vector $\vec{v}$ is $\|\vec{v}\| = \sqrt{v_x^2 + v_y^2}$ and you know the formula for the displacement vector is $\vec{v}_{PQ} = (Q_x - P_x, Q_y - P_y)$, so you combine these formulas to obtain the answer: $d(P,Q) = \|\vec{v}_{PQ}\| = \sqrt{(Q_x -$
One more win for the "don't worry and try it" strategy for solving math problems!

Vectors in three dimensions

A three-dimensional coordinate system consists of three axes: the x-axis, the y-axis, and the z-axis. The three axes point in perpendicular directions to each other, as illustrated in Figure 3.10. Look around you and find a corner of the room you're in where two walls and the floor meet. The x-axis and the y-axis are the edges where the floor meets the walls. The vertical edge where the two walls meet represents the z-axis.

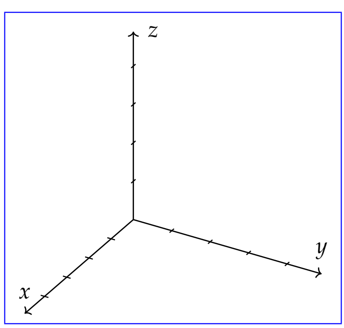


Figure 3.10: A three-dimensional coordinate system with x, y, and z axes.

The vector $\vec{v} = (v_x, v_y, v_z) \in \mathbb{R}^3$ describes the following displacement instructions: "Move v_x units in the direction of the x-axis, then move v_y along the y-axis, and finally move v_z in the direction of the z-axis." In three dimensions, there are three unit vectors ~~$\{\hat{\imath}, \hat{\jmath}, \hat{k}\}$ that point~~

~~along~~ that describe unit steps in the direction of each of the ~~three axis directions. Every vector is implicitly defined~~ axes:

$$\hat{\imath} = (1,0,0), \qquad \hat{\jmath} = (0,1,0) \quad \text{and} \quad \hat{k} = (0,0,1).$$

We can therefore describe the vector $\vec{v} = (v_x, v_y, v_z)$ in terms of ~~this coordinate system. When we 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~~ unit vectors as $\vec{v} = v_x\hat{\imath} + v_y\hat{\jmath} + v_z\hat{k}$.

High-dimensional vectors

The most common types of vectors you'll encounter in math and physics are two-dimensional and three-dimensional vectors. In other fields of science like genetics and machine learning, it's common to see vectors with many more dimensions. For example, in machine learning we often represent "rich data" like images, videos, and text as vectors with thousands of dimensions.

An example of an n-dimensional vector is

$$\vec{v} = (v_1, v_2, \ldots, v_n) \in \mathbb{R}^n.$$

The vector algebra operations you learned in this section also apply to these high-dimensional vectors.

Vectors and vector coordinates

One final point we need to clarify is the difference between real-world vector quantities like the velocity of a tennis ball $\vec{v}$ and its mathematical representation as a coordinate vector (v_x, v_y, v_z). If you know the coordinate vector (v_x, v_y, v_z) then you know what the real-world velocity is, right? Not quite.

Let's say you're doing a physics research project on tennis serves. You define an xyz-coordinate system for the tennis court, which allows you to represent the ball's velocity $\vec{v}$ as a triple of components (v_x, v_y, v_z) interpreted as: "The ball is moving with velocity v_x units in the x-direction, ~~then move 4~~ v_y units in the y-direction, and ~~finally move 2~~ v_z 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~~

Suppose you want to describe the velocity vector $\vec{v}$ to a fellow physicist via text message. Referring to your sheet of calculations, you find the values $\vec{v} = (60, 3, -2)$, which you know were measured in metres per second. You send this message:

```
\DIFadd{The velocity is (60,3,-2) measured in metres per second. }
```

A few minutes later the following reply comes back:

```
\DIFadd{Wait whaaat? What coordinate system are you using? }
```

Indeed the information you sent is incomplete. Vector components depend on the coordinate system in which the vectors are represented. The triple of numbers $(60, 3, -2)$ only makes sense once you know the directions of the axes in the xyz-coordinate system. Realizing your mistake, you send a text with all the required information:

```
\DIFadd{Using the coordinate system centred at the south post of
the net, with the x-axis pointing east along the court,
the y-axis pointing north along the net, and the z-axis
pointing up, the velocity is (60,3,-2) in metres per second. }
```

A few seconds later, you get the reply:

```
\DIFadd{OK got it now. Thx! }
```

This hypothetical situation illustrates the importance of the coordinate systems for describing vectors. If you don't know what the coordinate system is, knowing the coordinates (v_x, v_y, v_z) doesn't tell you much. Only when you know the directions of the unit vectors $\hat{\imath}$, $\hat{\jmath}$, and $\hat{k}$ can you interpret the instructions $\vec{v} = v_x\hat{\imath} + v_y\hat{\jmath} + v_z\hat{k}$.

It turns out, using the xyz-coordinate system ~~and the~~ with the three 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~~ coordinates (v_1, v_2, v_3) with respect to any *basis* $\{\hat{e}_1, \hat{e}_2, \hat{e}_3\}$ ~~as follows:~~ using the expression $\vec{v} = v_1\hat{e}_1 + v_2\hat{e}_2 + v_3\hat{e}_3$~~. What is~~, which corresponds to the instructions: "Move v_1 units in the direction of $\hat{e}_1$, move v_2 units in the direction of $\hat{e}_2$, and move v_3 units in the direction of $\hat{e}_3$."

What's a basis, you ask? I'm glad you asked, because this is the subject of the next section.

~~Exercises~~

~~Given the vectors $\vec{v}_1 = (2,1)$, $\vec{v}_2 = (2,-1)$, and $\vec{v}_3 = (3,3)$, calculate the following expressions: **a)** $\vec{v}_1 + \vec{v}_2$~~

~~**b)** $\vec{v}_2 - 2\vec{v}_1$~~

~~**c)** $\vec{v}_1 + \vec{v}_2 + \vec{v}_3$~~

~~**a)** $(4,0)$. **b)** $(-2,-3)$. **c)** $(7,3)$.~~

~~Express the following vectors as components: **a)** $\vec{v}_1 = 10\angle 30°$~~

~~**b)** $\vec{v}_2 = 12\angle -90°$~~

~~**c)** $\vec{v}_3 = 3\angle 170°$~~

~~**a)** $\vec{v}_1 = (5\sqrt{3}, 5) = (8.66, 5)$. **b)** $\vec{v}_2 = (0, -12)$. **c)** $\vec{v}_3 = (-2.95, 0.52)$.~~

~~Express the following vectors in length-and-direction notation:~~
~~**a)** $\vec{u}_1 = (4,0)$~~
~~**b)** $\vec{u}_2 = (1,1)$~~
~~**c)** $\vec{u}_3 = (-1,3)$~~
~~**a)** $\vec{u}_1 = 4\angle 0°$. **b)** $\vec{u}_2 = \sqrt{2}\angle 45°$. **c)** $\vec{u}_3 = \sqrt{10}\angle 108.43°$.~~

3.3 Basis

One of the most important concepts in the study of vectors is the concept of a *basis*. Consider the three-dimensional vector space $\mathbb{R}^3$. A *basis* for $\mathbb{R}^3$ is a set of vectors $\{\hat{e}_1, \hat{e}_2, \hat{e}_3\}$ that 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~~ coordinates (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 triple (v_x, v_y, v_z) with respect to the basis $\{\hat{\imath}, \hat{\jmath}, \hat{k}\}$ through the formula $\vec{v} = v_x\hat{\imath} + v_y\hat{\jmath} + v_z\hat{k}$. The whole point of this section is to let you know that other bases (coordinate systems) exist, and to get you into the habit of asking, "With respect to which coordinate system?" every time you see a coordinate vector (a, b, c).

An analogy

Let's start with a simple example of a basis. If you look at the HTML source code behind any web page, you're sure to find at least one mention of the colour stylesheet directive such as `color:#336699;`. The numbers should be interpreted as a triple 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 triple $(33, 66, 99)$ 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*~~*components* 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~~components:

$$(33, 66, 99)_{RGB} = \text{CoolBlue} = (222, 189, 156)_{CMY} = 222\mathbf{C} + 189\mathbf{M} + 156\mathbf{Y}.$$

The *same* colour CoolBlue is represented by a *different* set of ~~coefficients~~ components when the CMY colour basis is used.

Note that a triple of ~~coefficients by itself does not~~ components by itself doesn’t mean anything unless we know the basis being used. For example, if we were to interpret the triple of ~~coordinates~~ components $(33, 66, 99)$ with respect to the CMY basis, ~~will~~ we would obtain a completely different colour, which would not be cool at all.

A basis is required to convert mathematical objects like the triple (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 triple of ~~coefficients~~components. Writing $(222, 189, 156)_{CMY}$ and $(33, 66, 99)_{RGB}$ clarifies which basis to use for each triple of components.

Discussion

It ~~’s~~ would be 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~~start a new problem that involves vector calculations, the first thing you should do is ~~draw a coordinate system . Always keep in mind the coordinate system you’re using when computing~~ choose the coordinate system you want to use, and indicate it clearly in the diagram.

Using a non-standard coordinate system can sometimes simplify the equations you have to solve. For example, let’s say we want to study the motion of a block sliding down an incline with velocity $\vec{v}$, as illustrated in Figure 3.11. Using the standard xy-basis, the velocity vector is represented as $(v\cos\theta, -v\sin\theta)_{xy}$, which has components in both the x- and y-directions and requires using trigonometric functions. If instead you use the non-standard $x'y'$-basis, the components of ~~vectors~~the velocity will be $(v, 0)_{x'y'}$. Note the velocity only has a component along the x'-direction, which will simplify all subsequent calculations.

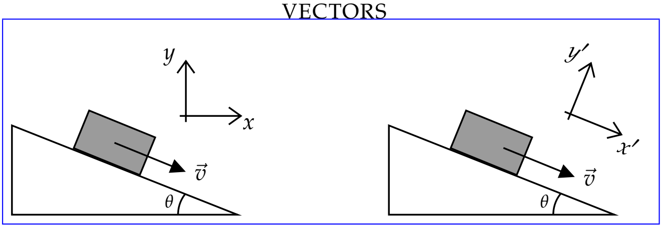


Figure 3.11: The vector $\vec{v}$ is described by the coordinates $(v\cos\theta, -v\sin\theta)_{xy}$ with respect to the standard basis xy. The same vector $\vec{v}$ is described by the coordinates $(v, 0)_{x'y'}$ with respect to the "tilted" basis $x'y'$.

Recall the polar coordinates representation we used to describe points $r\angle\theta$ and vectors $\|\vec{v}\|\angle\theta$ in two dimensions (see page 216). This is another example of an alternative coordinate system that's useful for describing rotations and circular motion. Note certain textbooks will write the polar coordinates of the vector $\vec{v} = \|\vec{v}\|\angle\theta$ using the bracket notation $(\|\vec{v}\|, \theta)$, which can easily be confused with the Cartesian coordinates of the vector (v_x, v_y). Indicating the coordinate system as a subscript after the bracket can avoid any confusion: $\vec{v} = (\|\vec{v}\|, \theta)_{r\theta} = (v_x, v_y)_{xy}$.

Links

[Vectors and vector operations explained by 3Blue1Brown]
`https://www.youtube.com/watch?v=fNk_zzaMoSs`

[More vector illustrations and definitions from Wikipedia]
`https://en.wikipedia.org/wiki/Euclidean_vector`

Exercises

E3.1 Given the vectors $\vec{v}_1 = (2, 1)$, $\vec{v}_2 = (2, -1)$, and $\vec{v}_3 = (3, 3)$, calculate the following expressions:

a) $\vec{v}_1 + \vec{v}_2$ **b)** $\vec{v}_2 - 2\vec{v}_1$ **c)** $\vec{v}_1 + \vec{v}_2 + \vec{v}_3$

E3.2 Express the following vectors as components:

a) $\vec{v}_1 = 10\angle 30°$ **b)** $\vec{v}_2 = 12\angle -90°$ **c)** $\vec{v}_3 = 3\angle 170°$

E3.3 Express the following vectors in length-and-direction notation:

a) $\vec{u}_1 = (4, 0)$ **b)** $\vec{u}_2 = (1, 1)$ **c)** $\vec{u}_3 = (-1, 3)$

3.4 Vector products

~~Since adding two vectors $\vec{v}$ and $\vec{w}$ corresponds to adding their components $(v_x + w_x, v_y + w_y, v_z + w_z)$, you might logically think thatthe product of two vectors also corresponds to the product of their components $(v_x w_x, v_y w_y, v_z w_z)$, but this type of product is not used. Instead, we'll~~ We'll now define the *dot product* and the *cross product*~~, which allow us to perform useful geometric operations with~~ : two geometric operations useful for working with three-dimensional vectors.

Dot product

The *dot product* takes two vectors as inputs and produces a single, real number as an output:

$$\cdot : \mathbb{R}^3 \times \mathbb{R}^3 \quad \to \quad \mathbb{R}.$$

The dot product between ~~two vectors~~ the vector $\vec{v} = (v_x, v_y, v_z)$ and the vector $\vec{w} = (w_x, w_y, w_z)$ can be computed using either the algebraic formula,

$$\vec{v} \cdot \vec{w} \cancel{\equiv} = v_x w_x + v_y w_y + v_z w_z,$$

or the geometric formula,

$$\vec{v} \cdot \vec{w} \cancel{\equiv} = \|\vec{v}\| \|\vec{w}\| \cos(\varphi),$$

where φ is the angle between the two vectors. Note the value of the dot product depends on the vectors' lengths and the cosine of the angle between them.

The name *dot product* comes from the symbol used to denote it. It is also known as the *scalar product*, since the result of the dot product is a scalar number—a number that does not change when the basis changes. The dot product is also sometimes called the *inner product*.

We can combine the algebraic and the geometric formulas for the dot product to obtain the formula,

$$\cos(\varphi) = \frac{\vec{v} \cdot \vec{w}}{\|\vec{v}\| \|\vec{w}\|} = \frac{v_x w_x + v_y w_y + v_z w_z}{\|\vec{v}\| \|\vec{w}\|} \quad \text{and} \quad \varphi = \cos^{-1}(\cos(\varphi)).$$

This formula makes it possible to find the angle between two vectors if we know their components.

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 $\cos(\varphi) = \cos(0°) = 1$, so $\vec{v} \cdot \vec{w} = \|\vec{v}\|\|\vec{w}\|$.
- If the vectors are perpendicular to each other, then $\cos(\varphi) = \cos(90°) = 0$, so $\vec{v} \cdot \vec{w} = 0$.
- If the vectors point in exactly opposite directions, then $\cos(\varphi) = \cos(180°) = -1$, so $\vec{v} \cdot \vec{w} = -\|\vec{v}\|\|\vec{w}\|$.

The dot product is defined for vectors of any dimension; as long as two vectors ~~have the same number of components~~are defined with respect to the same basis, we can compute ~~their dot product~~ the dot product between them.

Cross product

The *cross product* takes two vectors as inputs and produces another vector as the output:

$$\times : \mathbb{R}^3 \times \mathbb{R}^3 \quad \rightarrow \quad \mathbb{R}^3.$$

The cross product of two vectors is perpendicular to both vectors:

$$\vec{v} \times \vec{w} = \{ \text{ a vector perpendicular to both } \vec{v} \text{ and } \vec{w} \ \} \quad \in \mathbb{R}^3.$$

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: $\hat{\imath} \times \hat{\jmath} = \hat{k}$. The name *cross product* comes from the symbol used to denote it. It is also sometimes called the *vector product*, since the output of this operation is a vector.

The cross products of individual basis elements are defined as

$$\hat{\imath} \times \hat{\jmath} = \hat{k}, \qquad \hat{\jmath} \times \hat{k} = \hat{\imath}, \qquad \hat{k} \times \hat{\imath} = \hat{\jmath}.$$

Look at Figure 3.10 on page 223 and imagine the vectors $\hat{\imath}$, $\hat{\jmath}$, and $\hat{k}$ pointing along each axis. Try to visualize the three equations above.

The cross product is *anticommutative*, which means swapping the order of the inputs introduces a negative sign in the output:

$$\hat{\jmath} \times \hat{\imath} = -\hat{k}, \qquad \hat{k} \times \hat{\jmath} = -\hat{\imath}, \qquad \hat{\imath} \times \hat{k} = -\hat{\jmath}.$$

It's likely that, until now, 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 of two vectors is commutative $\vec{u} \cdot \vec{v} = \vec{v} \cdot \vec{u}$, but the cross product of two vectors is *anti*commutative $\vec{v} \times \vec{w} = -\vec{w} \times \vec{v}$.

~~For two arbitrary vectors $\vec{a} = (a_x, a_y, a_z)$ and $\vec{b} = (b_x, b_y, b_z)$, the~~ Given two vectors $\vec{a} = a_x\hat{\imath} + a_y\hat{\jmath} + a_z\hat{k}$ and $\vec{b} = b_x\hat{\imath} + b_y\hat{\jmath} + b_z\hat{k}$, their cross product is calculated as

$$\vec{a} \times \vec{b} = (a_y b_z - a_z b_y)\hat{\imath} + (a_z b_x - a_x b_z)\hat{\jmath} + (a_x b_y - a_y b_x)\hat{k}.$$

~~The cross product 's output has a length that is proportional to the sine of the angle between the vectors:~~

$$\cancel{\|\vec{a} \times \vec{b}\| = \|\vec{a}\|\|\vec{b}\| \sin(\varphi).}$$

~~The direction of the vector $(\vec{a} \times \vec{b})$~~ Computing the cross product requires a specific combination of multiplications and subtractions of the input vectors' components. The result of this combination is the vector $\vec{a} \times \vec{b}$ which is perpendicular to both $\vec{a}$ and $\vec{b}$.

The length of the cross product of two vectors is proportional to the sine of the angle between the two vectors:

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

The right-hand rule

Consider the plane formed by the vectors $\vec{a}$ and $\vec{b}$. There are actually *two* vectors perpendicular to this plane: one above the plane and one below the plane. We use the *right-hand rule* to figure out which of these vectors corresponds to the cross product $\vec{a} \times \vec{b}$.

Make a fist with your right hand and then extend your thumb, first finger, and middle finger. When your index finger points in the same direction as the vector $\vec{a}$ and your middle finger points in the direction of $\vec{b}$, your thumb will point in the direction of $\vec{a} \times \vec{b}$. The relationship encoded in the right-hand rule matches the relationship between the standard basis vectors: $\hat{\imath} \times \hat{\jmath} = \hat{k}$.

Links

[~~A nice illustration~~ Nice illustrations of the cross product]
`http://1ucasvb.tumblr.com/post/76812811092/`
`https://www.youtube.com/watch?v=eu6i7WJeinw`

The *phase*, also known as the *argument* of the complex number $z = a + bi$ is given by the formula

$$\varphi_z \equiv = \arg z = \texttt{atan2}(b,a) =^{\dagger} \tan^{-1}(b/a). \begin{cases} \tan^{-1}\left(\frac{b}{a}\right) & \text{if } a > 0, \\ \pi + \tan^{-1}\left(\frac{b}{a}\right) & \text{if } a < 0, \\ \frac{\pi}{2} & \text{if } a = 0 \text{ and } b > 0 \\ -\frac{\pi}{2} & \text{if } a = 0 \text{ and } b < 0 \end{cases}$$

The phase corresponds to the angle that z forms with the real axis. ~~Note the equality labelled † is true only when $a > 0$, because the function $\tan^{-1}$ always returns numbers in the range $[-\frac{\pi}{2}, \frac{\pi}{2}]$. Manual corrections of the output of $\tan^{-1}(b/a)$ are required for complex numbers with $a < 0$~~

We previously saw this complicated-looking formula with four cases when we talked about converting from Cartesian coordinates to polar coordinates for points (Section 1.17) and vectors (Section 3.2). When a certain formula comes up three times in a math book, this should tell you the author *really* wants you to know it. Seriously, do me a favour and revisit the exercise E1.31 (page 128) and the exercise E3.3 (page 228).

~~Some programming languages~~ Computer algebra systems provide the two-input ~~math function~~ inverse tangent function `atan2`~~(y,x) that correctly computes the angle that the vector (x, y) makes with the x-axis in all four quadrants. Because complex numbers behave like two-dimensional vectors, you can use~~ , which is the easiest way to calculate the phase φ_z for the complex number $z = a + bi$. The function `atan2` ~~to compute their phase~~ handles all four cases automatically and always computes the correct phase φ_z.

* * *

In addition to the vector-like ~~properties of~~ operations we can perform on complex numbers, like computing their magnitude and phase, we can also perform other operations ~~with~~ on complex numbers that are not defined for vectors. The set of complex numbers $\mathbb{C}$ is a *field*. This means, in addition to the addition and subtraction operations, we can also perform multiplication and division with complex numbers.

Multiplication

The product of two complex numbers is computed using the usual rules of algebra:

$$(a + bi)(c + di) = (ac - bd) + (ad + bc)i.$$

$$\begin{aligned}(a+bi)(c+di) &= a(c+di)+bi(c+di)\\ &= ac+adi+bci+bdi^2\\ &= (ac-bd)+(ad+bc)i.\end{aligned}$$

In the polar representation, the product formula is

$$(p\angle\phi)(q\angle\psi) = pq\angle(\phi+\psi).$$

To multiply two complex numbers, multiply their magnitudes and add their phases.

Example Verify that $z\,\bar{z} = a^2+b^2 = |z|^2$.

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 $|s|^2 = s\bar{s}$~~ $\bar{s}$ and $|s|^2 = s\bar{s}$ and then use

$$z/s = z\frac{\bar{s}}{|s|^2}\frac{\bar{s}}{|s|^2}.$$

You can think of $\frac{\bar{s}}{|s|^2}\frac{\bar{s}}{|s|^2}$ as being equivalent to s^{-1}.

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:

$$x_1 = 5+\sqrt{15}i \quad\text{and}\quad x_2 = 5-\sqrt{15}i."$$

When you add x_1+x_2 you obtain 10. When you multiply the two numbers the answer is

$$\begin{aligned}x_1x_2 &= \left(5+\sqrt{15}i\right)\left(5-\sqrt{15}i\right)\\ &= 25-5\sqrt{15}i+5\sqrt{15}i-\sqrt{15}^2i^2 = 25+15 = 40.\end{aligned}$$

Hence $5+\sqrt{15}i$ and $5-\sqrt{15}i$ are two numbers whose sum is 10 and whose product is 40.

Example 2 ~~Compute~~ Let's compute the product of ~~i and~~ -1 ~~. Both i and -1 have a magnitude of 1 but different phases. The phase of~~ and i. The answer is obviously $-i$, but let's look at this simple calculation geometrically. The polar representation of the number i is $1\angle\frac{\pi}{2}$. Multiplication of any complex number $z=|z|\angle\varphi_z$ by i corresponds to adding $\frac{\pi}{2}$ ~~(90°), while -1 has phase π (180°). The product of these two numbers is~~

$$(i)(-1) \;=\; (1\angle\tfrac{\pi}{2})(1\angle\pi) \;=\; 1\angle\left(\tfrac{\pi}{2}+\pi\right) \;=\; 1\angle\tfrac{3\pi}{2} \;=\; -i.$$

~~Multiplication~~ to the phase of the number:

$$zi=(|z|\angle\varphi_z)(1\angle\tfrac{\pi}{2})=(|z|\cdot 1)\angle(\varphi_z+\tfrac{\pi}{2})=|z|\angle(\varphi_z+\tfrac{\pi}{2}).$$

In other words, multiplication by i is ~~effectively a rotation by~~ equivalent to applying a $\frac{\pi}{2}$ (90°) ~~to the left~~counterclockwise rotation in the complex plane. We can therefore interpret the answer $(-1)(i)=-i$ as the complex number $-1=1\angle\pi$ experiencing a $\frac{\pi}{2}$ rotation to arrive at $1\angle(\pi+\frac{\pi}{2})=1\angle\frac{3\pi}{2}=-i$.

Example 3 Find the polar representation of $z=-3-i$ and compute z^6. Let's denote the polar representation of z by $z=r\angle\varphi$ as shown in Figure 3.14. We find $r=\sqrt{3^2+1^2}=\sqrt{10}$ and $\varphi=\tan^{-1}(\frac{1}{3})+\pi=0.322+\pi$. Using the polar representation, we can easily compute z^6:

$$z^6 = r^6\angle(6\varphi) = (\sqrt{10})^6\angle\, 6(0.322+\pi) = 10^3\angle 1.932+6\pi = 10^3\angle 1.932.$$

Note we can ignore multiples of 2π in the phase. ~~In component form,~~ We thus find the value of z^6 is ~~equal to~~ $1000\cos(1.932)+1000\sin(1.932)i=-353.4+935.5i$.

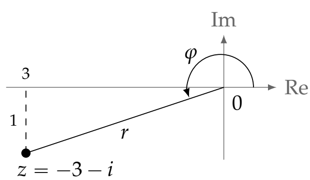


Figure 3.14: The complex number $z=3-i$ has magnitude $r=\sqrt{10}$ and phase $\varphi=0.322+\pi=3.463$.

Fundamental theorem of algebra

The ~~solutions to *any* polynomial equation $a_0 + a_1x + \cdots + a_nx^n = 0$ are of the form~~ fundamental theorem of algebra states that any polynomial of degree n,

$$z = a + bi.$$

~~In particular, any polynomial $P(x)$ of n^{th} degree~~ $P(x) = a_nx^n + \cdots + a_2x^2 + a_1x +$ can be written as

$$P(x) = a_n(x - z_1)(x - z_2)\cdots(x - z_n),$$

where $z_i \in \mathbb{C}$ are the polynomial's ~~*complex* roots.~~ roots. In other words, the equation $P(x) = 0$ has n solutions: the complex numbers $z_1, z_2, \ldots, z_n$. Before today, you might have said the equation ~~$x^2 + 1 = 0$~~ $x^2 + 1 = 0$ has no solutions. Now you know its solutions are the complex numbers $z_1 = i$ and $z_2 = -i$.

The theorem is "fundamental" because it tells us we~~won't ever~~'ll never need to invent ~~any~~ numbers "fancier" ~~set of~~ than the complex numbers to solve polynomial equations. ~~Recall~~ To understand why this is important, recall that each set of numbers is associated with a different class of equations. Figure 1.2 on page 8 shows the nested containment structure of the number sets $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$. The natural numbers $\mathbb{N}$ appear as solutions of the equation $m + n = x$, where m and n are natural numbers (denoted $m, n \in \mathbb{N}$). The integers $\mathbb{Z}$ are the solutions to equations of the form $x + m = n$, where $m, n \in \mathbb{N}$. The rational numbers $\mathbb{Q}$ are necessary to solve for x in $mx = n$, with $m, n \in \mathbb{Z}$. To find the solutions of $x^2 = 2$, we need the real numbers $\mathbb{R}$. ~~The process of requiring new types of numbers for solving more complicated types of equationsstops at~~ And in this section, we learned that the solutions to the equation $x^2 = -1$ are complex numbers $\mathbb{C}$~~; any polynomial equation—no~~. At this point you might be wondering if you're attending some sort of math party, where mathematicians write down complicated equations and—just for the fun of it—invent new sets of numbers to describe the solutions to these equations. Can this process continue indefinitely?

Nope. The party ends with $\mathbb{C}$. The fundamental theorem of algebra guarantees that any polynomial equation you could come up with—no matter how complicated it is—has solutions that are complex numbers $\mathbb{C}$.

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~~related to the functions sine and cosine. Lo and behold, we have ~~Euler's formula~~*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: $z = |z|\angle\varphi_z = |z|e^{i\varphi_z}$.

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~~ to obtain the equation $e^{i\pi} + 1 = 0$. The equation $e^{i\pi} + 1 = 0$ is called *Euler's identity*, and it 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?

One way to understand the equation $e^{i\pi} + 1 = 0$, is to think of $e^{i\pi}$ as the polar representation of the complex number $z = 1e^{i\pi} = 1\angle\pi$, which is the same as 1 rotated counterclockwise by π radians (180°) in the complex plane. We know $e^{i\pi} = 1\angle\pi = -1$ and so $e^{i\pi} + 1 = 0$.

De Moivre's formula

By replacing θ in Euler's formula with $n\theta$, we obtain de Moivre's formula:

$$(\cos\theta + i\sin\theta)^n = \cos n\theta + i\sin n\theta.$$

De Moivre's formula 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 (page 117) 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~~ Intuitive proof of the fundamental theorem of algebra]
https://www.youtube.com/watch?v=shEk8sz1oOw

P3.8 Find a unit vector that is perpendicular to both $\vec{u} = (1, 0, 1)$ and $\vec{v} = (1, 2, 0)$.
Hint: Use the cross product.

P3.9 Find a vector that is orthogonal to both $\vec{u}_1 = (1, 0, 1)$ and $\vec{u}_2 = (1, 3, 0)$, and whose dot product with the vector $\vec{v} = (1, 1, 0)$ is equal to 8.

P3.10 Compute the following expressions:

a) $\sqrt{-4}$ **b)** $\dfrac{2+3i}{2+2i}$ **c)** $e^{3i}(2+i)e^{-3i}$

P3.11 Solve for $x \in \mathbb{C}$ in the following equations:

a) $x^2 = -4$ **b)** $\sqrt{x} = 4i$

c) $x^2 + 2x + 2 = 0$ **d)** $x^4 + 4x^2 + 3 = 0$

Hint: To solve **d)**, use the substitution $u = x^2$.

P3.12 Given the numbers $z_1 = 2 + i$, $z_2 = 2 - i$, and $z_3 = -1 - i$, compute

a) $|z_1|$ **b)** $\dfrac{z_1}{z_3}$ **c)** $z_1 z_2 z_3$

P3.13 A real business is a business that is profitable. An imaginary business is an idea that is just turning around in your head. We can model the real-imaginary nature of a business project by representing the *project state* as a complex number $p \in \mathbb{C}$. For example, a business idea is described by the state $p_o = 100i$. In other words, it is 100% imaginary.

To bring an idea from the imaginary into the real, you must work on it. We'll model the work done on the project as a multiplication by the complex number $e^{-i\alpha h}$, where h is the number of hours of work and α is a constant that depends on the project. After h hours of work, the initial state of the project ~~is transformed as follows:~~ p_o has become $p_f = e^{-i\alpha h}p_o$. Working on the project for one hour "rotates" its state by $-\alpha$ ~~rad~~rad, making it ~~less imaginary and more real~~ more real and less imaginary.

If you start from an idea $p_o = 100i$ and the cumulative number of hours invested after t weeks of working on the project is $h(t) = 0.2t^2$, how long will it take for the project to become 100% real? Assume $\alpha = 2.904 \times 10^{-3}$.
Hint: A project is 100% real if $\text{Re}\{p\} = p$.

P3.14 A farmer with a passion for robotics has built a prototype of a robotic tractor. The tractor is programmed to move with a speed of 0.524 km/h and follow the direction of the hour-hand on a conventional watch. Assume the tractor starts at 12:00 p.m. (noon) and is left to roam about in a field until 6:00 p.m. What is the shape of the trajectory that the tractor will follow? What is the total distance travelled by the tractor after six hours?

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

$$v_t = r\omega = 4.5025 \times 10^6 \times 7.2921 \times 10^{-5} = 328.32[\text{m/s}],$$

which is equal to 1181.95[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 1200[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~~ and solve for accelerations a_r ~~,~~ and 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 that shows a cross section of the train in the $\hat{r}$ and $\hat{z}$ directions (see Figure 4.16). 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 $\hat{r}$

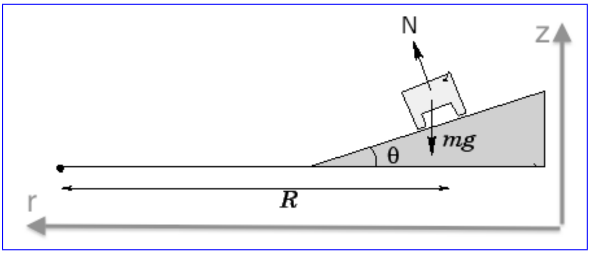


Figure 4.16: ~~Force diagram~~ Diagram showing the forces acting on a magnetically levitating train that is travelling on a circular racetrack of radius R and bank angle θ.

and $\hat{z}$ directions:

$$\begin{aligned} \sum F_r &= N\sin\theta = ma_r = m\frac{v_t^2}{R} &&\Rightarrow\quad N\sin\theta = m\frac{v_t^2}{R}, \\ \sum F_z &= N\cos\theta - mg = 0 &&\Rightarrow\quad N\cos\theta = mg. \end{aligned}$$

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 $\vec{a}$ acted in the direction of the velocity, causing a change in the magnitude of the velocity $\vec{v}$.

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

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 1.12 (page 75) 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 ~~, formulasor derivations~~and formulas, let's demonstrate how calculus is used in a real-world example.

use the variables u, t, and τ to denote the inputs. The function's output is denoted $f(x)$ and is usually identified with the y-coordinate in graphs.

The *derivative* function, denoted $f'(x)$, $\frac{d}{dx}f(x)$, $\frac{df}{dx}$, or $\frac{dy}{dx}$, 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 $(x_0, f(x_0))$.

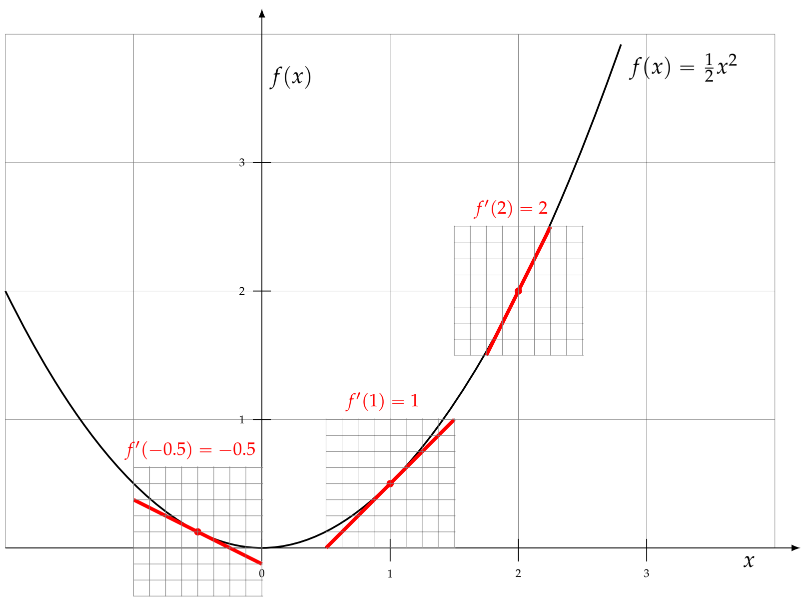


Figure 5.2: The diagram illustrates how to compute the derivative of the function $f(x) = \frac{1}{2}x^2$ at three different points on the graph of the function. To calculate the derivative of $f(x)$ at $x = 1$, we can "zoom in" near the point $(1, \frac{1}{2})$ and draw a line that has the same slope 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 $x = -\frac{1}{2}$ 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~~ Is there a general pattern that describes the slope of the graph for all x?

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, takes as input a function $f(x)$ and produces as output the derivative function $f'(x)$. The derivative operator notation is useful because it ~~show the derivative is an operation you do to a function :~~ makes it clear that differentiation is an *operation* you can apply to any function to obtain its derivative:

$$f'(x) \cancel{\equiv} = \frac{d}{dx} f(x).$$

The derivative operator $\frac{d}{dx}$ acts on the original function $f(x)$ to produce the derivative function $f'(x)$, which describes the rate of change of f for all x. Applying the derivative operator to a function is also called "taking the derivative" of a function.

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~~ We'll formally define the derivative operation in Section 5.6. Afterward, we'll develop various 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 364), you'll find a table of formulas for taking the derivatives of common functions. In Section 5.8, we'll learn the ~~basic~~ general rules for computing derivatives of sums, products, and compositions of ~~the basic~~ functions.

Sequences

So far, we've 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 ~~$\mathbb{N}_+ = \{1, 2, 3, 4, \ldots\}$~~ $\mathbb{N}^* = \{1, 2, 3, 4, \ldots\}$ 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 the 3^{rd}, the 4^{th}, and the 5^{th} elements ~~of~~ in the sequence a_n, we turn to summation notation:

$$a_3 + a_4 + a_5 \equiv = \sum_{3 \leqslant n \leqslant 5} a_n \equiv = \sum_{n=3}^{5} a_n.$$

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:

$$\sum a_n \cancel{\equiv} = S_\infty = \lim_{N\to\infty} S_N = \sum_{n=0}^{\infty} a_n = a_0 + a_1 + a_2 + a_3 + a_4 + \cdots.$$

Note this is an infinite sum.

Series 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 ~~$S_\infty \equiv \sum_{n\in\mathbb{N}} a_n$~~ $S_\infty = \sum_{n\in\mathbb{N}} a_n$ equals some finite number $L \in \mathbb{R}$~~.~~:

$$S_\infty = \sum_{n=0}^{\infty} a_n = L \qquad \Rightarrow \qquad \text{the series } \sum a_n \text{ converges.}$$

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~~ $S_\infty = \sum_{n\in\mathbb{N}} a_n$ does not converge, we say the series $\sum a_n$ *diverges*.

$$S_\infty = \sum_{n=0}^{\infty} a_n = \pm\infty \qquad \Rightarrow \qquad \text{the series } \sum a_n \text{ diverges.}$$

Examples of divergent series include series that "blow up" to infinity or negative infinity, or series that oscillate between different values and fail to "settle down" close to a single value L.

The main series technique you need to learn is how to spot the differences between series that converge and series that diverge. You'll learn how to perform different *convergence 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:

$$e = \sum_{n=0}^{\infty} \frac{1}{n!} = 1 + 1 + \frac{1}{2\cdot 1} + \frac{1}{3\cdot 2} + \frac{1}{4\cdot 3\cdot 2} + \frac{1}{5\cdot 4\cdot 3\cdot 2} + \cdots.$$

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 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~~ $N+1$ 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~~ accurate to 15 decimals, the computer scientist uses the parameter $N = 19$ ~~terms in the series:~~ in the general formula, and computes the summation

$$e_{19} = \sum_{n=0}^{19} \frac{1}{n!} = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{19!}\,\cancel{.}$$

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: $|e_{19} - e| \leqslant 10^{-15}$.

When the mathematician asks for a precision of $\epsilon' = 10^{-25}$, the computer ~~scientists takes~~ scientist sets the parameter to $N = 26$ ~~terms in the series~~ ~~to produce~~ formula:

$$e_{26} = \sum_{n=0}^{26} \frac{1}{n!} = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{19!} + \cdots + \frac{1}{26!}\,,$$

which satisfies $|e_{26} - e| \leqslant \epsilon'$. In the third step, the mathematician demands a precision $\epsilon'' = 10^{-50}$, and the ~~CS student computes~~ computer scientist uses $N = 42$ ~~terms in the series~~in her formula, to produce an approximation ~~satisfying~~ e_{42} that satisfies $|e_{42} - e| \leqslant \epsilon''$.

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

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~~ In words, this means that a function $f(x)$ is continuous at $x = a$ if the limit from the left $\lim_{x\to a^-} f(x)$ and the limit from the right $\lim_{x\to a^+} f(x)$ are both equal to the value of the function at $x = a$.

Take a moment to think about the mathematical definition of continuity ~~given in the equation above~~at a point. Can you ~~see how it connects~~ connect the math definition to the intuitive idea ~~of continuous functions as functions that~~ that functions are continuous if they 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 $f : \mathbb{R}\backslash\{0\} \to \mathbb{R}$ defined by

$$f(x) = \frac{|x-3|}{x-3} = \begin{cases} 1 & \text{if } x > 3, \\ -1 & \text{if } x < 3. \end{cases}$$

~~This function~~ The function f is continuous everywhere on the real line except at $x = 3$. Since this function f is "missing" only at a single point, we can try to "patch it" by filling in the missing value. Consider the function $g : \mathbb{R} \to \mathbb{R}$ defined as

$$g(x) = \begin{cases} 1 & \text{if } x > 3, \\ 1 & \text{if } x = 3, \\ -1 & \text{if } x < 3. \end{cases}$$

The function g is *continuous from the right* at the point $x = 3$, since ~~$\lim_{x\to 3^+} f(x) = 1 = f(3)$~~$\lim_{x\to 3^+} g(x) = 1 = g(3)$. However, taking the limit from the left, we find ~~$\lim_{x\to 3^-} f(x) = -1 \neq f(3)$. Therefore, the function~~ $\lim_{x\to 3^-} g(x) = -1 \neq g(3)$, which tells us g is not continuous ~~. The function $f(x)$ is continuous everywhere on the real line except~~ from the left. We say the function g has a *jump discontinuity* at $x = 3$.

Example 3 We can calculate the limit $\lim_{x\to 5} \dfrac{2x+1}{x}$ as follows:

$$\lim_{x\to 5} \frac{2x+1}{x} = \frac{2(5)+1}{5} = \frac{11}{5}.$$

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*:

$$\lim_{x\to a} \frac{f(x)}{g(x)} \overset{\text{H.R.}}{=} \lim_{x\to a} \frac{f'(x)}{g'(x)}.$$

You can find the derivative formulas you'll need for using L'Hopital's rule in the table of derivative formulas on page 364.

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 364, 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

[Visual explanation of the ϵ,δ-game for limits and L'Hopital's rule]
https://www.youtube.com/watch?v=kfF40MiS7zA&t=523

[See the Wikipedia page for more examples of limits]
https://en.wikipedia.org/wiki/Limit_of_a_function

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 the simple harmonic motion equation for a pendulum on page 315. 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~~ common 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.~~, so you'll get to know them very well.

$$\begin{array}{rcl}
f(x) & -\text{ derivative} \rightarrow & f'(x) \\
a & -\frac{d}{dx}\rightarrow & 0 \\
\alpha f(x) + \beta g(x) & -\frac{d}{dx}\rightarrow & \alpha f'(x) + \beta g'(x) \\
x & -\frac{d}{dx}\rightarrow & 1 \\
x^n & -\frac{d}{dx}\rightarrow & nx^{n-1} \\
\frac{1}{x} = x^{-1} & -\frac{d}{dx}\rightarrow & \frac{-1}{x^2} = -x^{-2} \\
\sqrt{x} = x^{\frac{1}{2}} & -\frac{d}{dx}\rightarrow & \frac{1}{2\sqrt{x}} = \frac{1}{2}x^{-\frac{1}{2}} \\
e^x & -\frac{d}{dx}\rightarrow & e^x \\
a^x & -\frac{d}{dx}\rightarrow & a^x \ln(a) \\
\ln(x) & -\frac{d}{dx}\rightarrow & \frac{1}{x} \\
\log_a(x) & -\frac{d}{dx}\rightarrow & (x\ln(a))^{-1} \\
\sin(x) & -\frac{d}{dx}\rightarrow & \cos(x) \\
\cos(x) & -\frac{d}{dx}\rightarrow & -\sin(x) \\
\tan(x) & -\frac{d}{dx}\rightarrow & \sec^2(x) = \cos^{-2}(x)
\end{array}$$

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:

$$\begin{aligned} &f(x+\delta)g(x+\delta)\overbrace{-f(x)g(x+\delta)+f(x)g(x+\delta)}^{=0} - f(x)g(x) \\ &= [f(x+\delta)-f(x)]g(x+\delta) + f(x)[g(x+\delta)-g(x)]. \end{aligned}$$

The expression for the derivative of the product becomes

$$[f(x)g(x)]' = \lim_{\delta\to 0}\left\{\frac{[f(x+\delta)-f(x)]}{\delta}g(x+\delta) + f(x)\frac{[g(x+\delta)-g(x)]}{\delta}\right\}.$$

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:

$$[f(x)g(x)]' = f'(x)g(x) + f(x)g'(x).$$

~~Proving the correctness of the chain rule for derivatives is a bit more complicated. Actually, it is *a lot* more complicated. The argument presented in the next section is the most technical part of this book, and it's totally fine if you're not able to follow all the details. It's my duty as your calculus teacher to prove to you that the formula~~ $\cancel{[f(g(x))]' = f'(g(x))g'(x)}$ ~~is correct, but the proof is included only for readers who insist on seeing the full, excruciating details. Other readers should feel free to skip the next section and continue reading on page **??**.~~

Derivation of the chain rule

Shortened the proof to skip the boring details...

Assume $f(x)$ and $g(x)$ are differentiable functions. We want to show that the derivative of $f(g(x))$ equals $f'(g(x))g'(x)$, which is the chain rule for derivatives~~:~~.

$$[f(g(x))]' = f'(g(x))g'(x).$$

~~Before we begin, I'd like to comment on the notation used to define derivatives. I happen to like the Greek letter δ (lowercase *delta*), so I defined the derivative of $f(x)$ as~~

$$f'(x) = \lim_{\delta \to 0} \frac{f(x+\delta) - f(x)}{\delta}.$$

~~Instead, we could also use the variable Δ (uppercase *delta*) and write~~

$$f'(x) \equiv \lim_{\Delta \to 0} \frac{f(x+\Delta) - f(x)}{\Delta}.$$

~~In fact~~Using the definition of the derivative, we can ~~use *any* variable for the limit expression. All that matters is that we divide by the *same* non-zero quantity as~~ write the derivative of the function $f(g(x))$ as follows:

$$[f(g(x))]' = \lim_{\delta \to 0} \frac{f(g(x+\delta)) - f(g(x))}{\delta}.$$

The next step in the proof is to split the expression $\frac{f(g(x+\delta))-f(g(x))}{\delta}$ into two factors F_1 and F_2, that will later turn into the ~~quantity added to x inside the function, and that this quantity goes to zero. If we're not careful with our choice of limit variable we could run into trouble. Specifically, the definition of a limit depends on a "small, nonzero number Δ," which is then used in the limit $\Delta \to 0$. The condition $\Delta \neq 0$ is essential because the expression $\frac{f(x+\Delta)-f(x)}{\Delta}$ is not well defined when $\Delta = 0$, since it leads to a~~ factors $f'(g(x))$ and $g'(x)$ in the chain rule formula, after we take the limit:

$$\frac{f(g(x+\delta)) - f(g(x))}{\delta} = \underbrace{F\big(g(x), g(x+\delta)\big)}_{F_1} \underbrace{\frac{g(x+\delta) - g(x)}{\delta}}_{F_2}.$$

The second term F_2 corresponds to the rise-over-run calculation for the function g at the input x. The first term F_1 corresponds to the rise-over-run calculation for the function f at the input $g(x)$, but we'll introduce a new quantity $F(a,b)$ in order to handle the case of the zero-run calculation correctly. We define $F(a,b)$ as follows:

$$F(a,b) \stackrel{\text{def}}{=} \begin{cases} \frac{f(b)-f(a)}{b-a} & \text{if } a \neq b, \\ f'(a) & \text{if } a = b. \end{cases}$$

In words, the quantity $F(a,b)$ tells us the rise-over-run calculation for the function f computed between the points $(a, f(a))$ and $(b, f(b))$.

We've just "patched" the formula so it doesn't result in a divide-by-zero ~~error.~~

~~In order to avoid any possibility of such errors, we define the following piecewise function:~~

$$R(y,b) \equiv \begin{cases} \frac{f(y)-f(b)}{y-b} & \text{if } y \neq b, \\ f'(b) & \text{if } y = b. \end{cases}$$

~~Observe the function $R(y,b)$ is continuous in y, when wetreat b as a constant. This follows from the definition of the derivative formula and the assumption that $f(x)$ is differentiable. Using the function $R(y,b)$, we can write the formula for the derivative of $f(x)$ as $f'(x) = \lim_{\Delta\to 0} R(x+\Delta, x)$. Note this formula is valid even~~ error in the case ~~$\Delta = 0$~~$a = b$.

~~To prove the chain rule~~For the sake of brevity, we'll ~~need the function $R(g(x+\delta), g(x))$, which is defined as follows:~~

$$R\big(g(x+\delta), g(x)\big) \equiv \begin{cases} \frac{f(g(x+\delta))-f(g(x))}{g(x+\delta)-g(x)} & \text{if } g(x+\delta) \neq g(x), \\ f'(g(x)) & \text{if } g(x+\delta) = g(x). \end{cases}$$

~~Okay, we're done with the preliminaries, so we can get back to proving the chain rule, $[f(g(x))]' = f'(g(x))g'(x)$. We start with the limit expression for the left-hand side of the equation:~~

$$[f(g(x))]' = \lim_{\delta\to 0} \frac{f(g(x+\delta)) - f(g(x))}{\delta}.$$

~~Observe that the fraction inside the limit can be written as~~

$$\frac{f(g(x+\delta)) - f(g(x))}{\delta} = R\big(g(x+\delta), g(x)\big)\frac{g(x+\delta)-g(x)}{\delta}.$$

~~This is the trickiest part of the proof, so let's analyze carefully why this equation holds. We must check that the equation holds in the two special cases in the definition of $R(g(x+\delta), g(x))$.~~

~~**Case A** Whenever $g(x+\delta) \neq g(x)$, we have:~~

$$\begin{aligned} \frac{f(g(x+\delta)) - f(g(x))}{\delta} &= \frac{f(g(x+\delta)) - f(g(x))}{\delta}\frac{g(x+\delta)-g(x)}{g(x+\delta)-g(x)} \\ &= \frac{f(g(x+\delta)) - f(g(x))}{g(x+\delta)-g(x)}\frac{g(x+\delta)-g(x)}{\delta} \\ &= R\big(g(x+\delta), g(x)\big)\frac{g(x+\delta)-g(x)}{\delta}. \end{aligned}$$

Case B ~~For points where $g(x+\delta) = g(x)$, we have:~~

$$\frac{f(g(x+\delta)) - f(g(x))}{\delta} = \frac{0}{\delta} = 0,$$

skip the detailed analysis of the function $F(a,b)$, and the calculations of the limits of the two factors $\lim_{\delta\to 0} F_1 = f'(g(x))$ and

$$R\big(g(x+\delta), g(x)\big)\frac{g(x+\delta) - g(x)}{\delta} = f'(g(x))\frac{0}{\delta} = 0.$$

~~Thus, the equation $\frac{f(g(x+\delta))-f(g(x))}{\delta} = R\big(g(x+\delta), g(x)\big)\frac{g(x+\delta)-g(x)}{\delta}$ holds in both cases.~~

~~We can now rewrite the limit expression for $[f(g(x))]'$ using the equation established above:~~

$$[f(g(x))]' = \lim_{\delta\to 0} \frac{f(g(x+\delta)) - f(g(x))}{\delta}$$

$$= \lim_{\delta\to 0}\left(\underbrace{R\big(g(x+\delta), g(x)\big)}_{F_1}\underbrace{\frac{g(x+\delta)-g(x)}{\delta}}_{F_2}\right).$$

~~We're trying to evaluate a limit expression that is~~ $\lim_{\delta\to 0} F_2 = g'(x)$. Let's jump directly to the final steps in the proof, when we evaluate $\lim_{\delta\to 0} F_1F_2$. In Section 5.5 (see page 359) we learned that the limit of the product of two factors ~~;~~ $\lim_{\delta\to 0} F_1F_2$ ~~. The limit of a product exists if~~ is equal to the product of the limits of ~~both factors—~~the individual factors $\lim_{\delta\to 0} F_1$ and $\lim_{\delta\to 0} F_2$~~—exist. Before we proceed, we must evaluate the limit $\delta \to 0$ for both factors to ensure the limits exist.~~ Using this property of limits we find:

$$\lim_{\delta\to 0} F_1F_2 = \big(\lim_{\delta\to 0} F_1\big)\big(\lim_{\delta\to 0} F_2\big) = f'(g(x))\, g'(x),$$

which completes the proof of the chain rule $[f(g(x))]' = f'(g(x))g'(x)$.

~~To obtain the limit of the first factor, we'll rely on the continuity of the functions $g(x)$ and $R(y,b)$:~~

$$\lim_{\delta\to 0} g(x+\delta) = g(x) \quad\text{and}\quad \lim_{\Delta\to 0} R\big(b+\Delta, b\big) = R\big(b,b\big) = f'(b).$$

~~We define the quantity $\Delta = g(x+\delta) - g(x)$ and using the continuity of $g(x)$, we can establish $\Delta \to 0$ as $\delta \to 0$. We are therefore allowed~~

~~to change the limit variable from δ to Δ, and evaluate the limit of the first factor as follows:~~

$$\lim_{\delta\to 0} F_1 = \lim_{\delta\to 0} R\big(g(x+\delta), g(x)\big)$$

$$= \lim_{\Delta\to 0} R\big(g(x)+\Delta, g(x)\big)$$

$$= R\big(g(x), g(x)\big) = f'(g(x)).$$

~~We also know the limit of the second factor exists because it corresponds to the derivative of $g(x)$:~~

$$\lim_{\delta\to 0} F_2 = \lim_{\delta\to 0} \frac{g(x+\delta)-g(x)}{\delta} = g'(x),$$

~~and, since we assumed $g(x)$ is differentiable, its derivative must exist~~Readers interested in learning the technical details of the parts of the proof we skipped can watch the video tutorial below.

~~Since the limits of both factors—$\lim_{\delta\to 0} F_1$ and $\lim_{\delta\to 0} F_2$—exist and are well defined, we can now complete the proof :~~

$$[f(g(x))]' = \lim_{\delta\to 0}\left(R\big(g(x+\delta), g(x)\big)\frac{g(x+\delta)-g(x)}{\delta}\right)$$

$$= \left(\lim_{\delta\to 0} R\big(g(x+\delta), g(x)\big)\right)\left(\lim_{\delta\to 0} \frac{g(x+\delta)-g(x)}{\delta}\right)$$

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

~~This establishes the validity~~ [Technical details of the proof of the chain rule ~~$[f(g(x))]' = f'(g(x))g'(x).$~~]
https://www.youtube.com/watch?v=ydjj0crm34w

~~Alternate~~ Alternative notation

The presence of ~~so~~ many primes and brackets can make derivative formulas difficult to read. As an alternative, we sometimes use the *Leibniz notation* for derivatives. The three rules of derivatives in Leibniz notation are written as follows:

- Linearity: ~~$\frac{d}{dx}(\alpha f(x) + \beta g(x)) = \alpha\frac{df}{dx} + \beta\frac{dg}{dx}$~~ $\frac{d}{dx}(\alpha f(x) + \beta g(x)) = \alpha\frac{df}{dx}$

- Product rule: $\frac{d}{dx}(f(x)g(x)) = \frac{df}{dx}g(x) + f(x)\frac{dg}{dx}$
- Chain rule: ~~$\frac{d}{dx}(f(g(x))) = \frac{df}{dg}\frac{dg}{dx}$~~ $\frac{d}{dx}(f(g(x))) = \frac{df}{dg}\frac{dg}{dx}$

Some authors prefer the notation $\frac{df}{dx}$ for the derivative of $f(x)$ because it is more evocative of a rise-over-run calculation.

Links

[Geometric explanations of derivative formulas by 3Blue1Brown]
`https://www.youtube.com/watch?v=S0_qX4VJhMQ`

[Visual explanations of the chain and product rules for derivatives]
`https://www.youtube.com/watch?v=YG15m2VwSjA`

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

$$f''(x) \equiv \stackrel{\text{def}}{=} \left[f'(x)\right]' \equiv = \frac{d}{dx}f'(x) \equiv = \frac{d^2}{dx^2}f(x).$$

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

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 $f^{(4)}(x) = f(x)$.

Links

[Visual explanation of the second derivative by 3Blue1Brown]
https://www.youtube.com/watch?v=BLkz5LGWihw

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.

Example

The boss of a large drug organization has recently run into problems with the authorities. The more drugs he sells, the more money he makes; but if he sells too much, the authorities will start to regulate his operations and he loses money. When you're in the drug business, the last thing you want is to attract undue attention!

Fed up with this situation, he decides to find the *optimal* amount of drugs to push: as much as possible, but not enough to run into trouble with the law. One day he tells all his advisors and underbosses to leave the room, he picks up a pencil and a piece of paper, takes a deep breath, and sits down to do some calculus.

- *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 $f(x) = x^3$ 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 $f(x^*) = M$.

Algorithm for finding extrema

Input: a function $f(x)$ and a constraint region $C = [x_i, x_f]$
Output: the locations and values 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~~ There can be multiple solutions. Make a list of them. We'll call this the list of *candidates*.

4. For each candidate x^* in the list, check to see whether it is a maximum, a minimum, or a saddle point:

   - If $f'(x^* - 0.1)$ is positive and $f'(x^* + 0.1)$ is negative, then the point x^* is a maximum. 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 minimum. 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

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 this implicit function $y(x)$. After isolating $\frac{dy}{dx}$, we obtain an expression that describes the slope of the circle at any point $P = (x_P, y_P)$. You can check that the slope predicted for the point at top of the circle $(0, R)$ is zero. Also note the slope is infinite at $(R, 0)$ since the tangent to the circle is vertical at that point.

Let's now look at an example involving implicit differentiation.

Example In the corporate world, an executive officer's ego E is related to the executive's salary S by the following equation:

$$E^2 = S^3.$$

Suppose both E and S are functions of time. What is the rate of change of the executive's ego when the executive's salary is 60k and the salary increases at a rate of 5k per year?

This is called a *related rates* problem. We know the relation $E^2 = S^3$ and the rate $\frac{dS}{dt} = 5000$ and we're asked to find the rate of change $\frac{dE}{dt}$ when $S = 60000$. First, take the implicit derivative of the salary-to-ego relation:

$$\frac{d}{dt}\left[E^2 \right] = \frac{d}{dt}\left[S^3 \right],$$
$$2E\frac{dE}{dt} = 3S^2\frac{dS}{dt}.$$

We're interested in the point where $S = 60000$. To find the ego points, solve for E in the relation $E^2 = S^3$; $E = \sqrt{60000^3} = 14696938.46$ ego points when $S = 60000$. Substituting all known values into the derivative of the relation, we find

$$2(14696938.46)\frac{dE}{dt} = 3(60000)^2(5000).$$

The executive'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 .~~ What are they good for again?

Total derivative

Consider again a relation $g(x, y) = 0$, but this time assume that both x and y are implicit functions of a third variable t. To compute the

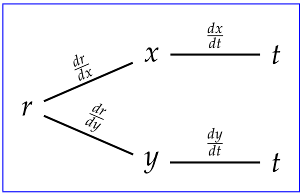


Figure 5.14: Computing the total derivative of $r(x(t), y(t))$ with respect to t.

TEXT FIX

When $t = 6$, ~~$x = 13$~~ $x = 121$ and $y = 207$, and the answer is $\frac{dr}{dt} = 65.346$.

Differentials

The *differential* of a quantity Q is the same as the derivative but without specifying the "with respect to" variable. The differential dQ represents the change in Q that will result for a given change in the variable (or variables) that Q depends on. The rules for computing differentials are analogous to the rules for computing derivatives:

$$Q = ax^n \qquad \Rightarrow \qquad dQ = nx^{n-1}dx.$$

You can think of differentials as incomplete derivatives: if we later discover that x depends on t, we can divide both sides of the above equation by dt to obtain the derivative expression $\frac{dQ}{dt} = nx^{n-1}\frac{dx}{dt}$.

Application of differentials to computing 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 in Figure 5.15. 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.

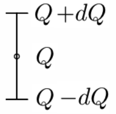


Figure 5.15: The error bars dQ are a visual representation of the uncertainty of the quantity Q.

Suppose the quantity Q depends on the variables x and y. We can express the dependence between the error in the measurement of Q

and the error in the measurement of x and y as the formula:

$$dQ = \frac{dQ}{dx}dx + \frac{dQ}{dy}dy.$$

This is the *total differential* of Q. Note the similarity of the total differential formula to the total derivative formula.

Example You want to calculate the kinetic energy of a particle using the formula $K = \frac{1}{2}mv^2$. You measure the particle's mass m with precision 3%, and the particle's velocity with precision 2%. What is the precision of your kinetic energy calculation?

We want to find $\frac{dK}{K}$ and we're told $\frac{dm}{m} = 0.03$ and $\frac{dv}{v} = 0.02$. The first step is to calculate the *total differential* of 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 used the product rule and the chain rule for derivatives. To obtain the relative error, divide both sides by K to obtain

$$\frac{dK}{K} = \frac{\frac{1}{2}v^2\,dm \;+\; mv\,dv}{\frac{1}{2}mv^2}\frac{\frac{1}{2}v^2\,dm \;+\; mv\,dv}{\frac{1}{2}mv^2} = \frac{dm}{m} + 2\frac{dv}{v}.$$

The precision of the kinetic energy calculation in your experiment 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, since it appears with exponent two in the formula.

Links

[Visual explanation of implicit differentiation by 3Blue1Brown]
https://www.youtube.com/watch?v=qb40J4N1fa4

Discussion

We have reached the half-point of the calculus chapter. We learned about derivatives and described applications of derivatives to optimization problems, finding tangent lines, related rates, etc.

Before you continue reading about integrals in the second half of the chapter, I highly recommend you attempt to solve some of the derivative problems starting on page 462.

Another thing I would recommend is to watch some of the lectures of the *Highlights of Calculus* course by Prof. Gilbert Strang (see https://youtube.com/playlist?list=PLBE9407EA64E2C318).

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

$$F(x) = x^4 \quad \xrightarrow{\frac{d}{dx}} \quad F'(x) = 4x^3 = f(x),$$

then it must be that

$$f(x) = 4x^3 \quad \xrightarrow{\int dx} \quad F(x) = \int 4x^3\,dx = x^4 + C.$$

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

  $$F(x) = \int \sin\theta\,d\theta = -\cos\theta + C,$$

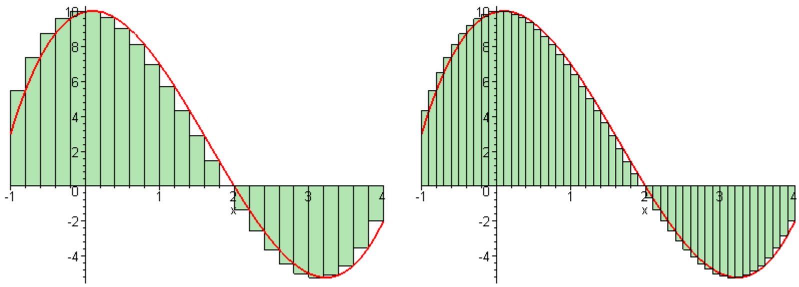
  since $\frac{d}{d\theta}[-\cos\theta] = \sin\theta$.
- The integral of $f(x) = x^n$ for any number $n \neq -1$ is

  $$F(x) = \int x^n\,dx = \frac{1}{n+1}x^{n+1} + C,$$

  since $\frac{d}{d\theta}x^n = nx^{n-1}$.
- The integral of $f(x) = x^{-1} = \frac{1}{x}$ is

  $$F(x) = \int \frac{1}{x}\,dx = \ln|x| + C, \quad \text{for } x \neq 0.$$

  When $x > 0$, we know $\int \frac{1}{x}\,dx = \ln x + C$, since $\frac{d}{dx}\ln x = \frac{1}{x}$. For $x < 0$, we can use the symmetry in the graph of $f(x) = \frac{1}{x}$ to obtain the formula $\int \frac{1}{x}\,dx = \ln(-x) + C$. The absolute value lets us combine these two special cases into a single formula.



(a) $n = 25$ (b) $n = 50$

Figure 5.19: 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~~ shown in Figure 5.20 corresponds to the approximation $S_{100}(a, b) = 12.7906$.

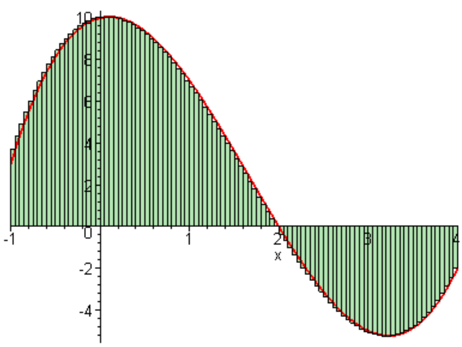


Figure 5.20: An approximation of the area under the function $f(x) = x^3 - 5x^2 + x + 10$ between $x = -1$ and $x = 4$ using $n = 100$ rectangles.

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 becomes *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:

$$\int_a^b f(x)\,dx \equiv \lim_{n\to\infty} \sum_{k=1}^{n} f(a + k\Delta x)\Delta x \equiv A(a,b).$$

$$\int_a^b f(x)\,dx \stackrel{\text{def}}{=} \lim_{n\to\infty} \sum_{k=1}^{n} f(a + k\Delta x)\Delta x = A(a,b).$$

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 infinitely 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
>>> xk = -1 + k*5/n
>>> sk = (xk**3-5*xk**2+xk+10)*(5/n)
>>> summation( sk, (k,1,n) ).evalf()
12.9166 541666563
>>> integrate( x**3-5*x**2+x+10, (x,-1,4) ).evalf()
12.9166 666666667
```

Formal definition of the integral

~~We rarely compute integrals using Riemann sums.~~ The Riemann sum is important as a *theoretical construct* like the rise-over-run calculation that we use to define the derivative operation:

$$f'(x) \equiv \stackrel{\text{def}}{=} \lim_{\delta\to 0} \frac{f(x+\delta) - f(x)}{\delta}.$$

The integral is defined as the approximation of the area under the curve with infinitely many rectangles:

$$\int_a^b f(x)\,dx \stackrel{\text{def}}{=} \lim_{n\to\infty}\sum_{k=1}^{n} f(a+k\Delta x)\Delta x, \quad \Delta x = \frac{b-a}{n}.$$

It is usually much easier to refer to a table of derivative formulas (see page 364) 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 364), rather than computing the 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

[A Riemann sum demonstration]
https://www.geogebra.org/m/jF23GzmS

[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 $A_0(x)$ that corresponds to the calculation of the area under $f(x)$ starting from $x = 0$:

$$A_0(x) \stackrel{\text{def}}{=} \int_0^x f(t)\,dt.$$

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, there 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.

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?~~

Links

[Nice visual explanations about integrals by 3Blue1Brown]
https://www.youtube.com/watch?v=rfG8ce4nNh0
https://www.youtube.com/watch?v=FnJqaIESC2s

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 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 course. 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 $u = x$~~a trivial substitution like $u = x + a$, where a is a constant.

To find the relation between the small step du and the small step dx, 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 ~~$\ln(u)$~~$\ln|u| + C$. Accounting for the -1 horizontal shift and the factor of 2 in the numerator, we obtain the answer:

$$\int \frac{1}{x-\sqrt{x}}\, dx = \int \frac{2}{u-1}\, du = 2\ln(|u-1|) = 2\ln\big(\big|\sqrt{x}-1\big)\big|.$$

Note in the last step, we changed back to the ~~x variable to give the final answer~~variable x. The variable u exists only in our calculation. We ~~invented it~~ invented u out of thin air when we said, "Let $u = \sqrt{x}$" in the beginning, so we must convert back to the original variable x when reporting the final answer.

Thanks to the substitution, the integral became simpler: we were able to eliminate the square roots. The extra u that came from the expression $dx = 2u\, du$ ~~canceled~~ cancelled with one of the us in the denominator, thus making the expression even simpler. In practice, substituting x with u 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

$$\int_4^9 \frac{1}{x-\sqrt{x}}\, dx = \int_2^3 \frac{2}{u-1}\, du = 2\ln(|u-1|)\Big|_2^3 = 2(\ln(2) - \ln(1)) = 2\ln(2).$$

Let's recap. Substitution involves three steps:

1. Replace all occurrences of $u(x)$ with u.

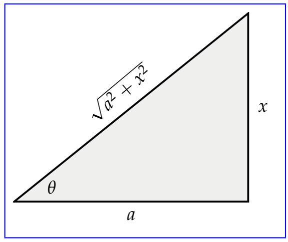


Figure 5.24: Triangle associated with the substitution $x = a\tan\theta$.

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

$$\frac{1}{y^2-6y+10}\,?$$

TEXT FIX

How do we integrate a this function? If there were no ~~$-2y$~~ $-6y$ 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.

Which substitution to use

There are three possible triangles you might need when applying the trigonometric substitution technique to compute an integral. See Figure 5.22, Figure 5.24, and Figure 5.25. In all cases, one side of the triangle corresponds to the constant a, another side corresponds to the variable x, and the length of the third side is a square-root expression involving a^2 and x^2. I wouldn't recommend trying to memorize the sides of these triangles. Instead, you can rely on your knowledge of trigonometry to choose the appropriate labels using trial and error. Draw a little triangle and label its sides so that the square-root expression corresponds to the integral you're trying to compute.

Interlude

By now, ~~things are~~ it's starting to get pretty ~~tight~~ difficult for your calculus teacher ~~. You are beginning to understand how to "handle" any kind of integral he~~ to surprise you with an integral problem. You know how to handle most kinds of integrals your teacher can throw at you: polynomials, fractions with x^2 ~~,~~ plus or minus a^2, and square roots. ~~He~~ Your teacher can't even fool you with dirty trigonometric tricks involving sin, cos, and tan substitutions, since you know about these, too. Are there any integrals left that ~~he can drop on the exam to trick you up~~ your teacher can use to trip you up on the exam?

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? How many more techniques could there possibly be? I know all these integration techniques that I've been throwing at you during the last 10 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 of the examples discussed: 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. When faced with a complicated integral prob-

Then apply a trig substitution $y = \sqrt{k}\tan\theta$ to obtain

$$\frac{1}{a}\int \frac{1}{y^2+k}\,dy = \frac{1}{a\sqrt{k}}\tan^{-1}\!\left(\frac{y}{\sqrt{k}}\right) = \frac{1}{a\sqrt{k}}\tan^{-1}\!\left(\frac{x-h}{\sqrt{k}}\right).$$

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:

$$\int \frac{1}{(x+1)(x+2)^2}dx = \ln(|x+1)| - \ln(|x+2)| + \frac{1}{x+2} + C.$$

The partial fractions technique for integrating rational functions is best understood using using a hands-on approach. Try solving the following exercises to see if you can apply the techniques.

will contain the factors required for the integral on the right-hand side of the integration by parts formula.

There's no general rule to follow when choosing which parts of the integral to label u and which parts to label dv. You just have to solve lots of practice problems to get the hang of it. My general advice would be to focus on the dv part. Look for parts of the integral that you know how to integrate and label them dv, then label whatever remains as u. It's very common to get the choice of u and dv wrong on the first attempt. If you apply the integration-by-parts substitutions and end up with an integral calculation that's more complicated than the original integral you started with, it's a sign you need to start over with a different choice of u and dv.

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 $\int x^2e^x\,dx$, we start by choosing the following substitutions:

$$\begin{aligned} u &= x^2 & dv &= e^x\,dx \\ du &= 2x\,dx & v &= e^x. \end{aligned}$$

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~~ , but we still end up with a finite total area.

Definition

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~~ we take the limit as the dummy variable goes 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 that 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 ~~$\mathbb{N} \equiv \{0,1,2,3,\ldots\}$~~ $\mathbb{N} = \{0,1,2,3,\ldots\}$
- ~~$\mathbb{N}_+ = \mathbb{N}\backslash\{0\}$~~ $\mathbb{N}^* = \mathbb{N}\backslash\{0\}$: the set of *strictly positive* natural numbers $\{1,2,3,\ldots\}$. The set ~~$\mathbb{N}_+$~~ $\mathbb{N}^*$ is the same as $\mathbb{N}$, except ~~$\mathbb{N}_+$~~ $\mathbb{N}^*$ 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:

$$\left(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \ldots\right)$$

$$a_n = \frac{1}{n}, \qquad n \in \mathbb{N}_{\cancel{+}}^{*}.$$

More generally, we ~~can define~~ refer to the sequences with terms like $a_n = \frac{1}{n}$, $a_n = \frac{1}{n^2}$, and $a_n = \frac{1}{n^3}$ as p-sequences. In a p-sequence~~in which~~, the index n appears in the denominator raised to the power p~~:~~. The terms in a p-sequence are described by

$$a_n = \frac{1}{n^p}, \qquad n \in \mathbb{N}_{\cancel{+}}^{*}.$$

Ratio convergence

The numbers in the Fibonacci sequence grow indefinitely large ($\lim_{n\to\infty} a_n = \infty$), 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~~

The usual derivative notion is a no-go ~~, because they depend~~ since it depends on the function $f(x)$ being *continuous*, and sequences are only defined for integer values. ~~We *can* take~~ Instead of derivatives, we can compute *finite differences*, which are sequences obtained by subtracting adjacent terms in the sequence. Given the sequence $(a_0, a_1, a_2, a_3, a_4, \ldots)$, the first differences sequence is the sequence $(a_1 - a_0, a_2 - a_1, a_3 - a_2, \ldots)$. Finite differences play a big role in the study of differential equations.

We can also compute 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, ~~, , , and~~ $\boxed{+}$, $\boxed{-}$, $\boxed{\times}$, and $\boxed{\div}$? I can tell you one way to do this; 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 $\ln(2) = 0.693147\ldots$~~ can compute the answer $\ln(2) = 0.69314718\ldots$ to *any* precision.

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)**(n+1)/n
 >>> sum([ an_ln2(n)  for n in range(1,100) ])
0.698172179310195

>>> sum([ an_ln2(n)  for n in range(1,1000) ])
0.6936474305598223

>>> sum([ an_ln2(n)  for n in range(1,1000000) ])
0.6931476805602526
```

Observe how the ~~approximation becomes~~ approximations become more accurate as more terms are ~~added in the sum.~~ used in the summation. The approximation with 100 terms is accurate to two decimals $0.69\ldots$, the approximation computed by summing 1000 terms is accurate to three decimals $0.693\ldots$, and the approximation with $1\,000\,000$ terms is accurate to six digits $0.693147\ldots$.

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

- ~~$\mathbb{N} = \{0, 1, 2, 3, 4, 5, 6, \ldots\}$~~ $\mathbb{N} = \{0, 1, 2, 3, 4, 5, 6, \ldots\}$: the set of natural numbers
- ~~$\mathbb{N}_+ = \mathbb{N}\backslash\{0\} = \{1, 2, 3, 4, 5, 6, \ldots\}$~~ $\mathbb{N}^* = \mathbb{N}\backslash\{0\} = \{1, 2, 3, 4, 5, 6, \ldots\}$: 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:

$$a_3 + a_4 + a_5 + a_6 + a_7 = \sum_{3 \leqslant n \leqslant 7} a_n = \sum_{n=3}^{7} a_n.$$

- $\sum a_n$: the *series* a_n is the sum of all terms in the sequence a_n:

$$S_\infty = \sum_{n=1}^{\infty} a_n = a_1 + a_2 + a_3 + a_4 + a_5 + a_6 + \cdots.$$

- $n!$: the *factorial* function $n! = n(n-1)(n-2)\cdots 3 \cdot 2 \cdot 1$, if $n \geqslant 1$. 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 $c_0 + c_1 x^1 + c_2 x^2 + c_3 x^3 + \ldots$ where the coefficients c_n are chosen so as to encode the properties of the function $f(x)$.

See problem **P5.119** for the derivations of these formulas. The sum of the first N terms in an arithmetic sequence is

$$\sum_{n=1}^{N}(a_0 + nd) = a_0 N + \frac{N(N+1)}{2}d.$$

It~~will be~~'s important to remember these formulas because they can occur in calculus problems. For example, ~~computing~~ these summation formulas are used to find the integral of the function $f(x) = ax^2 + bx + c$ ~~using an infinite Riemann sumrequires these formulas~~by computing the limit of a Riemann sum.

There are many other series whose infinite sum is described by an exact formula. Over the years, mathematicians have come up with various techniques for computing the values of infinite series. Below, you'll find some ~~known formulasfor the sums of certain infinite series~~of these formulas.

The p-*series* is the sum of all the terms in a p-sequence, which are described by the formula $a_n = \frac{1}{n^p}$, where p is the power (the exponent) in the denominator (see page 445). The p-series involving even values of p can be computed exactly:

$$\sum_{n=1}^{\infty}\frac{1}{n^2} = \frac{\pi^2}{6}, \qquad \sum_{n=1}^{\infty}\frac{1}{n^4} = \frac{\pi^4}{90}, \qquad \sum_{n=1}^{\infty}\frac{1}{n^6} = \frac{\pi^6}{945}.$$

Note you're not required to memorize these formulas. They are given here as examples of what is possible.

Other closed-form expressions for infinite series include:

$$\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n^2} = \frac{\pi^2}{12}, \qquad \sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n} = \ln(2), \qquad \sum_{n=1}^{\infty}\frac{1}{4n^2-1} = \frac{1}{2},$$

$$\sum_{n=0}^{\infty}\frac{(-1)^{n}}{2n+1} = \frac{\pi}{4}, \qquad \sum_{n=0}^{\infty}\frac{1}{(2n+1)^2} = \frac{\pi^2}{8}, \qquad \sum_{n=0}^{\infty}\frac{(-1)^{n}}{(2n+1)^3} = \frac{\pi^3}{32}.$$

Again, don't worry about memorizing all these formulas; just think of them as prizes in a trophy case—a representation of some mathematical success stories. Mathematicians experience great pride whenever they manage to make sense of some complicated, infinite sum expression by finding a simple formula to describe its value. In general most infinite series do not have such closed-form expressions, so you can understand mathematicians' excitement and why they'd want to build a trophy case of known formulas. The series formulas shown above are analogous to the "trophy case" of integral formulas on page 547.

Exercises

E5.7 Compute the values of the following summations using the formulas given above.

(a) $\displaystyle\sum_{n=1}^{N} c(a+bn)$ (b) $\displaystyle\sum_{n=1}^{N} c(a+bn)^2$ (c) $\displaystyle\sum_{n=1}^{\infty} \frac{6}{n^2}$

Convergence and divergence of series

Even when we can't compute an exact expression for the infinite sum of a series, it's important to distinguish series that converge from series that do not converge.

We say a series $\sum a_n$ *converges* if the infinite sum ~~$S_\infty = \sum_{n\in\mathbb{N}} a_n$~~ $S_\infty = \sum_{n\in\mathbb{N}} a_n$ 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 $S_\infty = \sum_{n\in\mathbb{N}} a_n$~~ We can also say "$\sum a_n$ is *convergent*," applying the infinite-sum-of-its-terms-is-a-finite-number property as an adjective. The opposite of a convergent series is a *divergent* series, which describes all series that are not convergent. A series can be divergent if it grows to infinity ~~, we say the series $\sum a_n$ *diverges*.~~

$$S_\infty = \sum_{n=0}^{\infty} a_n = \pm\infty \quad \Rightarrow \quad \text{the series } \sum a_n \text{ diverges.}$$

or if it jumps around between numbers. An example of a divergent series is the *harmonic series* $\sum_{n=1}^{\infty} \frac{1}{n} = \infty$. We say the harmonic series is divergent, because its infinite sum doesn't converge to a finite number but keeps growing indefinitely. Another example of a divergent series is $\sum_{n=0}(-1)^n = 1 - 1 + 1 - 1 + \ldots$, whose value alternates between 1 and 0 and never "settles down" around a single limit L, as is required for a convergent series.

Convergence of a series is not the same as convergence of the underlying sequence a_n, which we talked about in the previous section. The calculations with series are completely different. Consider the sequence of partial sums $S_N = \sum_{n=0}^{N} a_n$~~:~~

$$S_0, S_1, S_2, S_3, \ldots,$$

~~where each of the terms in the sequencecorresponds to~~. The terms in this sequence,

$$S_0,\ S_1,\ S_2,\ S_3,\ \ldots,$$

correspond to the following calculations:

$$a_0, \quad a_0+a_1, \quad a_0+a_1+a_2, \quad a_0+a_1+a_2+a_3, \quad \ldots \quad .$$

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 L.

Convergence tests

The main thing you need to know about series are the different *tests* you can perform to check whether a series converges or diverges.

Divergence 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~~

For example, consider the sequence a_n whose limit is some number $\ell \neq 0$. To find the limit of the series $\lim_{N\to\infty}\sum_{n=0}^{N} a_n$, we'll need to compute an infinite sum of ~~non-zero quantities add to a finite number?~~

~~Absolute convergence~~

numbers that are approximately equal to ℓ. 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~~ ℓ is nonzero then the quantity $N\ell$ "blows up" as N goes to infinity.

~~A sequence a_n for which $\sum_n |a_n|$ converges is called~~ Note the condition $\lim_{n\to\infty} a_n = 0$ for the terms is ~~*absolutely convergent*. *A sequence b_n for which $\sum_n b_n$ converges but $\sum_n |b_n|$ diverges is called*~~ *required* for the series $\sum_{n=0}^{\infty} a_n$ to converge, but it isn't ~~*conditionally convergent*~~*sufficient*. There are all kinds of sequences that satisfy $\lim_{n\to\infty} a_n = 0$ that have divergent series $\sum_{n=0}^{\infty} a_n$. In this section, we'll learn about some other tests that can tell us which series are convergent, and which series are divergent.

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 (~~$\lim a_n = 0$~~$\lim_{n\to\infty} a_n = 0$) 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,~~ Consider some function $f : \mathbb{R} \to \mathbb{R}$ that is nonnegative ($f(x) \geqslant 0$) and decreasing ($f(x) > f(x+\delta), \forall \delta > 0$) for all values of x, $x \geqslant 1$. If we evaluate this function only for inputs that are positive integers, then we can think of it as a sequence $f : \mathbb{N}^* \to \mathbb{R}$. The terms of the sequence will be $f(1), f(2), f(3), f(4)$, and so on.

Recall the definition of improper integrals we saw in Section 5.17. The improper integral $\int_1^{\infty} f(x)\,dx$ corresponds to an integral ~~then the series $\sum_n f(n)$ converges. If the integral $\int_a^{\infty} f(x)\,dx$~~ where one of the endpoints goes to infinity:

$$\int_1^{\infty} f(x)\,dx = \lim_{b\to\infty} \int_1^{b} f(x)\,dx.$$

This corresponds to the calculation of the area under $f(x)$ over the whole x-axis, all the way to infinity.

The *integral test* describes a connection between the convergence properties of the improper integral $\int_1^{\infty} f(x)\,dx$, and the convergence of the series $\sum_{n=1}^{\infty} f(n)$. Specifically, if the integral $\int_1^{\infty} f(x)\,dx$ diverges, then the series ~~$\sum_n f(n)$~~ $\sum_n f(n)$ also diverges.

~~The improper integral is defined as a limit expression:~~

$$\cancel{\int_a^{\infty} f(x)\,dx \equiv \lim_{b\to\infty} \int_a^{b} f(x)\,dx.}$$

If the integral $\int_1^\infty f(x)\,dx$ is finite, then the series $\sum_n f(n)$ converges.

The p-series converges if 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 $p \leqslant 1$. Note that $p = 1$ corresponds to the harmonic series $\sum_{n=1}^\infty \frac{1}{n}$ which diverges.

The convergence conditions for p-series, $a_n = \frac{1}{n^p}$, can be obtained using the integral test.

Direct comparison test

Often times we can understand the convergence properties of a series $\sum_n a_n$ by comparing it to another series $\sum_n b_n$ whose convergence properties are known. One approach is to directly compare the values of each term. In particular, we can draw the following conclusions for any two nonnegative sequences a_n and b_n:

- If $a_n \leqslant b_n$ for all n, and $\sum_n b_n$ converges, then $\sum_n a_n$ converges.
- If $a_n \geqslant b_n$ for all n, and ~~$\sum_n b_n$ diverges~~$\sum_n b_n = \infty$, then $\sum_n a_n$ diverges.

The first ~~conclusion~~ point follows from the squeezing principle: since b_n is always above a_n, and $\sum_n b_n$ converges, then so must $\sum_n a_n$. The second ~~conclusion~~ point uses this reasoning in reverse: since $\sum_n b_n = \infty$ and $a_n \geqslant b_n$, then we must also have $\sum_n a_n = \infty$.

Limit comparison test

We can also compare series by comparing the relative size of their n^{th} terms. 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 $\sum_n b_n$ converges, then $\sum_n a_n$ also converges.
- If $L = \infty$ and $\sum_n b_n$ diverges, then $\sum_n a_n$ also diverges.

The n$^{\text{th}}$ 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.

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* 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*.

Taylor series

The *Taylor series* of a function $f(x)$ approximates the function by an infinitely long polynomial:

$$f(x) = \sum_{n=0}^{\infty} c_n x^n = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + c_4 x^4 + \cdots.$$

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

$$\sin(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \frac{x^{11}}{11!} + \cdots.$$

$$\begin{aligned}\sin(x) &= \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1} \\ &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \frac{x^{11}}{11!} + \cdots.\end{aligned}$$

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~~ Taylor 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 $\cancel{c_n = \frac{f^{(n)}(0)}{n!} = \frac{1}{n!}.}$

$$c_n = \frac{f^{(n)}(0)}{n!} = \frac{1}{n!}.$$

The Taylor series of $f(x) = e^x$ is

$$e^x = \sum_{n=0}^{\infty} \frac{1}{n!}x^n = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \cdots$$

$$\begin{aligned} e^x &= \sum_{n=0}^{\infty} \frac{1}{n!}x^n \\ &= 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \cdots \end{aligned}$$

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 $x = a$, and its 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 456 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 (page 237) 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~~https://~~mathforum.org/~~mathimages.swarthmore.edu/index.php/Taylo

[Visual explanation of Taylor series by 3Blue1Brown]
https://www.youtube.com/watch?v=3d6DsjIBzJ4

[Good summary with many interesting examples]
http://en.wikipedia.org/wiki/Series_(mathematics)

[A comprehensive list of important math series]
http://en.wikipedia.org/wiki/List_of_mathematical_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 learned by trying to solve some calculus problems.

Calculus hasn't changed much in the last hundred years. It is testament to this fact that many of the problems presented here were adapted from the book "Calculus Made Easy" by Silvanus Thompson, originally published[1] in 1910. These problems remain as pertinent and interesting today as they were 100 years ago.

As much as calculus is about understanding things conceptually and seeing the big picture (abstraction), calculus is also about practice. There are more than 120 problems to solve in this section. The goal is to turn differentiation and integration into routine operations that you can carry out without stressing out. You should vanquish as many problems as you need to feel comfortable with the procedures of calculus.

Okay, enough prep talk. Let's get to the problems!

Limits problems

P5.1 Use the graph of the function $f(x)$ shown in Figure 5.33 to calculate the following limit expressions:

(1) $\lim_{x \to -5^-} f(x)$ (2) $\lim_{x \to -5^+} f(x)$ (3) $\lim_{x \to -5} f(x)$

(4) $\lim_{x \to 2^-} f(x)$ (5) $\lim_{x \to 2^+} f(x)$ (6) $\lim_{x \to 2} f(x)$

(7) $\lim_{x \to 5^-} f(x)$ (8) $\lim_{x \to 5^+} f(x)$ (9) $\lim_{x \to 5} f(x)$

(10) Is the function $f(x)$ continuous at $x = 5$?

(11) What are the intervals where the function $f(x)$ is continuous?

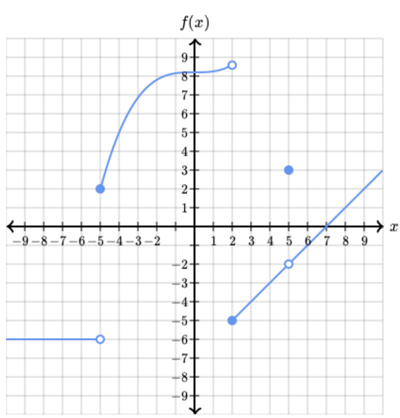


Figure 5.33: ~~The graph of a piecewise-continuous function $f(x)$.~~ The function $f(x)$ has two jump discontinuities at $x = -5$ and $x = 2$ and one removable discontinuity at $x = 5$.

[1] Full text is available at `http:/gutenberg.org/ebooks/33283` (public domain).

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 if you find a mistake or to let me know what you thought. You can reach me by email at `ivan@minireference.com`.

If you want to learn about other books in the ~~No bullshit guideseries and~~ **No Bullshit Guide** series or hear about the technology we're using at ~~`Minireference Publishing`~~ Minireference Co. to take over the textbook industry, check out the company blog at `minireference.com/blog/`. You can also find us on the twitter `@minireference` and on the facebook `fb.me/noBSguide`.

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 everything, 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

$\alpha T(\vec{v}_1) + \beta T(\vec{v}_2)$. Using the standard notation for functions, we write $T(\vec{x}) = \vec{y}$ to show the linear transformation T acting on an input vector $\vec{x} \in \mathbb{R}^n$ to produce the output vector $\vec{y} \in \mathbb{R}^m$. Every linear transformation T can be *represented* as a matrix $A_T \in \mathbb{R}^{m \times n}$, which is an array of numbers with m rows and n columns. Computing $T(\vec{x})$ is equivalent to computing the matrix-vector product $A_T\vec{x}$. Because of the equivalence between linear transformations and matrices, we can also say that linear algebra is the study of vectors and matrices.

Vectors and matrices are used all over the place! If your knowledge of high-school math gave you modelling superpowers, then linear algebra is the vector-upgrade that teaches you how to build models in multiple dimensions.

[VIDEO LECTURES] Gilbert Strang. *Linear Algebra*, MIT OpenCourseWare, 2010, online: `http://bit.ly/StrangLAlectures`.

[BOOK] Ivan Savov. *No bullshit guide to linear algebra*, Minireference Publishing, Second edition, 2017, ISBN 978-0-9920010-2-5.

General mathematics

Mathematics is a hugely broad field. There are all kinds of topics to learn about; some of them are fun, some of them are useful, and some of them are totally mind expanding.

The following ~~books~~ resources cover math topics of general interest and serve as a great overview of all areas of mathematics. I highly recommend you take a look at ~~both books for some easy and enlighteningreading~~these for further math enlightening.

[VIDEO] A map of mathematics that shows all the subfields of mathematics and their objects of study: `https://youtu.be/OmJ-4B-mS-Y`.

[VIDEOS] Video interviews and lessons by some of the best math educators in the world: `https://youtube.com/user/numberphile`.

[BOOK] Richard Elwes. *Mathematics 1001: Absolutely Everything*, Firefly Books, 2010, ISBN 1554077192.

[BOOK] Alfred North Whitehead. *An Introduction to Mathematics*, Williams & Norgate, 1911, `www.gutenberg.org/ebooks/41568`.

Probability

Probability distributions are a fundamental tool for modelling non-deterministic behaviour. A discrete random variable X is associated with a probability mass function $\cancel{p_X(x) = \Pr(\{X = x\})}$$p_X(x) \stackrel{\text{def}}{=} \Pr(\{X = x\})$, which assigns a "probability mass" to each of the possible outcomes of the random variable 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 distributions and random variables allow us to model random processes like the roll of a die. We can't predict the exact outcome when two dice X_1 and X_2 are rolled, but we can predict the probability of different outcomes. For example, the "pair of sixes" outcome is described by the event $\{X_1 + X_2 = 12\}$. Assuming the dice are fair, this outcome has probability $\Pr(\{X_1 + X_2 = 12\}) = \frac{1}{36}$.

Probability theory is used ~~all over the place~~in many places, including in gambling, risk analysis, statistics, machine learning, ~~quantum mechanics, gambling, and risk analysis~~and quantum mechanics.

[WEBSITE] A visual introduction to the basic ideas of probability theory: https://seeing-theory.brown.edu/basic-probability/.

[BOOK] Charles M. Grinstead and J. Laurie Snell. *Introduction to Probability*, Second Edition, AMS, 1997, ISBN 9780821894149. https://open.umn.edu/opentextbooks/textbooks/21.

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 and explains many more advanced topics.

[BOOK] Richard P. Feynman. *The Feynman Lectures on Physics, The Definitive and Extended Edition*, Addison Wesley, 2005, ISBN 0805390456. Read online at: http://feynmanlectures.caltech.edu

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 that 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.

Appendix A

Answers and solutions

Chapter 1 solutions

Answers to exercises

E1.1 a) $x = 3$; **b)** $x = 30$; **c)** $x = 2$; **d)** $x = -3$. **E1.2 a)** $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$; **b)** $\mathbb{C}$; **c)** $\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$; **d)** $\mathbb{Q}, \mathbb{R}, \mathbb{C}$; **e)** $\mathbb{R}, \mathbb{C}$. **E1.3 a)** 21; **b)** 0; **c)** $\frac{2}{27}$. **E1.4 a)** $\frac{5}{6}$; **b)** $\frac{13}{12} = 1\frac{1}{12}$; **c)** $\frac{31}{6} = 5\frac{1}{6}$. **E1.5 a)** $x = 2$; **b)** $x = 25$; **c)** $x = 100$. **E1.7 a)** $(x-1)(x-7)$; **b)** $(x+2)^2$; **c)** $(x+3)(x-3)$. **E1.8 a)** $\cancel{a^2 + 2ab + b^2}\, x^2 + 2x - 15 = (x+1)^2 - 16 = 0$, which has solutions $x = 3$ and $x = -5$; **b)** $\cancel{a^3 + 3a^2b + 3ab^2 + b^3}$; ~~**c)**~~ $\cancel{a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4}$; ~~**d)**~~ $\cancel{a^5 + 5a^4b + 10a^3b^2 + 10a^2b^3 + 5ab^4 + b^5}$. $x^2 + 4x + 1 = (x+2)^2 - 3 = 0$, with solutions $x = -2 + \sqrt{3}$ and $x = -2 - \sqrt{3}$. **E1.9** $x_1 = \frac{3}{2}$ and $x_2 = -1$. **E1.10** $x = \pm\sqrt{2}$. **E1.11 a)** 8; **b)** $a^{-1}b^{-2}c^{-3} = \frac{1}{ab^2c^3}$; **c)** $8\alpha^2$; **d)** a^6b^{-2}. **E1.12 a)** 3; **b)** 12; **c)** $\sqrt{3}$; **d)** $|a|$. **E1.13 a)** 2π; **b)** $4 + \frac{1}{4} = 4.25$; **c)** 1; **d)** x^2. **E1.14 a)** $x = \sqrt{a}$ and $x = -\sqrt{a}$; **b)** $x = \sqrt[3]{b}$; **c)** $x = \sqrt[4]{c}$ and $x = -\sqrt[4]{c}$; **d)** $x = \sqrt[5]{d}$. Bonus points if you can also solve $x^2 = -1$. We'll get to that in Section 3.5. **E1.15** $k_e = 8.988 \times 10^9$. **E1.16 a)** $\log(2xy)$. **b)** $-\log(z)$. **c)** $\log(y)$. **d)** 3. **e)** -3. **f)** 4. **E1.17** Domain: $\cancel{x \in \mathbb{R}}\,\mathbb{R}$. Image: $\cancel{f(x) \in [-2, 2]}\,[-2, 2]$. Roots: $\cancel{[\ldots, -\frac{\pi}{2}, \frac{\pi}{2}, \frac{3\pi}{2}, \frac{5\pi}{2}, \ldots]}\,\{\ldots, -\frac{\pi}{2}, \frac{\pi}{2}, \frac{3\pi}{2}, \frac{5\pi}{2}, \ldots\}$. **E1.18 a)** $p(x)$ is even and has degree 4. **b)** $q(x)$ is odd and has degree 7. **E1.19 a)** $x = 5$ and $x = -3$; **b)** $x = 1 + \sqrt{3}$ and $x = 1 - \sqrt{3}$. **E1.20 a)** $\cancel{(q \circ f)(x) = q(f(x)) = (x+5)^2}\,(q \circ f)(x) = q(f(x)) = (x+5)^2$; $q(x)$ shifted five units to the left. **b)** $(f \circ q)(x) = x^2 + 5$; $q(x)$ shifted upward by five units. **c)** $(q \circ g)(x) = (x-6)^2$; $q(x)$ shifted six units to the right. **d)** $(q \circ h)(x) = 49x^2$; $q(x)$ horizontally compressed by a factor of seven. **E1.21** $A = 5$, $\lambda = 0.1$, and $\phi = \frac{\pi}{8}$. **E1.22** $f(x) = x^2 - 2x + 5$. **E1.23** $g(x) = 2\sqrt{x-3} - 2$. **E1.24** $x = \sqrt{21}$. **E1.25** $V = 33.51$ and $A = 50.26$. **E1.26** Length of track $= 5C = 5\pi d = 11.47$ m. **E1.27** $x = 5\cos(45°) = 3.54$, $y = 5\sin(45°) = 3.54$; $C = 10\pi$. **E1.28 a)** $\frac{\pi}{6}$ ~~rad~~rad; **b)** $\frac{\pi}{4}$ ~~rad~~rad; **c)** $\frac{\pi}{3}$ ~~rad~~rad; **d)** $\frac{3\pi}{2}$ ~~rad~~rad. **E1.29 a)** -1; **b)** 1; **c)** 0. **E1.30 a)** 0; **b)** 1; **c)** $\frac{1}{2}$; **d)** 1. **E1.31** ~~Length of track $= 5C = 5\pi d = 11.47$m~~**a)** $3.16\angle 18.43°$; **b)** $2.24\angle 243.43° = 2.24\angle -116.57°$; **c)** $6\angle 270° = 6\angle -90°$; **d)** $(8.66, 5)$; **e)** $(9.66, 2.59)$; **f)** $(-5, 8.66)$. **E1.32** $\cancel{(x-1)^2 + (y-4)^2 = 9 \text{ or } \{(x,y) \in \mathbb{R}^2 \mid x = 1 + 3\cos\theta,\ y = 4 + 3\sin\theta,\ \theta \in [0, 2\pi)\}}\,y = 2$. **E1.33** $c = \sqrt{a^2 - b^2}$. **E1.34** $y = \frac{1}{4f}x^2$. **E1.35 a)** $(1, \frac{1}{2})$. **b)** $(1, 2)$. **c)** $(-2, 2)$. **E1.36** $x = 2$, $y = 3$. **E1.37** $x = 5$, $y = 6$, and $z = -3$.

E1.38 $p = 7$ and $q = 3$. **E1.39** **a)** \$53 974.14; **b)** \$59 209.77; **c)** \$65 948.79. **E1.40** \$32 563.11. **E1.41** **a)** $\{2,4,6,7\}$; **b)** $\{1,2,3,4,5,6\}$; **c)** $\{1,3,5\}$; **d)** $\varnothing$; **e)** $\{1,2,3,4,5,6,7\}$; **f)** $\{7\}$; **g)** $\{2,4,6,7\}$; **h)** $\varnothing$. **E1.42** **a)** ~~$x \in (-\infty, \frac{3}{2})$~~$(-\infty, \frac{3}{2})$; **b)** ~~$x \in (-\infty, -5]$~~$(-\infty, -5]$; **c)** ~~$x \in (-1, 4)$~~$(-1, 4)$; **d)** ~~$x \in (4, \infty)$~~$(4, \infty)$; **e)** ~~$x \in [\frac{14}{3}, \infty)$~~$[\frac{14}{3}, \infty)$; **f)** ~~$(-\infty, -4] \cup [2, \infty)$~~$(-\infty, -4] \cup [2, \infty)$.

Solutions to selected exercises

E1.4 **a)** To compute $\frac{1}{2} + \frac{1}{3}$, we rewrite both fractions using the common denominator 6, then compute the sum: $\frac{1}{2} + \frac{1}{3} = \frac{3}{6} + \frac{2}{6} = \frac{5}{6}$. **b)** You can use the answer from part (a), or compute the triple sum directly by setting all three fractions to a common denominator: $\frac{1}{2} + \frac{1}{3} + \frac{1}{4} = \frac{6}{12} + \frac{4}{12} + \frac{3}{12} = \frac{13}{12}$. **c)** Here we first rewrite $3\frac{1}{2}$ as $\frac{7}{2}$, then use the common denominator 6 for the computation: $\frac{7}{2} + 2 - \frac{1}{3} = \frac{21}{6} + \frac{12}{6} - \frac{2}{6} = \frac{31}{6}$.

E1.15 If you're using a very basic calculator, you should first compute the expression in the denominator, and then invert the fraction. Calculators that support scientific notation have an "`exp`" or "`E`" button, which allows you to enter ε_0 as `8.854e-12`. If your calculator supports expressions, you can type in the whole expression `1/(4*pi*8.854e-12)`. We report an answer with four significant digits because we started from a value of ε_0 with four significant digits of precision.

E1.19 **a)** Rewrite the equation putting all terms on the right-hand side: $0 = x^2 - 2x - 15$. We can factor this quadratic by inspection. Are there numbers a and b such that $a + b = -2$ and $ab = -15$? Yes, $a = -5$ and $b = 3$, so $0 = (x-5)(x+3)$. **b)** Rewrite the equation so all terms are on the left-hand side: $3x^2 - 6x - 6 = 0$. Nice, the cubic terms cancel! We'll use the quadratic formula to solve this equation $x = \frac{6 \pm \sqrt{(-6)^2 - 4(3)(-6)}}{6} = \frac{6 \pm 6\sqrt{3}}{6} = 1 \pm \sqrt{3}$.

E1.24 The cosine rule tells us ~~$x^2 = 4^2 + 5^2 - 2(4)(5)\cos(60°) = 16 + 25 - 40\frac{1}{2} = 21$~~$x^2 = 4^2 + 5^2 - 2$ Therefore $x = \sqrt{21}$.

E1.25 The volume of the sphere with radius $r = 2$ is $V = \frac{4}{3}\pi 2^3 = 33.51$. Its surface area is $A = 4\pi 2^2 = 50.26$.

E1.28 To convert an angle measure from degrees to radians we must multiply it by the conversion ratio $\frac{\pi}{180}$~~rad/°~~.

E1.32 Substitute the formula $\sin\theta = \frac{y}{r}$ into the equation to obtain $r = \frac{2r}{y}$, which simplifies to $y = 2$. The function $r(\theta) = \frac{2}{\sin\theta}$ in polar coordinates corresponds to the line with equation $y = 2$. See `www.desmos.com/calculator/5n5zzoal2t` for the graph.

E1.33 First define the vertex $V_2 = (a, 0)$ which corresponds to the right extremity of the ellipse. Considering the definition of the ellipse at the vertex V_2, we find $r_1 + r_2 = (c + a) + (a - c) = 2a$. Next, consider the vertex $V_3 = (0, b)$ at the top of the ellipse. The distances r_1 and r_2 from V_3 to the focal points F_1 and F_2 correspond to the hypotenuse of a triangle with base c and height b: $r_1 = r_2 = \sqrt{c^2 + b^2}$. Since $r_1 + r_2 = \text{const.}$ for all points on the ellipse, we can equate the results obtained from the length calculations for point V_2 and point V_3. We find $2a = 2\sqrt{c^2 + b^2}$, which we can solve for c to obtain $c = \sqrt{a^2 - b^2}$.

E1.34 For a parabola with focal length f, the focal point is at $F = (0, f)$ and the directrix is the line with equation $y = -f$. The distance from the focal point to an arbitrary point on the parabola is given by

$$r = d(P, F) = d\Big((x, y), (0, f)\Big) = \sqrt{x^2 + (y - f)^2}.$$

P1.48 The base of this triangle has length $2r$ and each side has length $R + r$. If you split this triangle through the middle, each half is a right triangle with an angle at the centre $\frac{360^\circ}{24} = 15^\circ$, hypotenuse $R + r$, and opposite side r. We therefore have $\sin 15^\circ = \frac{r}{R+r}$. After rearranging this equation, we find $\frac{R}{r} = \frac{1-\sin 15^\circ}{\sin 15^\circ} = 2.8637$.

P1.51 The tank's total capacity is $15 \times 6 \times 5 = 450\,\mathrm{m}^3$. If 30% of its capacity is spent, then 70% of the capacity remains: $315\,\mathrm{m}^3$. Knowing that $1\,\mathrm{m}^3 = 1000\,\mathrm{L}$, we find there are 315 000 L in the tank.

P1.52 The first tank contains $\frac{1}{4} \times 4000 = 1000\,\mathrm{L}$. The second tank contains three times more water, so 3000 L. The total is 4000 L.

P1.53 Let's define w and h to be the width and the height of the hole. Define d to be the distance from the hole to the sides of the lid. The statement of the problem dictates the following three equations must be satisfied: $w + 2d = 40$, $h + 2d = 30$, and $wh = 500$. After some manipulations, we find $w = 5(1 + \sqrt{21})$, $h = 5(\sqrt{21} - 1)$ and $d = \frac{1}{2}(35 - 5\sqrt{21})$.

P1.54 The amount of wood in a pack of wood is proportional to the area of a circle $A = \pi r^2$. The circumference of this circle is equal to the length of the rope $C = \ell$. Note the circumference is proportional to the radius $C = 2\pi r$. If we want double the area, we need the circle to have radius $\sqrt{2}r$, which means the circumference needs to be $\sqrt{2}$ times larger. If we want a pack with double the wood, we need to use a rope of length $\sqrt{2}\ell$.

P1.55 In 10 L of a 60% acid solution there are 6 L of acid and 4 L of water. A 20% acid solution will contain four times as much water as it contains acid, so 6 L acid and 24 L water. Since the 10 L we start from already contains 4 L of water, we must add 20 L.

P1.56 The document must have a 768/1004 aspect ratio, so its height must be $6 \times \frac{1004}{768} = 7.84375$ inches.

P1.57 If we rewrite $1 + 2 + 3 + \cdots + 98 + 99 + 100$ by pairing numbers, we obtain the sum $(1 + 100) + (2 + 99) + (3 + 98) + \cdots$. This list has 50 terms and each term has the value 101. Therefore $1 + 2 + 3 + \cdots + 100 = 50 \times 101 = 5050$.

P1.62 An nAPR of 12% means the monthly interest rate is $\frac{12\%}{12} = 1\%$. After 10 years you'll owe $\$5000(1.01)^{120} = \$16501.93$. Yikes!

P1.63 The graphs of the functions are shown in Figure ??A.1. Observe that $f(x)$ decreases to 37% of its initial value when $x = 2$. The increasing exponential $g(x)$ reaches 63% of its maximum value at $x = 2$.

P1.64 We're looking for the time t ~~such that~~ when $Q(t)/Q_o = \frac{1}{2}$, which is the same as $e^{-5t} = 0.5$. ~~Taking~~ Take logarithms of both sides ~~we~~ to find $-5t = \ln(0.5)$ ~~, and solving~~ and solve for t ~~we find~~ to get $t = 0.14\,\mathrm{s}$.

P1.65 We're told $T(24)/T_o = \frac{1}{2} = e^{-24/\tau}$, which we can rewrite as $\ln(\frac{1}{2}) = -24/\tau$. Solving for τ, we find $\tau = \frac{24}{\ln 2} = 34.625\,\mathrm{min}$. To find the time the body takes to reach 1% of its initial temperature, we must solve for t in $T(t)/T_o = 0.01 = e^{-t/34.625}$. We find $t = 159.45\,\mathrm{min}$.

P1.67 There exists at least one banker who is not a crook. Another way of saying the same thing is "not all bankers are crooks"—just *most* of them.

P1.68 Everyone steering the ship at Monsanto ought to burn in hell, forever.

P1.69 **a)** Investors with money but without connections. **b)** Investors with connections but no money. **c)** Investors with both money and connections.

P4.15 The period will decrease. **P4.16** $F_{\text{push}} = 4.71$[N]. **P4.17** The pulley with the larger radius R will spin faster and have more K_r. **P4.18** (1) $v_2 = -5.4$[m/s]. (2) Not elastic. **P4.19** (1) See solution. (2) $a_1 = 1$[m/s^2], $a_2 = 0$[m/s^2], and $a_3 = -1$[m/s^2]. (3) $F = 40$[N]. **P4.20** $v_{\text{obj}} = \sqrt{2}\,v$ [m/s]. **P4.21** (1) $\mathcal{T} = 4.9$[N m]. (2) $W = 9.6$[J]. **P4.22** $m = 1.46$[kg]. **P4.23** $\|\vec{F}\| = 3$[N]. **P4.24** (1) $h = d$. (2) $v_i = 19.8$[m/s] and $t_{\text{hit}} = 1.43$[s]. **P4.25** (1) $\mu_s \geqslant \frac{M}{m_1+m_2}$. (2) $\mu_s \geqslant 0.421$. (3) $a = \frac{Mg}{m_1+m_2+M}$. **P4.26** (1) $d = 4.41$[m]. (2) $\vec{v}_{2i} = 5\angle 90°$[m/s]. **P4.27** (1) $F_{\text{lift}} = 158 \times 10^3$[N]. (2) The vertical acceleration is zero so the plane will maintain a horizontal trajectory. **P4.28** $v = 10.2$[m/s]. **P4.29** (1) $v_i = (\frac{\sin 30}{\tan 60} + \cos 30)\sqrt{2dg\mu_k}$. (2) $v_i = 2.24$[m/s], $v_1 = 1.12$[m/s] and $v_2 = 1.94$[m/s]. (3) 0.233[m]. (4) The collision is elastic. **P4.30** (1) $\mathcal{T} = 3.6$[N m]. (2) 18.9 revolutions. **P4.31** The solid cylinder will reach the bottom first. **P4.32** $t_{\text{flight}} = 2t_{\text{top}} = 4.1$[s]. **P4.33** $\mu_k = \frac{M^2}{m^2}\frac{L}{d}$. **P4.34** Range is 0.65[m] greater on the summit than on the North Pole. **P4.35** $x(t) = 2t^2 + 10t + 20$ in metres. **P4.36** The slug loses contact at $R = \frac{0.4g}{\omega^2}$. **P4.37** upward $F_{fs} >$ stationary $F_{fs} >$ downward F_{fs}. **P4.38** The coin farthest from the centre will fly off first. **P4.39** (1) $F_f = 3000$[N] per wheel. (2) $\mathcal{T} = 180$[N m]. (3) 2.21 turns. (4) $2.\overline{7}$[m]. **P4.40** $\cancel{x_f = 2.09}$ $x_f = 2.85$[m]. **P4.41** $\cancel{y(x) = \ell \sin(\theta_{\max}) \cos((\omega/v)x)}$ $y(x) = \ell \sin\left(\theta_{\max} \cos\left(\frac{\sqrt{g}}{\sqrt{\ell}} x\right)\right)$.

FIX2 FIX1

Solutions to problems

P4.1 When the y-axis points up, $a_y = -g$ and v_{iy} is positive. The opposite applies when the y-axis is directed downward. The balloon moves at the same horizontal speed as the cat; the balloon is always directly above the cat, and splashes the cat when it comes back down. The cat is not happy about that.

P4.2 (1) $\vec{F}_4$ points right and is perpendicular to the left face of the block, (2) $\vec{F}_4$ points up and is perpendicular to the bottom face of the block, and (3) $\vec{F}_4$ points left and is perpendicular to the right face of the block. In each case, the sum of the forces produces an $\vec{a}_{\text{block}}$ in the desired direction.

P4.3 Calculate the momentum and energy using the formulas $\|\vec{p}\| = m\|\vec{v}\|$ and $K = \frac{1}{2}mv^2$. Observe that two objects moving with equal momentum can carry different amounts of kinetic energy; this problem shows momentum and energy are different quantities.

P4.4 In each case, the sum $\vec{p}_{\text{A}} + \vec{p}_{\text{B}}$ after the separation equals the momentum of the station before the compartments split apart: $2m\vec{v} = m\vec{v}_{\text{A}} + m\vec{v}_{\text{B}}$.

P4.5 When there is a velocity, there is kinetic energy K. When the spring is stretched, there is spring potential energy U_s. When the position of the mass is above or below $y = 0$, there is gravitational potential energy U_g.

P4.6 The ball's initial kinetic energy is the same on Earth and on the Moon. Because of conservation of energy, when the ball returns to ground level, it will have the same kinetic energy it had initially, regardless of the value of g.

P4.7 Define the zero potential-energy level to be at ground level. The bottom of the 10[m] pit has a lower potential energy on Earth because $g_{\text{Earth}} > g_{\text{Moon}}$. The ball will therefore gain more kinetic energy on Earth when it reaches the bottom of the pit and thus have a higher speed.

P4.8 The rotation of the mass M is at a constant angular velocity so the net torque on the mass is zero. Let us denote by L_{rod}, L_{mass}, and L_{sys} the angular momenta of the rod, the mass M, and the total angular momentum of the mass-on-a-rod system. Initially, $L_{\text{sys}} = L_{\text{rod}} + L_{\text{mass}}$. When the mass M detaches, its velocity $\vec{v}$ will remain the same as before the moment it detached. This means its angular momentum L_{mass} will remain the same after it detaches. This in turn implies the rod will also maintain its angular momentum, so its angular velocity will remain ω.

P4.30 Use $a = r\alpha$ to find α, then use $\mathcal{T} = I\alpha$ to find the torque. Use the angular equations of motion to find $\theta(4)$. The number of revolutions is $\frac{\theta(4)}{2\pi}$.

P4.32 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 to find t_{top}.

P4.33 First we use $U_i = K_f$ for the pendulum, obtaining $MgL = \frac{1}{2}Mv_{\text{in}}^2$ and thus $v_{\text{in}} = \sqrt{2gL}$. Next we use a momentum reasoning $\vec{p}_{\text{in}} = \vec{p}_{\text{out}}$ where the incoming momentum is that of the mass M and the outgoing momentum is that of the mass m. The conservation of momentum equation becomes $Mv_{\text{in}} + 0 = 0 + mv_{\text{out}}$, where v_{out} is the velocity of the mass m after the collision, and the momentum of the pendulum is zero after the collision since it doesn't bounce back. Solving for v_{out} we find $v_{\text{out}} = \frac{M}{m}\sqrt{2gL}$. Finally, we use an energy calculation $K_i = W_{\text{lost}}$, which becomes $\frac{1}{2}m\left(\frac{M}{m}\sqrt{2gL}\right)^2 = mg\mu_k d$. After some simplifications, we find $\mu_k = \frac{M^2}{m^2}\frac{L}{d}$.

P4.34 We want to find the range—how far the ball will reach after being kicked—in both situations. The first thing to calculate is the total time of flight by solving for t in $0 = 0 + v_{iy}t + \frac{1}{2}(-g)t^2$. The time of flight will be 4.347[s] on the Nevado Huascarán summit, and 4.316[s] on the North Pole. The range in each case corresponds to $d = v_{ix}4.347 = 92.21$[m] and $d = v_{ix}4.316 = 91.56$[m]. The difference in range is $92.21 - 91.56 = 0.65$[m].

P4.36 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, which happens at a distance of $R = \frac{0.4g}{\omega^2}$ from the centre.

P4.37 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). 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}.

P4.38 The coin farthest from the centre will be the first to fly off the spinning turntable because the centripetal force required to keep this coin turning is the largest. Recall that $F_r = ma_r$, $a_r = v^2/R$, and $v = \omega R$. If the turntable turns with angular velocity ω, the centripetal acceleration required to keep a coin turning in a radius R is $F_r = m\omega^2 R$. This centripetal force must be supplied by the static force of friction F_{fs} between the coin and the turntable. Larger Rs require more F_{fs}.

P4.39 (1) The friction force is proportional to the normal force. 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 wheels 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!

FIX1

P4.40 The energy equation $\sum E_i = \sum E_f$ in this case is $U_i = U_f + K_f$, 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 ~~$v = 4.48$~~$v = 6.345$[m/s]. Now for the projectile motion part. The initial velocity is ~~4.48~~6.345[m/s] at an angle of 10° with respect to the horizontal, so ~~$\vec{v}_i = (4.42, 0.778)$~~$\vec{v}_i = (6.24, 1.10)$[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 $t = 0.237$~~$0 = -4.9t^2 + 1.10t + 0.0911$ and find $t = 0.289$[s]. Tarzan will land at ~~$x_f = 6\sin(10) + 4.42t = 2.09$~~$x_f = 6\sin(10) + 6.24t = 2.85$[m].

FIX2 **P4.41** ~~We begin by writing~~ Based on the geometry of the setup, we see the y-displacement depends on the angle of the pendulum through the equation $y(\theta) = \ell\sin(\theta)$. We also know the general equation of motion for a pendulum ~~:~~ is $\theta(t) = \theta_{\max}\cos(\omega t)$, 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 walkwaycan 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~~Combining the two equations, we obtain a formula that describes the bucket's transversal displacement as a function of time, $y(t) = \ell\sin(\theta_{\max}\cos(\omega t))$. Enter the walkway, which is moving 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 $y(x) = \ell\sin(\theta_{\max})\cos((\omega/v)x)$~~with velocity v, meaning the x-position of the point where the paint is falling is given by $x = vt$. We want to rewrite $y(t)$ as a function of x, which we can do using the substitution $t = \frac{x}{v}$. We thus obtain the answer $y(x) = \ell\sin\left(\theta_{\max}\cos\left(\omega\frac{x}{v}\right)\right)$.

Chapter 5 solutions

Answers to exercises

E5.1 (a) 0. (b) 2. (c) ∞. Each limit expression describes what happens to the ratio of two functions for large values of the input variable. **E5.2** (a) 2. (b) ∞. (c) $-\infty$. (d) $\frac{3}{4}$. (e) 0. (f) 0. **E5.3** Max at $x = \frac{1}{3}$; $f\left(\frac{1}{3}\right) = \frac{4}{27}$. **E5.4** (a) $\frac{2}{x-3} + \frac{1}{x+4}$, $\int(\text{a})\,dx = 2\ln(x-3) + \ln(x+4)$. (b) $\frac{1}{x-1} + \frac{2}{x-2}$, $\int(\text{b})\,dx = \ln(x-1) + 2\ln(x-2)$. (c) $\frac{1}{4(x-1)} - \frac{1}{4(x+1)} + \frac{1}{2(x+1)^2}$, $\int(\text{c})\,dx = \frac{1}{4}\ln(x-1) - \frac{1}{4}\ln(x+1) - \frac{1}{2}\frac{1}{x+1}$. **E5.5** $\frac{\pi R^2 h}{3}$. **E5.6** $\frac{\pi R^2 h}{3}$. **E5.7** (a) $caN + cb\frac{N(N+1)}{2}$. (b) $ca^2N + cabN(N+1) + cb^2\frac{N(N+1)(2N+1)}{6}$. (c) π^2.

Answers to problems

P5.1 (1) -6. (2) 2. (3) Doesn't exist. (4) 8.6 (eyeballing it). (5) -5. (6) Doesn't exist. (7) -2. (8) -2. (9) -2. (10) No. (11)$[-10,-5)$, $[-5,2)$, $[2,5)$, $(5,10]$. **P5.2** (a) 4. (b) 6. (c) 5. **P5.4** (1) Doesn't exist. (2) 0. (3) Doesn't exist. (4) 0. (5) Doesn't exist. (6) 0. (7) 1. (8) 0. (9) 1. **P5.5** (1) Doesn't exist. (2) 3. (3) $2a$. **P5.7** (1) $\frac{dy}{dx} = 13x^{12}$. (2) $\frac{dy}{dx} = -\frac{3}{2}x^{-\frac{5}{2}}$. (3) $\frac{dy}{dx} = 2ax^{(2a-1)}$. (4) $\frac{du}{dx} = 2.4x^{1.4}$. (5) $\frac{dz}{dx} = \frac{1}{3}x^{-\frac{2}{3}}$. (6) $\frac{dy}{dx} = -\frac{5}{3}x^{-\frac{8}{3}}$. (7) $\frac{du}{dx} = -\frac{8}{5}x^{-\frac{13}{5}}$. (8) $\frac{dy}{dx} = 2ax^{a-1}$. (9) $\frac{dy}{dx} = \frac{3}{q}x^{\frac{3-q}{q}}$. **P5.8** (1) $\frac{dy}{dx} = 3ax^2$. (2) $\frac{dy}{dx} = 13 \times \frac{3}{2}x^{\frac{1}{2}}$. (3) $\frac{dy}{dx} = 6x^{-\frac{1}{2}}$. (4) $\frac{dy}{dx} = \frac{1}{2}c^{\frac{1}{2}}x^{-\frac{1}{2}}$. (5) $\frac{du}{dz} = \frac{an}{c}z^{n-1}$. (6) $\frac{dy}{dt} = 2.36t$. **P5.9** (a) $1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \ldots$; (b) $2ax + b$; (c) $3x^2 + 6ax + 3a^2$. **P5.10** (1) $\frac{dw}{dt} = a - bt$. (2) $\frac{dy}{dx} = 2x$. (3) $14110x^4 - 65404x^3 - 2244x^2 + 8192x + 1379$. (4) $\frac{dx}{dy} = 2y + 8$. (5) $185.9022654x^2 + 154.36334$. **P5.11** (1) $p'(x) = \frac{-5}{(3x+2)^2}$. (2) $q'(x) = \frac{6x^4+6x^3+9x^2}{(1+x+2x^2)^2}$. (3) $r'(x) = \frac{ad-bc}{(cx+d)^2}$. (4) $s'(x) = \frac{anx^{-n-1}+bnx^{n-1}+2nx^{-1}}{(x^{-n}+b)^2}$. **P5.12** (1) $\frac{x}{\sqrt{x^2+1}}$.

Appendix C

Constants, units, and conversion ratios

In this appendix you will find a number of tables of useful information that 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	~~$6.022\,141 \times 10^{23}$~~ $6.022\,14076 \times 10^{23}$	$\mathrm{mol^{-1}}$	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	$6.626\,069 \times 10^{-34}$	J s	Planck's constant

Appendix D

SymPy tutorial

Computers can be very useful for dealing with complicated math expressions or when slogging through tedious calculations. Throughout this book we used `SymPy` to illustrate several concepts from math and physics. We'll now review all the math and physics tools available through the `SymPy` command line. Don't worry if you're not a computer person; we'll only discuss concepts we covered in the book, and the computer commands we'll learn are very similar to the math operations you're already familiar with. This section also serves as a final review of the material covered in the book.

Introduction

You can use a computer algebra system (CAS) to compute complicated math expressions, solve equations, perform calculus procedures, and simulate physics systems.

All computer algebra systems offer essentially the same functionality, so it doesn't matter which system you use: there are free systems like `SymPy`, `Magma`, or `Octave`, and commercial systems like `Maple`, `MATLAB`, and `Mathematica`. This tutorial is an introduction to `SymPy`, which is a *symbolic* computer algebra system written in the programming language ~~`Python`~~Python. In a symbolic CAS, numbers and operations are represented symbolically, so the answers obtained are exact. For example, the number $\sqrt{2}$ is represented in `SymPy` as the object `Pow(2,1/2)`, whereas in *numerical* computer algebra systems like `Octave`, the number $\sqrt{2}$ is represented as the approximation 1.41421356237310 (a `float`). For most purposes the approximation is okay, but sometimes approximations can lead to problems: `float(sqrt(2))*float(sqrt(2)) = 2.00000000000000044` $\neq$ 2. Because `SymPy` uses exact representations, you'll never run into such problems: `Pow(2,1/2)*Pow(2,1/2)`$=2$.

This tutorial presents many explanations as ~~blocks of code~~ code snippets. Be sure to try the code examples on your own by typing the commands into SymPy. It's always important to verify for yourself!

Using SymPy

The easiest way to use SymPy, provided you're connected to the internet, is to visit http://live.sympy.org. You'll be presented with an interactive prompt into which you can enter your commands—right in your browser.

If you want to use SymPy on your own computer, you must ~~install Python and the python~~ first install Python and the Python package sympy. You can then open a command prompt and start a ~~SymPy~~ Python session using:

```
you@host> python
Python X.Y.Z
[GCC a.b.c (Build Info)] on platform
Type "help", "copyright", or "license" for more information.
>>>
\DIFdelbegin \DIFdel{from sympy import *
>>>
}\DIFdelend
```

The >>> prompt indicates you're in the ~~Python~~ Python shell which accepts ~~Python commands.~~ Python commands. Type the following in the Python shell:

```
\DIFadd{>>> from sympy import *
>>>
}
```

The command `from sympy import *` imports all the SymPy functions into the current namespace. All SymPy functions are now available to you. To exit the python shell press CTRL+D.

~~I highly recommend you also install ipython, which is an improved interactive python shell. If you have ipython and SymPy installed, you can start an ipython shell with SymPy pre-imported using the command isympy.~~ For an even better experience, you can try jupyter notebook, which is a web ~~frontend for the ipython shell.~~ interface for accessing the Python shell. Search the web for "jupyter notebook" and follow the installation instructions specific to your operating system. It's totally worth it!

Each section in this appendix begins with a python import statement for the functions used in that section. If you use the statement `from sympy import *` in the beginning of your code, you don't need to run these individual import statements, but I've included them so you'll know which SymPy vocabulary is covered in each section.

Fundamentals of mathematics

Let's begin by learning about the basic SymPy objects and the operations we can carry out on them. We'll learn the SymPy equivalents of many math verbs like: "to solve" (an equation), "to expand" (an expression), "to factor" (a polynomial).

Numbers

```
>>> from sympy import  sympify, S, evalf, N
```

In ~~Python~~Python, there are two types of number objects: `ints` and `floats`.

```
>>> 3
3                               # an int
>>> 3.0
3.0                             # a float
```

Integer objects in ~~Python~~ Python are a faithful representation of the set of integers $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$. Floating point numbers are approximate representations of the reals $\mathbb{R}$. ~~Regardless of its absolute size, a~~ A floating point number ~~is only accurate to~~ has 16 decimals of precision.

Special care is required when specifying rational numbers ~~,~~ ~~because integer division might not produce the answer you want~~ . ~~In other words, Python will not automatically convert the answer to~~ if you want to get exact answers. If you try to divide two numbers, Python will compute a floating point ~~number, but instead round the answer to the closest integer:~~

```
>>> 1/7
0                                  # int/int gives int
```

approximation:

~~To avoid this problem, you can force `float` division by using the number `1.0` instead of `1`:~~

```
>>> \DIFdelbegin \DIFdel{1.0}\DIFdelend \DIFaddbegin \DIFadd{1}\DIFaddend
0.14285714285714285           # \DIFdelbegin \DIFdel{float/int gives }\DIF
```

~~This result is better, but it's still only~~ The floating point number 0.14285714285714285 is an approximation of the exact number $\frac{1}{7} \in \mathbb{Q}$~~,~~ . ~~since a~~ The `float` approximation has 16 decimals while the decimal

The first statement ~~instructs python to convert 1/7 to 1.0/7 when dividing, potentially saving you from any int division confusion. The second statement~~ imports all the SymPy functions. The ~~remaining~~ other three statements define some generic symbols x, y, z, and t, and several other symbols with special properties.

Note the difference between the following two statements:

```
>>> x + 2
x + 2                # an Add expression
>>> p + 2
NameError: name 'p' is not defined
```

The name x is defined as a symbol, so SymPy knows that x + 2 is an expression; but the variable p is not defined, so SymPy doesn't know what to make of p + 2. To use p in expressions, you must first define it as a symbol:

```
>>> p = Symbol('p')   # the same as p = symbols('p')
>>> p + 2
p + 2                # = Add(Symbol('p'), Integer(2))
```

You can define a sequence of variables using the following notation:

```
>>> a0, a1, a2, a3 = symbols('a0:4')
```

You can use any name you want for a variable, but it's best if you avoid the letters Q,C,O,S,I,N and E because they have special uses in SymPy: I is the unit imaginary number ~~$i = \sqrt{-1}$~~$i \overset{\text{def}}{=} \sqrt{-1}$, E is the base of the natural logarithm, S() is the sympify function, N() is used to obtain numeric approximations, and O is used for big-O notation.

~~The underscore symbol~~ _ ~~is a special variable that contains the result of the last printed value. The variable~~ _ ~~is analogous to the ans button on certain calculators, and is useful in multi-step calculations:~~

```
>>> 3+3
6
>>> _*2
12
```

Expressions

```
>>> from sympy import simplify, factor, expand, collect
```

You define SymPy expressions by combining symbols with basic math operations and other functions:

```
>>> expr = 2*x + 3*x - sin(x) - 3*x + 42
>>> simplify(expr)                    \DIFaddbegin \DIFadd{# simplify the e
}\DIFaddend 2*x - sin(x) + 42
```

Recall that the roots of the polynomial $P(x)$ are defined as the solutions to the equation $P(x) = 0$. We can use the `solve` function to find the roots of the polynomial:

```
>>> roots = solve(P,x)
>>> roots
[1, 2, 3]
# let's check if P equals (x-1)(x-2)(x-3)
>>> simplify( P  -  (x-roots[0])*(x-roots[1])*(x-roots[2]) )
0
```

Equality checking

In the last example, we used the `simplify` function on the difference of two expressions to check whether ~~two expressions~~ they were equal. This way of checking equality works because $P = Q$ if and only if $P - Q = 0$. To know whether $P = Q$, we can calculate `simplify(P-Q)` and see if the result equals 0. This is the best way to check ~~if~~ whether two expressions are equal in `SymPy` because it attempts all possible simplifications when comparing the expressions. Below is a list of other ways to check whether two quantities are equal, with example cases where ~~they fail~~equality fails to be detected:

Trigonometry

```
from sympy import sin, cos, tan, trigsimp, expand_trig
```

The trigonometric functions `sin` and `cos` take inputs in radians:

```
>>> sin(pi/6)
1/2
>>> cos(pi/6)
sqrt(3)/2
```

For angles in degrees, you need a conversion factor of $\frac{\pi}{180}$[rad/°]:

```
>>> sin(30*pi/180)                    # 30 deg = pi/6 rads
1/2
```

The inverse trigonometric functions ~~$\sin^{-1}(x) \equiv \arcsin(x)$ and $\cos^{-1}(x) \equiv$ arccos~~ $\sin^{-1}(x) = \arcsin(x)$ and $\cos^{-1}(x) = \arccos(x)$ are used as follows:

```
>>> asin(1/2)
pi/6
>>> acos(sqrt(3)/2)
pi/6
```

Recall that ~~$\tan(x) \equiv \frac{\sin(x)}{\cos(x)}$~~$\tan(x) = \frac{\sin(x)}{\cos(x)}$. The inverse function of $\tan(x)$ is ~~$\tan^{-1}(x) \equiv \arctan(x) \equiv$~~ $\tan^{-1}(x) = \arctan(x) =$ `atan(x)`

```
>>> z = 4 + 3*I
>>> z
4 + 3*I
>>> re(z)
4
>>> im(z)
3
```

The *polar* representation of a complex number is ~~$z \equiv |z|\angle\theta \equiv |z|e^{i\theta}$~~ $z = |z|\angle\theta = |z|e^{i\theta}$. For a complex number $z = a + bi$, the quantity $|z| = \sqrt{a^2 + b^2}$ is known as the *absolute value* of z, and θ is its *phase* or its *argument*:

```
>>> Abs(z)
5
>>> arg(z)
atan(3/4)
```

The complex conjugate of $z = a + bi$ is the number ~~$\bar{z} = a - bi$~~ $\bar{z} = a - bi$, which has the same absolute value as z but opposite phase:

```
>>> conjugate( z )
4 - 3*I
```

Complex conjugation is important for computing the absolute value of z (~~$|z| \equiv \sqrt{z\bar{z}}$~~ $|z| = \sqrt{z\bar{z}}$) and for division by z (~~$\frac{1}{z} \equiv \frac{\bar{z}}{|z|^2}$~~ $\frac{1}{z} = \frac{\bar{z}}{|z|^2}$).

Euler's formula

```
>>> from sympy import expand, rewrite
```

Euler's formula shows an important relation between the exponential function e^x and the trigonometric functions $\sin(x)$ and $\cos(x)$:

$$e^{ix} = \cos x + i \sin x.$$

To obtain this result in SymPy, you must specify that the number x is real and also tell `expand` that you're interested in complex expansions:

```
>>> x = symbols('x', real=True)
>>> \DIFaddbegin \DIFadd{expand(}\DIFaddend exp(I*x)\DIFdelbegin \DIFdel{.
cos(x) + I*sin(x)
>>> re( exp(I*x) )
cos(x)
>>> im( exp(I*x) )
sin(x)
```

Basically, $\cos(x)$ is the real part of e^{ix}, and $\sin(x)$ is the imaginary part of e^{ix}. Whaaat? I know it's weird, but weird things are bound to happen when you input imaginary numbers to functions.

~~Euler's formula is often used to rewrite the functions sin and cos in terms of complex exponentials. For example,~~

~~»> (cos(x)).rewrite(exp) exp(I*x)/2 + exp(-I*x)/2~~

~~Compare this expression with the definition of hyperbolic cosine.~~

Vectors

A vector $\vec{v} \in \mathbb{R}^n$ is an n-tuple of real numbers. For example, consider a vector that has three components:

$$\vec{v} = (v_1, v_2, v_3) \in (\mathbb{R}, \mathbb{R}, \mathbb{R}) \equiv \mathbb{R}^3.$$

To specify the vector $\vec{v}$, we specify the values for its three components v_1, v_2, and v_3.

A matrix $A \in \mathbb{R}^{m \times n}$ is a rectangular array of real numbers with m rows and n columns. A vector is a special type of matrix; ~~we~~ you can think of ~~a vector $\vec{v} \in \mathbb{R}^n$ either as a row vector (~~the vector $\vec{v} \in \mathbb{R}^n$ as a $1 \times n$ matrix~~) or a column vector ($n \times 1$ matrix)~~. Because of this equivalence between vectors and matrices, ~~there is no need for a special vector object~~ in `SymPy` ~~, and~~ we use `Matrix` objects ~~are used for vectorsas well~~to represent vectors.

This is how we define vectors and compute their properties:

Dot product

The dot product of the 3-vectors $\vec{u}$ and $\vec{v}$ can be defined two ways:

$$\vec{u} \cdot \vec{v} \equiv \underbrace{u_x v_x + u_y v_y + u_z v_z}_{\text{algebraic def.}} \equiv \underbrace{\|\vec{u}\|\|\vec{v}\| \cos(\varphi)}_{\text{geometric def.}} \quad \in \mathbb{R},$$

where φ is the angle between the vectors $\vec{u}$ and $\vec{v}$. In `SymPy`,

```
>>> u = Matrix([ 4,5,6])
>>> v = Matrix([-1,1,2])
>>> u.dot(v)
13
```

We can combine the algebraic and geometric formulas for the dot product to obtain the cosine of the angle between the vectors

$$\cos(\varphi) = \frac{\vec{u} \cdot \vec{v}}{\|\vec{u}\|\|\vec{v}\|} = \frac{u_x v_x + u_y v_y + u_z v_z}{\|\vec{u}\|\|\vec{v}\|},$$

and use the `acos` function to find the angle measure:

```
>>> acos(u.dot(v)/(u.norm()*v.norm())).evalf()
0.921263115666387       # in radians  =  52.76 degrees
```

Just by looking at the coordinates of the vectors $\vec{u}$ and $\vec{v}$, it's difficult to determine their relative direction. Thanks to the dot product, however, we know the angle between the vectors is $52.76°$, which means they *kind of* point in the same direction. Vectors that are at an angle $\varphi = 90°$ are called *orthogonal*, meaning at right angles with each other. The dot product ~~of vectors for which $\varphi > 90°$ is negative because they point *mostly* in opposite directions~~between two vectors is negative when the angle between them is $\varphi > 90°$.

The notion of the "angle between vectors" applies more generally to vectors with any number of dimensions. The dot product for n-dimensional vectors is $\vec{u} \cdot \vec{v} = \sum_{i=1}^{n} u_i v_i$. This means we can talk about "the angle between" 1000-dimensional vectors. That's pretty crazy if you think about it—there is no way we could possibly "visualize" 1000-dimensional vectors, yet given two such vectors we can tell if they point mostly in the same direction, in perpendicular directions, or mostly in opposite directions.

The dot product is a commutative operation $\vec{u} \cdot \vec{v} = \vec{v} \cdot \vec{u}$:

```
>>> u.dot(v) == v.dot(u)
True
```

Cross product

The *cross product*, denoted $\times$, takes two vectors as inputs and produces a vector as output. The cross products of individual basis elements are defined as follows:

$$\hat{\imath} \times \hat{\jmath} = \hat{k}, \qquad \hat{\jmath} \times \hat{k} = \hat{\imath}, \qquad \hat{k} \times \hat{\imath} = \hat{\jmath}.$$

The cross product is defined by the following equation:

$$\vec{u} \times \vec{v} = \left(u_y v_z - u_z v_y,\ u_z v_x - u_x v_z,\ u_x v_y - u_y v_x\right).$$

Here~~is~~'s how to compute the cross product of two vectors~~in SymPy:~~:

```
>>> u = Matrix([ 4,5,6])
>>> v = Matrix([-1,1,2])
>>> u.cross(v)
[4, -14, 9]
```

The vector $\vec{u} \times \vec{v}$ is orthogonal to both $\vec{u}$ and $\vec{v}$. The norm of the cross product $\|\vec{u} \times \vec{v}\|$ is proportional to the lengths of the vectors and the sine of the angle between them:

```
(u.cross(v).norm()/(u.norm()*v.norm())).n()
0.796366206088088     # = sin(0.921..)
```

The ~~name "cross product" is well-suited for this operation since it is calculated by "cross-multiplying" the coefficients of the vectors:~~

$$\cancel{\vec{u} \times \vec{v} = \left(u_y v_z - u_z v_y,\ u_z v_x - u_x v_z,\ u_x v_y - u_y v_x\right).}$$

~~By defining individual symbols for the entries of two vectors, we can make SymPy show us the cross-product formula: »> u1,u2,u3 = symbols('u1:4') »> v1,v2,v3 = symbols('v1:4') »> Matrix(u1,u2,u3).cross(Matrix(v (u2*v3 - u3*v2), (-u1*v3 + u3*v1), (u1*v2 - u2*v1)~~

~~The~~ cross product is anticommutative, $\vec{u} \times \vec{v} = -\vec{v} \times \vec{u}$:

```
>>> u.cross(v)
[4, -14, 9]
>>> v.cross(u)
[-4, 14,-9]
```

Watch out for this, because it's a new thing. The product of two numbers ~~and the~~ a and b is commutative: $ab = ba$. The dot product of two vectors ~~are commutativeoperations. The cross product , however,~~ $\vec{u}$ and $\vec{v}$ is commutative: $\vec{u} \cdot \vec{v} = \vec{v} \cdot \vec{u}$. However the cross product is not commutative: $\vec{u} \times \vec{v} \neq \vec{v} \times \vec{u}$, it is anticommutative: $\vec{u} \times \vec{v} = -\vec{v} \times \vec{u}$.

Mechanics

The module called `sympy.physics.mechanics` contains elaborate tools for describing mechanical systems, manipulating reference frames, forces, and torques. These specialized functions are not necessary for a first-year mechanics course. The basic `SymPy` functions like `solve`, and the vector operations you learned in the previous sections are powerful enough for basic Newtonian mechanics.

Dynamics

The net force acting on an object is the sum of all the external forces acting on it $\vec{F}_{\text{net}} = \sum \vec{F}$. Since forces are vectors, we need to use vector addition to compute the net force.

Compute $\vec{F}_{\text{net}} = \vec{F}_1 + \vec{F}_2$, where $\vec{F}_1 = 4\hat{\imath}$[N] and $\vec{F}_2 = 5\angle 30^\circ$[N]:

```
>>> F_1 = Matrix( [4,0] )
>>> F_2 = Matrix( [5*cos(30*pi/180), 5*sin(30*pi/180)] )
>>> F_net = F_1 + F_2
>>> F_net
[4 + 5*sqrt(3)/2, 5/2]
\DIFdelbegin \DIFdel{# in Newtons
}\DIFdelend >>> F_net.evalf()
[8.33012701892219, 2.5]                    # in Newtons
```

To express the answer in length-and-direction notation, use `norm` to find the length of $\vec{F}_{\text{net}}$~~and~~ , and use the two-input inverse tangent function atan2 [1] ~~to find its~~ to compute the direction:

```
>>> F_net.norm().evalf()
8.69718438067042                                  # |F_net| in [N]
>>> (atan2(F_net[1],F_net[0])*180/pi).n()
16.7053138060100                                  # angle in degrees
```

The net force on the object is $\vec{F}_{\text{net}} = 8.697\angle 16.7°[\text{N}]$.

Kinematics

Let $x(t)$ denote the position of an object, $v(t)$ denote its velocity, and $a(t)$ denote its acceleration. Together $x(t)$, $v(t)$, and $a(t)$ are known as the *equations of motion* of the object.

Starting from the knowledge of $\vec{F}_{\text{net}}$, we can compute $a(t) = \frac{\vec{F}_{\text{net}}}{m}$, then obtain $v(t)$ by integrating $a(t)$, and finally obtain $x(t)$ by integrating $v(t)$:

$$\underbrace{\frac{\vec{F}_{\text{net}}}{m} = a(t)}_{\text{Newton's 2}^{\text{nd}}\text{ law}} \quad \underbrace{\overset{v_i+\int dt}{\longrightarrow} \quad v(t) \quad \overset{x_i+\int dt}{\longrightarrow} \quad x(t)}_{\text{kinematics}}.$$

Uniform acceleration motion (UAM)

Let's analyze the case where the net force on the object is constant. A constant force causes a constant acceleration $a = \frac{F}{m} = \text{constant}$. If the acceleration function is constant over time $a(t) = a$. We find $v(t)$ and $x(t)$ as follows:

```
>>> t, a, v_i, x_i = symbols('t a v_i x_i')
>>> v = v_i + integrate(a, (t,0,t))
>>> v
a*t + v_i
>>> x = x_i + integrate(v, (t,0,t))
>>> x
a*t**2/2 + v_i*t + x_i
```

You may remember these equations from Section 2.4 (page 196). They are the *uniform accelerated motion* (UAM) equations:

$$\begin{aligned} a(t) &= a, \\ v(t) &= v_i + at, \\ x(t) &= x_i + v_i t + \frac{1}{2}at^2. \end{aligned}$$

[1] ~~The function atan2(y,x) computes the correct direction for all vectors (x, y), unlike atan(y/x) which requires corrections for angles in the range $[\frac{\pi}{2}, \frac{3\pi}{2}]$.~~

$\dfrac{dy}{dx}$	$\longleftarrow \quad y \quad \longrightarrow$	$\int y\,dx$
Exponential and Logarithmic		
e^x	e^x	$e^x + C$
x^{-1}	$\ln x$	$x(\ln x - 1) + C$
$\frac{1}{\ln 10}\,x^{-1}$	$\log_{10} x$	$\frac{1}{\ln 10}\,x(\ln x - 1) + C$
$a^x \ln a$	a^x	$\dfrac{a^x}{\ln a} + C$
Trigonometric		
$\cos x$	$\sin x$	$-\cos x + C$
$-\sin x$	$\cos x$	$\sin x + C$
$\sec^2 x$	$\tan x$	$\cancel{-\ln \cos x + C}$ $-\ln\lvert\cos x\rvert + C$
Inverse trigonometric		
$\dfrac{1}{\sqrt{(1-x^2)}}$	$\sin^{-1}(x)$	$x\sin^{-1}(x) + \sqrt{1-x^2} + C$
$-\dfrac{1}{\sqrt{(1-x^2)}}$	$\cos^{-1}(x)$	$x\cos^{-1}(x) - \sqrt{1-x^2} + C$
$\dfrac{1}{1+x^2}$	$\tan^{-1}(x)$	$x\tan^{-1}(x) - \frac{1}{2}\ln(1+x^2) + C$
Hyperbolic		
$\cosh x$	$\sinh x$	$\cosh x + C$
$\sinh x$	$\cosh x$	$\sinh x + C$
$\operatorname{sech}^2 x$	$\tanh x$	$\cancel{\ln\cosh x + C}$ $\ln(\cosh x) + C$
Inverse hyperbolic		
$-\dfrac{x}{(a^2+x^2)^{\frac{3}{2}}}$	$\dfrac{1}{\sqrt{a^2+x^2}}$	$\sinh^{-1}(\frac{x}{a}) + C = \ln(x+\sqrt{a^2+x^2}) + C$