CLP 1

DIFFERENTIAL CALCULUS

FELDMAN   RECHNITZER   YEAGER

CLP-1 DIFFERENTIAL CALCULUS

Joel FELDMAN   Andrew RECHNITZER   Elyse YEAGER

↠ **Legal stuff**

- Links to the source files can be found at the text webpage

CONTENTS

THE BASICS

We won't make this section of the text too long — all we really want to do here is to take a short memory-jogging excursion through little bits and pieces you should remember about sets and numbers. The material in this chapter will not be (directly) examined.

0.1 ▴ Numbers

Before we do anything else, it is very important that we agree on the definitions and names of some important collections of numbers.

- Natural numbers — These are the "whole numbers" 1,2,3,...that we learn first at about the same time as we learn the alphabet. We will denote this collection of numbers by the symbol "$\mathbb{N}$". The symbol $\mathbb{N}$ is written in a type of bold-face font that we call "black-board bold" (and is definitely *not* the same symbol as *N*). You should become used to writing a few letters in this way since it is typically used to denote collections of important numbers. Unfortunately there is often some confusion as to whether or not zero should be included[1]. In this text the natural numbers does not include zero.

  Notice that the set of natural numbers is *closed* under addition and multiplication. This means that if you take any two natural numbers and add them you get another natural number. Similarly if you take any two natural numbers and multiply them you get another natural number. However the set is not closed under subtraction or division; we need negative numbers and fractions to make collections of numbers closed under subtraction and division.

  Two important subsets of natural numbers are:

---

[1] This lack of agreement comes from some debate over how "natural" zero is — "how can nothing be something?" It was certainly not used by the ancient Greeks who really first looked at proof and number. If you are a mathematician then generally 0 is not a natural number. If you are a computer scientist then 0 generally is.

- Prime numbers — a natural number is prime when the only natural numbers that divide it exactly are 1 and itself. Equivalently it cannot be written as the product of two natural numbers neither of which are 1. Note that 1 is not a prime number[2].

- Composite numbers — a natural number is a composite number when it is not prime.

Hence the number 7 is prime, but $6 = 3 \times 2$ is composite.

- Integers — all positive and negative numbers together with the number zero. We denote the collection of all integers by the symbol "$\mathbb{Z}$". Again, note that this is not the same symbol as "Z", and we must write it in the same black-board bold font. The $\mathbb{Z}$ stands for the German *Zahlen* meaning numbers[3]. Note that $\mathbb{Z}$ is closed under addition, subtraction and multiplication, but not division.

  Two important subsets of integers are:

  - Even numbers — an integer is even if it is exactly divisible by 2, or equivalently if it can be written as the product of 2 and another integer. This means that $-14, 6$ and 0 are all even.

  - Odd numbers — an integer is odd when it is not even. Equivalently it can be written as $2k + 1$ where k is another integer. Thus $11 = 2 \times 5 + 1$ and $-7 = 2 \times (-4) + 1$ are both odd.

- Rational numbers — this is all numbers that can be written as the ratio of two integers. That is, any rational number r can be written as p/q where p, q are integers. We denote this collection by $\mathbb{Q}$ standing for *quoziente* which is Italian for quotient or ratio. Now we finally have a set of numbers which is closed under addition, subtraction, multiplication and division (of course you still need to be careful not to divide by zero).

- Real numbers — generally we think of these numbers as numbers that can be written as decimal expansions and we denote it by $\mathbb{R}$. It is beyond the scope of this text to go into the details of how to give a precise definition of real numbers, and the notion that a real number can be written as a decimal expansion will be sufficient.

  It took mathematicians quite a long time to realise that there were numbers that

---

2   If you let 1 be a prime number then you have to treat $1 \times 2 \times 3$ and 2×3 as different factorisations of the number 6. This causes headaches for mathematicians, so they don't let 1 be prime.

3   Some schools (and even some provinces!!) may use "*I*" for integers, but this is extremely non-standard and they really should use correct notation.

could not be written as ratios of integers[4]. The first numbers that were shown to be not-rational are square-roots of prime numbers, like $\sqrt{2}$. Other well known examples are π and e. Usually the fact that some numbers cannot be represented as ratios of integers is harmless because those numbers can be approximated by rational numbers to any desired precision.

The reason that we can approximate real numbers in this way is the surprising fact that between any two real numbers, one can always find a rational number. So if we are interested in a particular real number we can always find a rational number that is extremely close. Mathematicians refer to this property by saying that $\mathbb{Q}$ is *dense* in $\mathbb{R}$.

So to summarise

> **Definition 0.1.1** (Sets of numbers).
>
> This is not really a definition, but you should know these symbols
>
> - $\mathbb{N}$ = the natural numbers,
>
> - $\mathbb{Z}$ = the integers,
>
> - $\mathbb{Q}$ = the rationals, and
>
> - $\mathbb{R}$ = the reals.

⤠ More on Real Numbers

In the preceding paragraphs we have talked about the decimal expansions of real numbers and there is just one more point that we wish to touch on. The decimal expansions of rational numbers are always *periodic*, that is the expansion eventually starts to repeat itself. For example

$$\frac{2}{15} = 0.133333333\ldots$$
$$\frac{5}{17} = 0.\underline{2941176470588235}\underline{2941176470588235}\underline{2941176470588235}2941176470588823\ldots$$

---

[4]   The existence of such numbers caused mathematicians (particularly the ancient Greeks) all sorts of philosophical problems. They thought that the natural numbers were somehow fundamental and beautiful and "natural". The rational numbers you can get very easily by taking "ratios" — a process that is still somehow quite sensible. There were quite influential philosophers (in Greece at least) called Pythagoreans (disciples of Pythagoras originally) who saw numbers as almost mystical objects explaining all the phenomena in the universe, including beauty — famously they found fractions in musical notes etc and "numbers constitute the entire heavens". They believed that everything could be explained by whole numbers and their ratios. But soon after Pythagoras' theorem was discovered, so were numbers that are not rational. The first proof of the existence of irrational numbers is sometimes attributed to Hippasus in around 400BCE (not really known). It seems that his philosopher "friends" were not very happy about this and essentially exiled him. Some accounts suggest that he was drowned by them.

where we have underlined some of the last example to make the period clearer. On the other hand, irrational numbers, such as $\sqrt{2}$ and π, have expansions that never repeat.

If we want to think of real numbers as their decimal expansions, then we need those expansions to be unique. That is, we don't want to be able to write down two different expansions, each giving the same real number. Unfortunately there are an infinite set of numbers that do not have unique expansions. Consider the number 1. We usually just write "1", but as a decimal expansion it is

$$1.00000000000\ldots$$

that is, a single 1 followed by an infinite string of 0's. Now consider the following number

$$0.99999999999\ldots$$

This second decimal expansions actually represents the same number — the number 1. Let's prove this. First call the real number this represents q, then

$$q = 0.99999999999\ldots$$

Let's use a little trick to get rid of the long string of trailing 9's. Consider $10q$:

$$q = 0.99999999999\ldots$$
$$10q = 9.99999999999\ldots$$

If we now subtract one from the other we get

$$9q = 9.0000000000\ldots$$

and so we are left with $q = 1.0000000\ldots$. So both expansions represent the same real number.

Thankfully this sort of thing only happens with rational numbers of a particular form — those whose denominators are products of 2s and 5s. For example

$$\frac{3}{25} = 0.1200000\cdots = 0.119999999\ldots$$
$$-\frac{7}{32} = -0.2187500000\cdots = -0.2187499999\ldots$$
$$\frac{9}{20} = 0.45000000\cdots = 0.4499999\ldots$$

We can formalise this result in the following theorem (which we haven't proved in general, but it's beyond the scope of the text to do so):

> **Theorem 0.1.2.**
>
> Let x be a real number. Then x must fall into one of the following two categories,
>
> - x has a unique decimal expansion, or
>
> - x is a rational number of the form $\frac{a}{2^k 5^\ell}$ where $a \in \mathbb{Z}$ and k, l are non-negative integers.
>
> In the second case, x has exactly two expansions, one that ends in an infinite string of 9's and the other ending in an infinite string of 0's.

When we do have a choice of two expansions, it is usual to avoid the one that ends in an infinite string of 9's and write the other instead (omitting the infinite trailing string of 0's).

0.2 ▲ Sets

All of you will have done some basic bits of set-theory in school. Sets, intersection, unions, Venn diagrams etc etc. Set theory now appears so thoroughly throughout mathematics that it is difficult to imagine how Mathematics could have existed without it. It is really quite surprising that set theory is a much newer part of mathematics than calculus. Mathematically rigorous set theory was really only developed in the 19th Century — primarily by Georg Cantor[5]. Mathematicians were using sets before then (of course), however they were doing so without defining things too rigorously and formally.

In mathematics (and elsewhere, including "real life") we are used to dealing with collections of things. For example

- a family is a collection of relatives.

- hockey team is a collection of hockey players.

- shopping list is a collection of items we need to buy.

Generally when we give mathematical definitions we try to make them very formal and rigorous so that they are as clear as possible. We need to do this so that when we come across a mathematical object we can decide with complete certainty whether or not it satisfies the definition.

Unfortunately, it is the case that giving a completely rigorous definition of "set" would take up far more of our time than we would really like[6].

> **Definition 0.2.1** (A not-so-formal definition of set).
>
> A "set" is a collection of distinct objects. The objects are referred to as "elements" or "members" of the set.

Now — just a moment to describe some conventions. There are many of these in mathematics. These are not firm mathematical rules, but just traditions. It makes it much easier for people reading your work to understand what you are trying to say.

---

5   An extremely interesting mathematician who is responsible for much of our understanding of infinity. Arguably his most famous results are that there are more real numbers than integers, and that there are an infinite number of different infinities. His work, though now considered to be extremely important, was not accepted by his peers, and he was labelled "a corrupter of youth" for teaching it. For some reason we know that he spent much of his honeymoon talking and doing mathematics with Richard Dedekind.

6   The interested reader is invited to google (or whichever search engine you prefer — DuckDuckGo?) "Russell's paradox", "Axiomatic set theory" and "Zermelo-Fraenkel set theory" for a more complete and *far* more detailed discussion of the basics of sets and why, when you dig into them a little, they are not so basic.

- Use capital letters to denote sets, A, B, C, X, Y etc.

- Use lower case letters to denote elements of the sets a, b, c, x, y.

So when you are writing up homework, or just describing what you are doing, then if you stick with these conventions people reading your work (including the person marking your exams) will know — "Oh A is that set they are talking about" and "a is an element of that set.". On the other hand, if you use any old letter or symbol it is correct, but confusing for the reader. Think of it as being a bit like spelling — if you don't spell words correctly people can usually still understand what you mean, but it is much easier if you spell words the same way as everyone else.

We will encounter more of these conventions as we go — another good one is

- The letters i, j, k, l, m, n usually denote integers (like $1, 2, 3, -5, 18, \dots$).

- The letters x, y, z, w usually denote real numbers (like $1.4323, \pi, \sqrt{2}, 6.0221415 \times 10^{23} \dots$ and so forth).

So now that we have defined sets, what can we do with them? There is only thing we can ask of a set

"Is this object in the set?"

and the set will answer

"yes" or "no"

For example, if A is the set of even numbers we can ask "Is 4 in A?" We get back the answer "yes". We write this as

$$4 \in A$$

While if we ask "Is 3 in A?", we get back the answer "no". Mathematically we would write this as

$$3 \notin A$$

So this symbol "$\in$" is mathematical shorthand for "is an element of", while the same symbol with a stroke through it "$\notin$" is shorthand for "is not an element of".

Notice that both of these statements, though they are written down as short strings of three symbols, are really complete sentences. That is, when we read them out we have

| "$4 \in A$" | is read as | "Four is an element of A." |
| "$3 \notin A$" | is read as | "Three is not an element of A." |

The mathematical symbols like "$+$", "$=$" and "$\in$" are shorthand[7] and mathematical statements like "$4 + 3 = 7$" are complete sentences.

---

7   Precise definitions aside, by "shorthand" we mean a collection of accepted symbols and abbreviations to allow us to write more quickly and hopefully more clearly. People have been using various systems of shorthand as long as people have been writing. Many of these are used and understood only by the individual, but if you want people to be able to understand what you have written, then you need to use shorthand that is commonly understood.

This is an important point — mathematical writing is just like any other sort of writing. It is very easy to put a bunch of symbols or words down on the page, but if we would like it to be easy to read and understand, then we have to work a bit harder. When you write mathematics you should keep in mind that someone else should be able to read it and understand it.

> *Easy reading is damn hard writing.*

> Nathaniel Hawthorne, but possibly also a few others like Richard Sheridan.

We will come across quite a few different sets when doing mathematics. It must be completely clear from the definition how to answer the question "Is this object in the set or not?"

- "Let A be the set of even integers between 1 and 13." — nice and clear.

- "Let B be the set of tall people in this class room." — not clear.

More generally if there are only a small number of elements in the set we just list them all out

- "Let $C = \{1, 2, 3\}$."

When we write out the list we put the elements inside braces "$\{\cdot\}$". Note that the order we write things in doesn't matter

$$C = \{1, 2, 3\} = \{2, 1, 3\} = \{3, 2, 1\}$$

because the only thing we can ask is "Is this object an element of C?" We cannot ask more complex questions like "What is the third element of C?" — we require more sophisticated mathematical objects to ask such questions[8]. Similarly, it doesn't matter how many times we write the same object in the list

$$C = \{1, 1, 1, 2, 3, 3, 3, 3, 1, 2, 1, 2, 1, 3\} = \{1, 2, 3\}$$

because all we ask is "Is $1 \in C$?". Not "how many times is 1 in C?".

Now — if the set is a bit bigger then we might write something like this

- $C = \{1, 2, 3, \dots, 40\}$ the set of all integers between 1 and 40 (inclusive).

- $A = \{1, 4, 9, 16, \dots\}$ the set of all perfect squares[9]

The "$\dots$" is again shorthand for the missing entries. You have to be careful with this as you can easily confuse the reader

- $B = \{3, 5, 7, \dots\}$ — is this all odd primes, or all odd numbers bigger than 1 or ?? What is written is not sufficient for us to have a firm idea of what the writer intended.

Only use this where it is completely clear by context. A few extra words can save the reader (and yourself) a lot of confusion.

> *Always think about the reader.*

---

8   The interested reader is invited to look at "lists", "multisets", "totally ordered sets" and "partially ordered sets" amongst many other mathematical objects that generalise the basic idea of sets.

9   i.e. integers that can be written as the square of another integer.

0.3 ▴ Other Important Sets

We have seen a few important sets above — namely $\mathbb{N}, \mathbb{Z}, \mathbb{Q}$ and $\mathbb{R}$. However, arguably the most important set in mathematics is the empty set.

> **Definition 0.3.1** (Empty set).
>
> The empty set (or null set or void set) is the set which contains no elements. It is denoted $\varnothing$. For any object x, we always have $x \notin \varnothing$; hence $\varnothing = \{\}$.

Note that it is important to realise that the empty set is not *nothing*; think of it as an empty bag. Also note that with quite a bit of hard work you can actually define the natural numbers in terms of the empty set. Doing so is very formal and well beyond the scope of this text.

When a set does not contain too many elements it is fine to specify it by listing out its elements. But for infinite sets or even just big sets we can't do this and instead we have to give the defining rule. For example the set of all perfect square numbers we write as

$$S = \{x \text{ s.t. } x = k^2 \text{ where } k \in \mathbb{Z}\}$$

Notice we have used another piece of shorthand here, namely s.t. , which stands for "such that" or "so that". We read the above statement as "S is the set of elements x such that x equals k-squared where k is an integer". This is the standard way of writing a set defined by a rule, though there are several shorthands for "such that". We shall use two them:

$$P = \{p \text{ s.t. } p \text{ is prime}\} = \{p \mid p \text{ is prime}\}$$

Other people also use ":" as shorthand for "such that". You should recognise all three of these shorthands.

> Example 0.3.2 (examples of sets)
>
> Even more examples...
>
> - Let $A = \{2, 3, 5, 7, 11, 13, 17, 19\}$ and let
>
> $$B = \{a \in A \mid a < 8\} = \{2, 3, 5, 7\}$$
>
>   the set of elements of A that are strictly less than 8.
>
> - Even and odd integers
>
> $$\begin{aligned} E &= \{n \mid n \text{ is an even integer}\} \\ &= \{n \mid n = 2k \text{ for some } k \in \mathbb{Z}\} \\ &= \{2n \mid n \in \mathbb{Z}\}, \end{aligned}$$
>
>   and similarly
>
> $$\begin{aligned} O &= \{n \mid n \text{ is an odd integer}\} \\ &= \{2n + 1 \mid n \in \mathbb{Z}\}. \end{aligned}$$

- Square integers

$$S = \{n^2 | n \in \mathbb{Z}\}.$$

The set[10] $S' = \{n^2 | n \in \mathbb{N}\}$ is not the same as S because S' does not contain the number 0, which is definitely a square integer and 0 is in S. We could also write $S = \{n^2 | n \in \mathbb{Z}, n \geqslant 0\}$ and $S = \{n^2 | n = 0, 1, 2, \dots\}$.

Example 0.3.2

The sets A and B in the above example illustrate an important point. Every element in B is an element in A, and so we say that B is a subset of A

Definition 0.3.3.

Let A and B be sets. We say "A is a subset of B" if every element of A is also an element of B. We denote this $A \subseteq B$ (or $B \supseteq A$). If A is a subset of B and A and B are not the same , so that there is some element of B that is not in A then we say that A is a proper subset of B. We denote this by $A \subset B$ (or $B \supset A$).

Two things to note about subsets:

- Let A be a set. It is always the case that $\varnothing \subseteq A$.

- If A is not a subset of B then we write $A \nsubseteq B$. This is the same as saying that there is some element of A that is not in B. That is, there is some $a \in A$ such that $a \notin B$.

Example 0.3.4 (subsets)

Let $S = \{1, 2\}$. What are all the subsets of S? Well — each element of S can either be in the subset or not (independent of the other elements of the set). So we have $2 \times 2 = 4$ possibilities: neither 1 nor 2 is in the subset, 1 is but 2 is not, 2 is but 1 is not, and both 1 and 2 are. That is

$$\varnothing, \{1\}, \{2\}, \{1, 2\} \subseteq S$$

This argument can be generalised with a little work to show that a set that contains exactly n elements has exactly 2^n subsets.

Example 0.3.4

In much of our work with functions later in the text we will need to work with subsets of real numbers, particularly segments of the "real line". A convenient and standard way of representing such subsets is with interval notation.

10　Notice here we are using another common piece of mathematical short-hand. Very often in mathematics we will be talking or writing about some object, like the set S above, and then we will create a closely related object. Rather than calling this new object by a new symbol (we could have used T or R or...), we instead use the same symbol but with some sort of accent — such as the little single quote mark we added to the symbol S to make S' (read "S prime"). The point of this is to let the reader know that this new object is related to the original one, but not the same. You might also see $\dot{S}, \hat{S}, \bar{S}, \tilde{S}$ and others.

Definition 0.3.5 (Open and closed intervals of $\mathbb{R}$).

Let $a, b \in \mathbb{R}$ such that $a < b$. We name the subset of all numbers between a and b in different ways depending on whether or not the ends of the interval (a and b) are elements of the subset.

- The closed interval $[a, b] = \{x \in \mathbb{R} : a \leqslant x \leqslant b\}$ — both end points are included.

- The open interval $(a, b) = \{x \in \mathbb{R} : a < x < b\}$ — neither end point is included.

We also define half-open[11] intervals which contain one end point but not the other:

$$(a, b] = \{x \in \mathbb{R} : a < x \leqslant b\} \qquad [a, b) = \{x \in \mathbb{R} : a \leqslant x < b\}$$

We sometimes also need unbounded intervals

$$[a, \infty) = \{x \in \mathbb{R} : a \leqslant x\} \qquad (a, \infty) = \{x \in \mathbb{R} : a < x\}$$
$$(-\infty, b] = \{x \in \mathbb{R} : x \leqslant b\} \qquad (-\infty, b) = \{x \in \mathbb{R} : x < b\}$$

These unbounded intervals do not include "$\pm\infty$", so that end of the interval is always open[12].

⤳ More on Sets

So we now know how to say that one set is contained within another. We will now define some other operations on sets. Let us also start to be a bit more precise with our definitions and set them out carefully as we get deeper into the text.

Definition 0.3.6.

Let A and B be sets. We define the union of A and B, denoted $A \cup B$, to be the set of all elements that are in at least one of A or B.

$$A \cup B = \{x | x \in A \text{ or } x \in B\}$$

---

11 Also called "half-closed". The preference for one term over the other may be related to whether a 500ml glass containing 250ml of water is half-full or half-empty.

12 Infinity is not a real number. As mentioned in an earlier footnote, Cantor proved that there are an infinite number of different infinities and so it is incorrect to think of ∞ as being a single number. As such it cannot be an element in an interval of the real line. We suggest that the reader that wants to learn more about how mathematics handles infinity look up transfinite numbers and transfinite arithmetic. Needless to say these topics are beyond the scope of this text.

It is important to realise that we are using the word "or" in a careful mathematical sense. We mean that x belongs to A or x belongs to B *or both*. Whereas in normal everyday English "or" is often used to be "exclusive or" — A or B but not both[13].

We also start the definition by announcing "Definition" so that the reader knows "We are about to define something important". We should also make sure that everything is (reasonably) self-contained — we are not assuming the reader already knows A and B are sets.

It is vital that we make our definitions clear otherwise anything we do with the definitions will be very difficult to follow. As writers we must try to be nice to our readers[14].

Definition 0.3.7.

Let A and B be sets. We define the intersection of A and B, denoted $A \cap B$, to be the set of elements that belong to both A and B.

$$A \cap B = \{x \mid x \in A \text{ and } x \in B\}$$

Again note that we are using the word "and" in a careful mathematical sense (which is pretty close to the usual use in English).

Example 0.3.8 (Union and intersection)

Let $A = \{1,2,3,4\}$, $B = \{p : p \text{ is prime}\}$, $C = \{5,7,9\}$ and $D = \{\text{even positive integers}\}$. Then

$$A \cap B = \{2,3\}$$
$$B \cap D = \{2\}$$
$$A \cup C = \{1,2,3,4,5,7,9\}$$
$$A \cap C = \varnothing$$

In this last case we see that the two sets have no elements in common — they are said to be *disjoint*.

Example 0.3.8

0.4 ▲ Functions

Now that we have reviewed basic ideas about sets we can start doing more interesting things with them — functions.

---

13 When you are asked for your dining preferences on a long flight you are usually asked something like "Chicken or beef?" — you get one or the other, but not both. Unless you are way at the back near the toilets in which case you will be presented with which ever meal was less popular. Probably fish.
14 If you are finding this text difficult to follow then please complain to us authors and we will do our best to improve it.

When we are introduced to functions in mathematics, it is almost always as formulas. We take a number x and do some things to it to get a new number y. For example,

$$y = f(x) = 3x - 7$$

Here, we take a number x, multiply it by 3 and then subtract seven to get the result.

This view of functions — a function is a formula — was how mathematicians defined them up until the 19th century. As basic ideas of sets became better defined, people revised ideas surrounding functions. The more modern definition of a function between two sets is that it is a rule which assigns to each element of the first set a unique element of the second set.

Consider the set of days of the week, and the set containing the alphabet

$A = \{$Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday$\}$

$B = \{$a,b,c,d,e,...,x,y,z$\}$

We can define a function f that takes a day (that is, an element of A) and turns it into the first letter of that day (that is, an element of B). This is a valid function, though there is no formula. We can draw a picture of the function as

Figure 0.4.1.
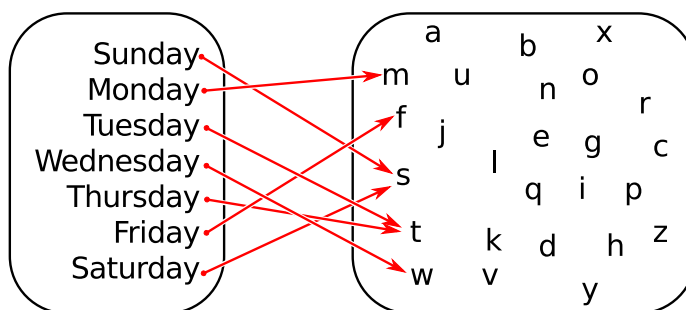
Clearly such pictures will work for small sets, but will get very messy for big ones. When we shift back to talking about functions on real numbers, then we will switch to using graphs of functions on the Cartesian plane.

This example is pretty simple, but this serves to illustrate some important points. If our function gives us a rule for taking elements in A and turning them into elements from B then

- the function must be defined for all elements of A — that is, no matter which element of A we choose, the function must be able to give us an answer. Every function must have this property.

- on the other hand, we don't have to "hit" every element from B. In the above example, we miss almost all the letters in B. A function that does reach every element of B is said to be "surjective" or "onto".

- a given element of B may be reached by more than one element of A. In the above example, the days "Tuesday" and "Thursday" both map to the letter T and similarly the letters S is mapped to by both "Sunday" and "Saturday". A function which does not do this, that is, every element in A maps to a different element in B is called "injective" or "one-to-one" — again we will come back to this later when we discuss inverse function in Section 0.6.

Summarising this more formally, we have

Definition 0.4.1.

Let A, B be non-empty sets. A function f from A to B, is a rule or formula that takes elements of A as inputs and returns elements of B as outputs. We write this as

$$f : A \to B$$

and if f takes $a \in A$ as an input and returns $b \in B$ then we write this as $f(a) = b$. Every function must satisfy the following two conditions

- The function must be defined on every possible input from the set A. That is, no matter which element $a \in A$ we choose, the function must return an element $b \in B$ so that $f(a) = b$.

- The function is only allowed to return one result for each input[15]. So if we find that $f(a) = b_1$ and $f(a) = b_2$ then the only way that f can be a function is if b_1 is exactly the same as b_2.

We must include the input and output sets A and B in the definition of the function. This is one of the reasons that we should not think of functions as just formulas. The input and output sets have proper mathematical names, which we give below:

---

15   You may have learned this in the context of plotting functions on the Cartesian plane, as "the vertical line test". If the graph intersects a vertical line twice, then the same x-value will give two y-values and so the graph does not represent a function.

Definition 0.4.2.

Let $f : A \to B$ be a function. Then

- the set A of inputs to our function is the "domain" of f,

- the set B which contains all the results is called the codomain,

- We read "$f(a) = b$" as "f of a is b", but sometimes we might say "f maps a to b" or "b is the image of a".

- The codomain must contain all the possible results of the function, but it might also contain a few other elements. The subset of B that is exactly the outputs of A is called the "range" of f. We define it more formally by

$$\text{range of } f = \{b \in B \mid \text{there is some } a \in A \text{ so that } f(a) = b\}$$
$$= \{f(a) \in B \mid a \in A\}$$

  The only elements allowed in that set are those elements of B that are the images of elements in A.

Example 0.4.3 (domains and ranges)

Let us go back to the "days of the week" function example that we worked on above, we can define the domain, codomain and range:

- The domain, A, is the set of days of the week.

- The codomain, B, is the 26 letters of the alphabet.

- The range is the set $\{F, M, T, S, W\}$ — no other elements of B are images of inputs from A.

Example 0.4.3

Example 0.4.4 (more domains and ranges)

A more numerical example — let $g : \mathbb{R} \to \mathbb{R}$ be defined by the formula $g(x) = x^2$. Then

- the domain and codomain are both the set of all real numbers, but

- the range is the set $[0, \infty)$.

Now — let $h : [0, \infty) \to [0, \infty)$ be defined by the formula $h(x) = \sqrt{x}$. Then

- the domain and codomain are both the set $[0, \infty)$, that is all non-negative real numbers, and

- in this case the range is equal to the codomain, namely $[0, \infty)$.

Example 0.4.4

Example 0.4.5 (piece-wise function)

Yet another numerical example.

$$V : [-1, 1] \to \mathbb{R} \qquad \text{defined by } V(t) = \begin{cases} 0 & \text{if } -1 \leqslant t < 0 \\ 120 & \text{if } 0 \leqslant t \leqslant 1 \end{cases}$$

This is an example of a "piece-wise" function — that is, one that is not defined by a single formula, but instead defined piece-by-piece. This function has domain $[-1, 1]$ and its range is $\{0, 120\}$. We could interpret this function as measuring the voltage across a switch that is flipped on at time $t = 0$.

Example 0.4.5

Almost all the functions we look at from here on will be formulas. However it is important to note, that we have to include the domain and codomain when we describe the function. If the domain and codomain are not stated explicitly then we should assume that both are $\mathbb{R}$.

0.5 ▲ Parsing Formulas

Consider the formula

$$f(x) = \frac{1 + x}{1 + 2x - x^2}$$

This is an example of a simple rational function — that is, the ratio of two polynomials. When we start to examine these functions later in the text, it is important that we are able to understand how to evaluate such functions at different values of x. For example

$$f(5) = \frac{1 + 5}{1 + 10 - 25} = \frac{6}{-14} = -\frac{3}{7}$$

More important, however, is that we understand how we decompose this function into simpler pieces. Since much of your calculus course will involve creating and studying complicated functions by building them up from simple pieces, it is important that you really understand this point.

Now to get there we will take a small excursion into what are called parse-trees. You already implicitly use these when you evaluate the function at a particular value of x, but our aim here is to formalise this process a little more.

We can express the steps used to evaluate the above formula as a tree-like diagram[16]. We can decompose this formula as the following tree-like diagram

16   Such trees appear in many areas of mathematics and computer science. The reason for the name is that they look rather like trees — starting from their base they grow and branch out towards their many leaves. For some reason, which remains mysterious, they are usually drawn upside down.

Figure 0.5.1.
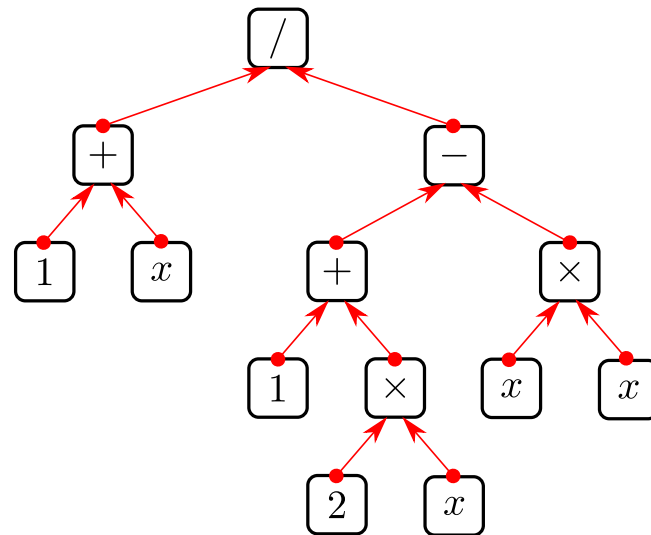
A parse tree of the function $\frac{1+x}{1+2x-x^2}$.

Let us explain the pieces here.

- The picture consists of boxes and arrows which are called "nodes" and "edges" respectively.

- There are two types of boxes, those containing numbers and the variable x, and those containing arithmetic operations "+","−","×" and "/".

- If we wish to represent the formula $3 + 5$, then we can draw this as the following cherry-like configuration
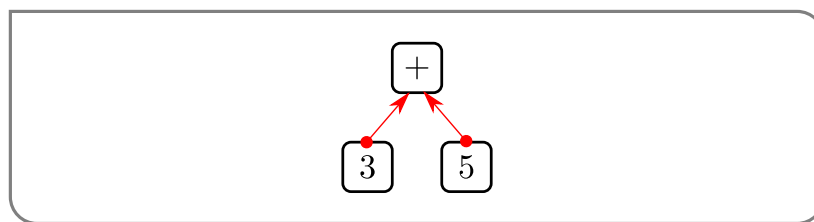
which tells us to take the numbers "3" and "5" and add them together to get 8.

evaluates to
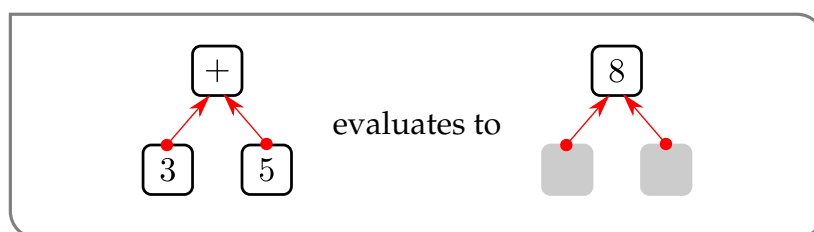
- By stringing such little "cherries" together we can describe more complicated formulas. For example, if we compute "$(3 + 5) \times 2$", we first compute "$(3 + 5)$" and then multiply the result by 2. The corresponding diagrams are
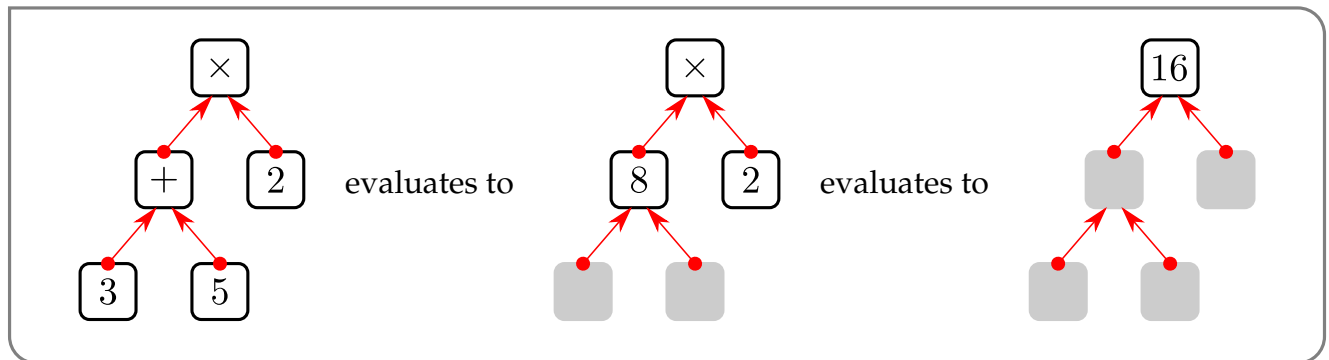
The tree we drew in Figure 0.5.1 above representing our formula has x in some of the boxes, and so when we want to compute the function at a particular value of x — say at $x = 5$ — then we replace those "x"s in the tree by that value and then compute back up the tree. See the example below

Figure 0.5.2.

Start

$\mapsto$

$\mapsto$

$\mapsto$

$\mapsto$

and we are done.
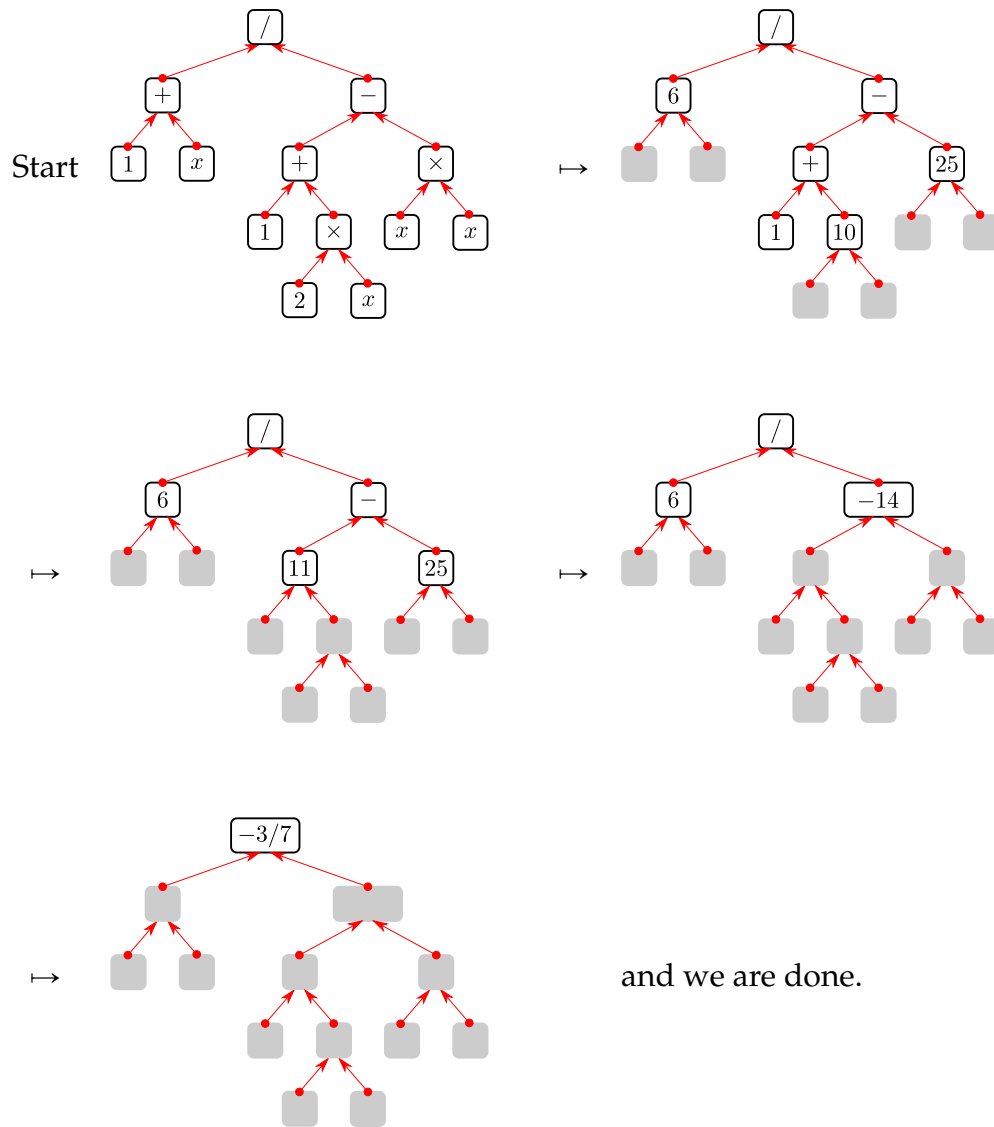
This is not the only parse tree associated with the formula for $f(x)$; we could also decompose it as

Figure 0.5.3.
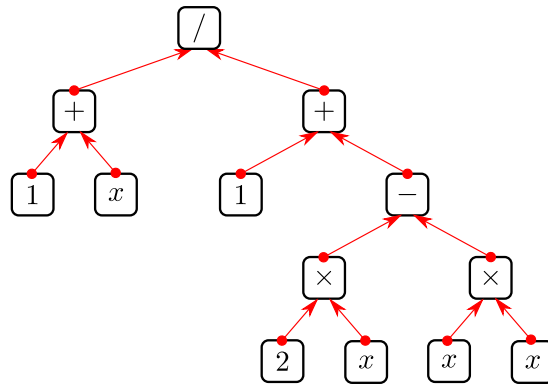
We are able to do this because when we compute the denominator $1 + 2x - x^2$, we can compute it as

$$1 + 2x - x^2 = \text{ either } (1 + 2x) - x^2 \text{ or } = 1 + (2x - x^2).$$

Both[17] are correct because addition is "associative". Namely

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

Multiplication is also associative:

$$a \times b \times c = (a \times b) \times c = a \times (b \times c).$$

Example 0.5.1 (parsing a formula)

Consider the formula

$$g(t) = \left(\frac{t + \pi}{t - \pi} \right) \cdot \sin \left(\frac{t + \pi}{2} \right).$$

This introduces a new idea — we have to evaluate $\frac{t+\pi}{2}$ and then compute the sine of that number. The corresponding tree can be written as
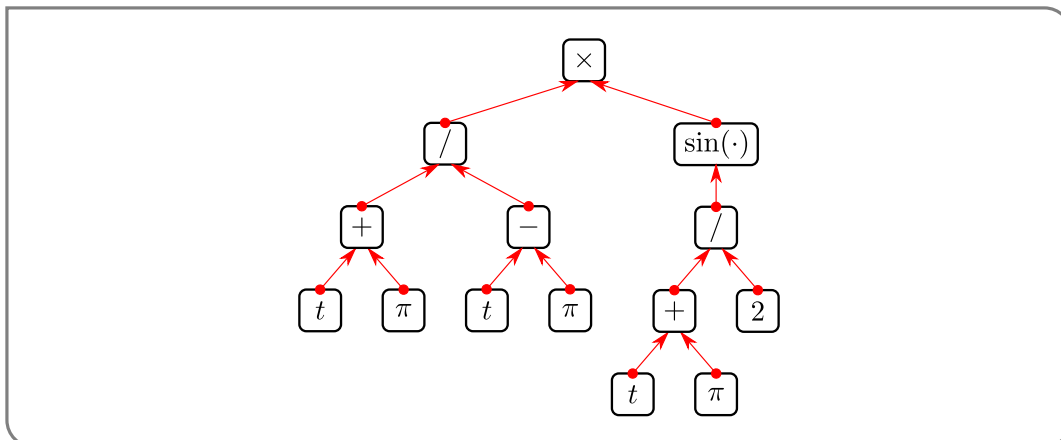
17   We could also use, for example, $1 + 2x - x^2 = (1 - x^2) + 2x$.

If we want to evaluate this at $t = \pi/2$ then we get the following...

Example 0.5.1
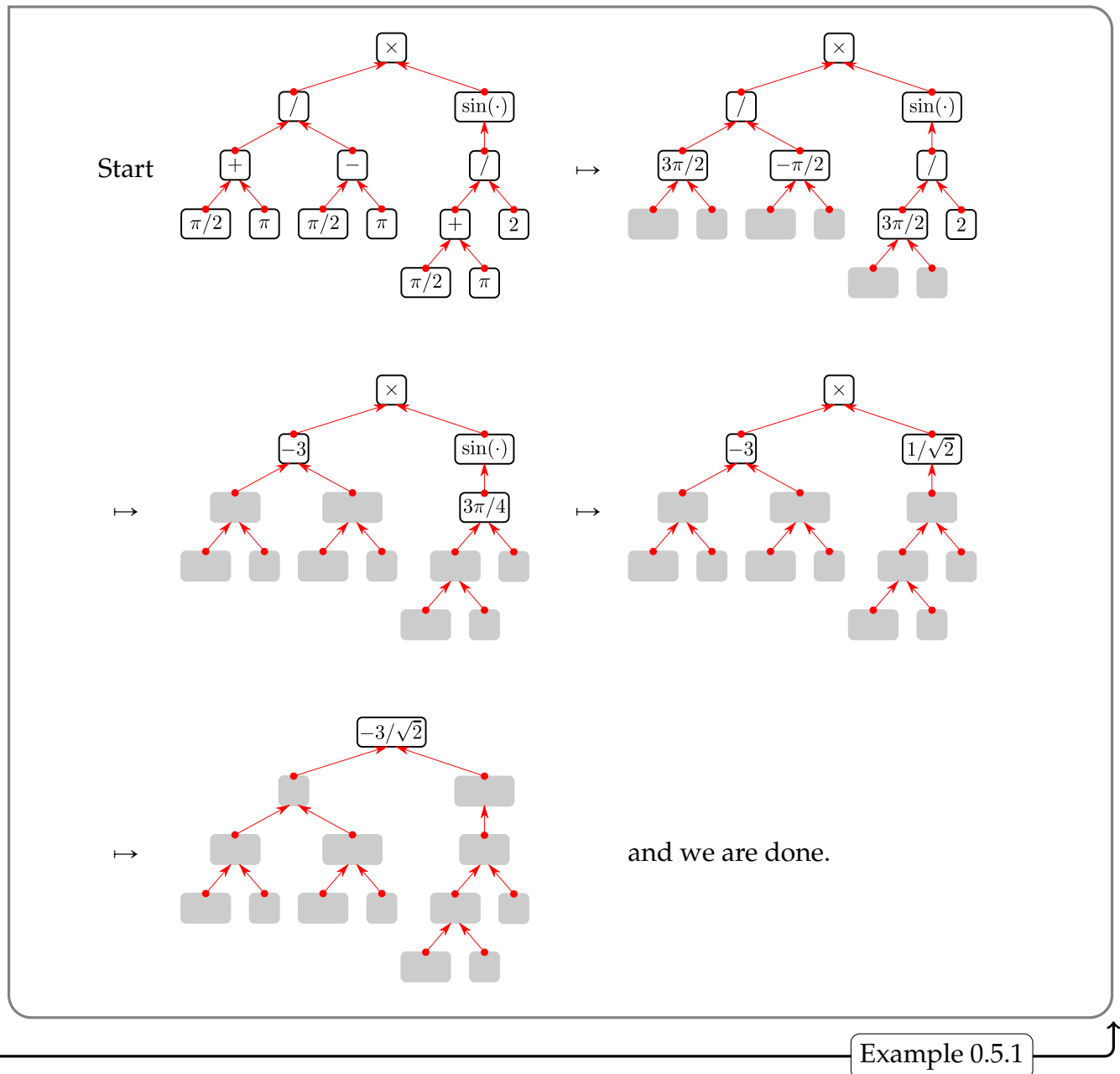
It is highly unlikely that you will ever need to explicitly construct such a tree for any problem in the remainder of the text. The main point of introducing these objects and working through a few examples is to realise that all the functions that we will examine are constructed from simpler pieces. In particular we have constructed all the above examples from simple "building blocks"

- constants — fixed numbers like $1, \pi$ and so forth

- variables — usually x or t, but sometimes other symbols

- standard functions — like trigonometric functions (sine, cosine and tangent), exponentials and logarithms.

These simple building blocks are combined using arithmetic

- addition and subtraction — $a + b$ and $a - b$

- multiplication and division — $a \cdot b$ and a/b

- raising to a power — a^n

- composition — given two functions $f(x)$ and $g(x)$ we form a new function $f(g(x))$ by evaluating $y = g(x)$ and then evaluating $f(y) = f(g(x))$.

During the rest of the course when we learn how to compute limits and derivatives, our computations require us to understand the way we construct functions as we have just described.

That is, in order to compute the derivative[18] of a function we have to see how to construct the function from these building blocks (i.e. the constants, variables and standard functions) using arithmetic operations. We will then construct the derivative by following these same steps. There will be simple rules for finding the derivatives of the simpler pieces and then rules for putting them together following the arithmetic used to construct the function.

0.6 ▲ Inverse Functions

There is one last thing that we should review before we get into the main material of the course and that is inverse functions. As we have seen above functions are really just rules for taking an input (almost always a number), processing it somehow (usually by a formula) and then returning an output (again, almost always a number).

$$\text{input number } x \quad \mapsto \quad f \text{ does "stuff" to } x \quad \mapsto \quad \text{return number } y$$

In many situations it will turn out to be very useful if we can undo whatever it is that our function has done. ie

$$\text{take output } y \quad \mapsto \quad \text{do "stuff" to } y \quad \mapsto \quad \text{return the original } x$$

When it exists, the function "which undoes" the function $f(x)$ is found by solving $y = f(x)$ for x as a function of y and is called the inverse function of f. It turns out that it is not always possible to solve $y = f(x)$ for x as a function of y. Even when it is possible, it can be really hard to do[19].

For example — a particle's position, s, at time t is given by the formula $s(t) = 7t$ (sketched below). Given a calculator, and any particular number t, you can quickly work out the corresponding positions s. However, if you are asked the question "When does the particle reach $s = 4$?" then to answer it we need to be able to "undo" $s(t) = 4$ to

---

18   We get to this in Chapter 2 — don't worry about exactly what it is just now.
19   Indeed much of encryption exploits the fact that you can find functions that are very quick to do, but very hard to undo. For example — it is very fast to multiply two large prime numbers together, but very hard to take that result and factor it back into the original two primes. The interested reader should look up trapdoor functions.

isolate t. In this case, because $s(t)$ is always increasing, we can always undo $s(t)$ to get a unique answer:

$$s(t) = 7t = 4 \qquad \text{if and only if} \qquad t = \frac{4}{7}.$$

However, this question is not always so easy. Consider the sketch of $y = \sin(x)$ below; when is $y = \frac{1}{2}$? That is, for which values x is $\sin(x) = \frac{1}{2}$? To rephrase it again, at which values of x does the curve $y = \sin x$ (which is sketched in the right half of Figure 0.6.1) cross the horizontal straight line $y = \frac{1}{2}$ (which is also sketched in the same figure)?

Figure 0.6.1.
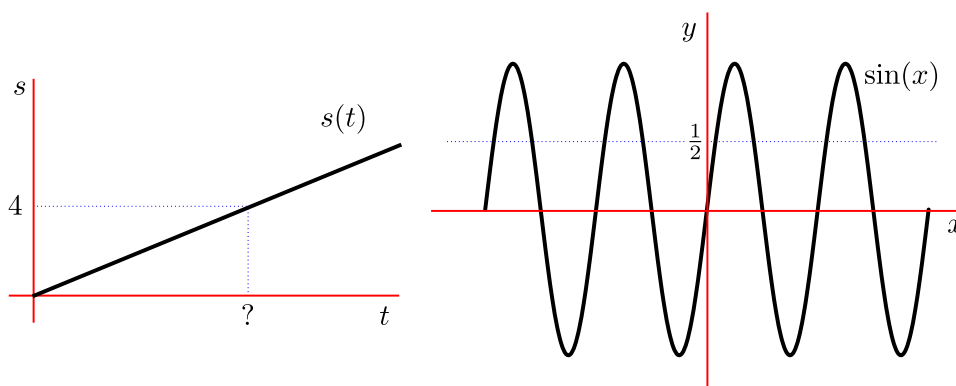
We can see that there are going to be an infinite number of x-values that give $y = \sin(x) = \frac{1}{2}$; there is no unique answer.

Recall (from Definition 0.4.1) that for any given input, a function must give a unique output. So if we want to find a *function* that undoes $s(t)$, then things are good — because each s-value corresponds to a unique t-value. On the other hand, the situation with $y = \sin x$ is problematic — any given y-value is mapped to by many different x-values. So when we look for an *unique* answer to the question "When is $\sin x = \frac{1}{2}$?" we cannot answer it.

This "uniqueness" condition can be made more precise:

Definition 0.6.1.

A function f is one-to-one (injective) when it never takes the same y value more than once. That is

$$\text{if } x_1 \neq x_2 \text{ then } f(x_1) \neq f(x_2)$$

There is an easy way to test this when you have a plot of the function — the horizontal line test.

Definition 0.6.2 (Horizontal line test).

A function is one-to-one if and only if no horizontal line $y = c$ intersects the graph $y = f(x)$ more than once.

i.e. every horizontal line intersects the graph either zero or one times. Never twice or more. This test tell us that $y = x^3$ is one-to-one, but $y = x^2$ is not. However note that if we restrict the domain of $y = x^2$ to $x \geqslant 0$ then the horizontal line test is passed. This is one of the reasons we have to be careful to consider the domain of the function.

Figure 0.6.2.
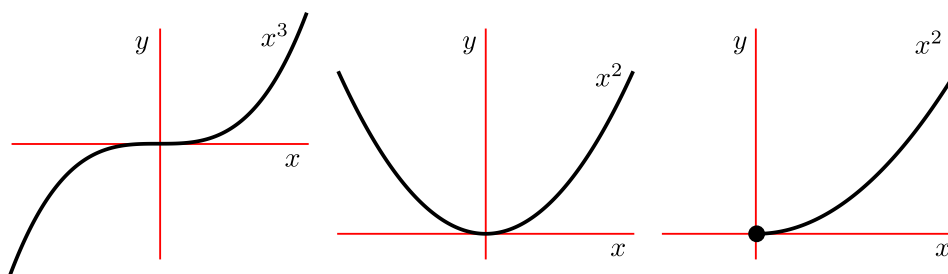
When a function is one-to-one then it has an inverse function.

Definition 0.6.3.

Let f be a one-to-one function with domain A and range B. Then its inverse function is denoted f^{-1} and has domain B and range A. It is defined by

$$f^{-1}(y) = x \qquad \text{whenever} \qquad f(x) = y$$

for any $y \in B$.
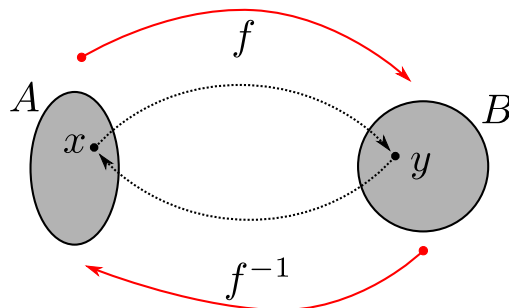
So if f maps x to y, then f^{-1} maps y back to x. That is f^{-1} "undoes" f. Because of this

we have

$$f^{-1}(f(x)) = x \qquad\qquad \text{for any } x \in A$$
$$f(f^{-1}(y)) = y \qquad\qquad \text{for any } y \in B$$

We have to be careful not to confuse $f^{-1}(x)$ with $\dfrac{1}{f(x)}$. The "-1" is not an exponent.

Example 0.6.4

Let $f(x) = x^5 + 3$ on domain $\mathbb{R}$. To find its inverse we do the following

- Write $y = f(x)$; that is $y = x^5 + 3$.

- Solve for x in terms of y (this is not always easy) — $x^5 = y - 3$, so $x = (y-3)^{1/5}$.

- The solution is $f^{-1}(y) = (y-3)^{1/5}$.

- Recall that the "y" in $f^{-1}(y)$ is a dummy variable. That is, $f^{-1}(y) = (y-3)^{1/5}$ means that if you feed the number y into the function f^{-1} it outputs the number $(y-3)^{1/5}$. You may call the input variable anything you like. So if you wish to call the input variable "x" instead of "y" then just replace every y in $f^{-1}(y)$ with an x.

- That is $f^{-1}(x) = (x-3)^{1/5}$.

Example 0.6.4

Example 0.6.5

Let $g(x) = \sqrt{x-1}$ on the domain $x \geqslant 1$. We can find the inverse in the same way:

$$y = \sqrt{x-1}$$
$$y^2 = x - 1$$
$$x = y^2 + 1 = f^{-1}(y) \qquad\qquad \text{or, writing input variable as "x":}$$
$$f^{-1}(x) = x^2 + 1.$$

Example 0.6.5

Let us now turn to finding the inverse of $\sin(x)$ — it is a little more tricky and we have to think carefully about domains.

Example 0.6.6

We have seen (back in Figure 0.6.1) that $\sin(x)$ takes each value y between -1 and $+1$ for infinitely many different values of x (see the left-hand graph in the figure below). Consequently $\sin(x)$, with domain $-\infty < x < \infty$ does not have an inverse function.
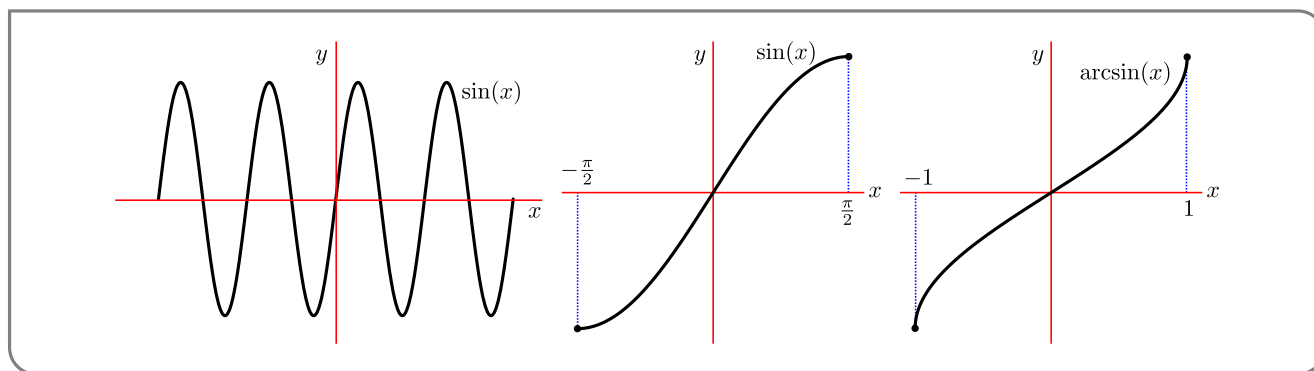
But notice that as x runs from $-\frac{\pi}{2}$ to $+\frac{\pi}{2}$, $\sin(x)$ increases from -1 to $+1$. (See the middle graph in the figure above.) In particular, $\sin(x)$ takes each value $-1 \leqslant y \leqslant 1$ for exactly one $-\frac{\pi}{2} \leqslant x \leqslant \frac{\pi}{2}$. So if we restrict $\sin x$ to have domain $-\frac{\pi}{2} \leqslant x \leqslant \frac{\pi}{2}$, it does have an inverse function, which is traditionally called arcsine (see Appendix A.9).

That is, by definition, for each $-1 \leqslant y \leqslant 1$, $\arcsin(y)$ is the unique $-\frac{\pi}{2} \leqslant x \leqslant \frac{\pi}{2}$ obeying $\sin(x) = y$. Equivalently, exchanging the dummy variables x and y throughout the last sentence gives that for each $-1 \leqslant x \leqslant 1$, $\arcsin(x)$ is the unique $-\frac{\pi}{2} \leqslant y \leqslant \frac{\pi}{2}$ obeying $\sin(y) = x$.

Example 0.6.6

It is an easy matter to construct the graph of an inverse function from the graph of the original function. We just need to remember that

$$Y = f^{-1}(X) \iff f(Y) = X$$

which is $y = f(x)$ with x renamed to Y and y renamed to X.

Start by drawing the graph of f, labelling the x– and y–axes and labelling the curve $y = f(x)$.
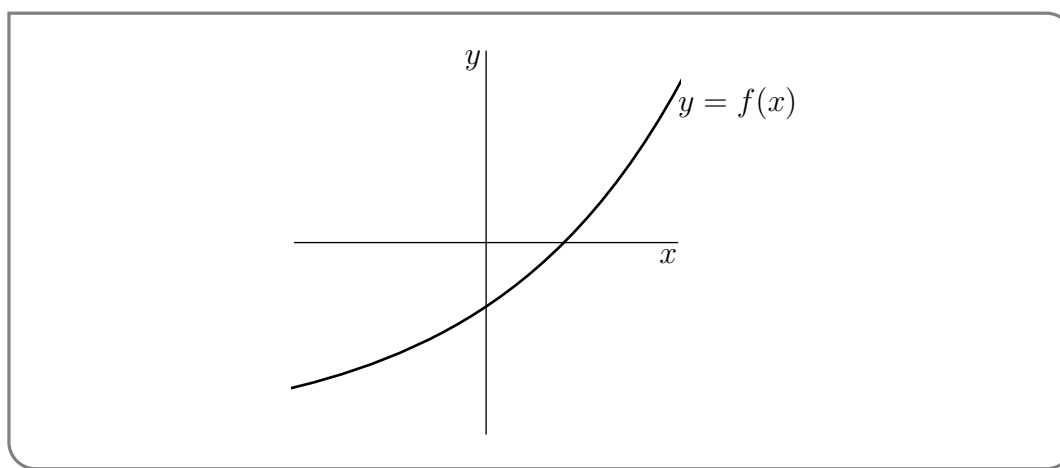
Now replace each x by Y and each y by X and replace the resulting label $X = f(Y)$ on the curve by the equivalent $Y = f^{-1}(X)$.
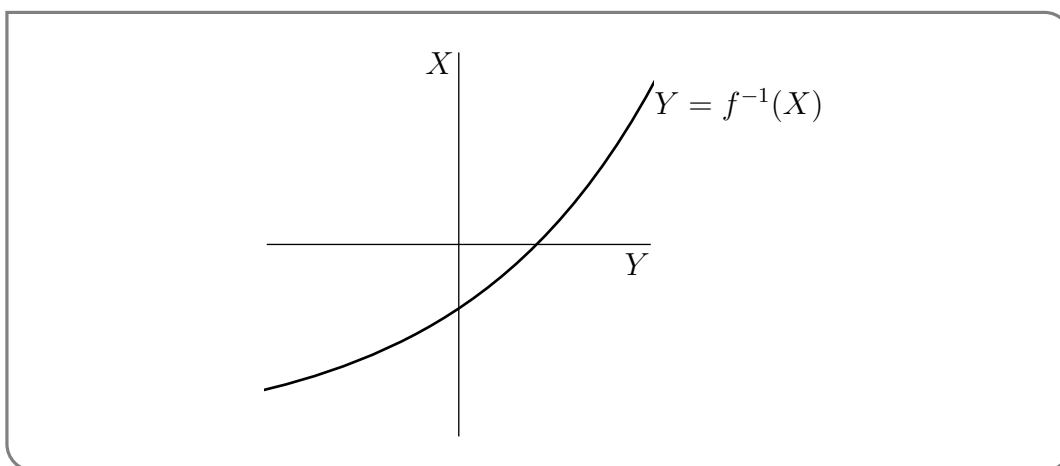
Finally we just need to redraw the sketch with the Y axis running vertically (with Y increasing upwards) and the X axis running horizontally (with X increasing to the right). To do so, pretend that the sketch is on a transparency or on a very thin piece of paper that you can see through. Lift the sketch up and flip it over so that the Y axis runs vertically and the X axis runs horizontally. If you want, you can also convert the upper case X into a lower case x and the upper case Y into a lower case y.
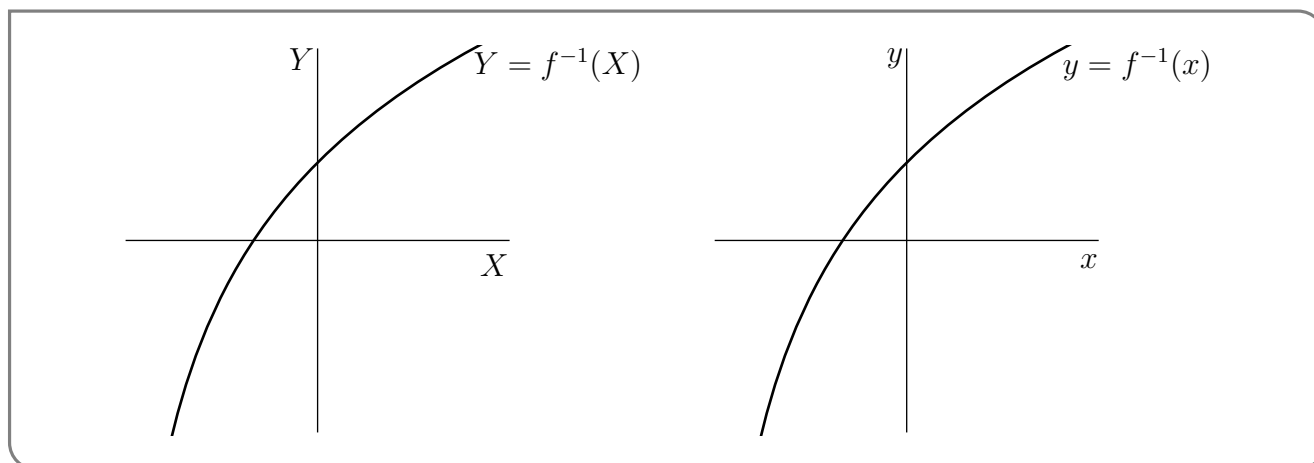
Another way to say "flip the sketch over so as to exchange the x– and y–axes" is "reflect in the line $y = x$". In the figure below the blue "horizontal" elliptical disk that is centred on (a, b) has been reflected in the line $y = x$ to give the red "vertical" elliptical disk centred on (b, a).
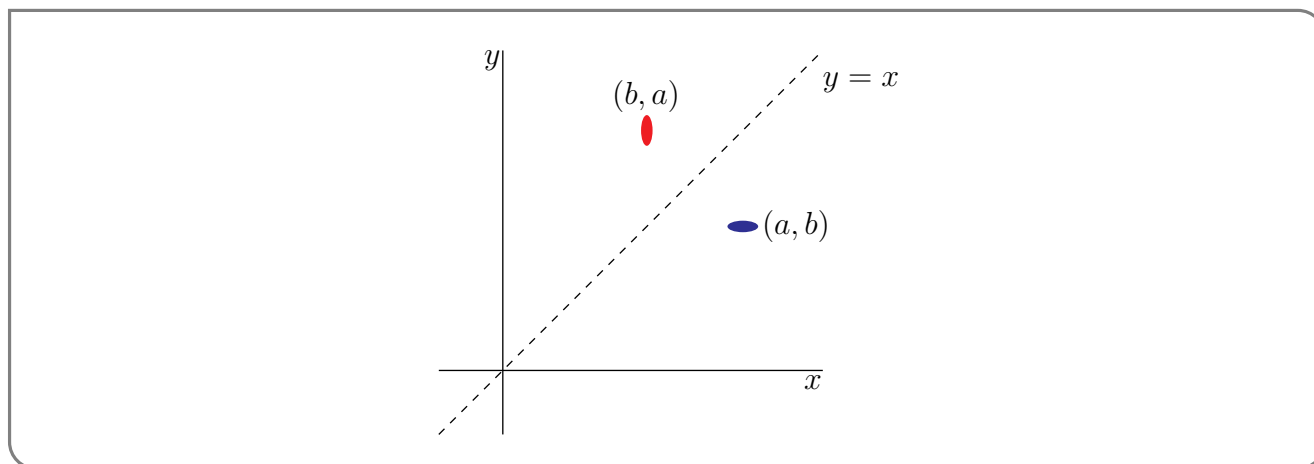
Example 0.6.7

As an example, let $f(x) = x^2$ with domain $0 \leqslant x < \infty$.

- When $x = 0$, $f(x) = 0^2 = 0$.
- As x increases, x^2 gets bigger and bigger.
- When x is very large and positive, x^2 is also very large and positive. (For example, think $x = 100$.)

The graph of $y = f(x) = x^2$ is the blue curve below. By definition, $Y = f^{-1}(X)$ if $X = f(Y) = Y^2$. That is, if $Y = \sqrt{X}$. (Remember that, to be in the domain of f, we must have $Y \geqslant 0$.) So the inverse function of "square" is "square root". The graph of f^{-1} is the red curve below. The red curve is the reflection of the blue curve in the line $y = x$.
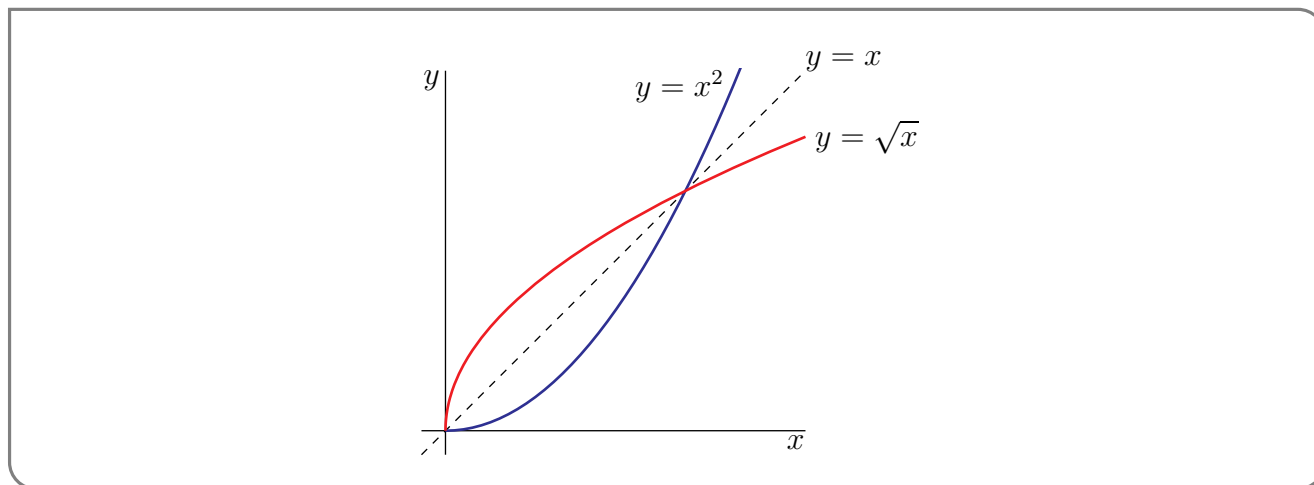
Example 0.6.7

LIMITS

So very roughly speaking, "Differential Calculus" is the study of how a function changes as its input changes. The mathematical object we use to describe this is the "derivative" of a function. To properly describe what this thing is we need some machinery; in particular we need to define what we mean by "tangent" and "limit". We'll get back to defining the derivative in Chapter 2.

1.1 ▴ Drawing Tangents and a First Limit

Our motivation for developing "limit" — being the title and subject of this chapter — is going to be two related problems of drawing tangent lines and computing velocity.

Now — our treatment of limits is not going to be completely mathematically rigorous, so we won't have too many formal definitions. There will be a few mathematically precise definitions and theorems as we go, but we'll make sure there is plenty of explanation around them.

Let us start with the "tangent line" problem. Of course, we need to define "tangent", but we won't do this formally. Instead let us draw some pictures.

Figure 1.1.1.
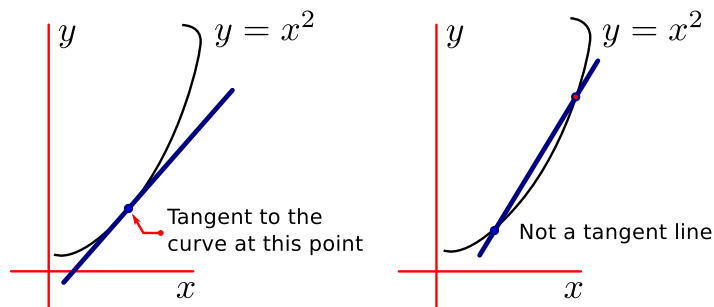
Here we have drawn two very rough sketches of the curve $y = x^2$ for $x \geqslant 0$. These are not very good sketches for a couple of reasons

- The curve in the figure does not pass through $(0,0)$, even though $(0,0)$ lies on $y = x^2$.

- The top-right end of the curve doubles back on itself and so fails the vertical line test that all functions must satisfy[1] — for each x-value there is exactly one y-value for which (x,y) lies on the curve $y = x^2$.

So let's draw those more carefully.

Figure 1.1.2.
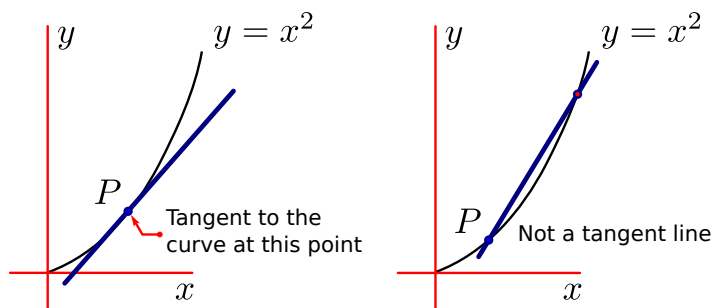
Sketches of the curve $y = x^2$. (left) shows a tangent line, while (right) shows a line that is not a tangent.

These are better. In both cases we have drawn $y = x^2$ (carefully) and then picked a point on the curve — call it P. Let us zoom in on the "good" example:

Figure 1.1.3.
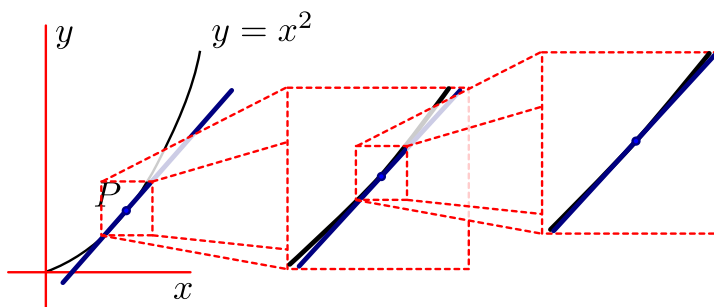
We see that, the more we zoom in on the point P, the more the graph of the function (drawn in black) looks like a straight line — that line is the tangent line (drawn in blue).

We see that as we zoom in on the point P, the graph of the function looks more and more like a straight line. If we kept on zooming in on P then the graph of the function would be indistinguishable from a straight line. That line is the tangent line (which we

1   Take a moment to go back and reread Definition 0.4.1.

have drawn in blue). A little more precisely, the blue line is "the tangent line to the function at P". We have to be a little careful, because if we zoom in at a different point, then we will find a different tangent line.

Now let's zoom in on the "bad" example we see that the blue line looks very different from the function; because of this, the blue line is not the tangent line at P.

Figure 1.1.4.
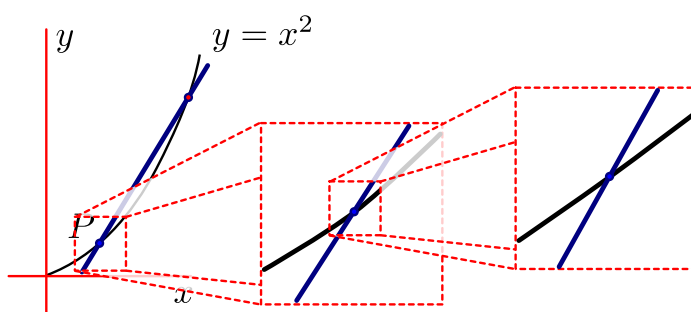
Zooming in on P we see that the function (drawn in black) looks more and more like a straight line — however it is not the same line as that drawn in blue. Because of this the blue line is not the tangent line.

Here are a couple more examples of tangent lines

Figure 1.1.5.
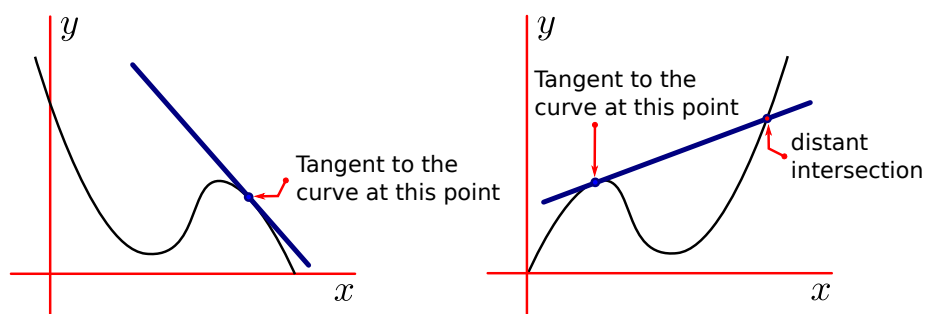
More examples of tangent lines.

The one on the left is very similar to the good example on $y = x^2$ that we saw above, while the one on the right is different — it looks a little like the "bad" example, in that it crosses our function the curve at some distant point. Why is the line in Figure 1.1.5(right) a tangent while the line in Figure 1.1.2(right) not a tangent? To see why, we should again zoom in close to the point where we are trying to draw the tangent.

Figure 1.1.6.
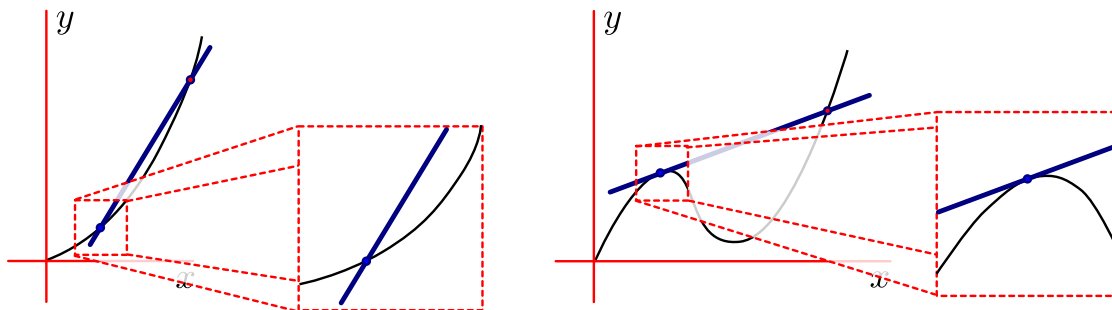
As we saw above in Figure 1.1.4, when we zoom in around our example of "not a tangent line" we see that the straight line looks very different from the curve at the "point of tangency" — i.e. where we are trying to draw the tangent. The line drawn in Figure 1.1.5(right) looks more and more like the function as we zoom in.

This example raises an important point — when we are trying to draw a tangent line, we don't care what the function does a long way from the point; the tangent line to the curve at a particular point P, depends only on what the function looks like close to that point P.

To illustrate this consider the sketch of the function $y = \sin(x)$ and its tangent line at $(x, y) = (0, 0)$:

Figure 1.1.7.
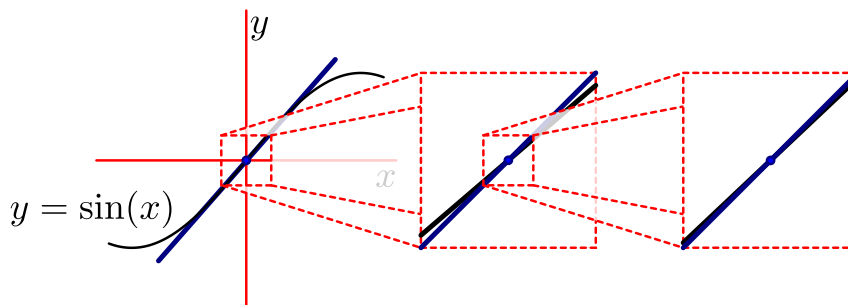
As we zoom in, the graph of $\sin(x)$ looks more and more like a straight line — in fact it looks more and more like the line $y = x$. We have also sketched this tangent line. What makes this example a little odd is that the tangent line crosses the function. In the examples above, our tangent lines just "kissed" the curve and did not cross it (or at least did not cross it nearby).

Using this idea of zooming in at a particular point, drawing a tangent line is not too hard. However, finding the equation of the tangent line presents us with a few challenges. Rather than leaping into the general theory, let us do a specific example. Let us find

the equation of the tangent line to the curve $y = x^2$ at the point P with coordinates[2] $(x, y) = (1, 1)$.

To find the equation of a line we either need

- the slope of the line and a point on the line, or

- two points on the line, from which we can compute the slope via the formula

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

and then write down the equation for the line via a formula such as

$$y = m \cdot (x - x_1) + y_1.$$

We cannot use the first method because we do not know what the slope of the tangent line should be. To work out the slope we need calculus — so we'll be able to use this method once we get to the next chapter on "differentiation".

It is not immediately obvious how we can use the second method, since we only have one point on the curve, namely $(1, 1)$. However we can use it to "sneak up" on the answer. Let's approximate the tangent line, by drawing a line that passes through $(1, 1)$ and some nearby point — call it Q. Here is our recipe:

- We are given the point $P = (1, 1)$ and we are told

  Find the tangent line to the curve $y = x^2$ that passes through $P = (1, 1)$.

- We don't quite know how to find a line given just 1 point, however we do know how to find a line passing through 2 points. So pick another point on the curves whose coordinates are very close to P. Now rather than picking some actual numbers, I am going to write our second point as $Q = (1 + h, (1 + h)^2)$. That is, a point Q whose x-coordinate is equal to that of P plus a little bit — where the little bit is some small number h. And since this point lies on the curve $y = x^2$, and Q's x-coordinate is $1 + h$, Q's y-coordinate must be $(1 + h)^2$.

  If having h as a variable rather than a number bothers you, start by thinking of h as 0.1.

- A picture of the situation will help.

---

2   Note that the *coordinates* (x, y) is an ordered pair of two numbers x and y. Traditionally the first number is called the *abscissa* while the second is the *ordinate*, but these terms are a little archaic. It is now much more common to hear people refer to the first number as the *x-coordinate* and the second as *y-coordinate*.

Figure 1.1.8.
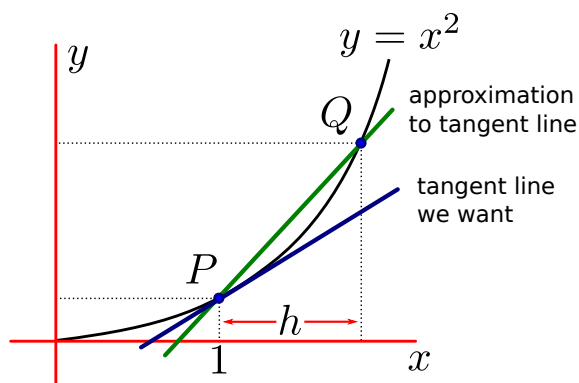
- This line that passes through the curve in two places P and Q is called a "secant line".

- The slope of the line is then

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$
$$= \frac{(1+h)^2 - 1}{(1+h) - 1} = \frac{1 + 2h + h^2 - 1}{h} = \frac{2h + h^2}{h} = 2 + h$$

where we have expanded $(1+h)^2 = 1 + 2h + h^2$ and then cleaned up a bit.

Now this isn't our tangent line because it passes through 2 nearby points on the curve — however it is a reasonable approximation of it. Now we can make that approximation better and so "sneak up" on the tangent line by considering what happens when we move this point Q closer and closer to P. i.e. make the number h closer and closer to zero.

Figure 1.1.9.
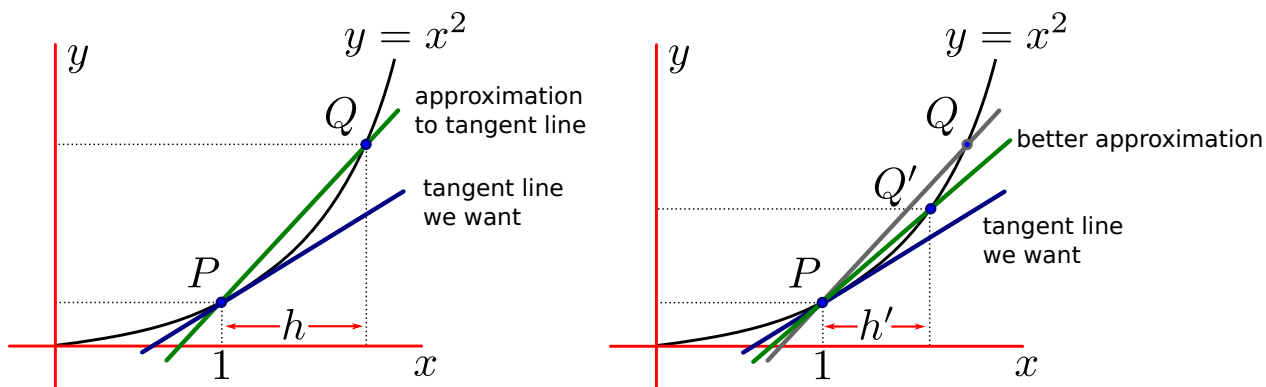
First look at the picture. The original choice of Q is on the left, while on the right we have drawn what happens if we choose h' to be some number a little smaller than h, so

that our point Q becomes a new point Q' that is a little closer to P. The new approximation is better than the first.

So as we make h smaller and smaller, we bring Q closer and closer to P, and make our secant line a better and better approximation of the tangent line. We can observe what happens to the slope of the line as we make h smaller by plugging some numbers into our formula $m = 2 + h$:

$$h = 0.1 \qquad\qquad\qquad m = 2.1$$
$$h = 0.01 \qquad\qquad\qquad m = 2.01$$
$$h = 0.001 \qquad\qquad\qquad m = 2.001.$$

So again we see that as this difference in x becomes smaller and smaller, the slope appears to be getting closer and closer to 2. We can write this more mathematically as

$$\lim_{h \to 0} \frac{(1+h)^2 - 1}{h} = 2$$

This is read as

The limit, as h approaches 0, of $\frac{(1+h)^2-1}{h}$ is 2.

This is our first limit! Notice that we can see this a little more clearly with a quick bit of algebra:

$$\begin{aligned}
\frac{(1+h)^2 - 1}{h} &= \frac{(1 + 2h + h^2) - 1}{h} \\
&= \frac{2h + h^2}{h} \\
&= 2 + h
\end{aligned}$$

So it is not unreasonable to expect that

$$\lim_{h \to 0} \frac{(1+h)^2 - 1}{h} = \lim_{h \to 0}(2 + h) = 2.$$

Our tangent line can be thought of as the end of this process — namely as we bring Q closer and closer to P, the slope of the secant line comes closer and closer to that of the tangent line we want. Since we have worked out what the slope is — that is the limit we saw just above — we now know the slope of the tangent line is 2. Given this, we can work out the equation for the tangent line.

- The equation for the line is $y = mx + c$. We have 2 unknowns m and c — so we need 2 pieces of information to find them.

- Since the line is tangent to $P = (1,1)$ we know the line must pass through $(1,1)$. From the limit we computed above, we also know that the line has slope 2.

- Since the slope is 2 we know that $m = 2$. Thus the equation of the line is $y = 2x + c$.

- We know that the line passes through $(1,1)$, so that $y = 2x + c$ must be 1 when $x = 1$. So $1 = 2 \cdot 1 + c$, which forces $c = -1$.

So our tangent line is $y = 2x - 1$.

1.2 ▴ Another Limit and Computing Velocity

Computing tangent lines is all very well, but what does this have to do with applications or the "Real World"? Well - at least initially our use of limits (and indeed of calculus) is going to be a little removed from real world applications. However as we go further and learn more about limits and derivatives we will be able to get closer to real problems and their solutions.

So stepping just a little closer to the real world, consider the following problem. You drop a ball from the top of a very very tall building. Let t be elapsed time measured in seconds, and $s(t)$ be the distance the ball has fallen in metres. So $s(0) = 0$.

Quick aside: there is quite a bit going on in the statement of this problem. We have described the general picture — tall building, ball, falling — but we have also introduced notation, variables and units. These will be common first steps in applications and are necessary in order to translate a real world problem into mathematics in a clear and consistent way.

Galileo[3] worked out that $s(t)$ is a quadratic function:

$$s(t) = 4.9t^2.$$

The question that is posed is

> How fast is the ball falling after 1 second?

Now before we get to answering this question, we should first be a little more precise. The wording of this question is pretty sloppy for a couple of reasons:

- What we do mean by "after 1 second"? We know the ball will move faster and faster as time passes, so after 1 second it does not fall at one fixed speed.

- As it stands a reasonable answer to the question would be just "really fast". If the person asking the question wants a numerical answer it would be better to ask "At what speed" or "With what velocity".

We should also be careful using the words "speed" and "velocity" — they are not interchangeable.

- Speed means the distance travelled per unit time and is always a non-negative number. An unmoving object has speed 0, while a moving object has positive speed.

- Velocity, on the other hand, also specifies the direction of motion. In this text we will almost exclusively deal with objects moving along straight lines. Because of this

---

3   Perhaps one of the most famous experiments in all of physics is Galileo's leaning tower of Pisa experiment, in which he dropped two balls of different masses from the top of the tower and observed that the time taken to reach the ground was independent of their mass. This disproved Aristotle's assertion that heavier objects fall faster. It is quite likely that Galileo did not actually perform this experiment. Rather it was a thought-experiment. However a quick glance at Wikipedia will turn up some wonderful footage from the Apollo 15 mission showing a hammer and feather being dropped from equal height hitting the moon's surface at the same time. Finally, Galileo determined that the speed of falling objects increases at a constant rate, which is equivalent to the formula stated here, but it is unlikely that he wrote down an equation exactly as it is here.

velocities will be positive or negative numbers indicating which direction the object is moving along the line. We will be more precise about this later[4].

A better question is

What is the velocity of the ball precisely 1 second after it is dropped?

or even better:

What is the velocity of the ball at the 1 second mark?

This makes it very clear that we want to know what is happening at exactly 1 second after the ball is dropped.

There is something a little subtle going on in this question. In particular, what do we mean by the velocity at $t = 1$?. Surely if we freeze time at $t = 1$ second, then the object is not moving at all? This is definitely *not* what we mean.

If an object is moving at a constant velocity[5] in the positive direction, then that velocity is just the distance travelled divided by the time taken. That is

$$v = \frac{\text{distance moved}}{\text{time taken}}$$

An object moving at constant velocity that moves 27 metres in 3 seconds has velocity

$$v = \frac{27m}{3s} = 9m/s.$$

When velocity is constant everything is easy.

However, in our falling object example, the object is being acted on by gravity and its speed is definitely not constant. Instead of asking for *THE* velocity, let us examine the "average velocity" of the object over a certain window of time. In this case the formula is very similar

$$\text{average velocity} = \frac{\text{distance moved}}{\text{time taken}}$$

But now I want to be more precise, instead write

$$\text{average velocity} = \frac{\text{difference in distance}}{\text{difference in time}}$$

Now in spoken English we haven't really changed much — the distance moved is the difference in position, and the time taken is just the difference in time — but the latter is more mathematically precise, and is easy to translate into the following equation

$$\text{average velocity} = \frac{s(t_2) - s(t_1)}{t_2 - t_1}.$$

---

4   Getting the sign of velocity wrong is a very common error — you should be careful with it.
5   Newton's first law of motion states that an object in motion moves with constant velocity unless a force acts on it — for example gravity or friction.

This is the formula for the average velocity of our object between time t_1 and t_2. The denominator is just the difference between these times and the numerator is the difference in position — i.e. position at time t_1 is just $s(t_1)$ and position at time t_2 is just $s(t_2)$.

So what is the average velocity of the falling ball between 1 and 1.1 seconds? All we need to do now is plug some numbers into our formula

$$
\begin{aligned}
\text{average velocity} &= \frac{\text{difference in position}}{\text{difference in time}} \\
&= \frac{s(1.1) - s(1)}{1.1 - 1} \\
&= \frac{4.9(1.1)^2 - 4.9(1)}{0.1} = \frac{4.9 \times 0.21}{0.1} = 10.29 m/s
\end{aligned}
$$

And we have our average velocity. However there is something we should notice about this formula and it is easier to see if we sketch a graph of the function $s(t)$

Figure 1.2.1.
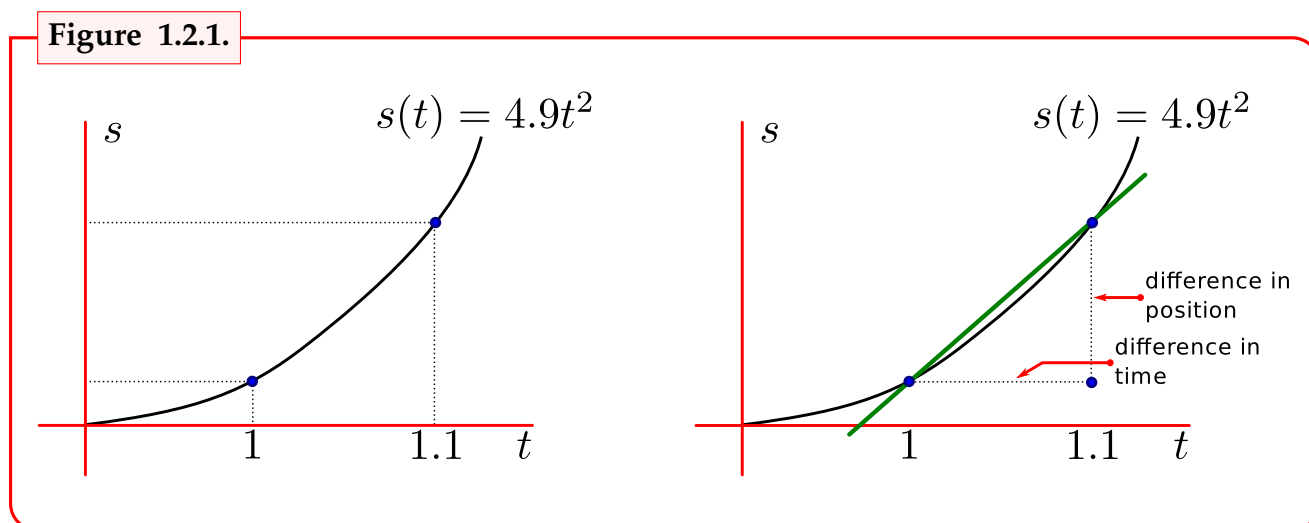
So on the left I have drawn the graph and noted the times $t = 1$ and $t = 1.1$. The corresponding positions on the axes and the two points on the curve. On the right I have added a few more details. In particular I have noted the differences in position and time, and the line joining the two points. Notice that the slope of this line is

$$
\text{slope} = \frac{\text{change in } y}{\text{change in } x} = \frac{\text{difference in } s}{\text{difference in } t}
$$

which is precisely our expression for the average velocity.

Let us examine what happens to the average velocity as we look over smaller and smaller time-windows.

time window	average velocity
$1 \leqslant t \leqslant 1.1$	10.29
$1 \leqslant t \leqslant 1.01$	9.849
$1 \leqslant t \leqslant 1.001$	9.8049
$1 \leqslant t \leqslant 1.0001$	9.80049

As we make the time interval smaller and smaller we find that the average velocity is getting closer and closer to 9.8. We can be a little more precise by finding the average velocity between $t = 1$ and $t = 1 + h$ — this is very similar to what we did for tangent lines.

$$\begin{aligned} \text{average velocity} &= \frac{s(1+h) - s(1)}{(1+h) - 1} \\ &= \frac{4.9(1+h)^2 - 4.9}{h} \\ &= \frac{9.8h + 4.9h^2}{h} \\ &= 9.8 + 4.9h \end{aligned}$$

Now as we squeeze this window between $t = 1$ and $t = 1 + h$ down towards zero, the average velocity becomes the "instantaneous velocity" — just as the slope of the secant line becomes the slope of the tangent line. This is our second limit

$$v(1) = \lim_{h \to 0} \frac{s(1+h) - s(1)}{h} = 9.8$$

More generally we define the instantaneous velocity at time $t = a$ to be the limit

$$v(a) = \lim_{h \to 0} \frac{s(a+h) - s(a)}{h}$$

We read this as

The velocity at time a is equal to the limit as h goes to zero of $\frac{s(a+h) - s(a)}{h}$.

While we have solved the problem stated at the start of this section, it is clear that if we wish to solve similar problems that we will need to understand limits in a more general and systematic way.

1.3 ▲ The Limit of a Function

Before we come to definitions, let us start with a little notation for limits.

> **Notation 1.3.1.**
>
> We will often write
>
> $$\lim_{x \to a} f(x) = L$$
>
> which should be read as
>
> The limit of $f(x)$ as x approaches a is L.

The notation is just shorthand — we don't want to have to write out long sentences as we do our mathematics. Whenever you see these symbols you should think of that sentence.

This shorthand also has the benefit of being mathematically precise (we'll see this later), and (almost) independent of the language in which the author is writing. A mathematician who does not speak English can read the above formula and understand exactly what it means.

In mathematics, like most languages, there is usually more than one way of writing things and we can also write the above limit as

$$f(x) \to L \text{ as } x \to a$$

This can also be read as above, but also as

$f(x)$ goes to L as x goes to a

They mean exactly the same thing in mathematics, even though they might be written, read and said a little differently.

To arrive at the definition of limit, we want to start[6] with a very simple example.

Example 1.3.2

Consider the following function.

$$f(x) = \begin{cases} 2x & x < 3 \\ 9 & x = 3 \\ 2x & x > 3 \end{cases}$$

This is an example of a piece-wise function[7]. That is, a function defined in several pieces, rather than as a single formula. We evaluate the function at a particular value of x on a case-by-case basis. Here is a sketch of it
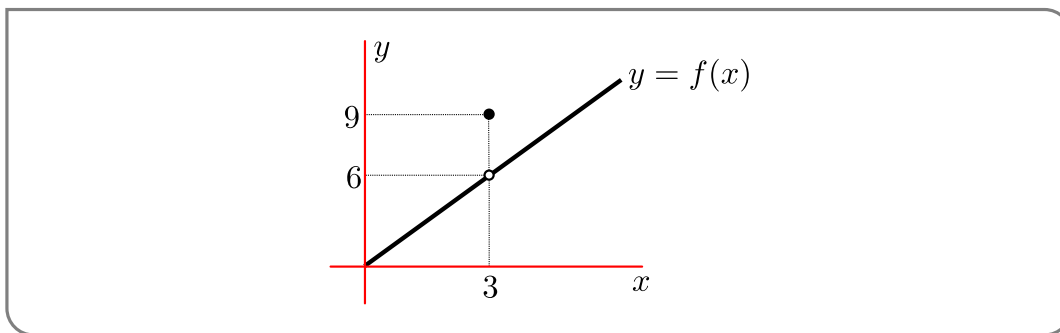
Notice the two circles in the plot. One is open, ∘ and the other is closed •.

- A filled circle has quite a precise meaning — a filled circle at (x, y) means that the function takes the value $f(x) = y$.

---

6   Well, we had two limits in the previous sections, so perhaps we really want to "restart" with a very simple example.

7   We saw another piecewise function back in Example 0.4.5.

- An open circle is a little harder — an open circle at $(3, 6)$ means that the point $(3, 6)$ is not on the graph of $y = f(x)$, i.e. $f(3) \neq 6$. We should only use the open circle where it is absolutely necessary in order to avoid confusion.

This function is quite contrived, but it is a very good example to start working with limits more systematically. Consider what the function does close to $x = 3$. We already know what happens exactly at $3 - f(x) = 9 -$ but I want to look at how the function behaves very close to $x = 3$. That is, what does the function do as we look at a point x that gets closer and closer to $x = 3$.

If we plug in some numbers very close to 3 (but not exactly 3) into the function we see the following:

x	2.9	2.99	2.999	∘	3.001	3.01	3.1
$f(x)$	5.8	5.98	5.998	∘	6.002	6.02	6.2

So as x moves closer and closer to 3, without being exactly 3, we see that the function moves closer and closer to 6. We can write this as

$$\lim_{x \to 3} f(x) = 6$$

That is

The limit as x approaches 3 of $f(x)$ is 6.

So for x very close to 3, without being exactly 3, the function is very close to 6 — which is a long way from the value of the function exactly at 3, $f(3) = 9$. Note well that the behaviour of the function as x gets very close to 3 *does not* depend on the value of the function *at* 3.

Example 1.3.2

We now have enough to make an informal definition of a limit, which is actually sufficient for most of what we will do in this text.

Definition 1.3.3 (Informal definition of limit).

We write

$$\lim_{x \to a} f(x) = L$$

if the value of the function $f(x)$ is sure to be arbitrarily close to L whenever the value of x is close enough to a, without[8] being exactly a.

---

8 You may find the condition "without being exactly a" a little strange, but there is a good reason for it. One very important application of limits, indeed the main reason we teach the topic, is in the definition of derivatives (see Definition 2.2.1 in the next chapter). In that definition we need to compute the limit $\lim_{x \to a} \dfrac{f(x) - f(a)}{x - a}$. In this case the function whose limit is being taken, namely $\dfrac{f(x) - f(a)}{x - a}$, is not defined at all at $x = a$.

In order to make this definition more mathematically correct, we need to make the idea of "closer and closer" more precise — we do this in Section 1.7. It should be emphasised that the formal definition and the contents of that section are optional material.

For now, let us use the above definition to examine a more substantial example.

Example 1.3.4

Let $f(x) = \frac{x-2}{x^2+x-6}$ and consider its limit as $x \to 2$.

- We are really being asked

$$\lim_{x \to 2} \frac{x-2}{x^2+x-6} = \text{ what?}$$

- Now if we try to compute $f(2)$ we get $0/0$ which is undefined. The function is not defined at that point — this is a good example of why we need limits. We have to sneak up on these places where a function is not defined (or is badly behaved).

- **VERY IMPORTANT POINT:** the fraction $\frac{0}{0}$ is *not* ∞ and it is not 1, it is not defined. We cannot ever divide by zero in normal arithmetic and obtain a consistent and mathematically sensible answer. If you learned otherwise in high-school, you should quickly unlearn it.

- Again, we can plug in some numbers close to 2 and see what we find

x	1.9	1.99	1.999	$\circ$	2.001	2.01	2.1
$f(x)$	0.20408	0.20040	0.20004	$\circ$	0.19996	0.19960	0.19608

- So it is reasonable to suppose that

$$\lim_{x \to 2} \frac{x-2}{x^2+x-6} = 0.2$$

Example 1.3.4

The previous two examples are nicely behaved in that the limits we tried to compute actually exist. We now turn to two nastier examples[9] in which the limits we are interested in do not exist.

Example 1.3.5 (A bad example)

Consider the following function $f(x) = \sin(\pi/x)$. Find the limit as $x \to 0$ of $f(x)$.

We should see something interesting happening close to $x = 0$ because $f(x)$ is undefined there. Using your favourite graph-plotting software you can see that the graph looks roughly like

---

9   Actually, they are good examples, but the functions in them are nastier.
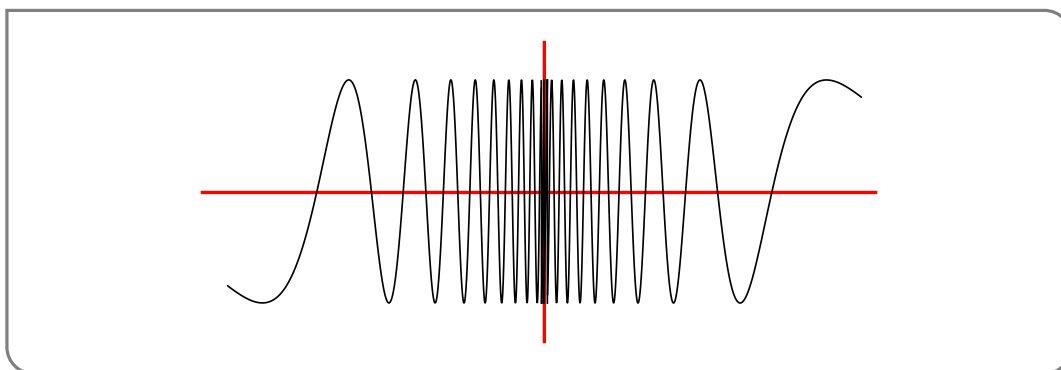
How to explain this? As x gets closer and closer to zero, π/x becomes larger and larger (remember what the plot of $y = 1/x$ looks like). So when you take sine of that number, it oscillates faster and faster the closer you get to zero. Since the function does not approach a single number as we bring x closer and closer to zero, the limit does not exist.

We write this as

$$\lim_{x \to 0} \sin\left(\frac{\pi}{x}\right) \text{ does not exist}$$

It's not very inventive notation, however it is clear. We frequently abbreviate "does not exist" to "DNE" and rewrite the above as

$$\lim_{x \to 0} \sin\left(\frac{\pi}{x}\right) = \text{DNE}$$

Example 1.3.5

In the following example, the limit we are interested in does not exist. However the way in which things go wrong is quite different from what we just saw.

Example 1.3.6

Consider the function

$$f(x) = \begin{cases} x & x < 2 \\ -1 & x = 2 \\ x + 3 & x > 2 \end{cases}$$

- The plot of this function looks like this
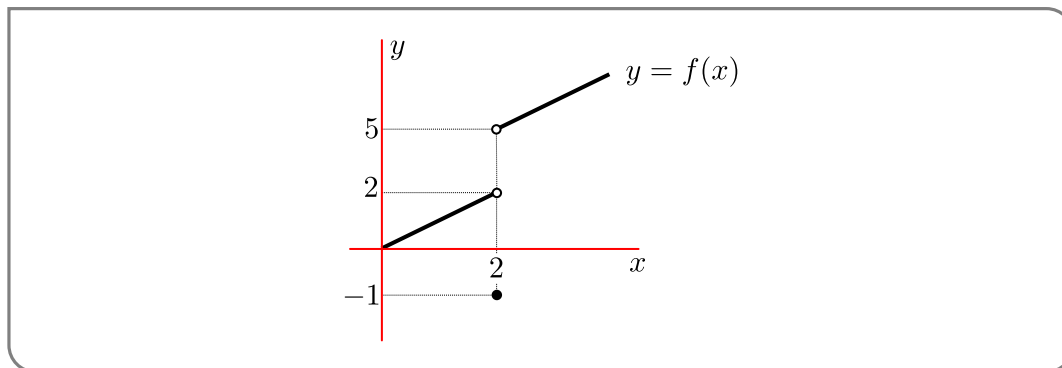
- So let us plug in numbers close to 2.

x	1.9	1.99	1.999	$\circ$	2.001	2.01	2.1
$f(x)$	1.9	1.99	1.999	$\circ$	5.001	5.01	5.1

- This isn't like before. Now when we approach from below, we seem to be getting closer to 2, but when we approach from above we seem to be getting closer to 5. Since we are not approaching the same number the limit does not exist.

$$\lim_{x \to 2} f(x) = \text{DNE}$$

Example 1.3.6

While the limit in the previous example does not exist, the example serves to introduce the idea of "one-sided limits". For example, we can say that

As x moves closer and closer to two *from below* the function approaches 2.

and similarly

As x moves closer and closer to two *from above* the function approaches 5.

Definition 1.3.7 (Informal definition of one-sided limits).

We write

$$\lim_{x \to a^-} f(x) = K$$

when the value of $f(x)$ gets closer and closer to K when $x < a$ and x moves closer and closer to a. Since the x-values are always less than a, we say that x approaches a *from below*. This is also often called the left-hand limit since the x-values lie to the left of a on a sketch of the graph.
We similarly write

$$\lim_{x \to a^+} f(x) = L$$

when the value of $f(x)$ gets closer and closer to L when $x > a$ and x moves closer and closer to a. For similar reasons we say that x approaches a from above, and sometimes refer to this as the right-hand limit.

Note — be careful to include the superscript $+$ and $-$ when writing these limits. You might also see the following notations:

$$\lim_{x \to a^+} f(x) = \lim_{x \to a+} f(x) = \lim_{x \downarrow a} f(x) = \lim_{x \searrow a} f(x) = L \qquad \text{right-hand limit}$$

$$\lim_{x \to a^-} f(x) = \lim_{x \to a-} f(x) = \lim_{x \uparrow a} f(x) = \lim_{x \nearrow a} f(x) = L \qquad \text{left-hand limit}$$

but please use with the notation in Definition 1.3.7 above.

Given these two similar notions of limits, when are they the same? The following theorem tell us

> **Theorem 1.3.8** (Limits and one sided limits).
>
> $$\lim_{x \to a} f(x) = L \qquad \text{if and only if} \qquad \lim_{x \to a^-} f(x) = L \text{ and } \lim_{x \to a^+} f(x) = L$$

Notice that this is really two separate statements because of the "if and only if"

- If the limit of $f(x)$ as x approaches a exists and is equal to L, then both the left-hand and right-hand limits exist and are equal to L. AND,

- If the left-hand and right-hand limits as x approaches a exist and are equal, then the limit as x approaches a exists and is equal to the one-sided limits.

That is — the limit of $f(x)$ as x approaches a will only exist if it doesn't matter which way we approach a (either from left or right) *AND* if we get the same one-sided limits when we approach from left and right, then the limit exists.

We can rephrase the above by writing the contrapositives[10] of the above statements.

- If either of the left-hand and right-hand limits as x approaches a fail to exist, or if they both exist but are different, then the limit as x approaches a does not exist. AND,

- If the limit as x approaches a does not exist, then the left-hand and right-hand limits are either different or at least one of them does not exist.

Here is another limit example

Example 1.3.9

Consider the following two functions and compute their limits and one-sided limits as x approaches 1:

---

10   Given a statement of the form "If A then B", the contrapositive is "If not B then not A". They are logically equivalent — if one is true then so is the other. We must take care not to confuse the contrapositive with the converse. Given "If A then B", the converse is "If B then A". These are definitely not the same.
To see this consider the statement "If he is Shakespeare then he is dead." The converse is "If he is dead then he is Shakespeare" — clearly garbage since there are plenty of dead people who are not Shakespeare. The contrapositive is "If he is not dead then he is not Shakespeare" — which makes much more sense.
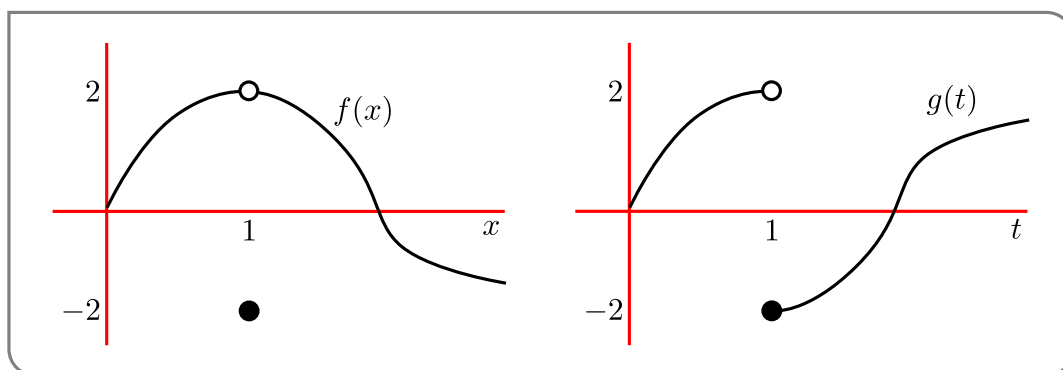
These are a little different from our previous examples, in that we do not have formulas, only the sketch. But we can still compute the limits.

- Function on the left — $f(x)$:

$$\lim_{x\to 1^-} f(x) = 2 \qquad\qquad \lim_{x\to 1^+} f(x) = 2$$

  so by the previous theorem

$$\lim_{x\to 1} f(x) = 2$$

- Function on the right — $g(t)$:

$$\lim_{t\to 1^-} g(t) = 2 \qquad\qquad \text{and} \ \lim_{t\to 1^+} g(t) = -2$$

  so by the previous theorem

$$\lim_{t\to 1} g(t) = \text{DNE}$$

Example 1.3.9

We have seen 2 ways in which a limit does not exist — in one case the function oscillated wildly, and in the other there was some sort of "jump" in the function, so that the left-hand and right-hand limits were different.

There is a third way that we must also consider. To describe this, consider the following four functions:

Figure 1.3.1.
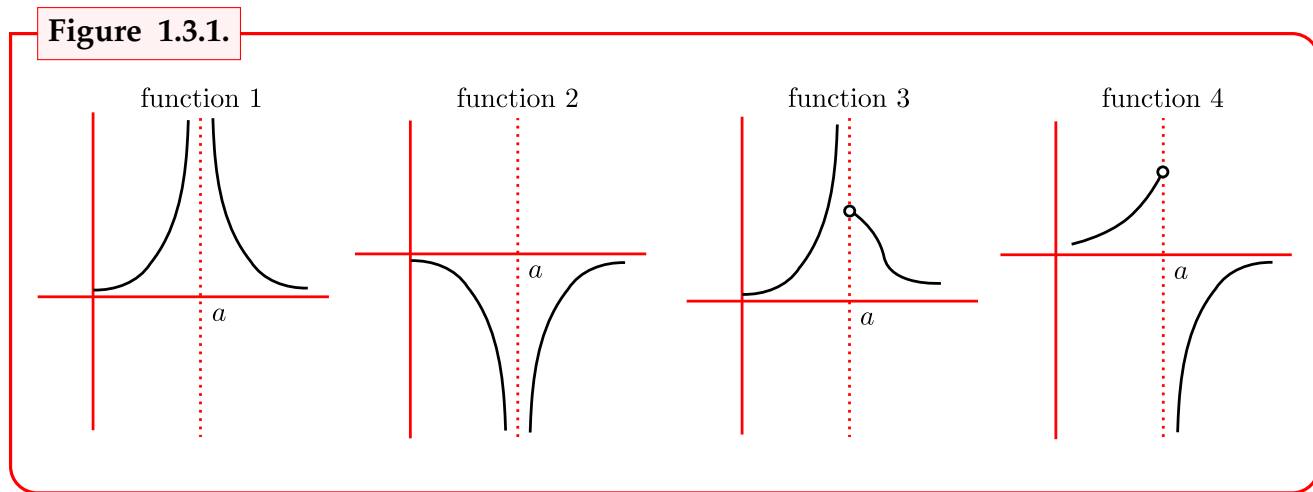
None of these functions are defined at $x = a$, nor do the limits as x approaches a exist. However we can say more than just "the limits do not exist".

Notice that the value of function 1 can be made bigger and bigger as we bring x closer and closer to a. Similarly the value of the second function can be made arbitrarily large and negative (i.e. make it as big a negative number as we want) by bringing x closer and closer to a. Based on this observation we have the following definition.

Definition 1.3.10.

We write

$$\lim_{x \to a} f(x) = +\infty$$

when the value of the function $f(x)$ becomes arbitrarily large and positive as x gets closer and closer to a, without being exactly a.
Similarly, we write

$$\lim_{x \to a} f(x) = -\infty$$

when the value of the function $f(x)$ becomes arbitrarily large and negative as x gets closer and closer to a, without being exactly a.

A good examples of the above is

$$\lim_{x \to 0} \frac{1}{x^2} = +\infty \qquad\qquad \lim_{x \to 0} -\frac{1}{x^2} = -\infty$$

IMPORTANT POINT: Please do not think of "$+\infty$" and "$-\infty$" in these statements as numbers. You should think of $\lim_{x \to a} f(x) = +\infty$ and $\lim_{x \to a} f(x) = -\infty$ as special cases of $\lim_{x \to a} f(x) = $ DNE. The statement

$$\lim_{x \to a} f(x) = +\infty$$

does not say "the limit of $f(x)$ as x approaches a is positive infinity". It says "the function $f(x)$ becomes arbitrarily large as x approaches a". These are different statements; remember that ∞ is not a number[11].

Now consider functions 3 and 4 in Figure 1.3.1. Here we can make the value of the function as big and positive as we want (for function 3) or as big and negative as we want (for function 4) but only when x approaches a from one side. With this in mind we can construct similar notation and a similar definition:

---

11 One needs to be very careful making statements about infinity. At some point in our lives we get around to asking ourselves "what is the biggest number", and we realise there isn't one. That is, we can go on counting integer after integer, for ever and not stop. Indeed the set of integers is the first infinite thing we really encounter. It is an example of a *countably infinite* set. The set of real-numbers is actually much bigger and is *uncountably infinite*. In fact there are an infinite number of different sorts of infinity! Much of the theory of infinite sets was developed by Georg Cantor; we mentioned him back in Section 0.2 and he is well worth googling.

Definition 1.3.11.

We write

$$\lim_{x \to a^+} f(x) = +\infty$$

when the value of the function $f(x)$ becomes arbitrarily large and positive as x gets closer and closer to a from above (equivalently — from the right), without being exactly a.
Similarly, we write

$$\lim_{x \to a^+} f(x) = -\infty$$

when the value of the function $f(x)$ becomes arbitrarily large and negative as x gets closer and closer to a from above (equivalently — from the right), without being exactly a.
The notation

$$\lim_{x \to a^-} f(x) = +\infty \qquad\qquad \lim_{x \to a^-} f(x) = -\infty$$

has a similar meaning except that limits are approached from below / from the left.

So for function 3 we have

$$\lim_{x \to a^-} f(x) = +\infty \qquad\qquad \lim_{x \to a^+} f(x) = \text{some positive number}$$

and for function 4

$$\lim_{x \to a^-} f(x) = \text{some positive number} \qquad\qquad \lim_{x \to a^+} f(x) = -\infty$$

More examples:

Example 1.3.12

Consider the function

$$g(x) = \frac{1}{\sin(x)}$$

Find the one-sided limits of this function as $x \to \pi$.

Probably the easiest way to do this is to first plot the graph of $\sin(x)$ and $1/x$ and then think carefully about the one-sided limits:
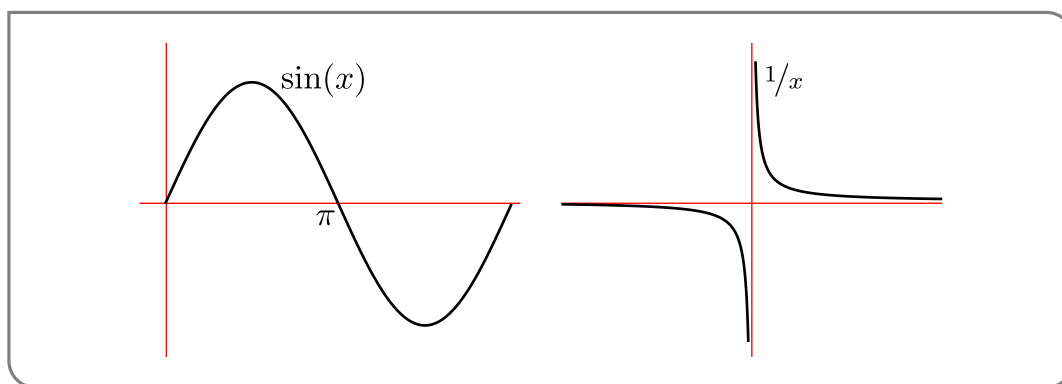
- As $x \to \pi$ from the left, $\sin(x)$ is a small positive number that is getting closer and closer to zero. That is, as $x \to \pi^-$, we have that $\sin(x) \to 0$ through positive numbers (i.e. from above). Now look at the graph of $1/x$, and think what happens as we move $x \to 0^+$, the function is positive and becomes larger and larger.

  So as $x \to \pi$ from the left, $\sin(x) \to 0$ from above, and so $1/\sin(x) \to +\infty$.

- By very similar reasoning, as $x \to \pi$ from the right, $\sin(x)$ is a small negative number that gets closer and closer to zero. So as $x \to \pi$ from the right, $\sin(x) \to 0$ through negative numbers (i.e. from below) and so $1/\sin(x)$ to $-\infty$.

Thus

$$\lim_{x \to \pi^-} \frac{1}{\sin(x)} = +\infty \qquad\qquad \lim_{x \to \pi^+} \frac{1}{\sin(x)} = -\infty$$

Example 1.3.12

Again, we can make Definitions 1.3.10 and 1.3.11 into mathematically precise formal definitions using techniques very similar to those in the optional Section 1.7. This is not strictly necessary for this course.

Up to this point we explored limits by sketching graphs or plugging values into a calculator. This was done to help build intuition, but it is not really the basis of a systematic method for computing limits. We have also avoided more formal approaches[12] since we do not have time in the course to go into that level of detail and (arguably) we don't need that detail to achieve the aims of the course. Thankfully we can develop a more systematic approach based on the idea of building up complicated limits from simpler ones by examining how limits interact with the basic operations of arithmetic.

1.4 ▴ Calculating Limits with Limit Laws

Think back to the functions you know and the sorts of things you have been asked to draw, factor and so on. Then they are all constructed from simple pieces, such as

---

12 The formal approaches are typically referred to as "epsilon-delta limits" or "epsilon-delta proofs" since the symbols ϵ and δ are traditionally used throughout. Take a peek at Section 1.7 to see.

- constants — c

- monomials — x^n

- trigonometric functions — $\sin(x), \cos(x)$ and $\tan(x)$

These are the building blocks from which we construct functions. Soon we will add a few more functions to this list, especially the exponential function and various inverse functions.

We then take these building blocks and piece them together using arithmetic

- addition and subtraction — $f(x) = g(x) + h(x)$ and $f(x) = g(x) - h(x)$

- multiplication — $f(x) = g(x) \cdot h(x)$

- division — $f(x) = \frac{g(x)}{h(x)}$

- substitution — $f(x) = g(h(x))$ — this is also called the composition of g with h.

The idea of building up complicated functions from simpler pieces was discussed in Section 0.5.

What we will learn in this section is how to compute the limits of the basic building blocks and then how we can compute limits of sums, products and so forth using "limit laws". This process allows us to compute limits of complicated functions, using very simple tools and without having to resort to "plugging in numbers" or "closer and closer" or "$\epsilon - \delta$ arguments".

In the examples we saw above, almost all the *interesting* limits happened at points where the underlying function was badly behaved — where it jumped, was not defined or blew up to infinity. In those cases we had to be careful and think about what was happening. Thankfully most functions we will see do not have too many points at which these sorts of things happen.

For example, polynomials do not have any nasty jumps and are defined everywhere and do not "blow up". If you plot them, they look smooth[13]. Polynomials and limits behave very nicely together, and for any polynomial $P(x)$ and any real number a we have that

$$\lim_{x \to a} P(x) = P(a)$$

That is — to evaluate the limit we just plug in the number. We will build up to this result over the next few pages.

Let us start with the two easiest limits[14]

> **Theorem 1.4.1** (Easiest limits).
>
> Let $a, c \in \mathbb{R}$. The following two limits hold
>
> $$\lim_{x \to a} c = c \qquad \text{and} \qquad \lim_{x \to a} x = a.$$

---

13 We have used this term in an imprecise way, but it does have a precise mathematical meaning.
14 Though it lies outside the scope of the course, you can find the formal ϵ-δ proof of this result at the end of Section 1.7.

Since we have not seen too many theorems yet, let us examine it carefully piece by piece.

- **Let** $a, c \in \mathbb{R}$ — just as was the case for definitions, we start a theorem by defining terms and setting the scene. There is not too much scene to set: the symbols a and c are real numbers.

- **The following two limits hold** — this doesn't really contribute much to the statement of the theorem, it just makes it easier to read.

- $\lim_{x \to a} c = c$ — when we take the limit of a constant function (for example think of $c = 3$), the limit is (unsurprisingly) just that same constant.

- $\lim_{x \to a} x = a$ — as we noted above for general polynomials, the limit of the function $f(x) = x$ as x approaches a given point a, is just a. This says something quite obvious — as x approaches a, x approaches a (if you are not convinced then sketch the graph).

Armed with only these two limits, we cannot do very much. But combining these limits with some arithmetic we can do quite a lot. For a moment, take a step back from limits for a moment and think about how we construct functions. To make the discussion a little more precise think about how we might construct the function

$$h(x) = \frac{2x - 3}{x^2 + 5x - 6}$$

If we want to compute the value of the function at $x = 2$, then we would

- compute the numerator at $x = 2$

- compute the denominator at $x = 2$

- compute the ratio

Now to compute the numerator we

- take x and multiply it by 2

- subtract 3 to the result

While for the denominator

- multiply x by x

- multiply x by 5

- add these two numbers and subtract 6

This sequence of operations can be represented pictorially as the tree shown in Figure 1.4.1 below.

Figure 1.4.1.
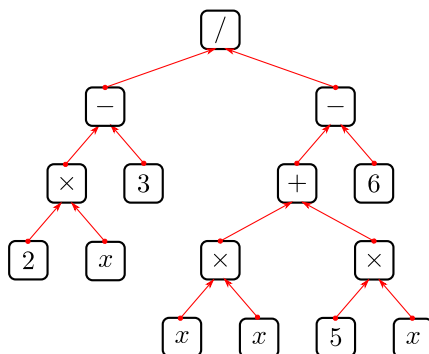
Such trees were discussed in Section 0.5 (now is not a bad time to quickly review that section before proceeding). The point here is that in order to compute the value of the function we just repeatedly add, subtract, multiply and divide constants and x.

To compute the limit of the above function at $x = 2$ we can do something very similar. From the previous theorem we know how to compute

$$\lim_{x \to 2} c = c \qquad \text{and} \qquad \lim_{x \to 2} x = 2$$

and the next theorem will tell us how to stitch together these two limits using the arithmetic we used to construct the function.

Theorem 1.4.2 (Arithmetic of limits).

Let $a, c \in \mathbb{R}$, let $f(x)$ and $g(x)$ be defined for all x's that lie in some interval about a (but f, g need not be defined exactly at a).

$$\lim_{x \to a} f(x) = F \qquad\qquad \lim_{x \to a} g(x) = G$$

exist with $F, G \in \mathbb{R}$. Then the following limits hold

- $\lim_{x \to a}(f(x) + g(x)) = F + G$ — limit of the sum is the sum of the limits.

- $\lim_{x \to a}(f(x) - g(x)) = F - G$ — limit of the difference is the difference of the limits.

- $\lim_{x \to a} c f(x) = cF$.

- $\lim_{x \to a}(f(x) \cdot g(x)) = F \cdot G$ — limit of the product is the product of limits.

- If $G \neq 0$ then $\lim_{x \to a} \dfrac{f(x)}{g(x)} = \dfrac{F}{G}$ and, in particular, $\lim_{x \to a} \dfrac{1}{g(x)} = \dfrac{1}{G}$.

Note — be careful with this last one — the denominator cannot be zero.

The above theorem shows that limits interact very simply with arithmetic. If you are asked to find the limit of a sum then the answer is just the sum of the limits. Similarly the limit of a product is just the product of the limits.

How do we apply the above theorem to the rational function $h(x)$ we defined above? Here is a warm-up example:

Example 1.4.3

You are given two functions f, g (not explicitly) which have the following limits as x approaches 1:

$$\lim_{x \to 1} f(x) = 3 \qquad \text{and} \qquad \lim_{x \to 1} g(x) = 2$$

Using the above theorem we can compute

$$\lim_{x \to 1} 3f(x) = 3 \times 3 = 9$$
$$\lim_{x \to 1} 3f(x) - g(x) = 3 \times 3 - 2 = 7$$
$$\lim_{x \to 1} f(x)g(x) = 3 \times 2 = 6$$
$$\lim_{x \to 1} \frac{f(x)}{f(x) - g(x)} = \frac{3}{3 - 2} = 3$$

Example 1.4.3

Another simple example

Example 1.4.4

Find $\lim_{x \to 3} 4x^2 - 1$

We use the arithmetic of limits:

$$\begin{aligned}
\lim_{x \to 3} 4x^2 - 1 &= \left(\lim_{x \to 3} 4x^2 \right) - \lim_{x \to 3} 1 && \text{difference of limits} \\
&= \left(\lim_{x \to 3} 4 \cdot \lim_{x \to 3} x^2 \right) - \lim_{x \to 3} 1 && \text{product of limits} \\
&= 4 \cdot \left(\lim_{x \to 3} x^2 \right) - 1 && \text{limit of constant} \\
&= 4 \cdot \left(\lim_{x \to 3} x \right) \cdot \left(\lim_{x \to 3} x \right) - 1 && \text{product of limits} \\
&= 4 \cdot 3 \cdot 3 - 1 && \text{limit of } x \\
&= 36 - 1 \\
&= 35
\end{aligned}$$

Example 1.4.4

This is an excruciating level of detail, but when you first use this theorem and try some

examples it is a good idea to do things step by step by step until you are comfortable with it.

Example 1.4.5

Yet another limit — compute $\lim\limits_{x \to 2} \dfrac{x}{x-1}$.

To apply the arithmetic of limits, we need to examine numerator and denominator separately and make sure the limit of the denominator is non-zero. Numerator first:

$$\lim_{x \to 2} x = 2 \qquad\qquad\qquad\qquad\qquad \text{limit of } x$$

and now the denominator:

$$\lim_{x \to 2} x - 1 = \left(\lim_{x \to 2} x \right) - \left(\lim_{x \to 2} 1 \right) \qquad\qquad \text{difference of limits}$$

$$= 2 - 1 \qquad\qquad \text{limit of } x \text{ and limit of constant} = 1$$

Since the limit of the denominator is non-zero we can put it back together to get

$$\lim_{x \to 2} \frac{x}{x-1} = \frac{\lim\limits_{x \to 2} x}{\lim\limits_{x \to 2} (x-1)}$$

$$= \frac{2}{1}$$

$$= 2$$

Example 1.4.5

In the next example we show that many different things can happen if the limit of the denominator is zero.

Example 1.4.6 (Be careful with limits of ratios)

We must be careful when computing the limit of a ratio — it is the ratio of the limits except when the limit of the denominator is zero. When the limit of the denominator is zero Theorem 1.4.2 **does not apply** and a few interesting things can happen

- If the limit of the numerator is non-zero then the limit of the ratio does not exist

$$\lim_{x \to a} \frac{f(x)}{g(x)} = DNE \qquad\qquad \text{when } \lim_{x \to a} f(x) \neq 0 \text{ and } \lim_{x \to a} g(x) = 0$$

  For example, $\lim\limits_{x \to 0} \dfrac{1}{x^2} = DNE$.

- If the limit of the numerator is zero then the above theorem does not give us enough information to decide whether or not the limit exists. It is possible that

  – the limit does not exist, eg. $\lim\limits_{x \to 0} \dfrac{x}{x^2} = \lim\limits_{x \to 0} \dfrac{1}{x} = DNE$

- the limit is $\pm\infty$, eg. $\lim\limits_{x\to0}\dfrac{x^2}{x^4} = \lim\limits_{x\to0}\dfrac{1}{x^2} = +\infty$ or $\lim\limits_{x\to0}\dfrac{-x^2}{x^4} = \lim\limits_{x\to0}\dfrac{-1}{x^2} = -\infty$.

- the limit is zero, eg. $\lim\limits_{x\to0}\dfrac{x^2}{x} = 0$

- the limit exists and is non-zero, eg. $\lim\limits_{x\to0}\dfrac{x}{x} = 1$

Now while the above examples are very simple and a little contrived they serve to illustrate the point we are trying to make — be careful if the limit of the denominator is zero.

Example 1.4.6

We now have enough theory to return to our rational function and compute its limit as x approaches 2.

Example 1.4.7

Let $h(x) = \dfrac{2x - 3}{x^2 + 5x - 6}$ and find its limit as x approaches 2.

Since this is the limit of a ratio, we compute the limit of the numerator and denominator separately. Numerator first:

$$
\begin{aligned}
\lim_{x\to2} 2x - 3 &= \left(\lim_{x\to2} 2x\right) - \left(\lim_{x\to2} 3\right) && \text{difference of limits} \\
&= 2\cdot\left(\lim_{x\to2} x\right) - 3 && \text{product of limits and limit of constant} \\
&= 2\cdot 2 - 3 && \text{limits of } x \\
&= 1
\end{aligned}
$$

Denominator next:

$$
\begin{aligned}
\lim_{x\to2} x^2 + 5x - 6 &= \left(\lim_{x\to2} x^2\right) + \left(\lim_{x\to2} 5x\right) - \left(\lim_{x\to2} 6\right) && \text{sum of limits} \\
&= \left(\lim_{x\to2} x\right)\cdot\left(\lim_{x\to2} x\right) + 5\cdot\left(\lim_{x\to2} x\right) - 6 && \text{product of limits and limit of constant} \\
&= 2\cdot 2 + 5\cdot 2 - 6 && \text{limits of } x \\
&= 8
\end{aligned}
$$

Since the limit of the denominator is non-zero, we can obtain our result by taking the ratio of the separate limits.

$$
\lim_{x\to2}\frac{2x - 3}{x^2 + 5x - 6} = \frac{\lim\limits_{x\to2} 2x - 3}{\lim\limits_{x\to2} x^2 + 5x - 6} = \frac{1}{8}
$$

The above works out quite simply. However, if we were to take the limit as $x \to 1$ then things are a bit harder. The limit of the numerator is:

$$
\lim_{x\to1} 2x - 3 = 2\cdot 1 - 3 = -1
$$

(we have not listed all the steps). And the limit of the denominator is

$$\lim_{x \to 1} x^2 + 5x - 6 = 1 \cdot 1 + 5 - 6 = 0$$

Since the limit of the numerator is non-zero, while the limit of the denominator is zero, the limit of the ratio does not exist.

$$\lim_{x \to 1} \frac{2x - 3}{x^2 + 5x - 6} = DNE$$

Example 1.4.7

It is **IMPORTANT TO NOTE** that it is not correct to write

$$\lim_{x \to 1} \frac{2x - 3}{x^2 + 5x - 6} = \frac{-1}{0} = DNE$$

Because we can only write

$$\lim_{x \to a} \frac{f(x)}{g(x)} = \frac{\lim_{x \to a} f(x)}{\lim_{x \to a} g(x)} = \text{something}$$

when the limit of the denominator is non-zero (see Example 1.4.6 above).

With a little care you can use the arithmetic of limits to obtain the following rules for limits of powers of functions and limits of roots of functions:

Theorem 1.4.8 (More arithmetic of limits — powers and roots).

Let n be a positive integer, let $a \in \mathbb{R}$ and let f be a function so that

$$\lim_{x \to a} f(x) = F$$

for some real number F. Then the following holds

$$\lim_{x \to a} (f(x))^n = \left(\lim_{x \to a} f(x) \right)^n = F^n$$

so that the limit of a power is the power of the limit. Similarly, if

- n is an even number and $F > 0$, or

- n is an odd number and F is any real number

then

$$\lim_{x \to a} (f(x))^{1/n} = \left(\lim_{x \to a} f(x) \right)^{1/n} = F^{1/n}$$

More generally[15], if $F > 0$ and p is any real number,

$$\lim_{x \to a} (f(x))^p = \left(\lim_{x \to a} f(x) \right)^p = F^p$$

Notice that we have to be careful when taking roots of limits that might be negative numbers. To see why, consider the case $n = 2$, the limit

$$\lim_{x \to 4} x^{1/2} = 4^{1/2} = 2$$

$$\lim_{x \to 4} (-x)^{1/2} = (-4)^{1/2} = \text{not a real number}$$

In order to evaluate such limits properly we need to use complex numbers which are beyond the scope of this text.

Also note that the notation $x^{1/2}$ refers to the *positive* square root of x. While 2 and (-2) are both square-roots of 4, the notation $4^{1/2}$ means 2. This is something we must be careful of[16].

So again — let us do a few examples and carefully note what we are doing.

Example 1.4.9

$$\lim_{x \to 2} (4x^2 - 3)^{1/3} = \left(\left(\lim_{x \to 2} 4x^2 \right) - \left(\lim_{x \to 2} 3 \right) \right)^{1/3}$$

$$= \left(4 \cdot 2^2 - 3 \right)^{1/3}$$

$$= (16 - 3)^{1/3}$$

$$= 13^{1/3}$$

Example 1.4.9

By combining the last few theorems we can make the evaluation of limits of polynomials and rational functions much easier:

Theorem 1.4.10 (Limits of polynomials and rational functions).

Let $a \in \mathbb{R}$, let $P(x)$ be a polynomial and let $R(x)$ be a rational function. Then

$$\lim_{x \to a} P(x) = P(a)$$

and provided $R(x)$ is defined at $x = a$ then

$$\lim_{x \to a} R(x) = R(a)$$

If $R(x)$ is not defined at $x = a$ then we are not able to apply this result.

---

15 You may not know the definition of the power b^p when p is not a rational number, so here it is. If $b > 0$ and p is any real number, then b^p is the limit of b^r as r approaches p through rational numbers. We won't do so here, but it is possible to prove that the limit exists.

16 Like ending sentences in prepositions — "This is something up with which we will not put." This quote is attributed to Churchill though there is some dispute as to whether or not he really said it.

So the previous examples are now much easier to compute:

$$\lim_{x \to 2} \frac{2x - 3}{x^2 + 5x - 6} = \frac{4 - 3}{4 + 10 - 6} = \frac{1}{8}$$

$$\lim_{x \to 2}(4x^2 - 1) = 16 - 1 = 15$$

$$\lim_{x \to 2} \frac{x}{x - 1} = \frac{2}{2 - 1} = 2$$

It is clear that limits of polynomials are very easy, while those of rational functions are easy except when the denominator might go to zero. We have seen examples where the resulting limit does not exist, and some where it does. We now work to explain this more systematically. The following example demonstrates that it is sometimes possible to take the limit of a rational function to a point at which the denominator is zero. Indeed we must be able to do exactly this in order to be able to define derivatives in the next chapter.

Example 1.4.11

Consider the limit

$$\lim_{x \to 1} \frac{x^3 - x^2}{x - 1}.$$

If we try to apply the arithmetic of limits then we compute the limits of the numerator and denominator separately

$$\lim_{x \to 1} x^3 - x^2 = 1 - 1 = 0 \tag{1.4.1}$$

$$\lim_{x \to 1} x - 1 = 1 - 1 = 0 \tag{1.4.2}$$

Since the denominator is zero, we cannot apply our theorem and we are, for the moment, stuck. However, there is more that we can do here — the hint is that the numerator and denominator *both* approach zero as x approaches 1. This means that there might be something we can cancel.

So let us play with the expression a little more before we take the limit:

$$\frac{x^3 - x^2}{x - 1} = \frac{x^2(x - 1)}{x - 1} = x^2 \qquad\qquad \text{provided } x \neq 1.$$

So what we really have here is the following function

$$\frac{x^3 - x^2}{x - 1} = \begin{cases} x^2 & x \neq 1 \\ \text{undefined} & x = 1 \end{cases}$$

If we plot the above function the graph looks exactly the same as $y = x^2$ except that the function is not defined at $x = 1$ (since at $x = 1$ both numerator and denominator are zero).
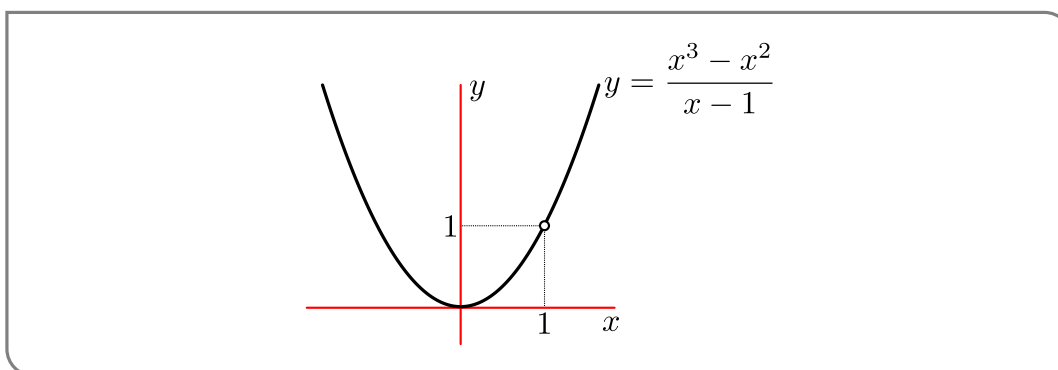
When we compute a limit as $x \to a$, the value of the function exactly at $x = a$ is irrelevant. We only care what happens to the function as we bring x very close to a. So for the above problem we can write

$$\frac{x^3 - x^2}{x - 1} = x^2 \qquad \text{when } x \text{ is close to 1 but not at } x = 1$$

So the limit as $x \to 1$ of the function is the same as the limit $\lim\limits_{x \to 1} x^2$ since the functions are the same except exactly at $x = 1$. By this reasoning we get

$$\lim_{x \to 1} \frac{x^3 - x^2}{x - 1} = \lim_{x \to 1} x^2 = 1$$

Example 1.4.11

The reasoning in the above example can be made more general:

Theorem 1.4.12.

If $f(x) = g(x)$ except when $x = a$ then $\lim\limits_{x \to a} f(x) = \lim\limits_{x \to a} g(x)$ provided the limit of g exists.

How do we know when to use this theorem? The big clue is that when we try to compute the limit in a naive way, we end up with $\frac{0}{0}$. We know that $\frac{0}{0}$ does not make sense, but it is an indication that there might be a common factor between numerator and denominator that can be cancelled. In the previous example, this common factor was $(x - 1)$.

Example 1.4.13

Using this idea compute

$$\lim_{h \to 0} \frac{(1+h)^2 - 1}{h}$$

- First we should check that we cannot just substitute $h = 0$ into this — clearly we cannot because the denominator would be 0.

- But we should also check the numerator to see if we have $\frac{0}{0}$, and we see that the numerator gives us $1 - 1 = 0$.

- Thus we have a hint that there is a common factor that we might be able to cancel. So now we look for the common factor and try to cancel it.

$$
\begin{aligned}
\frac{(1+h)^2 - 1}{h} &= \frac{1 + 2h + h^2 - 1}{h} & \text{expand} \\
&= \frac{2h + h^2}{h} = \frac{h(2+h)}{h} & \text{factor and then cancel} \\
&= 2 + h
\end{aligned}
$$

- Thus we really have that

$$
\frac{(1+h)^2 - 1}{h} = \begin{cases} 2+h & h \neq 0 \\ \text{undefined} & h = 0 \end{cases}
$$

and because of this

$$
\begin{aligned}
\lim_{h \to 0} \frac{(1+h)^2 - 1}{h} &= \lim_{h \to 0} 2 + h \\
&= 2
\end{aligned}
$$

Example 1.4.13

Of course — we have written everything out in great detail here and that is way more than is required for a solution to such a problem. Let us do it again a little more succinctly.

Example 1.4.14

Compute the following limit:

$$
\lim_{h \to 0} \frac{(1+h)^2 - 1}{h}
$$

If we try to use the arithmetic of limits, then we see that the limit of the numerator and the limit of the denominator are both zero. Hence we should try to factor them and cancel any common factor. This gives

$$
\begin{aligned}
\lim_{h \to 0} \frac{(1+h)^2 - 1}{h} &= \lim_{h \to 0} \frac{1 + 2h + h^2 - 1}{h} \\
&= \lim_{h \to 0} 2 + h \\
&= 2
\end{aligned}
$$

Example 1.4.14

Notice that even though we did this example carefully above, we have still written some text in our working explaining what we have done. You should always think about the

reader and if in doubt, put in more explanation rather than less. We could make the above example even more terse

Example 1.4.15

Compute the following limit:

$$\lim_{h \to 0} \frac{(1+h)^2 - 1}{h}$$

Numerator and denominator both go to zero as $h \to 0$. So factor and simplify:

$$\lim_{h \to 0} \frac{(1+h)^2 - 1}{h} = \lim_{h \to 0} \frac{1 + 2h + h^2 - 1}{h}$$
$$= \lim_{h \to 0} 2 + h = 2$$

Example 1.4.15

A slightly harder one now

Example 1.4.16

Compute the limit

$$\lim_{x \to 0} \frac{x}{\sqrt{1+x} - 1}$$

If we try to use the arithmetic of limits we get

$$\lim_{x \to 0} x = 0$$
$$\lim_{x \to 0} \sqrt{1+x} - 1 = \sqrt{\lim_{x \to 0} 1 + x} - 1 = 1 - 1 = 0$$

So doing the naive thing we'd get $0/0$. This suggests a common factor that can be cancelled. Since the numerator and denominator are not polynomials we have to try other tricks[17] . We can simplify the denominator $\sqrt{1+x} - 1$ a lot, and in particular eliminate

---

17 While these tricks are useful (and even cute[18]), Taylor polynomials (see Section 3.4) give us a more systematic way of approaching this problem.

18 Mathematicians tend to have quite strong opinions on the beauty of mathematics. For example, Paul Erdös[19]said "Why are numbers beautiful? It's like asking why is Beethoven's Ninth Symphony beautiful. If you don't see why, someone can't tell you. I know numbers are beautiful. If they aren't beautiful, nothing is.".

19 Arguably the most prolific mathematician of the 20th century — definitely worth a google. The authors do not know his opinion on nested footnotes[20].

20 Nested footnotes are generally frowned upon, since they can get quite contorted; see XKCD-1208 and also the novel "House of Leaves" by Mark Z. Danielewski.

the square root, by multiplying it by its conjugate $\sqrt{1+x}+1$.

$$
\begin{aligned}
\frac{x}{\sqrt{1+x}-1} &= \frac{x}{\sqrt{1+x}-1} \times \frac{\sqrt{1+x}+1}{\sqrt{1+x}+1} && \text{multiply by } \frac{\text{conjugate}}{\text{conjugate}} = 1 \\
&= \frac{x\left(\sqrt{1+x}+1\right)}{\left(\sqrt{1+x}-1\right)\left(\sqrt{1+x}+1\right)} && \text{bring things together} \\
&= \frac{x\left(\sqrt{1+x}+1\right)}{\left(\sqrt{1+x}\right)^2 - 1\cdot 1} && \text{since } (a-b)(a+b) = a^2 - b^2 \\
&= \frac{x\left(\sqrt{1+x}+1\right)}{1+x-1} && \text{clean up a little} \\
&= \frac{x\left(\sqrt{1+x}+1\right)}{x} \\
&= \sqrt{1+x}+1 && \text{cancel the } x
\end{aligned}
$$

So now we have

$$
\begin{aligned}
\lim_{x\to 0}\frac{x}{\sqrt{1+x}-1} &= \lim_{x\to 0}\sqrt{1+x}+1 \\
&= \sqrt{1+0}+1 = 2
\end{aligned}
$$

Example 1.4.16

How did we know what to multiply by? Our function was of the form

$$
\frac{a}{\sqrt{b}-c}
$$

so, to eliminate the square root from the denominator, we employ a trick — we multiply by 1. Of course, multiplying by 1 doesn't do anything. But if you multiply by 1 carefully you can leave the value the same, but change the form of the expression. More precisely

$$
\begin{aligned}
\frac{a}{\sqrt{b}-c} &= \frac{a}{\sqrt{b}-c}\cdot 1 \\
&= \frac{a}{\sqrt{b}-c}\cdot \underbrace{\frac{\sqrt{b}+c}{\sqrt{b}+c}}_{=1} \\
&= \frac{a\left(\sqrt{b}+c\right)}{\left(\sqrt{b}-c\right)\left(\sqrt{b}+c\right)} && \text{expand denominator carefully} \\
&= \frac{a\left(\sqrt{b}+c\right)}{\sqrt{b}\cdot\sqrt{b} - c\sqrt{b} + c\sqrt{b} - c\cdot c} && \text{do some cancellation} \\
&= \frac{a\left(\sqrt{b}+c\right)}{b-c^2}
\end{aligned}
$$

Now the numerator contains roots, but the denominator is just a polynomial.

Before we move on to limits at infinity, there is one more theorem to see. While the scope of its application is quite limited, it can be extremely useful. It is called a sandwich theorem or a squeeze theorem for reasons that will become apparent.

Sometimes one is presented with an unpleasant ugly function such as

$$f(x) = x^2 \sin(\pi/x)$$

It is a fact of life, that not all the functions that are encountered in mathematics will be elegant and simple; this is especially true when the mathematics gets applied to real world problems. One just has to work with what one gets. So how can we compute

$$\lim_{x \to 0} x^2 \sin(\pi/x)?$$

Since it is the product of two functions, we might try

$$\lim_{x \to 0} x^2 \sin(\pi/x) = \left(\lim_{x \to 0} x^2 \right) \cdot \left(\lim_{x \to 0} \sin(\pi/x) \right)$$

$$= 0 \cdot \left(\lim_{x \to 0} \sin(\pi/x) \right)$$

$$= 0$$

But we just cheated — we cannot use the arithmetic of limits theorem here, because the limit

$$\lim_{x \to 0} \sin(\pi/x) = DNE$$

does not exist. Now we did see the function $\sin(\pi/x)$ before (in Example 1.3.5), so you should go back and look at it again. Unfortunately the theorem "the limit of a product is the product of the limits" only holds when the limits you are trying to multiply together actually exist. So we cannot use it.

However, we do see that the function naturally decomposes into the product of two pieces — the functions x^2 and $\sin(\pi/x)$. We have sketched the two functions in the figure on the left below.

Figure 1.4.2.
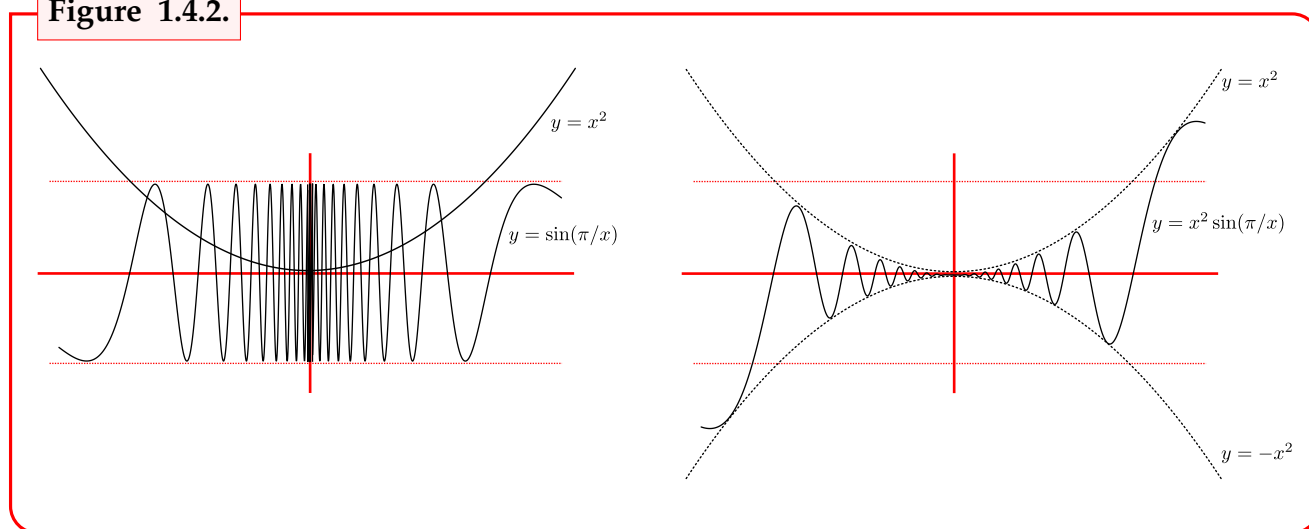
While x^2 is a very well behaved function and we know quite a lot about it, the function $\sin(\pi/x)$ is quite ugly. One of the few things we can say about it is the following

$$-1 \leqslant \sin(\pi/x) \leqslant 1 \qquad\qquad \text{provided } x \neq 0$$

But if we multiply this expression by x^2 we get (because $x^2 \geqslant 0$)

$$-x^2 \leqslant x^2 \sin(\pi/x) \leqslant x^2 \qquad\qquad \text{provided } x \neq 0$$

and we have sketched the result in the figure above (on the right). So the function we are interested in is *squeezed* or *sandwiched* between the functions x^2 and $-x^2$.

If we focus in on the picture close to $x = 0$ we see that x approaches 0, the functions x^2 and $-x^2$ both approach 0. Further, because $x^2 \sin(\pi/x)$ is sandwiched between them, it seems that it also approaches 0.

The following theorem tells us that this is indeed the case:

> **Theorem 1.4.17** (Squeeze theorem (or sandwich theorem or pinch theorem)).
>
> Let $a \in \mathbb{R}$ and let f, g, h be three functions so that
>
> $$f(x) \leqslant g(x) \leqslant h(x)$$
>
> for all x in an interval around a, except possibly exactly at $x = a$. Then if
>
> $$\lim_{x \to a} f(x) = \lim_{x \to a} h(x) = L$$
>
> then it is also the case that
>
> $$\lim_{x \to a} g(x) = L$$

Using the above theorem we can compute the limit we want and write it up nicely

Example 1.4.18

Compute the limit

$$\lim_{x \to 0} x^2 \sin(\pi/x)$$

Since $-1 \leqslant \sin(\theta) \leqslant 1$ for all real numbers θ, we have

$$-1 \leqslant \sin(\pi/x) \leqslant 1 \qquad\qquad \text{for all } x \neq 0$$

Multiplying the above by x^2 we see that

$$-x^2 \leqslant x^2 \sin(\pi/x) \leqslant x^2 \qquad\qquad \text{for all } x \neq 0$$

Since $\lim_{x \to 0} x^2 = \lim_{x \to 0}(-x^2) = 0$ by the sandwich (or squeeze or pinch) theorem we have

$$\lim_{x \to 0} x^2 \sin(\pi/x) = 0$$

Notice how we have used "words". We have remarked on this several times already in the text, but we will keep mentioning it. It is okay to use words in your answers to maths problems — and you should do so! These let the reader know what you are doing and help you understand what you are doing.

Another sandwich theorem example

Example 1.4.19

Let $f(x)$ be a function such that $1 \leqslant f(x) \leqslant x^2 - 2x + 2$. What is $\lim_{x \to 1} f(x)$?

We are already supplied with an inequality, so it is likely that it is going to help us. We should examine the limits of each side to see if they are the same:

$$\lim_{x \to 1} 1 = 1$$

$$\lim_{x \to 1} x^2 - 2x + 2 = 1 - 2 + 2 = 1$$

So we see that the function $f(x)$ is trapped between two functions that both approach 1 as $x \to 1$. Hence by the sandwich / pinch / squeeze theorem, we know that

$$\lim_{x \to 1} f(x) = 1$$

To get some intuition as to why the squeeze theorem is true, consider when x is very very close to a. In particular, consider when x is sufficiently close to a that we know $h(x)$ is within 10^{-6} of L and that $f(x)$ is also within 10^{-6} of L. That is

$$|h(x) - L| < 10^{-6} \qquad \text{and} \qquad |f(x) - L| < 10^{-6}.$$

This means that

$$L - 10^{-6} < f(x) \leqslant h(x) < L + 10^{-6}$$

since we know that $f(x) \leqslant h(x)$.

But now by the hypothesis of the squeeze theorem we know that $f(x) \leqslant g(x) \leqslant h(x)$ and so we have

$$L - 10^{-6} < f(x) \leqslant g(x) \leqslant h(x) < L + 10^{-6}$$

And thus we know that

$$L - 10^{-6} \leqslant g(x) \leqslant L + 10^{-6} \tag{1.4.3}$$

That is $g(x)$ is also within 10^{-6} of L.

In this argument our choice of 10^{-6} was arbitrary, so we can really replace 10^{-6} with any small number we like. Hence we know that we can force $g(x)$ as close to L as we like, by bringing x sufficiently close to a. We give a more formal and rigorous version of this argument at the end of Section 1.9.

1.5 ▲ Limits at Infinity

Up until this point we have discussed what happens to a function as we move its input x closer and closer to a particular point a. For a great many applications of limits we need to understand what happens to a function when its input becomes extremely large — for example what happens to a population at a time far in the future.

The definition of a limit at infinity has a similar flavour to the definition of limits at finite points that we saw above, but the details are a little different. We also need to distinguish between positive and negative infinity. As x becomes very large and positive it moves off towards $+\infty$ but when it becomes very large and negative it moves off towards $-\infty$.

Again we give an informal definition; the full formal definition can be found in (the optional) Section 1.8 near the end of this chapter.

Definition 1.5.1 (Limits at infinity — informal).

We write

$$\lim_{x\to\infty} f(x) = L$$

when the value of the function $f(x)$ gets closer and closer to L as we make x larger and larger and positive.
Similarly we write

$$\lim_{x\to-\infty} f(x) = L$$

when the value of the function $f(x)$ gets closer and closer to L as we make x larger and larger and negative.

Example 1.5.2

Consider the two functions depicted below
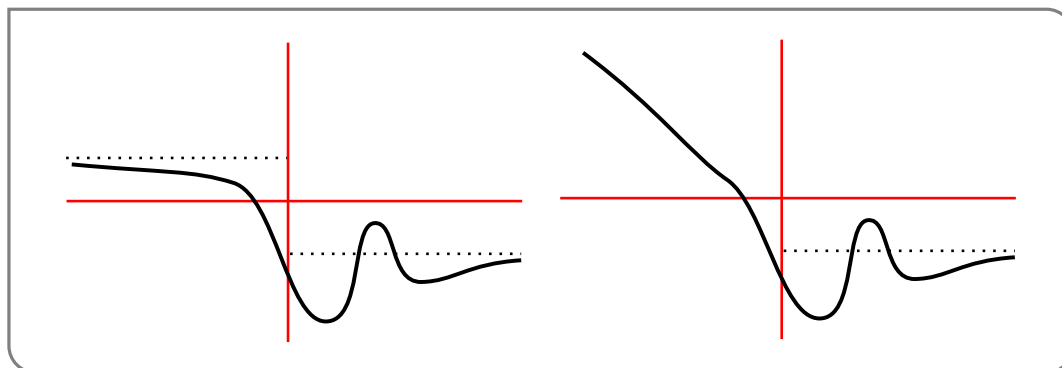
The dotted horizontal lines indicate the behaviour as x becomes very large. The function on the left has limits as $x \to \infty$ and as $x \to -\infty$ since the function "settles down" to a

particular value. On the other hand, the function on the right does not have a limit as $x \to -\infty$ since the function just keeps getting bigger and bigger.

Example 1.5.2

Just as was the case for limits as $x \to a$ we will start with two very simple building blocks and build other limits from those.

Theorem 1.5.3.

Let $c \in \mathbb{R}$ then the following limits hold

$$\lim_{x \to \infty} c = c \qquad\qquad \lim_{x \to -\infty} c = c$$

$$\lim_{x \to \infty} \frac{1}{x} = 0 \qquad\qquad \lim_{x \to -\infty} \frac{1}{x} = 0$$

Again, these limits interact nicely with standard arithmetic:

Theorem 1.5.4 (Arithmetic of limits at infinity).

Let $f(x), g(x)$ be two functions for which the limits

$$\lim_{x \to \infty} f(x) = F \qquad\qquad \lim_{x \to \infty} g(x) = G$$

exist. Then the following limits hold

$$\lim_{x \to \infty} f(x) \pm g(x) = F \pm G$$

$$\lim_{x \to \infty} f(x)g(x) = FG$$

$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = \frac{F}{G} \qquad\qquad\qquad \text{provided } G \neq 0$$

and for real numbers p

$$\lim_{x \to \infty} f(x)^p = F^p \qquad \text{provided } F^p \text{ and } f(x)^p \text{ are defined for all } x$$

The analogous results hold for limits to $-\infty$.

Note that, as was the case in Theorem 1.4.8, we need a little extra care with powers of functions. We must avoid taking square roots of negative numbers, or indeed any even root of a negative number[21].

---

21   To be more precise, there is no real number x so that $x^{\text{even power}}$ is a negative number. Hence we cannot take the even-root of a negative number and express it as a real number. This is precisely what complex numbers allow us to do, but alas there is not space in the course for us to explore them.

Hence we have for all rational $r > 0$

$$\lim_{x \to \infty} \frac{1}{x^r} = 0$$

but we have to be careful with

$$\lim_{x \to -\infty} \frac{1}{x^r} = 0$$

This is only true if the denominator of r is not an even number[22].
    For example

- $\lim_{x \to \infty} \frac{1}{x^{1/2}} = 0$, but $\lim_{x \to -\infty} \frac{1}{x^{1/2}}$ does not exist, because $x^{1/2}$ is not defined for $x < 0$.

- On the other hand, $x^{4/3}$ is defined for negative values of x and $\lim_{x \to -\infty} \frac{1}{x^{4/3}} = 0$.

Our first application of limits at infinity will be to examine the behaviour of a rational function for very large x. To do this we use a "trick".

Example 1.5.5

Compute the following limit:

$$\lim_{x \to \infty} \frac{x^2 - 3x + 4}{3x^2 + 8x + 1}$$

As x becomes very large, it is the x^2 term that will dominate in both the numerator and denominator and the other bits become irrelevant. That is, for very large x, x^2 is much much larger than x or any constant. So we pull out these dominant parts

$$\frac{x^2 - 3x + 4}{3x^2 + 8x + 1} = \frac{x^2 \left(1 - \frac{3}{x} + \frac{4}{x^2}\right)}{x^2 \left(3 + \frac{8}{x} + \frac{1}{x^2}\right)}$$

$$= \frac{1 - \frac{3}{x} + \frac{4}{x^2}}{3 + \frac{8}{x} + \frac{1}{x^2}} \qquad \text{remove the common factors}$$

$$\lim_{x \to \infty} \frac{x^2 - 3x + 4}{3x^2 + 8x + 1} = \lim_{x \to \infty} \frac{1 - \frac{3}{x} + \frac{4}{x^2}}{3 + \frac{8}{x} + \frac{1}{x^2}}$$

$$= \frac{\lim_{x \to \infty} \left(1 - \frac{3}{x} + \frac{4}{x^2}\right)}{\lim_{x \to \infty} \left(3 + \frac{8}{x} + \frac{1}{x^2}\right)} \qquad \text{arithmetic of limits}$$

$$= \frac{\lim_{x \to \infty} 1 - \lim_{x \to \infty} \frac{3}{x} + \lim_{x \to \infty} \frac{4}{x^2}}{\lim_{x \to \infty} 3 + \lim_{x \to \infty} \frac{8}{x} + \lim_{x \to \infty} \frac{1}{x^2}} \qquad \text{more arithmetic of limits}$$

$$= \frac{1 + 0 + 0}{3 + 0 + 0} = \frac{1}{3}$$

22 where we write $r = \frac{p}{q}$ with p, q integers with no common factors. For example, $r = \frac{6}{14}$ should be written as $r = \frac{3}{7}$ when considering this rule.

Example 1.5.5

The following one gets a little harder

Example 1.5.6

Find the limit as $x \to \infty$ of $\frac{\sqrt{4x^2+1}}{5x-1}$

We use the same trick — try to work out what is the biggest term in the numerator and denominator and pull it to one side.

- The denominator is dominated by $5x$.

- The biggest contribution to the numerator comes from the $4x^2$ inside the square-root. When we pull x^2 outside the square-root it becomes x, so the numerator is dominated by $x \cdot \sqrt{4} = 2x$

- To see this more explicitly rewrite the numerator

$$\sqrt{4x^2+1} = \sqrt{x^2(4+1/x^2)} = \sqrt{x^2}\sqrt{4+1/x^2} = x\sqrt{4+1/x^2}.$$

- Thus the limit as $x \to \infty$ is

$$\lim_{x\to\infty} \frac{\sqrt{4x^2+1}}{5x-1} = \lim_{x\to\infty} \frac{x\sqrt{4+1/x^2}}{x(5-1/x)}$$
$$= \lim_{x\to\infty} \frac{\sqrt{4+1/x^2}}{5-1/x}$$
$$= \frac{2}{5}$$

Example 1.5.6

Now let us also think about the limit of the same function, $\frac{\sqrt{4x^2+1}}{5x-1}$, as $x \to -\infty$. There is something subtle going on because of the square-root. First consider the function[23]

$$h(t) = \sqrt{t^2}$$

Evaluating this at $t = 7$ gives

$$h(7) = \sqrt{7^2} = \sqrt{49} = 7$$

We'll get much the same thing for any $t \geqslant 0$. For any $t \geqslant 0$, $h(t) = \sqrt{t^2}$ returns exactly t. However now consider the function at $t = -3$

$$h(-3) = \sqrt{(-3)^2} = \sqrt{9} = 3 = -(-3)$$

23 Just to change things up let's use t and $h(t)$ instead of the ubiquitous x and $f(x)$.

that is the function is returning -1 times the input.

This is because when we defined $\sqrt{}$, we defined it to be the *positive* square-root. i.e. the function $\sqrt{t}$ can never return a negative number. So being more careful

$$h(t) = \sqrt{t^2} = |t|$$

Where the $|t|$ is the absolute value of t. You are perhaps used to thinking of absolute value as "remove the minus sign", but this is not quite correct. Let's sketch the function

Figure 1.5.1.
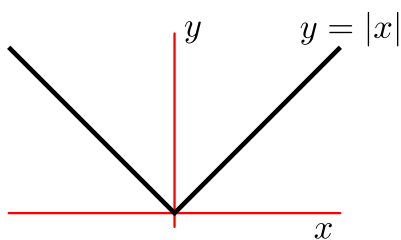
It is a piecewise function defined by

$$|x| = \begin{cases} x & x \geqslant 0 \\ -x & x < 0 \end{cases}$$

Hence our function $h(t)$ is really

$$h(t) = \sqrt{t^2} = \begin{cases} t & t \geqslant 0 \\ -t & t < 0 \end{cases}$$

So that when we evaluate $h(-7)$ it is

$$h(-7) = \sqrt{(-7)^2} = \sqrt{49} = 7 = -(-7)$$

We are now ready to examine the limit as $x \to -\infty$ in our previous example. Mostly it is copy and paste from above.

Example 1.5.7

Find the limit as $x \to -\infty$ of $\frac{\sqrt{4x^2+1}}{5x-1}$

We use the same trick — try to work out what is the biggest term in the numerator and denominator and pull it to one side. Since we are taking the limit as $x \to -\infty$ we should think of x as a large negative number.

- The denominator is dominated by $5x$.

- The biggest contribution to the numerator comes from the $4x^2$ inside the square-root. When we pull the x^2 outside a square-root it becomes $|x| = -x$ (since we are taking the limit as $x \to -\infty$), so the numerator is dominated by $-x \cdot \sqrt{4} = -2x$

- To see this more explicitly rewrite the numerator

$$
\begin{aligned}
\sqrt{4x^2 + 1} &= \sqrt{x^2(4 + 1/x^2)} = \sqrt{x^2}\sqrt{4 + 1/x^2} \\
&= |x|\sqrt{4 + 1/x^2} \qquad\qquad \text{and since } x < 0 \text{ we have} \\
&= -x\sqrt{4 + 1/x^2}
\end{aligned}
$$

- Thus the limit as $x \to -\infty$ is

$$
\begin{aligned}
\lim_{x\to-\infty} \frac{\sqrt{4x^2 + 1}}{5x - 1} &= \lim_{x\to-\infty} \frac{-x\sqrt{4 + 1/x^2}}{x(5 - 1/x)} \\
&= \lim_{x\to-\infty} \frac{-\sqrt{4 + 1/x^2}}{5 - 1/x} \\
&= -\frac{2}{5}
\end{aligned}
$$

Example 1.5.7

So the limit as $x \to -\infty$ is almost the same but we gain a minus sign. This **is definitely not** the case in general — you have to think about each example separately.

Here is a sketch of the function in question.

Figure 1.5.2.

$$y = \frac{\sqrt{4x^2+1}}{5x-1}$$
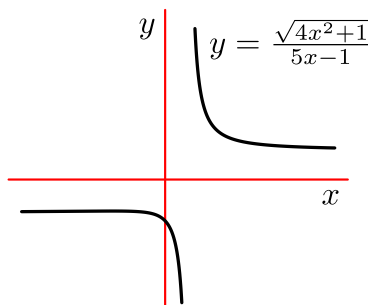
Example 1.5.8

Compute the following limit:

$$
\lim_{x\to\infty} \left(x^{7/5} - x \right)
$$

In this case we cannot use the arithmetic of limits to write this as

$$
\begin{aligned}
\lim_{x\to\infty} \left(x^{7/5} - x \right) &= \left(\lim_{x\to\infty} x^{7/5} \right) - \left(\lim_{x\to\infty} x \right) \\
&= \infty - \infty
\end{aligned}
$$

because the limits do not exist. We can only use the limit laws when the limits exist. So we should go back and think some more.

When x is very large, $x^{7/5} = x \cdot x^{2/5}$ will be much larger than x, so the $x^{7/5}$ term will dominate the x term. So factor out $x^{7/5}$ and rewrite it as

$$x^{7/5} - x = x^{7/5}\left(1 - \frac{1}{x^{2/5}}\right)$$

Consider what happens to each of the factors as $x \to \infty$

- For large x, $x^{7/5} > x$ (this is actually true for any $x > 1$). In the limit as $x \to +\infty$, x becomes arbitrarily large and positive, and $x^{7/5}$ must be bigger still, so it follows that

$$\lim_{x\to\infty} x^{7/5} = +\infty.$$

- On the other hand, $(1 - x^{-2/5})$ becomes closer and closer to 1 — we can use the arithmetic of limits to write this as

$$\lim_{x\to\infty}\left(1 - x^{-2/5}\right) = \lim_{x\to\infty} 1 - \lim_{x\to\infty} x^{-2/5} = 1 - 0 = 1$$

So the product of these two factors will be come larger and larger (and positive) as x moves off to infinity. Hence we have

$$\lim_{x\to\infty} x^{7/5}\left(1 - 1/x^{2/5}\right) = +\infty$$

Example 1.5.8

But remember $+\infty$ and $-\infty$ are not numbers; the last equation in the example is shorthand for "the function becomes arbitrarily large".

In the previous section we saw that finite limits and arithmetic interact very nicely (see Theorems 1.4.2 and 1.4.8). This enabled us to compute the limits of more complicated function in terms of simpler ones. When limits of functions go to plus or minus infinity we are quite a bit more restricted in what we can deduce. The next theorem states some results concerning the sum, difference, ratio and product of infinite limits — unfortunately in many cases we cannot make general statements and the results will depend on the details of the problem at hand.

Theorem 1.5.9 (Arithmetic of infinite limits).

Let $a, c, H \in \mathbb{R}$ and let f, g, h be functions defined in an interval around a (but they need not be defined at $x = a$), so that

$$\lim_{x \to a} f(x) = +\infty \qquad \lim_{x \to a} g(x) = +\infty \qquad \lim_{x \to a} h(x) = H$$

- $\lim_{x \to a} (f(x) + g(x)) = +\infty$

- $\lim_{x \to a} (f(x) + h(x)) = +\infty$

- $\lim_{x \to a} (f(x) - g(x))$ undetermined

- $\lim_{x \to a} (f(x) - h(x)) = +\infty$

- $\lim_{x \to a} cf(x) = \begin{cases} +\infty & c > 0 \\ 0 & c = 0 \\ -\infty & c < 0 \end{cases}$

- $\lim_{x \to a} (f(x) \cdot g(x)) = +\infty.$

- $\lim_{x \to a} f(x)h(x) = \begin{cases} +\infty & H > 0 \\ -\infty & H < 0 \\ \text{undetermined} & H = 0 \end{cases}$

- $\lim_{x \to a} \dfrac{f(x)}{g(x)}$ undetermined

- $\lim_{x \to a} \dfrac{f(x)}{h(x)} = \begin{cases} +\infty & H > 0 \\ -\infty & H < 0 \\ \text{undetermined} & H = 0 \end{cases}$

- $\lim_{x \to a} \dfrac{h(x)}{f(x)} = 0$

- $\lim_{x \to a} f(x)^p = \begin{cases} +\infty & p > 0 \\ 0 & p < 0 \\ 1 & p = 0 \end{cases}$

Note that by "undetermined" we mean that the limit may or may not exist, but cannot be determined from the information given in the theorem. See Example 1.4.6 for an example of what we mean by "undetermined". Additionally consider the following example.

Example 1.5.10

Consider the following 3 functions:

$$f(x) = x^{-2} \qquad\qquad g(x) = 2x^{-2} \qquad\qquad h(x) = x^{-2} - 1.$$

Their limits as $x \to 0$ are:

$$\lim_{x \to 0} f(x) = +\infty \qquad\qquad \lim_{x \to 0} g(x) = +\infty \qquad\qquad \lim_{x \to 0} h(x) = +\infty.$$

Say we want to compute the limit of the difference of two of the above functions as $x \to 0$. Then the previous theorem cannot help us. This is not because it is too weak, rather it is because the difference of two infinite limits can be, either plus infinity, minus infinity or some finite number depending on the details of the problem. For example,

$$\lim_{x \to 0} (f(x) - g(x)) = \lim_{x \to 0} -x^{-2} = -\infty$$
$$\lim_{x \to 0} (f(x) - h(x)) = \lim_{x \to 0} 1 = 1$$
$$\lim_{x \to 0} (g(x) - h(x)) = \lim_{x \to 0} x^{-2} + 1 = +\infty$$

Example 1.5.10

1.6 ▴ Continuity

We have seen that computing the limits some functions — polynomials and rational functions — is very easy because

$$\lim_{x \to a} f(x) = f(a).$$

That is, the limit as x approaches a is just $f(a)$. Roughly speaking, the reason we can compute the limit this way is that these functions do not have any abrupt jumps near a.

Many other functions have this property, $\sin(x)$ for example. A function with this property is called "continuous" and there is a precise mathematical definition for it. If you do not recall interval notation, then now is a good time to take a quick look back at Definition 0.3.5.

Definition 1.6.1.

A function $f(x)$ is continuous at a if

$$\lim_{x \to a} f(x) = f(a).$$

If a function is not continuous at a then it is said to be discontinuous at a.
When we write that f is continuous without specifying a point, then typically this means that f is continuous at a for all $a \in \mathbb{R}$.
When we write that $f(x)$ is continuous on the open interval (a, b) then the function is continuous at every point c satisfying $a < c < b$.

So if a function is continuous at $x = a$ we immediately know that

- $f(a)$ exists

- $\lim\limits_{x \to a^-}$ exists and is equal to $f(a)$, and

- $\lim\limits_{x \to a^+}$ exists and is equal to $f(a)$.

⤷ Quick Aside — One-sided Continuity

Notice in the above definition of continuity on an interval (a, b) we have carefully avoided saying anything about whether or not the function is continuous at the endpoints of the interval — i.e. is $f(x)$ continuous at $x = a$ or $x = b$. This is because talking of continuity at the endpoints of an interval can be a little delicate.

In many situations we will be given a function $f(x)$ defined on a closed interval $[a, b]$. For example, we might have:

$$f(x) = \frac{x+1}{x+2} \qquad\qquad \text{for } x \in [0, 1].$$

For any $0 \leqslant x \leqslant 1$ we know the value of $f(x)$. However for $x < 0$ or $x > 1$ we know nothing about the function — indeed it has not been defined.

So now, consider what it means for $f(x)$ to be continuous at $x = 0$. We need to have

$$\lim_{x \to 0} f(x) = f(0),$$

however this implies that the one-sided limits

$$\lim_{x \to 0^+} f(x) = f(0) \qquad\qquad \text{and} \qquad\qquad \lim_{x \to 0^-} f(x) = f(0)$$

Now the first of these one-sided limits involves examining the behaviour of $f(x)$ for $x > 0$. Since this involves looking at points for which $f(x)$ is defined, this is something we can do. On the other hand the second one-sided limit requires us to understand the behaviour of $f(x)$ for $x < 0$. This we cannot do because the function hasn't been defined for $x < 0$.

One way around this problem is to generalise the idea of continuity to one-sided continuity, just as we generalised limits to get one-sided limits.

Definition 1.6.2.

A function $f(x)$ is continuous from the right at a if

$$\lim_{x \to a^+} f(x) = f(a).$$

Similarly a function $f(x)$ is continuous from the left at a if

$$\lim_{x \to a^-} f(x) = f(a)$$

Using the definition of one-sided continuity we can now define what it means for a function to be continuous on a closed interval.

> **Definition 1.6.3.**
>
> A function $f(x)$ is continuous on the closed interval $[a, b]$ when
>
> - $f(x)$ is continuous on (a, b),
>
> - $f(x)$ is continuous from the right at a, and
>
> - $f(x)$ is continuous from the left at b.
>
> Note that the last two conditions are equivalent to
>
> $$\lim_{x \to a^+} f(x) = f(a) \qquad \text{and} \qquad \lim_{x \to b^-} f(x) = f(b).$$

▸▸ Back to the Main Text

We already know from our work above that polynomials are continuous, and that rational functions are continuous at all points in their domains — i.e. where their denominators are non-zero. As we did for limits, we will see that continuity interacts "nicely" with arithmetic. This will allow us to construct complicated continuous functions from simpler continuous building blocks (like polynomials).

But first, a few examples. . .

Example 1.6.4

Consider the functions drawn below
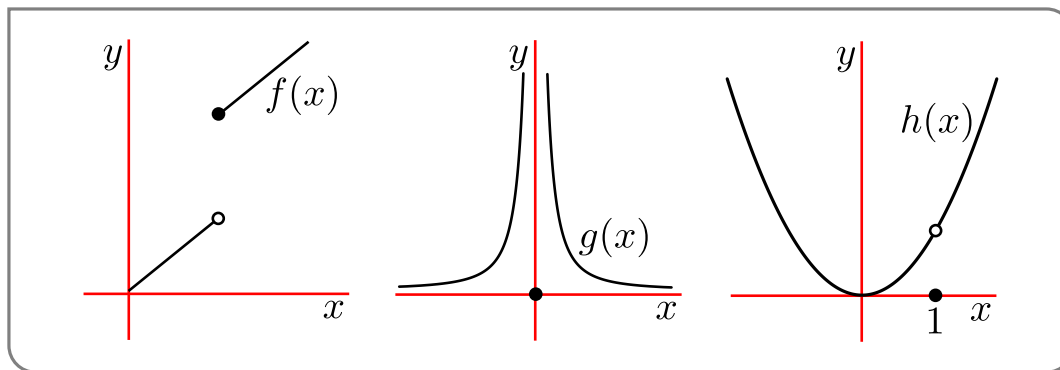
These are

$$f(x) = \begin{cases} x & x < 1 \\ x + 2 & x \geqslant 1 \end{cases} \qquad g(x) = \begin{cases} 1/x^2 & x \neq 0 \\ 0 & x = 0 \end{cases} \qquad h(x) = \begin{cases} \frac{x^3 - x^2}{x - 1} & x \neq 1 \\ 0 & x = 1 \end{cases}$$

Determine where they are continuous and discontinuous:

- When $x < 1$ then $f(x)$ is a straight line (and so a polynomial) and so it is continuous at every point $x < 1$. Similarly when $x > 1$ the function is a straight line and so it is

continuous at every point $x > 1$. The only point which might be a discontinuity is at $x = 1$. We see that the one sided limits are different. Hence the limit at $x = 1$ does not exist and so the function is discontinuous at $x = 1$.

But note that that $f(x)$ is continuous from one side — which?

- The middle case is much like the previous one. When $x \neq 0$ the $g(x)$ is a rational function and so is continuous everywhere on its domain (which is all reals except $x = 0$). Thus the only point where $g(x)$ might be discontinuous is at $x = 0$. We see that neither of the one-sided limits exist at $x = 0$, so the limit does not exist at $x = 0$. Hence the function is discontinuous at $x = 0$.

- We have seen the function $h(x)$ before. By the same reasoning as above, we know it is continuous except at $x = 1$ which we must check separately.

  By definition of $h(x)$, $h(1) = 0$. We must compare this to the limit as $x \to 1$. We did this before.

  $$\frac{x^3 - x^2}{x - 1} = \frac{x^2(x - 1)}{x - 1} = x^2$$

  So $\lim_{x \to 1} \frac{x^3 - x^2}{x - 1} = \lim_{x \to 1} x^2 = 1 \neq h(1)$. Hence h is discontinuous at $x = 1$.

Example 1.6.4

This example illustrates different sorts of discontinuities:

- The function $f(x)$ has a "jump discontinuity" because the function "jumps" from one finite value on the left to another value on the right.

- The second function, $g(x)$, has an "infinite discontinuity" since $\lim f(x) = +\infty$.

- The third function, $h(x)$, has a "removable discontinuity" because we could make the function continuous at that point by redefining the function at that point. i.e. setting $h(1) = 1$. That is

  $$\text{new function } h(x) = \begin{cases} \frac{x^3 - x^2}{x - 1} & x \neq 1 \\ 1 & x = 1 \end{cases}$$

Showing a function is continuous can be a pain, but just as the limit laws help us compute complicated limits in terms of simpler limits, we can use them to show that complicated functions are continuous by breaking them into simpler pieces.

> **Theorem 1.6.5** (Arithmetic of continuity).
>
> Let $a, c \in \mathbb{R}$ and let $f(x)$ and $g(x)$ be functions that are continuous at a. Then the following functions are also continuous at $x = a$:
>
> - $f(x) + g(x)$ and $f(x) - g(x)$,
>
> - $cf(x)$ and $f(x)g(x)$, and
>
> - $\frac{f(x)}{g(x)}$ provided $g(a) \neq 0$.

Above we stated that polynomials and rational functions are continuous (being careful about domains of rational functions — we must avoid the denominators being zero) without making it a formal statement. This is easily fixed...

> **Lemma 1.6.6.**
>
> Let $c \in \mathbb{R}$. The functions
>
> $$f(x) = x \qquad\qquad g(x) = c$$
>
> are continuous everywhere on the real line

This isn't quite the result we wanted (that's a couple of lines below) but it is a small result that we can combine with the arithmetic of limits to get the result we want. Such small helpful results are called "lemmas" and they will arise more as we go along.

Now since we can obtain any polynomial and any rational function by carefully adding, subtracting, multiplying and dividing the functions $f(x) = x$ and $g(x) = c$, the above lemma combines with the "arithmetic of continuity" theorem to give us the result we want:

> **Theorem 1.6.7** (Continuity of polynomials and rational functions).
>
> Every polynomial is continuous everywhere. Similarly every rational function is continuous except where its denominator is zero (i.e. on all its domain).

With some more work this result can be extended to wider families of functions:

> **Theorem 1.6.8.**
>
> The following functions are continuous everywhere in their domains
>
> - polynomials, rational functions
>
> - roots and powers
>
> - trig functions and their inverses
>
> - exponential and the logarithm

We haven't encountered inverse trigonometric functions, nor exponential functions or logarithms, but we will see them in the next chapter. For the moment, just file the information away.

Using a combination of the above results you can show that many complicated functions are continuous except at a few points (usually where a denominator is equal to zero).

Example 1.6.9

Where is the function $f(x) = \frac{\sin(x)}{2+\cos(x)}$ continuous?

We just break things down into pieces and then put them back together keeping track of where things might go wrong.

- The function is a ratio of two pieces — so check if the numerator is continuous, the denominator is continuous, and if the denominator might be zero.

- The numerator is $\sin(x)$ which is "continuous on its domain" according to one of the above theorems. Its domain is all real numbers[24], so it is continuous everywhere. No problems here.

- The denominator is the sum of 2 and $\cos(x)$. Since 2 is a constant it is continuous everywhere. Similarly (we just checked things for the previous point) we know that $\cos(x)$ is continuous everywhere. Hence the denominator is continuous.

- So we just need to check if the denominator is zero. One of the facts that we should know[25] is that

$$-1 \leqslant \cos(x) \leqslant 1$$

and so by adding 2 we get

$$1 \leqslant 2 + \cos(x) \leqslant 3$$

Thus no matter what value of x, $2 + \cos(x) \geqslant 1$ and so cannot be zero.

---

24 Remember that sin and cos are defined on all real numbers, so $\tan(x) = \sin(x)/\cos(x)$ is continuous everywhere except where $\cos(x) = 0$. This happens when $x = \frac{\pi}{2} + n\pi$ for any integer n. If you cannot remember where $\tan(x)$ "blows up" or $\sin(x) = 0$ or $\cos(x) = 0$ then you should definitely revise trigonometric functions. Come to think of it — just revise them anyway.

25 If you do not know this fact then you should revise trigonometric functions. See the previous footnote.

- So the numerator is continuous, the denominator is continuous and nowhere zero, so the function is continuous everywhere.

If the function were changed to $\dfrac{\sin(x)}{x^2 - 5x + 6}$ much of the same reasoning can be used. Being a little terse we could answer with:

- Numerator and denominator are continuous.

- Since $x^2 - 5x + 6 = (x-2)(x-3)$ the denominator is zero when $x = 2, 3$.

- So the function is continuous everywhere except possibly at $x = 2, 3$. In order to verify that the function really is discontinuous at those points, it suffices to verify that the numerator is non-zero at $x = 2, 3$. Indeed we know that $\sin(x)$ is zero only when $x = n\pi$ (for any integer n). Hence $\sin(2), \sin(3) \neq 0$. Thus the numerator is non-zero, while the denominator is zero and hence $x = 2, 3$ really are points of discontinuity.

Note that this example raises a subtle point about checking continuity when numerator and denominator are *simultaneously* zero. There are quite a few possible outcomes in this case and we need more sophisticated tools to adequately analyse the behaviour of functions near such points. We will return to this question later in the text after we have developed Taylor expansions (see Section 3.4).

Example 1.6.9

So we know what happens when we add subtract multiply and divide, what about when we compose functions? Well - limits and compositions work nicely when things are continuous.

Theorem 1.6.10 (Compositions and continuity).

If f is continuous at b and $\lim\limits_{x \to a} g(x) = b$ then $\lim\limits_{x \to a} f(g(x)) = f(b)$. I.e.

$$\lim_{x \to a} f\left(g(x)\right) = f\left(\lim_{x \to a} g(x)\right)$$

Hence if g is continuous at a and f is continuous at $g(a)$ then the composite function $(f \circ g)(x) = f(g(x))$ is continuous at a.

So when we compose two continuous functions we get a new continuous function. We can put this to use

Example 1.6.11

Where are the following functions continuous?

$$f(x) = \sin\left(x^2 + \cos(x)\right)$$
$$g(x) = \sqrt{\sin(x)}$$

Our first step should be to break the functions down into pieces and study them. When we put them back together we should be careful of dividing by zero, or falling outside the domain.

- The function $f(x)$ is the composition of $\sin(x)$ with $x^2 + \cos(x)$.

- These pieces, $\sin(x), x^2, \cos(x)$ are continuous everywhere.

- So the sum $x^2 + \cos(x)$ is continuous everywhere

- And hence the composition of $\sin(x)$ and $x^2 + \cos(x)$ is continuous everywhere.

The second function is a little trickier.

- The function $g(x)$ is the composition of $\sqrt{x}$ with $\sin(x)$.

- $\sqrt{x}$ is continuous on its domain $x \geqslant 0$.

- $\sin(x)$ is continuous everywhere, but it is negative in many places.

- In order for $g(x)$ to be defined and continuous we must restrict x so that $\sin(x) \geqslant 0$.

- Recall the graph of $\sin(x)$:
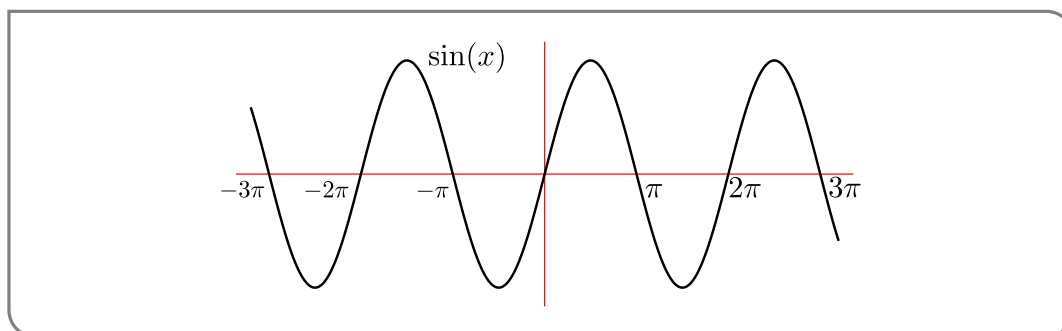
Hence $\sin(x) \geqslant 0$ when $x \in [0, \pi]$ or $x \in [2\pi, 3\pi]$ or $x \in [-2\pi, -\pi]$ or.... To be more precise $\sin(x)$ is positive when $x \in [2n\pi, (2n+1)\pi]$ for any integer n.

- Hence $g(x)$ is continuous when $x \in [2n\pi, (2n+1)\pi]$ for any integer n.

Example 1.6.11

Continuous functions are very nice (mathematically speaking). Functions from the "real world" tend to be continuous (though not always). The key aspect that makes them nice is the fact that they don't jump about.

The absence of such jumps leads to the following theorem which, while it can be quite confusing on first glance, actually says something very natural — obvious even. It says, roughly speaking, that, as you draw the graph $y = f(x)$ starting at $x = a$ and ending at $x = b$, y changes continuously from $y = f(a)$ to $y = f(b)$, with no jumps, and consequently y must take every value between $f(a)$ and $f(b)$ at least once. We'll start by just giving the precise statement and then we'll explain it in detail.

> **Theorem 1.6.12** (Intermediate value theorem (IVT)).
>
> Let $a < b$ and let f be a function that is continuous at all points $a \leqslant x \leqslant b$. If Y is any number between $f(a)$ and $f(b)$ then there exists some number $c \in [a, b]$ so that $f(c) = Y$.

Like the $\epsilon - \delta$ definition of limits[26], we should break this theorem down into pieces. Before we do that, keep the following pictures in mind.

Figure 1.6.1.
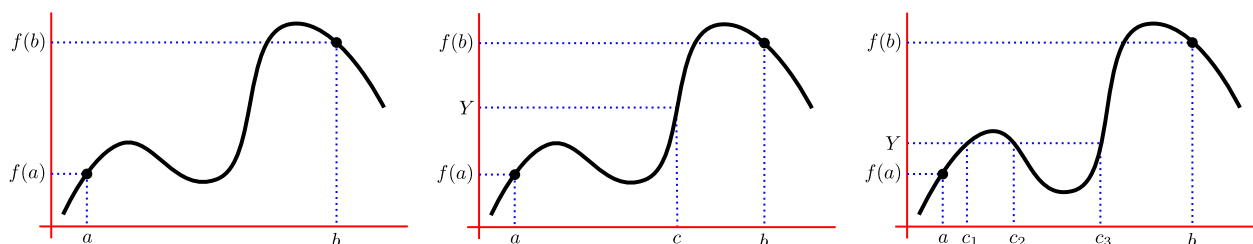
Now the break-down

- **Let $a < b$ and let f be a function that is continuous at all points $a \leqslant x \leqslant b$.** — This is setting the scene. We have a, b with $a < b$ (we can safely assume these to be real numbers). Our function must be continuous at all points between a and b.

- **if Y is any number between $f(a)$ and $f(b)$** — Now we need another number Y and the only restriction on it is that it lies between $f(a)$ and $f(b)$. That is, if $f(a) \leqslant f(b)$ then $f(a) \leqslant Y \leqslant f(b)$. Or if $f(a) \geqslant f(b)$ then $f(a) \geqslant Y \geqslant f(b)$. So notice that Y could be equal to $f(a)$ or $f(b)$ — if we wanted to avoid that possibility, then we would normally explicitly say $Y \neq f(a), f(b)$ or we would write that Y is *strictly* between $f(a)$ and $f(b)$.

- **there exists some number $c \in [a, b]$ so that $f(c) = Y$** — so if we satisfy all of the above conditions, then there has to be some real number c lying between a and b so that when we evaluate $f(c)$ it is Y.

So that breaks down the theorem statement by statement, but what does it actually mean?

- Draw any continuous function you like between a and b — it must be continuous.

- The function takes the value $f(a)$ at $x = a$ and $f(b)$ at $x = b$ — see the left-hand figure above.

---

26   The interested student is invited to take a look at the optional Section 1.7

- Now we can pick any Y that lies between $f(a)$ and $f(b)$ — see the middle figure above. The IVT[27] tells us that there must be some x-value that when plugged into the function gives us Y. That is, there is some c between a and b so that $f(c) = Y$. We can also interpret this graphically; the IVT tells us that the horizontal straight line $y = Y$ must intersect the graph $y = f(x)$ at some point (c, Y) with $a \leqslant c \leqslant b$.

- Notice that the IVT does not tell us how many such c-values there are, just that there is at least one of them. See the right-hand figure above. For that particular choice of Y there are three different c values so that $f(c_1) = f(c_2) = f(c_3) = Y$.

This theorem says that if $f(x)$ is a continuous function on all of the interval $a \leqslant x \leqslant b$ then as x moves from a to b, $f(x)$ takes every value between $f(a)$ and $f(b)$ at least once. To put this slightly differently, if f were to avoid a value between $f(a)$ and $f(b)$ then f cannot be continuous on $[a, b]$.

It is not hard to convince yourself that the continuity of f is crucial to the IVT. Without it one can quickly construct examples of functions that contradict the theorem. See the figure below for a few non-continuous examples:

Figure 1.6.2.
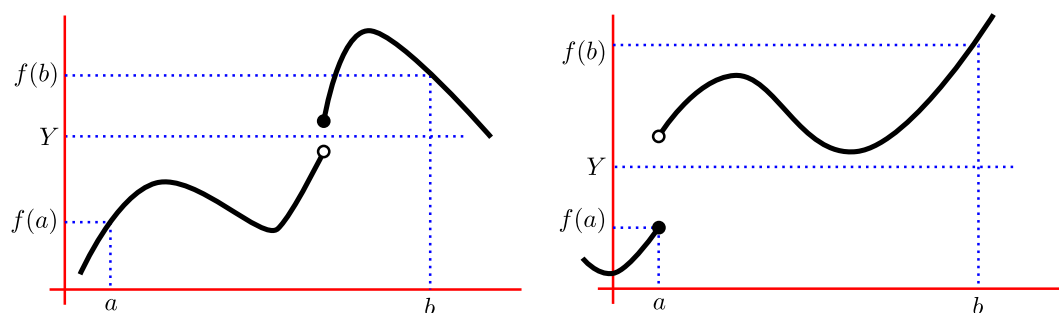
In the left-hand example we see that a discontinuous function can "jump" over the Y-value we have chosen, so there is no x-value that makes $f(x) = Y$. The right-hand example demonstrates why we need to be be careful with the ends of the interval. In particular, a function must be continuous over the whole interval $[a, b]$ *including* the endpoints of the interval. If we only required the function to be continuous on (a, b) (so strictly between a and b) then the function could "jump" over the Y-value at a or b.

If you are still confused then here is a "real-world" example

Example 1.6.13

You are climbing the Grouse-grind[28] with a friend — call him Bob. Bob was eager and

---

27 Often with big important useful theorems like this one, writing out the full name again and again becomes tedious, so we abbreviate it. Such abbreviations are okay provided the reader knows this is what you are doing, so the first time you use an abbreviation you should let the reader know. Much like we are doing here in this footnote: "IVT" stands for "intermediate value theorem", which is Theorem 1.6.12.

28 If you don't know it then google it.

started at 9am. Bob, while very eager, is also very clumsy; he sprained his ankle some-where along the path and has stopped moving at 9:21am and is just sitting[29] enjoying the view. You get there late and start climbing at 10am and being quite fit you get to the top at 11am. The IVT implies that at some time between 10am and 11am you meet up with Bob.

You can translate this situation into the form of the IVT as follows. Let t be time and let $a = 10$am and $b = 11$am. Let $g(t)$ be your distance along the trail. Hence[30] $g(a) = 0$ and $g(b) = 2.9km$. Since you are a mortal, your position along the trail is a continuous function — no helicopters or teleportation or... We have no idea where Bob is sitting, except that he is somewhere between $g(a)$ and $g(b)$, call this point Y. The IVT guarantees that there is some time c between a and b (so between 10am and 11am) with $g(c) = Y$ (and your position will be the same as Bob's).

Example 1.6.13

Aside from finding Bob sitting by the side of the trail, one of the most important ap-plications of the IVT is determining where a function is zero. For quadratics we know (or should know) that

$$ax^2 + bx + c = 0 \qquad\qquad \text{when } x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

While the Babylonians could (mostly, but not quite) do the above, the corresponding for-mula for solving a cubic is uglier and that for a quartic is uglier still. One of the most famous results in mathematics demonstrates that no such formula exists for quintics or higher degree polynomials[31].

So even for polynomials we cannot, in general, write down explicit formulae for their zeros and have to make do with numerical approximations — i.e. write down the root as a decimal expansion to whatever precision we desire. For more complicated functions we have no choice — there is no reason that the zeros should be expressible as nice neat little formulas. At the same time, finding the zeros of a function:

$$f(x) = 0$$

or solving equations of the form[32]

$$g(x) = h(x)$$

can be a crucial step in many mathematical proofs and applications.

---

29 Hopefully he remembered to carry something warm.
30 It's amazing what facts you can find on Wikipedia.
31 The similar (but uglier) formula for solving cubics took until the 15th century and the work of del Ferro and Cardano (and Cardano's student Ferrari). A similar (but even uglier) formula for quartics was also found by Ferrari. The extremely famous Abel-Ruffini Theorem (nearly by Ruffini in the late 18th century and completely by Abel in early 19th century) demonstrates that a similar formula for the zeros of a quintic does not exist. Note that the theorem does *not* say that quintics do not have zeros; rather it says that the zeros cannot in general be expressed using a finite combination of addition, multiplication, division, powers and roots. The interested student should also look up Évariste Galois and his contributions to this area.
32 In fact both of these are the same because we can write $f(x) = g(x) - h(x)$ and then the zeros of $f(x)$ are exactly when $g(x) = h(x)$.

For this reason there is a considerable body of mathematics which focuses just on finding the zeros of functions. The IVT provides a very simple way to "locate" the zeros of a function. In particular, if we know a continuous function is negative at a point $x = a$ and positive at another point $x = b$, then there must (by the IVT) be a point $x = c$ between a and b where $f(c) = 0$.

Figure 1.6.3.
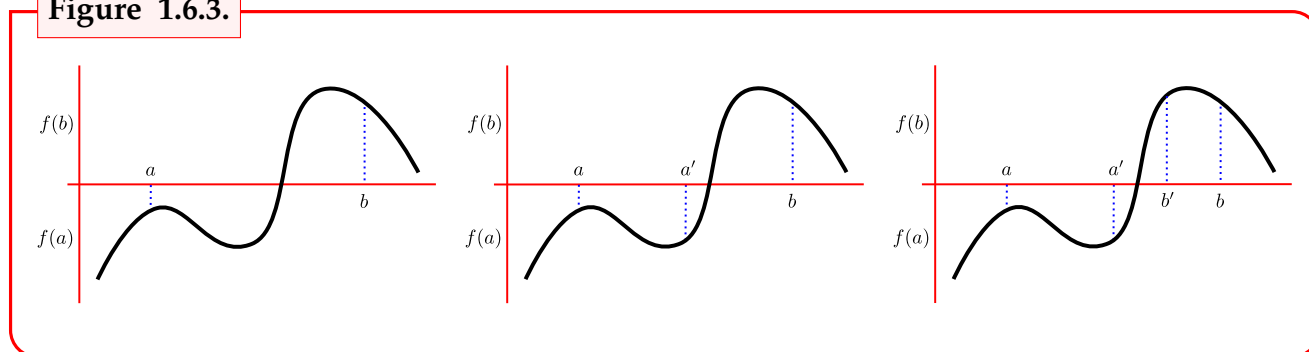
Consider the leftmost of the above figures. It depicts a continuous function that is negative at $x = a$ and positive at $x = b$. So choose $Y = 0$ and apply the IVT — there must be some c with $a \leqslant c \leqslant b$ so that $f(c) = Y = 0$. While this doesn't tell us c exactly, it does give us bounds on the possible positions of at least one zero — there must be at least one c obeying $a \leqslant c \leqslant b$.

See middle figure. To get better bounds we could test a point half-way between a and b. So set $a' = \frac{a+b}{2}$. In this example we see that $f(a')$ is negative. Applying the IVT again tells us there is some c between a' and b so that $f(c) = 0$. Again — we don't have c exactly, but we have halved the range of values it could take.

Look at the rightmost figure and do it again — test the point half-way between a' and b. In this example we see that $f(b')$ is positive. Applying the IVT tells us that there is some c between a' and b' so that $f(c) = 0$. This new range is a quarter of the length of the original. If we keep doing this process the range will halve each time until we know that the zero is inside some tiny range of possible values. This process is called the bisection method.

Consider the following zero-finding example

Example 1.6.14

Show that the function $f(x) = x - 1 + \sin(\pi x/2)$ has a zero in $0 \leqslant x \leqslant 1$.

This question has been set up nicely to lead us towards using the IVT; we are already given a nice interval on which to look. In general we might have to test a few points and experiment a bit with a calculator before we can start narrowing down a range.

Let us start by testing the endpoints of the interval we are given

$$f(0) = 0 - 1 + \sin(0) = -1 < 0$$
$$f(1) = 1 - 1 + \sin(\pi/2) = 1 > 0$$

So we know a point where f is positive and one where it is negative. So by the IVT there is a point in between where it is zero.

BUT in order to apply the IVT we have to show that the function is continuous, and we cannot simply write

it is continuous

We need to explain to the reader *why* it is continuous. That is — we have to prove it.

So to write up our answer we can put something like the following — keeping in mind we need to tell the reader what we are doing so they can follow along easily.

- We will use the IVT to prove that there is a zero in $[0, 1]$.

- First we must show that the function is continuous.

    - Since $x - 1$ is a polynomial it is continuous everywhere.
    - The function $\sin(\pi x/2)$ is a trigonometric function and is also continuous everywhere.
    - The sum of two continuous functions is also continuous, so $f(x)$ is continuous everywhere.

- Let $a = 0, b = 1$, then

$$f(0) = 0 - 1 + \sin(0) = -1 < 0$$
$$f(1) = 1 - 1 + \sin(\pi/2) = 1 > 0$$

- The function is negative at $x = 0$ and positive at $x = 1$. Since the function is continuous we know there is a point $c \in [0, 1]$ so that $f(c) = 0$.

Notice that though we have not used full sentences in our explanation here, we are still using words. Your mathematics, unless it is very straight-forward computation, should contain words as well as symbols.

Example 1.6.14

The zero is actually located at about $x = 0.4053883559$.

The bisection method is really just the idea that we can keep repeating the above reasoning (with a calculator handy). Each iteration will tell us the location of the zero more precisely. The following example illustrates this.

Example 1.6.15

Use the bisection method to find a zero of

$$f(x) = x - 1 + \sin(\pi x/2)$$

that lies between 0 and 1.

So we start with the two points we worked out above:

- $a = 0, b = 1$ and

$$f(0) = -1$$
$$f(1) = 1$$

- Test the point in the middle $x = \frac{0+1}{2} = 0.5$

$$f(0.5) = 0.2071067813 > 0$$

- So our new interval will be $[0, 0.5]$ since the function is negative at $x = 0$ and positive at $x = 0.5$

Repeat

- $a = 0, b = 0.5$ where $f(0) < 0$ and $f(0.5) > 0$.

- Test the point in the middle $x = \frac{0+0.5}{2} = 0.25$

$$f(0.25) = -0.3673165675 < 0$$

- So our new interval will be $[0.25, 0.5]$ since the function is negative at $x = 0.25$ and positive at $x = 0.5$

Repeat

- $a = 0.25, b = 0.5$ where $f(0.25) < 0$ and $f(0.5) > 0$.

- Test the point in the middle $x = \frac{0.25+0.5}{2} = 0.375$

$$f(0.375) = -0.0694297669 < 0$$

- So our new interval will be $[0.375, 0.5]$ since the function is negative at $x = 0.375$ and positive at $x = 0.5$

Below is an illustration of what we have observed so far together with a plot of the actual function.
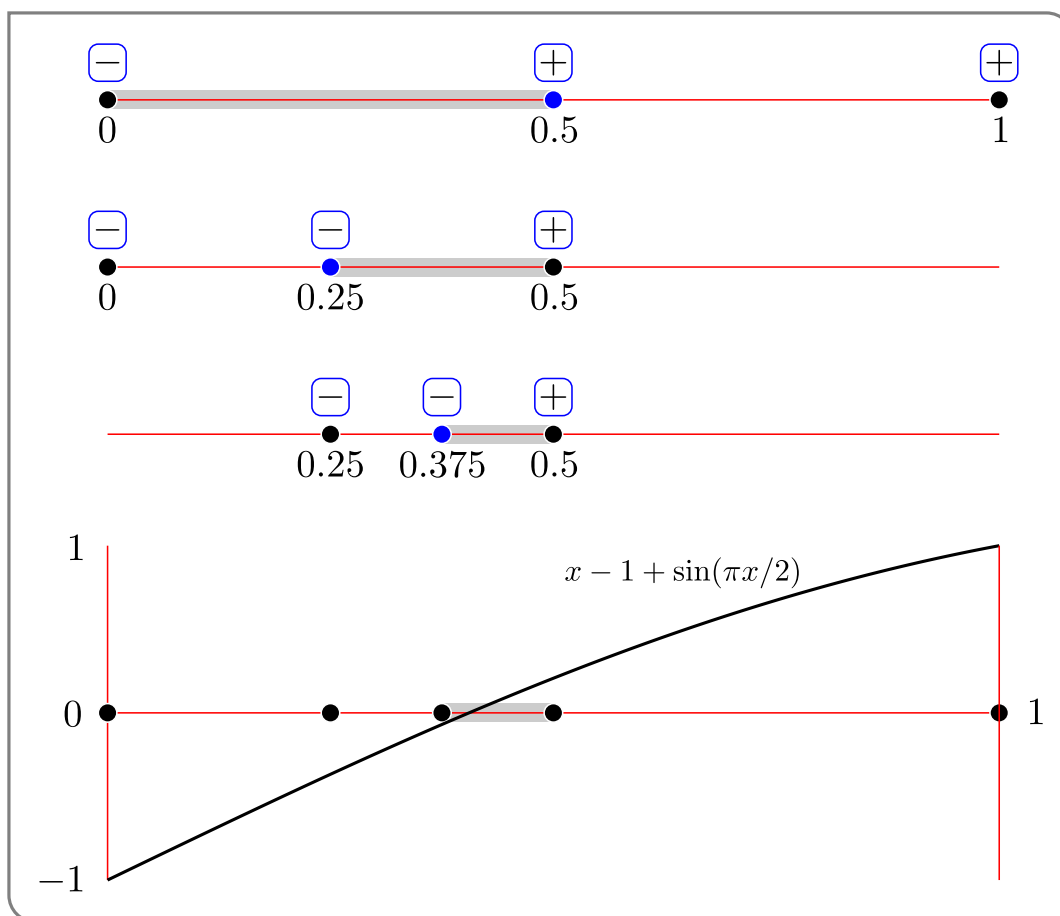
And one final iteration:

- $a = 0.375, b = 0.5$ where $f(0.375) < 0$ and $f(0.5) > 0$.

- Test the point in the middle $x = \frac{0.375+0.5}{2} = 0.4375$

$$f(0.4375) = 0.0718932843 > 0$$

- So our new interval will be $[0.375, 0.4375]$ since the function is negative at $x = 0.375$ and positive at $x = 0.4375$

So without much work we know the location of a zero inside a range of length $0.0625 = 2^{-4}$. Each iteration will halve the length of the range and we keep going until we reach the precision we need, though it is much easier to program a computer to do it.

Example 1.6.15

1.7 ▲ (Optional) — Making the Informal a Little More Formal

As we noted above, the definition of limits that we have been working with was quite informal and not mathematically rigorous. In this (optional) section we will work to understand the rigorous definition of limits.

Here is the formal definition — we will work through it all very slowly and carefully afterwards, so do not panic.

Definition 1.7.1.

Let $a \in \mathbb{R}$ and let $f(x)$ be a function defined everywhere in a neighbourhood of a, except possibly at a. We say that

the limit as x approaches a of $f(x)$ is L

or equivalently

as x approaches a, $f(x)$ approaches L

and write

$$\lim_{x \to a} f(x) = L$$

if and only if for every $\epsilon > 0$ there exists $\delta > 0$ so that

$$|f(x) - L| < \epsilon \text{ whenever } 0 < |x - a| < \delta$$

Note that an equivalent way of writing this very last statement is

$$\text{if } 0 < |x - a| < \delta \text{ then } |f(x) - L| < \epsilon.$$

This is quite a lot to take in, so let us break it down into pieces.

> **Definition 1.7.2** (The typical 3 pieces of a definition).
>
> Usually a definition can be broken down into three pieces.
>
> - Scene setting — define symbols and any restrictions on the objects that we are talking about.
>
> - Naming — state the name and any notation for the property or object that the definition is about.
>
> - Properties and restrictions — this is the heart of the definition where we explain to the reader what it is that the object (in our case a function) has to do in order to satisfy the definition.

Let us go back to the definition and look at each of these pieces in turn.

- Setting things up — The first sentence of the definition is really just setting up the picture. It is telling us what the definition is about and sorting out a few technical details.

    - **Let $a \in \mathbb{R}$** — This simply tells us that the symbol "a" is a real number[33].

    - **Let $f(x)$ be a function** — This is just setting the scene so that we understand all of the terms and symbols.

    - **defined everywhere in a neighbourhood of a, except possibly at a** — This is just a technical requirement; we need our function to be defined in a little region[34] around a. The function doesn't have to be defined everywhere, but it must be defined for all x-values a little less than a and a little more than a. The definition does not care about what the function does outside this little window, nor does it care what happens exactly at a.

- Names, phrases and notation — The next part of the definition is simply naming the property we are discussing and tells us how to write it down. i.e. we are talking about "limits" and we write them down using the symbols indicated.

- The heart of things — we explain this at length below, but for now we will give a quick explanation. **Work on these two points. They are hard.**

    - **for all $\epsilon > 0$ there exists $\delta > 0$** — It is important we read this in order. It means that we can pick any positive number ϵ we want and there will always be another positive number δ that is going to make what ever follows be true.

---

33 The symbol "$\in$" is read as "is an element of" — it is definitely not the same as e or ϵ or ε. If you do not recognise "$\mathbb{R}$" or understand the difference between $\mathbb{R}$ and R, then please go back and read Chapter 0 carefully.

34 The term "neighbourhood of a" means a small open interval around a — for example $(a - 0.01, a + 0.01)$. Typically we don't really care how big this little interval is.

- **if** $0 < |x - a| < \delta$ **then** $|f(x) - L| < \epsilon$ — From the previous point we have our two numbers — any $\epsilon > 0$ then based on that choice of ϵ we have a positive number δ. The current statement says that whenever we have chosen x so that it is very close to a, then $f(x)$ has to be very close to L. How close is "very close"? Well $0 < |x - a| < \delta$ means that x has to be within a distance δ of a (but not exactly a) and similarly $|f(x) - L| < \epsilon$ means that $f(x)$ has to be within a distance ϵ of L.

That is the definition broken up into pieces which hopefully now make more sense, but what does it actually *mean*? Consider a function we saw earlier

$$f(x) = \begin{cases} 2x & x \neq 3 \\ 9 & x = 3 \end{cases}$$

and sketch it again:

Figure 1.7.1.
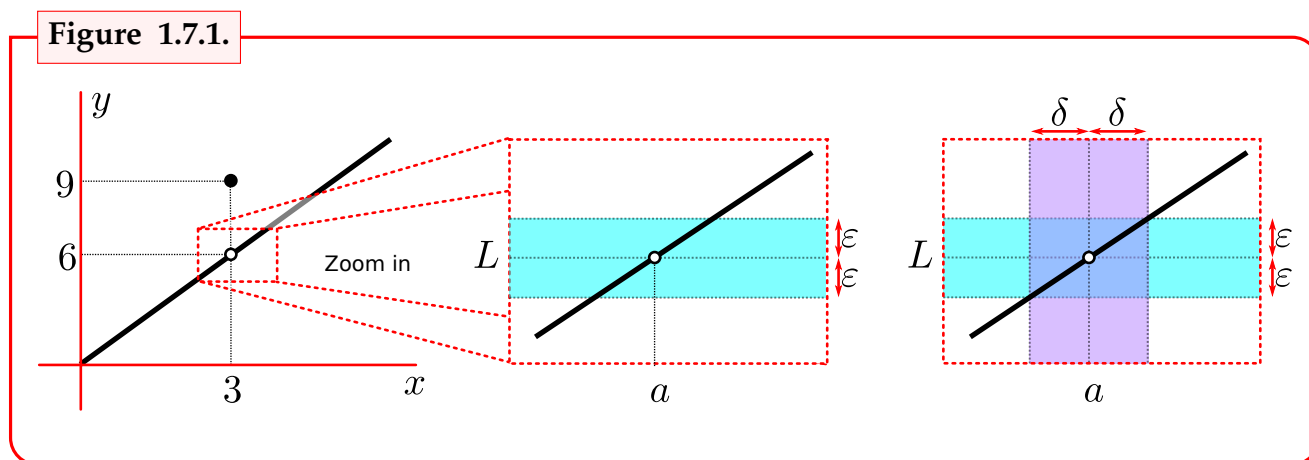
We know (from our earlier work) that $\lim_{x \to 3} f(x) = 6$, so zoom in around $(x, y) = (3, 6)$. To make this look more like our definition, we have $a = 3$ and $L = 6$.

- Pick some small number $\epsilon > 0$ and highlight the horizontal strip of all points (x, y) for which $|y - L| < \epsilon$. This means all the y-values have to satisfy $L - \epsilon < y < L + \epsilon$.

- You can see that the graph of the function passes through this strip for some x-values close to a. What we need to be able to do is to pick a vertical strip of x-values around a so that the function lies inside the horizontal strip.

- That is, we must find a small number $\delta > 0$ so that for any x-value inside the vertical strip $a - \delta < x < a + \delta$, *except exactly at* $x = a$, the value of the function lies inside the horizontal strip, namely $L - \epsilon < y = f(x) < L + \epsilon$.

- We see (pictorially) that we can do this. If we were to choose a smaller value of ϵ making the horizontal strip narrower, it is clear that we can choose the vertical strip to be narrower. Indeed, it doesn't matter how small we make the horizontal strip, we will always be able to construct the second vertical strip.

The above is a pictorial argument, but we can quite easily make it into a mathematical one. We want to show the limit is 6. That means for any ϵ we need to find a δ so that when

$$3 - \delta < x < 3 + \delta \text{ with } x \neq 3 \qquad \text{we have} \qquad 6 - \epsilon < f(x) < 6 + \epsilon$$

Now we note that when $x \neq 3$, we have $f(x) = 2x$ and so

$$6 - \epsilon < f(x) < 6 + \epsilon \qquad \text{implies that} \qquad 6 - \epsilon < 2x < 6 + \epsilon$$

this nearly specifies a range of x values, we just need to divide by 2

$$3 - \epsilon/2 < x < 3 + \epsilon/2$$

Hence if we choose $\delta = \epsilon/2$ then we get the desired inequality

$$3 - \delta < x < 3 + \delta$$

i.e. — no matter what $\epsilon > 0$ is chosen, if we put $\delta = \epsilon/2$ then when $3 - \delta < x < 3 + \delta$ with $x \neq 3$ we will have $6 - \epsilon < f(x) < 6 + \epsilon$. This is exactly what we need to satisfy the definition of "limit" above.

The above work gives us the argument we need, but it still needs to be written up properly. We do this below.

Example 1.7.3

Find the limit as $x \to 3$ of the following function

$$f(x) = \begin{cases} 2x & x \neq 3 \\ 9 & x = 3 \end{cases}$$

Proof. We will show that the limit is equal to 6. Let $\epsilon > 0$ and $\delta = \epsilon/2$. It remains to show that $|f(x) - 6| < \epsilon$ whenever $|x - 3| < \delta$.

So assume that $|x - 3| < \delta$, and so

$$3 - \delta < x < 3 + \delta \qquad \text{multiply both sides by 2}$$
$$6 - 2\delta < 2x < 6 + 2\delta$$

Recall that $f(x) = 2x$ and that since $\delta = \epsilon/2$

$$6 - \epsilon < f(x) < 6 + \epsilon.$$

We can conclude that $|f(x) - 6| < \epsilon$ as required. $\qquad\qquad\qquad\qquad\qquad$ □

Example 1.7.3

Because of the ϵ and δ in the definition of limits, we need to have ϵ and δ in the proof. While ϵ and δ are just symbols playing particular roles, and could be replaced with other symbols, this style of proof is usually called ϵ–δ proof.

In the above example everything works, but it can be very instructive to see what happens in an example that doesn't work.

Example 1.7.4

Look again at the function

$$f(x) = \begin{cases} x & x < 2 \\ -1 & x = 2 \\ x + 3 & x > 2 \end{cases}$$

and let us see why, according to the definition of the limit, that $\lim_{x \to 2} f(x) \neq 2$. Again, start by sketching a picture and zooming in around $(x, y) = (2, 2)$:
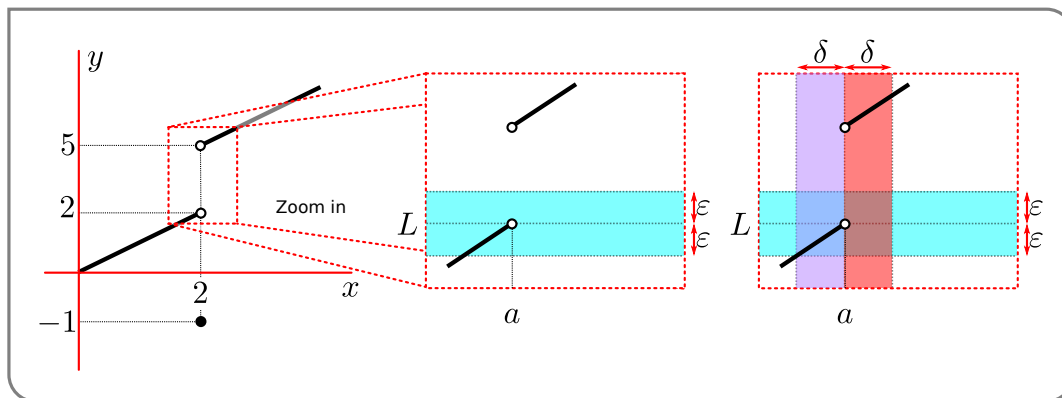
Try to proceed through the same steps as before:

- Pick some small number $\epsilon > 0$ and highlight a horizontal strip that contains all y-values with $|y - L| < \epsilon$. This means all the y-values have to satisfy $L - \epsilon < y < L + \epsilon$.

- You can see that the graph of the function passes through this strip for some x-values close to a. To the left of a, we can always find some x-values that make the function sit inside the horizontal-ϵ-strip. However, unlike the previous example, there is a problem to the right of a. Even for x-values just a little larger than a, the value of $f(x)$ lies well outside the horizontal-ϵ-strip.

- So given this choice of ϵ, we can find a $\delta > 0$ so that for x inside the vertical strip $a - \delta < x < a$, the value of the function sits inside the horizontal-ϵ-strip.

- Unfortunately, there is no way to choose a $\delta > 0$ so that for x inside the vertical strip $a < x < a + \delta$ (with $x \neq a$) the value of the function sits inside the horizontal-*epsilon*-strip.

- So it is impossible to choose δ so that for x inside the vertical strip $a - \delta < x < a + \delta$ the value of the function sits inside the horizontal strip $L - \epsilon < y = f(x) < L + \epsilon$.

- Thus the limit of $f(x)$ as $x \to 2$ is not 2.

Example 1.7.4

Doing things formally with ϵ's and δ's is quite painful for general functions. It is far better to make use of the arithmetic of limits (Theorem 1.4.2) and some basic building

blocks (like those in Theorem 1.4.1). Thankfully for most of the problems we deal with in calculus (at this level at least) can be approached in exactly this way.

This does leave the problem of proving the arithmetic of limits and the limits of the basic building blocks. The proof of the Theorem 1.4.2 is quite involved and we leave it to the very end of this Chapter. Before we do that we will prove Theorem 1.4.1 by a formal ϵ–δ proof. Then in the next section we will look at the formal definition of limits at infinity and prove Theorem 1.5.3. The proof of the Theorem 1.5.9, the arithmetic of infinite limits, is very similar to that of Theorem 1.4.2 and so we do not give it.

So let us now prove Theorem 1.4.1 in which we stated two simple limits:

$$\lim_{x \to a} c = c \qquad \text{and} \qquad \lim_{x \to a} x = a.$$

Here is the formal ϵ–δ proof:

Proof of Theorem 1.4.1. Since there are two limits to prove, we do each in turn. Let a, c be real numbers.

- Let $\epsilon > 0$ and set $f(x) = c$. Choose $\delta = 1$, then for any x satisfying $|x - a| < \delta$ (or indeed any real number x at all) we have $|f(x) - c| = 0 < \epsilon$. Hence $\lim_{x \to a} c = c$ as required.

- Let $\epsilon > 0$ and set $f(x) = x$. Choose $\delta = \epsilon$, then for any x satisfying $|x - a| < \delta$ we have

$$a - \delta < x < a + \delta \text{ but } f(x) = x \text{ and } \delta = \epsilon \text{ so}$$
$$a - \epsilon < f(x) < a + \epsilon$$

Thus we have $|f(x) - a| < \epsilon$. Hence $\lim_{x \to a} x = a$ as required.

This completes the proof.   □

1.8 ▲ (Optional) — Making Infinite Limits a Little More Formal

For those of you who made it through the formal $\epsilon - \delta$ definition of limits we give the formal definition of limits involving infinity:

> **Definition 1.8.1** (Limits involving infinity — formal).
>
> (a) Let f be a function defined on the whole real line. We say that
>
> > the limit as x approaches ∞ of $f(x)$ is L
>
> or equivalently
>
> > $f(x)$ converges to L as x goes to ∞
>
> and write
>
> $$\lim_{x \to \infty} f(x) = L$$
>
> if and only if for every $\epsilon > 0$ there exists $M \in \mathbb{R}$ so that $|f(x) - L| < \epsilon$ whenever $x > M$.
>
> Similarly we write
>
> $$\lim_{x \to -\infty} f(x) = K$$
>
> if and only if for every $\epsilon > 0$ there exists $N \in \mathbb{R}$ so that $|f(x) - K| < \epsilon$ whenever $x < N$.
>
> (b) Let a be a real number and $f(x)$ be a function defined for all $x \neq a$. We write
>
> $$\lim_{x \to a} f(x) = \infty$$
>
> if and only if for every $P > 0$ there exists $\delta > 0$ so that $f(x) > P$ whenever $0 < |x - a| < \delta$.
>
> (c) Let f be a function defined on the whole real line. We write
>
> $$\lim_{x \to \infty} f(x) = \infty$$
>
> if and only if for every $P > 0$ there exists $M > 0$ so that $f(x) > P$ whenever $x > M$.

Note that we can loosen the above requirements on the domain of definition of f — for example, in part (a) all we actually require is that $f(x)$ be defined for all x larger than some value. It would be sufficient to require "there is some $x_0 \in \mathbb{R}$ so that $f(x)$ is defined for all $x > x_0$". Also note that there are obvious variations of parts (b) and (c) with ∞ replaced by $-\infty$.

For completeness let's prove Theorem 1.5.3 using this formal definition. The layout of the proof will be very similar to our proof of Theorem 1.4.1.

Proof of Theorem 1.5.3. There are four limits to prove in total and we do each in turn. Let $c \in \mathbb{R}$.

- Let $\epsilon > 0$ and set $f(x) = c$. Choose $M = 0$, then for any x satisfying $x > M$ (or indeed any real number x at all) we have $|f(x) - c| = 0 < \epsilon$. Hence $\lim\limits_{x\to\infty} c = c$ as required.

- The proof that $\lim\limits_{x\to-\infty} c = c$ is nearly identical. Again, let $\epsilon > 0$ and set $f(x) = c$. Choose $N = 0$, then for any x satisfying $x < N$ we have $|f(x) - c| = 0 < \epsilon$. Hence $\lim\limits_{x\to-\infty} c = c$ as required.

- Let $\epsilon > 0$ and set $f(x) = x$. Choose $M = \frac{1}{\epsilon}$. Then when $x > M$ we have

$$0 < M < x \qquad\qquad\qquad \text{divide through by } xM \text{ to get}$$
$$0 < \frac{1}{x} < \frac{1}{M} = \epsilon$$

  Since $x > 0$, $1/x = |1/x| = |1/x - 0| < \epsilon$ as required.

- Again, the proof in the limit to $-\infty$ is similar but we have to be careful of signs. Let $\epsilon > 0$ and set $f(x) = x$. Choose $N = -\frac{1}{\epsilon}$. Then when $x < N$ we have

$$0 > N > x \qquad\qquad\qquad \text{divide through by } xN \text{ to get}$$
$$0 > \frac{1}{x} > \frac{1}{N} = -\epsilon$$

  Notice that by assumption both $x, N < 0$, so $xN > 0$. Now since $x < 0$, $1/x = -|1/x| = |1/x - 0| < \epsilon$ as required.

This completes the proof.

$\square$

1.9 ▲ (Optional) — Proving the Arithmetic of Limits

Perhaps the most useful theorem of this chapter is Theorem 1.4.2 which shows how limits interact with arithmetic. In this (optional) section we will prove both the arithmetic of limits Theorem 1.4.2 and the Squeeze Theorem 1.4.17. Before we get to the proofs it is very helpful to prove three technical lemmas that we'll need. The first is a very general result about absolute values of numbers:

> **Lemma 1.9.1** (The triangle inequality).
>
> For any $x, y \in \mathbb{R}$
>
> $$|x + y| \leqslant |x| + |y|$$

Proof. Notice that for any real number x, we always have $-x, x \leqslant |x|$ and either $|x| = x$ or $|x| = -x$. So now let $x, y \in \mathbb{R}$. Then we must have either

$$|x + y| = x + y \quad \leqslant |x| + |y|$$

or

$$|x + y| = -x - y \leqslant |x| + |y|$$

In both cases we end up with $|x + y| \leqslant |x| + |y|$. □

The second lemma is more specialised. It proves that if we have a function $f(x) \to F$ as $x \to a$ then there must be a small window around $x = a$ where the function $f(x)$ must only take values not far from F. In particular it tells us that $|f(x)|$ cannot be bigger than $|F| + 1$ when x is very close to a.

Lemma 1.9.2.

Let $a \in \mathbb{R}$ and let f be a function so that $\lim\limits_{x \to a} f(x) = F$. Then there exists a $\delta > 0$ so that if $0 < |x - a| < \delta$ then we also have $|f(x)| \leqslant |F| + 1$.

The proof is mostly just manipulating the ϵ–δ definition of a limit with $\epsilon = 1$.

Proof. Let $\epsilon = 1$. Then since $f(x) \to F$ as $x \to a$, there exists a $\delta > 0$ so that when $0 < |x - a| < \delta$, we also have $|f(x) - F| \leqslant \epsilon = 1$. So now assume $0 < |x - a| < \delta$. Then

$$-\epsilon \leqslant f(x) - F \leqslant \epsilon \qquad \text{rearrange a little}$$
$$-\epsilon + F \leqslant f(x) \leqslant \epsilon + F$$

Now $\epsilon + F \leqslant \epsilon + |F|$ and $-\epsilon + F \geqslant -\epsilon - |F|$, so

$$-\epsilon - |F| \leqslant f(x) \leqslant \epsilon + |F|$$

Hence we have $|f(x)| \leqslant \epsilon + |F| = |F| + 1$. □

Finally our third technical lemma gives us a bound in the other direction; it tells us that when x is close to a, the value of $|f(x)|$ cannot be much smaller than $|F|$.

Lemma 1.9.3.

Let $a \in \mathbb{R}$ and $F \neq 0$ and let f be a function so that $\lim\limits_{x \to a} f(x) = F$. Then there exists $\delta > 0$ so that when $0 < |x - a| < \delta$, we have $|f(x)| > |F|/2$.

Proof. Set $\epsilon = |F|/2 > 0$. Then since $f(x) \to F$, we know there exists a $\delta > 0$ so that when $0 < |x - a| < \delta$ we have $|f(x) - F| < \epsilon$. So now assume $0 < |x - a| < \delta$ so that $|f(x) - F| < \epsilon = |F|/2$. Then

$$|F| = |F - f(x) + f(x)| \qquad \text{sneaky trick}$$
$$\leqslant |f(x) - F| + |f(x)| \qquad \text{but } |f(x) - F| < \epsilon$$
$$< \epsilon + |f(x)|$$

Hence $|f(x)| > |F| - \epsilon = |F|/2$ as required. □

Now we are in a position to prove Theorem 1.4.2. The proof has more steps than the previous $\epsilon - \delta$ proofs we have seen. This is mostly because we do not have specific functions $f(x)$ and $g(x)$ and instead must play with them in the abstract — and make good use of the formal definition of limits.

We will break the proof into three pieces. The minimum that is required is to prove that

$$\lim_{x \to a} (f(x) + g(x)) = F + G$$

$$\lim_{x \to a} f(x) \cdot g(x) = F \cdot G$$

$$\lim_{x \to a} 1/g(x) = 1/G \quad \text{if } G \neq 0.$$

From these three we can prove that

$$\lim_{x \to a} f(x) \cdot c = F \cdot c$$

$$\lim_{x \to a} (f(x) - g(x)) = F - G$$

$$\lim_{x \to a} f(x)/g(x) = F/G \quad \text{if } G \neq 0.$$

The first follows by setting $g(x) = c$ and using $\lim f(x) \cdot g(x)$. The second follows by setting $c = -1$, putting $h(x) = (-1) \cdot g(x)$ and then applying both $\lim f(x) \cdot g(x)$ and $\lim f(x) + g(x)$. The third follows by setting $h(x) = 1/g(x)$ and then using $\lim f(x) \cdot h(x)$.

Starting with addition, in order to satisfy the definition of limit, we are going to have to show that

$$|(f(x) + g(x)) - (F + G)| \text{ is small}$$

when we know that $|f(x) - F|, |g(x) - G|$ are small. To do this we use the triangle inequality above showing that

$$|(f(x) + g(x)) - (F + G)| = |(f(x) - F) + (g(x) - G)| \leqslant |f(x) - F| + |g(x) - G|$$

This is the key technical piece of the proof. So if we want the LHS of the above to be size ϵ, we need to make sure that each term on the RHS is of size $\epsilon/2$. The rest of the proof is setting up facts based on the definition of limits and then rearranging facts to reach the conclusion.

Proof. Proof of Theorem 1.4.2 — limit of a sum. Let $a \in \mathbb{R}$ and assume that

$$\lim_{x \to a} f(x) = F \qquad \text{and} \qquad \lim_{x \to a} g(x) = G.$$

We wish to show that

$$\lim_{x \to a} f(x) + g(x) = F + G.$$

Let $\epsilon > 0$ — we have to find a $\delta > 0$ so that when $|x - a| < \delta$ we have $|(f(x) + g(x)) - (F + G)| < \epsilon$.

Let $\epsilon > 0$ and set $\epsilon_1 = \epsilon_2 = \epsilon/2$. By the definition of limits, because $f(x) \to F$ there exists some $\delta_1 > 0$ so that whenever $|x - a| < \delta_1$, we also have $|f(x) - F| < \epsilon_1$. Similarly

there exists $\delta_2 > 0$ so that if $|x - a| < \delta_2$, then we must have $|g(x) - G| < \epsilon_2$. So now choose $\delta = \min\{\delta_1, \delta_2\}$ and assume $|x - a| < \delta$. Then we must have that $|x - a| < \delta_1, \delta_2$ and so we also have

$$|f(x) - F| < \epsilon_1 \qquad\qquad |g(x) - G| < \epsilon_2$$

Now consider $|(f(x) + g(x)) - (F + G)|$ and rearrange the terms:

$$
\begin{aligned}
|(f(x) + g(x)) - (F + G)| &= |(f(x) - F) + (g(x) - G)| &&\text{now apply triangle inequality} \\
&\leqslant |f(x) - F| + |g(x) - G| &&\text{use facts from above} \\
&< \epsilon_1 + \epsilon_2 \\
&= \epsilon.
\end{aligned}
$$

Hence we have shown that for any $\epsilon > 0$ there exists some $\delta > 0$ so that when $|x - a| < \delta$ we also have $|(f(x) + g(x)) - (F + G)| < \epsilon$. Which is exactly the formal definition of the limit we needed to prove. $\qquad\square$

Let us do similarly for the limit of a product. Some of the details of the proof are very similar, but there is a little technical trick in the middle to make it work. In particular we need to show that

$$|f(x) \cdot g(x) - F \cdot G| \text{ is small}$$

when we know that $|f(x) - F|$ and $|g(x) - G|$ are both small. Notice that

$$
\begin{aligned}
f(x) \cdot g(x) - F \cdot G &= f(x) \cdot g(x) - F \cdot G + \underbrace{f(x) \cdot G - f(x) \cdot G}_{=0} \\
&= f(x) \cdot g(x) - f(x) \cdot G + f(x) \cdot G - F \cdot G \\
&= f(x) \cdot (g(x) - G) + (f(x) - F) \cdot G
\end{aligned}
$$

So if we know $|f(x) - F|$ is small and $|g(x) - G|$ is small then we are done — except that we also need to know that $f(x)$ doesn't become really large near a — this is exactly why we needed to prove Lemma 1.9.2.

As was the case in the previous proof, we want the LHS to be of size at most ϵ, so we want, for example, the two terms on the RHS to be of size at most $\epsilon/2$. This means

- we need $|G| \cdot |f(x) - F|$ to be of size at most $\epsilon/2$, and

- we need $|g(x) - G|$ to be of size at most $\epsilon/2(|F|+1)$ since we know that $|f(x)| \leqslant |F| + 1$ when x is close to a.

Armed with these tricks we turn to the proofs.

Proof. Proof of Theorem 1.4.2 — limit of a product. Let $a \in \mathbb{R}$ and assume that

$$\lim_{x \to a} f(x) = F \qquad\qquad \text{and} \qquad\qquad \lim_{x \to a} g(x) = G.$$

We wish to show that

$$\lim_{x \to a} f(x) \cdot g(x) = F \cdot G.$$

Let $\epsilon > 0$. Set $\epsilon_1 = \frac{\epsilon}{2(|G|+1)}$ (the extra $+1$ in the denominator is just there to make sure that ϵ_1 is well–defined even if $G = 0$), and $\epsilon_2 = \frac{\epsilon}{2(|F|+1)}$. From this we establish the existence of $\delta_1, \delta_2, \delta_3$ which we need below.

- By assumption $f(x) \to F$ so there exists $\delta_1 > 0$ so that whenever $|x - a| < \delta_1$, we also have $|f(x) - F| < \epsilon_1$.

- Similarly because $g(x) \to G$, there exists $\delta_2 > 0$ so that whenever $|x - a| < \delta_2$, we also have $|g(x) - G| < \epsilon_2$.

- By Lemma 1.9.2 there exists $\delta_3 > 0$ so that whenever $|x - a| < \delta_3$, we also have $|f(x)| \leqslant |F| + 1$.

Let $\delta = \min\{\delta_1, \delta_2, \delta_3\}$, assume $|x - a| < \delta$ and consider $|f(x) \cdot g(x) - F \cdot G|$. Rearrange the terms as we did above:

$$|f(x) \cdot g(x) - F \cdot G| = |f(x) \cdot (g(x) - G) + (f(x) - F) \cdot G|$$
$$\leqslant |f(x)| \cdot |g(x) - G| + |G| \cdot |f(x) - F|$$

By our three dot-points above we know that $|f(x) - F| < \epsilon_1$ and $|g(x) - G| < \epsilon_2$ and $|f(x)| \leqslant |F| + 1$, so we have

$$|f(x) \cdot g(x) - F \cdot G| < |f(x)| \cdot \epsilon_2 + |G| \cdot \epsilon_1 \qquad \text{sub in } \epsilon_1, \epsilon_2 \text{ and bound on } f(x)$$
$$< (|F| + 1) \cdot \frac{\epsilon}{2(|F| + 1)} + |G| \cdot \frac{\epsilon}{2(|G| + 1)}$$
$$\leqslant \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Thus we have shown that for any $\epsilon > 0$ there exists $\delta > 0$ so that when $|x - a| < \delta$ we also have $|f(x) \cdot g(x) - F \cdot G| < \epsilon$. Hence $f(x) \cdot g(x) \to F \cdot G$. $\qquad \square$

Finally we can prove the limit of a reciprocal. Notice that

$$\frac{1}{g(x)} - \frac{1}{G} = \frac{G - g(x)}{g(x) \cdot G}$$

We need to show the LHS is of size at most ϵ when x is close enough to a, so if $G - g(x)$ is small we are done — except if $g(x)$ or G are close to zero. By assumption (go back and read Theorem 1.4.2) we have $G \neq 0$, and we know from Lemma 1.9.3 that $|g(x)|$ cannot be smaller than $|G|/2$. Together these imply that the denominator on the RHS cannot be zero and indeed must be of magnitude at least $|G|^2/2$. Thus we need $|G - g(x)|$ to be of size at most $\epsilon \cdot |G|^2/2$.

Proof. Proof of Theorem 1.4.2 — limit of a reciprocal. Let $\epsilon > 0$ and set $\epsilon_1 = \epsilon |G|^2 \cdot \frac{1}{2}$. We now use this and Lemma 1.9.3 to establish the existence of δ_1, δ_2.

- Since $g(x) \to G$ we know that there exists $\delta_1 > 0$ so that when $|x - a| < \delta_1$ we also have $|g(x) - G| < \epsilon_1$.

- By Lemma 1.9.3 there exists δ_2 so that when $|x - a| < \delta_2$ we also have $|g(x)| > |G|/2$. Equivalently, when $|x - a| < \delta_2$ we also have $\left| \frac{G}{2g(x)} \right| < 1$.

Set $\delta = \min\{\delta_1, \delta_2\}$ and assume $|x - a| < \delta$. Then

$$
\left| \frac{1}{g(x)} - \frac{1}{G} \right| = \left| \frac{G - g(x)}{g(x) \cdot G} \right|
$$

$$
= |g(x) - G| \cdot \frac{1}{|G| \cdot |g(x)|} \qquad \text{by assumption}
$$

$$
< \frac{\epsilon_1}{|G| \cdot |g(x)|} \qquad \text{sub in } \epsilon_1
$$

$$
= \epsilon \cdot \frac{|G|}{2|g(x)|} \qquad \text{since } \left| \frac{G}{2g(x)} \right| < 1
$$

$$
< \epsilon
$$

Thus we have shown that for any $\epsilon > 0$ there exists $\delta > 0$ so that when $|x - a| < \delta$ we also have $|1/g(x) - 1/G| < \epsilon$. Hence $1/g(x) \to 1/G$. □

Proof. Proof of Theorem 1.4.17 — Squeeze / sandwich / pinch. In the squeeze theorem, we are given three functions $f(x)$, $g(x)$ and $h(x)$ and are told that

$$
f(x) \leqslant g(x) \leqslant h(x) \quad \text{and} \quad \lim_{x \to a} f(x) = \lim_{x \to a} h(x) = L
$$

and we must conclude from this that $\lim_{x \to a} g(x) = L$ too. That is, we are given some fixed, but unspecified, $\epsilon > 0$ and it is up to us to find a $\delta > 0$ with the property that $|g(x) - L| < \epsilon$ whenever $|x - a| < \delta$. Now because we have been told that f and h both converge to L, there exist $\delta_1 > 0$ and $\delta_2 > 0$ such that

- $|f(x) - L| < \epsilon$, i.e. $L - \epsilon < f(x) < L + \epsilon$, whenever $|x - a| < \delta_1$, and

- $|h(x) - L| < \epsilon$, i.e. $L - \epsilon < h(x) < L + \epsilon$, whenever $|x - a| < \delta_2$

So set $\delta = \min\{\delta_1, \delta_2\}$ and assume $|x - a| < \delta$. Then both $L - \epsilon < f(x) < L + \epsilon$ and $L - \epsilon < h(x) < L + \epsilon$ so that

$$
L - \epsilon < f(x) \leqslant g(x) \leqslant h(x) < L + \epsilon \qquad \text{which implies that}
$$
$$
L - \epsilon < g(x) < L + \epsilon \qquad \text{which in turn gives us}
$$
$$
|g(x) - L| < \epsilon
$$

as desired. □

DERIVATIVES

Calculus is built on two operations — differentiation, which is used to analyse instantaneous rate of change, and integration, which is used to analyse areas. Understanding differentiation and using it to compute derivatives of functions is one of the main aims of this course.

We had a glimpse of derivatives in the previous chapter on limits — in particular Sections 1.1 and 1.2 on tangents and velocities introduced derivatives in disguise. One of the main reasons that we teach limits is to understand derivatives. Fortunately, as we shall see, while one does need to understand limits in order to correctly understand derivatives, one does not need the full machinery of limits in order to *compute and work with* derivatives. The other main part of calculus, integration, we (mostly) leave until a later course.

The derivative finds many applications in many different areas of the sciences. Indeed the reason that calculus is taken by so many university students is so that they may then use the ideas both in subsequent mathematics courses and in other fields. In almost any field in which you study quantitative data you can find calculus lurking somewhere nearby.

Its development[1] came about over a very long time, starting with the ancient Greek geometers. Indian, Persian and Arab mathematicians made significant contributions from around the 6th century. But modern calculus really starts with Newton and Leibniz in the 17th century who developed independently based on ideas of others including Descartes. Newton applied his work to many physical problems (including orbits of moons and planets) but didn't publish his work. When Leibniz subsequently published his "calculus", Newton accused him of plagiarism — this caused a huge rift between British and continental-European mathematicians which wasn't closed for another century.

2.1 ▲ Revisiting Tangent Lines

By way of motivation for the definition of the derivative, we return to the discussion of tangent lines that we started in the previous chapter on limits. We consider, in Exam-

---

1   A quick google will turn up many articles on the development and history of calculus. Wikipedia has a good one.

ples 2.1.2 and 2.1.3, below, the problem of finding the slope of the tangent line to a curve at a point. But let us start by recalling, in Example 2.1.1, what is meant by the slope of a straight line.

> **Example 2.1.1**
>
> In this example, we recall what is meant by the slope of the straight line
>
> $$y = \tfrac{1}{2}x + \tfrac{3}{2}$$
>
> - We claim that if, as we walk along this straight line, our x–coordinate changes by an amount Δx, then our y–coordinate changes by exactly $\Delta y = \tfrac{1}{2}\Delta x$.
>
> - For example, in the figure on the left below, we move from the point
>
> $$(x_0, y_0) = (1\,,\, 2 = \tfrac{1}{2} \times 1 + \tfrac{3}{2})$$
>
> on the line to the point
>
> $$(x_1, y_1) = (5\,,\, 4 = \tfrac{1}{2} \times 5 + \tfrac{3}{2})$$
>
> on the line. In this move our x–coordinate changes by
>
> $$\Delta x = 5 - 1 = 4$$
>
> and our y–coordinate changes by
>
> $$\Delta y = 4 - 2 = 2$$
>
> which is indeed $\tfrac{1}{2} \times 4 = \tfrac{1}{2}\Delta x$, as claimed.
>
>
>
> - In general, when we move from the point
>
> $$(x_0, y_0) = (x_0, \tfrac{1}{2}x_0 + \tfrac{3}{2})$$
>
> on the line to the point
>
> $$(x_1, y_1) = (x_1, \tfrac{1}{2}x_1 + \tfrac{3}{2})$$
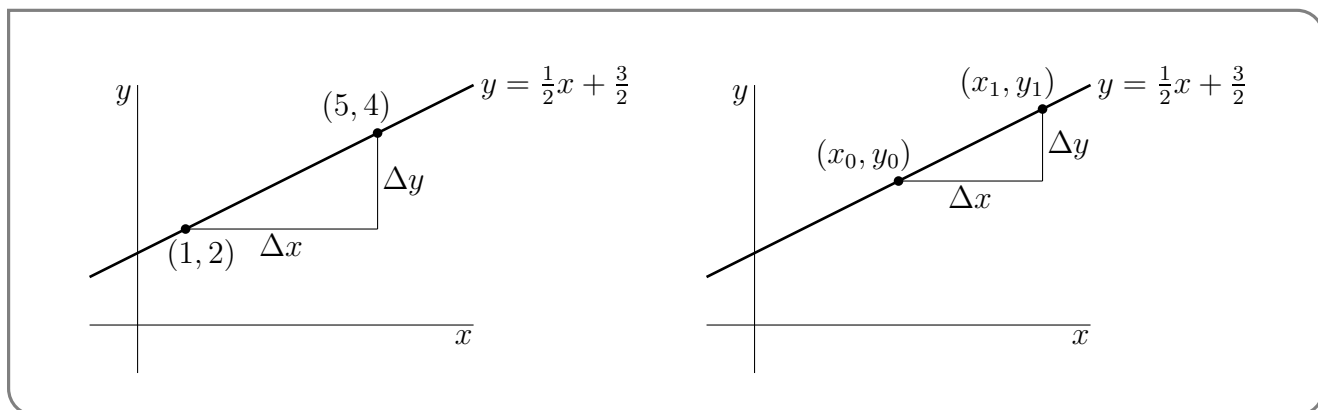
on the line, our x–coordinate changes by

$$\Delta x = x_1 - x_0$$

and our y–coordinate changes by

$$\Delta y = y_1 - y_0$$
$$= \left[\tfrac{1}{2}x_1 + \tfrac{3}{2}\right] - \left[\tfrac{1}{2}x_0 + \tfrac{3}{2}\right]$$
$$= \tfrac{1}{2}(x_1 - x_0)$$

which is indeed $\tfrac{1}{2}\Delta x$, as claimed.

- So, for the straight line $y = \tfrac{1}{2}x + \tfrac{3}{2}$, the ratio $\frac{\Delta y}{\Delta x} = \frac{y_1 - y_0}{x_1 - x_0}$ always takes the value $\tfrac{1}{2}$, regardless of the choice of initial point (x_0, y_0) and final point (x_1, y_1). This constant ratio is the slope of the line $y = \tfrac{1}{2}x + \tfrac{3}{2}$.

Example 2.1.1

Straight lines are special in that for each straight line, there is a fixed number m, called the slope of the straight line, with the property that if you take *any* two different points, (x_0, y_0) and (x_1, y_1), on the line, the ratio $\frac{\Delta y}{\Delta x} = \frac{y_1 - y_0}{x_1 - x_0}$, which is called the rate of change of y per unit rate of change[2] of x, always takes the value m. This is the property that distinguishes lines from other curves.

Other curves do not have this property. In the next two examples we illustrate this point with the parabola $y = x^2$. Recall that we studied this example back in Section 1.1. In Example 2.1.2 we find the slope of the tangent line to $y = x^2$ at a particular point. We generalise this in Example 2.1.3, to show that we can define "the slope of the curve $y = x^2$" at an arbitrary point $x = x_0$ by considering $\frac{\Delta y}{\Delta x} = \frac{y_1 - y_0}{x_1 - x_0}$ with (x_1, y_1) very close to (x_0, y_0).

Example 2.1.2

In this example, let us fix (x_0, y_0) to be the point $(2, 4)$ on the parabola $y = x^2$. Now let $(x_1, y_1) = (x_1, x_1^2)$ be some other point on the parabola; that is, a point with $x_1 \neq x_0$.

- Draw the straight line through (x_0, y_0) and (x_1, y_1) — this is a secant line and we saw these in Chapter 1 when we discussed tangent lines[3].

- The following table gives the slope, $\frac{y_1 - y_0}{x_1 - x_0}$, of the secant line through $(x_0, y_0) = (2, 4)$ and (x_1, y_1), for various different choices of $(x_1, y_1 = x_1^2)$.

x_1	1	1.5	1.9	1.99	1.999	∘	2.001	2.01	2.1	2.5	3
$y_1 = x_1^2$	1	2.25	3.61	3.9601	3.9960	∘	4.0040	4.0401	4.41	6.25	9
$\frac{y_1 - y_0}{x_1 - x_0} = \frac{y_1 - 4}{x_1 - 2}$	3	3.5	3.9	3.99	3.999	∘	4.001	4.01	4.1	4.5	5

2   In the "real world" the phrase "rate of change" usually refers to rate of change per unit time. In science it used more generally.

3   If you do not remember this, then please revisit the first couple of sections of Chapter 1.

- So now we have a big table of numbers — what do we do with them? Well, there are messages we can take away from this table.

  - Different choices of x_1 give different values for the slope, $\frac{y_1 - y_0}{x_1 - x_0}$, of the secant through (x_0, y_0) and (x_1, y_1). This is illustrated in Figure 2.1.1 below — the slope of the secant through (x_0, y_0) and (x_1, y_1) is different from the slope of the secant through (x_0, y_0) and (x_1', y_1').

Figure 2.1.1.
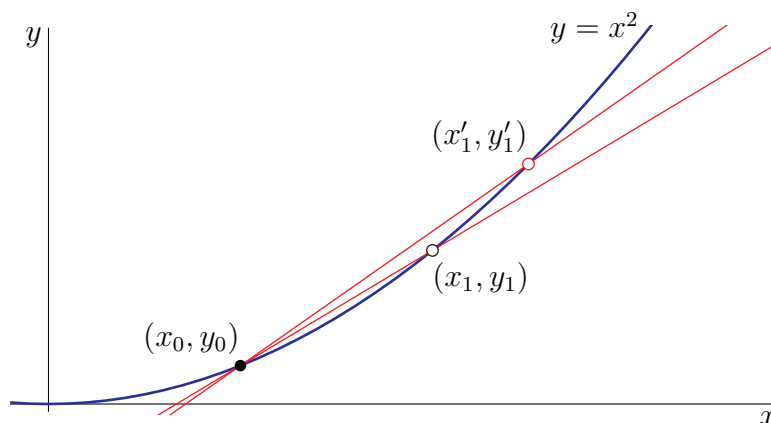
For a curvy curve, different secants have different slopes.

If the parabola were a straight line this would not be the case — the secant through any two different points on a line is always identical to the line itself and so always has exactly the same slope as the line itself, as is illustrated in Figure 2.1.2 below — the (yellow) secant through (x_0, y_0) and (x_1, y_1) lies exactly on top of the (red) line $y = \frac{1}{2}x + \frac{3}{2}$.

Figure 2.1.2.
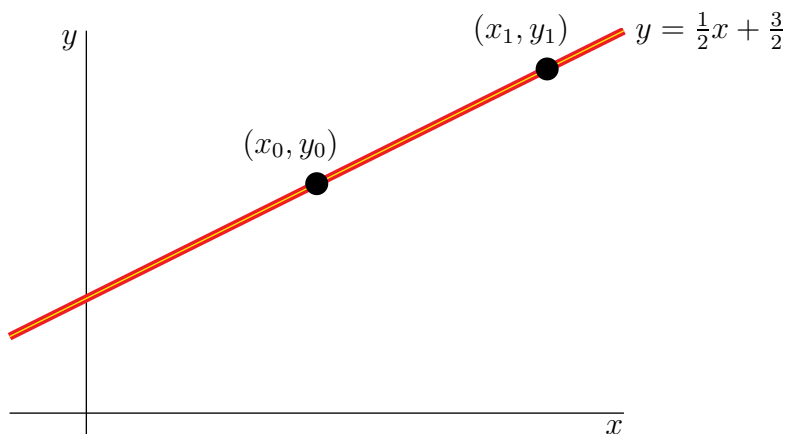
For a straight line, all secants have the same slope.

– Now look at the columns of the table closer to the middle. As x_1 gets closer and closer to $x_0 = 2$, the slope, $\frac{y_1 - y_0}{x_1 - x_0}$, of the secant through (x_0, y_0) and (x_1, y_1) appears to get closer and closer to the value 4.

Example 2.1.2

Example 2.1.3

It is very easy to generalise what is happening in Example 2.1.2.

• Fix any point (x_0, y_0) on the parabola $y = x^2$. If (x_1, y_1) is any other point on the parabola $y = x^2$, then $y_1 = x_1^2$ and the slope of the secant through (x_0, y_0) and (x_1, y_1) is

$$
\begin{aligned}
\text{slope} = \frac{y_1 - y_0}{x_1 - x_0} &= \frac{x_1^2 - x_0^2}{x_1 - x_0} && \text{since } y = x^2 \\
&= \frac{(x_1 - x_0)(x_1 + x_0)}{x_1 - x_0} && \text{remember } a^2 - b^2 = (a - b)(a + b) \\
&= x_1 + x_0
\end{aligned}
$$

You should check the values given in the table of Example 2.1.2 above to convince yourself that the slope $\frac{y_1 - y_0}{x_1 - x_0}$ of the secant line really is $x_0 + x_1 = 2 + x_1$ (since we set $x_0 = 2$).

• Now as we move x_1 closer and closer to x_0, the slope should move closer and closer to $2x_0$. Indeed if we compute the limit carefully — we now have the technology to do this — we see that in the limit as $x_1 \to x_0$ the slope becomes $2x_0$. That is

$$
\begin{aligned}
\lim_{x_1 \to x_0} \frac{y_1 - y_0}{x_1 - x_0} &= \lim_{x_1 \to x_0} (x_1 + x_0) && \text{by the work we did just above} \\
&= 2x_0
\end{aligned}
$$

Taking this limit gives us our first derivative. Of course we haven't yet given the definition of a derivative, so we perhaps wouldn't recognise it yet. We rectify this in the next section.

Figure 2.1.3.

Secants approaching a tangent line
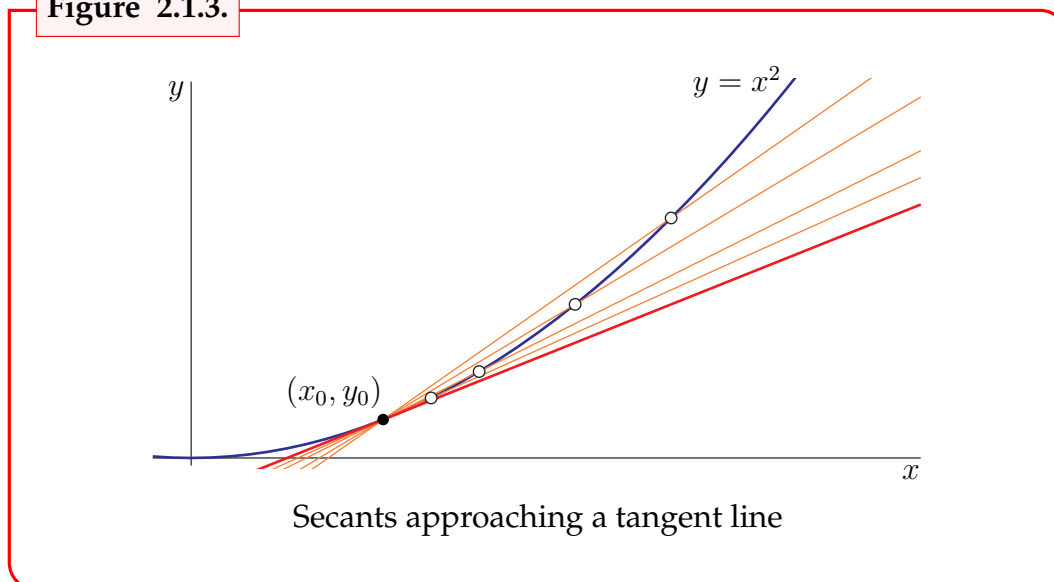
- So it is reasonable to say "as x_1 approaches x_0, the secant through (x_0, y_0) and (x_1, y_1) approaches the tangent line to the parabola $y = x^2$ at (x_0, y_0)". This is what we did back in Section 1.1.

  The figure above shows four different secants through (x_0, y_0) for the curve $y = x^2$. The four hollow circles are four different choices of (x_1, y_1). As (x_1, y_1) approaches (x_0, y_0), the corresponding secant does indeed approach the tangent to $y = x^2$ at (x_0, y_0), which is the heavy (red) straight line in the figure.

  Using limits we determined the slope of the tangent line to $y = x^2$ at x_0 to be $2x_0$. Often we will be a little sloppy with our language and instead say "the slope of the parabola $y = x^2$ at (x_0, y_0) is $2x_0$" — where we really mean the slope of the line tangent to the parabola at x_0.

Example 2.1.3

2.2 ▲ Definition of the Derivative

We now define the "derivative" explicitly, based on the limiting slope ideas of the previous section. Then we see how to compute some simple derivatives.

Let us now generalise what we did in the last section so as to find "the slope of the curve $y = f(x)$ at (x_0, y_0)" for any smooth enough[4] function $f(x)$.

As before, let (x_0, y_0) be any point on the curve $y = f(x)$. So we must have $y_0 = f(x_0)$. Now let (x_1, y_1) be any other point on the same curve. So $y_1 = f(x_1)$ and $x_1 \neq x_0$. Think of (x_1, y_1) as being pretty close to (x_0, y_0) so that the difference

$$\Delta x = x_1 - x_0$$

---

4   The idea of "smooth enough" can be made quite precise. Indeed the word "smooth" has a very precise meaning in mathematics, which we won't cover here. For now think of "smooth" as meaning roughly just "smooth".

in x–coordinates is pretty small. In terms of this Δx we have

$$x_1 = x_0 + \Delta x \qquad \text{and} \qquad y_1 = f(x_0 + \Delta x)$$

We can construct a secant line through (x_0, y_0) and (x_1, y_1) just as we did for the parabola above. It has slope

$$\frac{y_1 - y_0}{x_1 - x_0} = \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

If $f(x)$ is reasonably smooth[5], then as x_1 approaches x_0, i.e. as Δx approaches 0, we would expect the secant through (x_0, y_0) and (x_1, y_1) to approach the tangent line to the curve $y = f(x)$ at (x_0, y_0), just as happened in Figure 2.1.3. And more importantly, the slope of the secant through (x_0, y_0) and (x_1, y_1) should approach the slope of the tangent line to the curve $y = f(x)$ at (x_0, y_0).

Thus we would expect[6] the slope of the tangent line to the curve $y = f(x)$ at (x_0, y_0) to be

$$\lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

When we talk of the "slope of the curve" at a point, what we really mean is the slope of the tangent line to the curve at that point. So "the slope of the curve $y = f(x)$ at (x_0, y_0)" is also the limit[7] expressed in the above equation. The derivative of $f(x)$ at $x = x_0$ is also defined to be this limit. Which leads[8] us to the most important definition in this text:

---

5   Again the term "reasonably smooth" can be made more precise.
6   Indeed, we don't have to expect — it is!
7   This is of course under the assumption that the limit exists — we will talk more about that below.
8   We will rename "x_0" to "a" and "Δx" to "h".

> **Definition 2.2.1** (Derivative at a point).
>
> Let $a \in \mathbb{R}$ and let $f(x)$ be defined on an open interval[9] that contains a.
>
> - The derivative of $f(x)$ at $x = a$ is denoted $f'(a)$ and is defined by
>
> $$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$
>
> if the limit exists.
>
> - When the above limit exists, the function $f(x)$ is said to be differentiable at $x = a$. When the limit does not exist, the function $f(x)$ is said to be not differentiable at $x = a$.
>
> - We can equivalently define the derivative $f'(a)$ by the limit
>
> $$f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a}.$$
>
> To see that these two definitions are the same, we set $x = a + h$ and then the limit as h goes to 0 is equivalent to the limit as x goes to a.

Lets now compute the derivatives of some very simple functions. This is our first step towards building up a toolbox for computing derivatives of complicated functions — this process will very much parallel what we did in Chapter 1 with limits. The two simplest functions we know are $f(x) = c$ and $g(x) = x$.

Example 2.2.2 (Derivative of $f(x) = c$)

Let $a, c \in \mathbb{R}$ be two constants. Compute the derivative of the constant function $f(x) = c$ at $x = a$.

We compute the desired derivative by just substituting the function of interest into the formal definition of the derivative.

$$
\begin{aligned}
f'(a) &= \lim_{h \to 0} \frac{f(a+h) - f(a)}{h} && \text{(the definition)} \\
&= \lim_{h \to 0} \frac{c - c}{h} && \text{(substituted in the function)} \\
&= \lim_{h \to 0} 0 && \text{(simplified things)} \\
&= 0
\end{aligned}
$$

Example 2.2.2

---

9   Recall, from Definition 0.3.5, that the open interval (c, d) is just the set of all real numbers obeying $c < x < d$.

That was easy! What about the next most complicated function — arguably it's this one:

Example 2.2.3 (Derivative of $g(x) = x$)

Let $a \in \mathbb{R}$ and compute the derivative of $g(x) = x$ at $x = a$.

Again, we compute the derivative of g by just substituting the function of interest into the formal definition of the derivative and then evaluating the resulting limit.

$$
\begin{aligned}
g'(a) &= \lim_{h \to 0} \frac{g(a + h) - g(a)}{h} && \text{(the definition)} \\
&= \lim_{h \to 0} \frac{(a + h) - a}{h} && \text{(substituted in the function)} \\
&= \lim_{h \to 0} \frac{h}{h} && \text{(simplified things)} \\
&= \lim_{h \to 0} 1 && \text{(simplified a bit more)} \\
&= 1
\end{aligned}
$$

Example 2.2.3

That was a little harder than the first example, but still quite straight forward — start with the definition and apply what we know about limits.

Thanks to these two examples, we have our first theorem about derivatives:

Theorem 2.2.4 (Easiest derivatives).

Let $a, c \in \mathbb{R}$ and let $f(x) = c$ be the constant function and $g(x) = x$. Then

$$f'(a) = 0$$

and

$$g'(a) = 1.$$

To ratchet up the difficulty a little bit more, let us redo the example we have already done a few times $f(x) = x^2$. To make it a little more interesting let's change the names of the function and the variable so that it is not exactly the same as Examples 2.1.2 and 2.1.3.

Example 2.2.5 (Derivative of $h(t) = t^2$)

Compute the derivative of

$$h(t) = t^2 \qquad\qquad \text{at } t = a$$

- This function isn't quite like the ones we saw earlier — it's a function of t rather than x. Recall that a function is a rule which assigns to each input value an output value. So far, we have usually called the input value x. But this "x" is just a dummy

variable representing a generic input value. There is nothing wrong with calling a generic input value t instead. Indeed, from time to time you will see functions that are not written as formulas involving x, but instead are written as formulas in t (for example representing time — see Section 1.2), or z (for example representing height), or other symbols.

- So let us write the definition of the derivative

$$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

and then translate it to the function names and variables at hand:

$$h'(a) = \lim_{h \to 0} \frac{h(a+h) - h(a)}{h}$$

- But there is a problem — "h" plays two roles here — it is both the function name and the small quantity that is going to zero in our limit. It is extremely dangerous to have a symbol represent two different things in a single computation. We need to change one of them. So let's rename the small quantity that is going to zero in our limit from "h" to "Δt":

$$h'(a) = \lim_{\Delta t \to 0} \frac{h(a + \Delta t) - h(a)}{\Delta t}$$

- Now we are ready to begin. Substituting in what the function h is,

$$
\begin{aligned}
h'(a) &= \lim_{\Delta t \to 0} \frac{(a + \Delta t)^2 - a^2}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{a^2 + 2a\,\Delta t + \Delta t^2 - a^2}{\Delta t} && \text{(just squared out } (a + \Delta t)^2\text{)} \\
&= \lim_{\Delta t \to 0} \frac{2a\,\Delta t + \Delta t^2}{\Delta t} \\
&= \lim_{\Delta t \to 0} (2a + \Delta t) \\
&= 2a
\end{aligned}
$$

- You should go back check that this is what we got in Example 2.1.3 — just some names have been changed.

Example 2.2.5

▸▸ An Important Point (and Some Notation)

Notice here that the answer we get depends on our choice of a — if we want to know the derivative at $a = 3$ we can just substitute $a = 3$ into our answer $2a$ to get the slope is 6. If we want to know at $a = 1$ (like at the end of Section 1.1) we substitute $a = 1$ and get

the slope is 2. The important thing here is that we can move from the derivative being computed at a specific point to the derivative being a function itself — input any value of a and it returns the slope of the tangent line to the curve at the point $x = a$, $y = h(a)$. The variable a is a dummy variable. We can rename a to anything we want, like x, for example. So we can replace every a in

$$h'(a) = 2a \qquad \text{by } x \text{, giving} \qquad h'(x) = 2x$$

where all we have done is replaced the symbol a by the symbol x.

We can do this more generally and tweak the derivative at a specific point a to obtain the derivative as a function of x. We replace

$$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

with

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

which gives us the following definition

> **Definition 2.2.6** (Derivative as a function).
>
> Let $f(x)$ be a function.
>
> - The derivative of $f(x)$ with respect to x is
>
> $$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$
>
> provided the limit exists.
>
> - If the derivative $f'(x)$ exists for all $x \in (a, b)$ we say that f is differentiable on (a, b).
>
> - Note that we will sometimes be a little sloppy with our discussions and simply write "f is differentiable" to mean "f is differentiable on an interval we are interested in" or "f is differentiable everywhere".

Notice that we are no longer thinking of tangent lines, rather this is an operation we can do on a function. For example:

> Example 2.2.7 $\left(\text{The derivative of } f(x) = \frac{1}{x} \right)$
>
> Let $f(x) = \frac{1}{x}$ and compute its derivative with respect to x — think carefully about where the derivative exists.
>
> - Our first step is to write down the definition of the derivative — at this stage, we know of no other strategy for computing derivatives.
>
> $$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \qquad \text{(the definition)}$$

- And now we substitute in the function and compute the limit.

$$
\begin{aligned}
f'(x) &= \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} && \text{(the definition)} \\
&= \lim_{h \to 0} \frac{1}{h} \left[\frac{1}{x+h} - \frac{1}{x} \right] && \text{(substituted in the function)} \\
&= \lim_{h \to 0} \frac{1}{h} \frac{x - (x+h)}{x(x+h)} && \text{(wrote over a common denominator)} \\
&= \lim_{h \to 0} \frac{1}{h} \frac{-h}{x(x+h)} && \text{(started cleanup)} \\
&= \lim_{h \to 0} \frac{-1}{x(x+h)} \\
&= -\frac{1}{x^2}
\end{aligned}
$$

- Notice that the original function $f(x) = \frac{1}{x}$ was not defined at $x = 0$ and the derivative is also not defined at $x = 0$. This does happen more generally — if $f(x)$ is not defined at a particular point $x = a$, then the derivative will not exist at that point either.

Example 2.2.7

So we now have two slightly different ideas of derivatives:

- The derivative $f'(a)$ at a specific point $x = a$, being the slope of the tangent line to the curve at $x = a$, and

- The derivative as a function, $f'(x)$ as defined in Definition 2.2.6.

Of course, if we have $f'(x)$ then we can always recover the derivative at a specific point by substituting $x = a$.

As we noted at the beginning of the chapter, the derivative was discovered independently by Newton and Leibniz in the late 17$^{\text{th}}$ century. Because their discoveries were independent, Newton and Leibniz did not have exactly the same notation. Stemming from this, and from the many different contexts in which derivatives are used, there are quite a few alternate notations for the derivative:

Notation 2.2.8.

The following notations are all used for "the derivative of $f(x)$ with respect to x"

$$f'(x) \qquad \frac{\mathrm{d}f}{\mathrm{d}x} \qquad \frac{\mathrm{d}}{\mathrm{d}x}f(x) \qquad \dot{f}(x) \qquad Df(x) \qquad D_x f(x),$$

while the following notations are all used for "the derivative of $f(x)$ at $x = a$"

$$f'(a) \qquad \frac{\mathrm{d}f}{\mathrm{d}x}(a) \qquad \frac{\mathrm{d}}{\mathrm{d}x}f(x)\Big|_{x=a} \qquad \dot{f}(a) \qquad Df(a) \qquad D_x f(a).$$

Some things to note about these notations:

- We will generally use the first three, but you should recognise them all. The notation $f'(a)$ is due to Lagrange, while the notation $\frac{\mathrm{d}f}{\mathrm{d}x}(a)$ is due to Leibniz. They are both very useful. Neither can be considered "better".

- Leibniz notation writes the derivative as a "fraction" — however it is definitely not a fraction and should not be thought of in that way. It is just shorthand, which is read as "the derivative of f with respect to x".

- You read $f'(x)$ as "f–prime of x", and $\frac{\mathrm{d}f}{\mathrm{d}x}$ as "dee–f–dee–x", and $\frac{\mathrm{d}}{\mathrm{d}x}f(x)$ as "dee-by–dee–x of f".

- Similarly you read $\frac{\mathrm{d}f}{\mathrm{d}x}(a)$ as "dee–f–dee–x at a", and $\frac{\mathrm{d}}{\mathrm{d}x}f(x)\Big|_{x=a}$ as "dee-by-dee x of f at x equals a".

- The notation $\dot{f}$ is due to Newton. In physics, it is common to use $\dot{f}(t)$ to denote the derivative of f with respect to time.

⇥ Back to Computing Some Derivatives

At this point we could try to start working out how derivatives interact with arithmetic and make an "Arithmetic of derivatives" theorem just like the one we saw for limits (Theorem 2). We will get there shortly, but before that it is important that we become more comfortable with computing derivatives using limits and then understanding what the derivative actually means. So — more examples.

Example 2.2.9 $\left(\frac{\mathrm{d}}{\mathrm{d}x}\sqrt{x} \right)$

Compute the derivative, $f'(a)$, of the function $f(x) = \sqrt{x}$ at the point $x = a$ for any $a > 0$.

- So again we start with the definition of derivative and go from there:

$$f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a} = \lim_{x \to a} \frac{\sqrt{x} - \sqrt{a}}{x - a}$$

- As x tends to a, the numerator and denominator both tend to zero. But $\frac{0}{0}$ is not defined. So to get a well defined limit we need to exhibit a cancellation between the numerator and denominator — just as we saw in Examples 1.4.11 and 1.4.16. Now there are two equivalent ways to proceed from here, both based on a similar "trick".

- For the first, review Example 1.4.16, which concerned taking a limit involving square-roots, and recall that we used "multiplication by the conjugate" there:

$$
\begin{aligned}
\frac{\sqrt{x} - \sqrt{a}}{x - a} &= \frac{\sqrt{x} - \sqrt{a}}{x - a} \times \frac{\sqrt{x} + \sqrt{a}}{\sqrt{x} + \sqrt{a}} \qquad \left(\text{multiplication by } 1 = \frac{\text{conjugate}}{\text{conjugate}}\right) \\
&= \frac{(\sqrt{x} - \sqrt{a})(\sqrt{x} + \sqrt{a})}{(x - a)(\sqrt{x} + \sqrt{a})} \\
&= \frac{x - a}{(x - a)(\sqrt{x} + \sqrt{a})} \qquad \left(\text{since } (A - B)(A + B) = A^2 - B^2\right) \\
&= \frac{1}{\sqrt{x} + \sqrt{a}}
\end{aligned}
$$

- Alternatively, we can arrive at $\frac{\sqrt{x} - \sqrt{a}}{x - a} = \frac{1}{\sqrt{x} + \sqrt{a}}$ by using almost the same trick to factor the denominator. Just set $A = \sqrt{x}$ and $B = \sqrt{a}$ in $A^2 - B^2 = (A - B)(A + B)$ to get

$$
x - a = (\sqrt{x} - \sqrt{a})(\sqrt{x} + \sqrt{a})
$$

and then substitute this little fact into our expression

$$
\begin{aligned}
\frac{\sqrt{x} - \sqrt{a}}{x - a} &= \frac{\sqrt{x} - \sqrt{a}}{(\sqrt{x} - \sqrt{a})(\sqrt{x} + \sqrt{a})} \qquad \text{(now cancel common factors)} \\
&= \frac{1}{(\sqrt{x} + \sqrt{a})}
\end{aligned}
$$

- Once we know that $\frac{\sqrt{x} - \sqrt{a}}{x - a} = \frac{1}{\sqrt{x} + \sqrt{a}}$, we can take the limit we need:

$$
\begin{aligned}
f'(a) &= \lim_{x \to a} \frac{\sqrt{x} - \sqrt{a}}{x - a} \\
&= \lim_{x \to a} \frac{1}{\sqrt{x} + \sqrt{a}} \\
&= \frac{1}{2\sqrt{a}}
\end{aligned}
$$

- We should think about the domain of f' here — that is, for which values of a is $f'(a)$ defined? The original function $f(x)$ was defined for all $x \geqslant 0$, however the derivative $f'(a) = \frac{1}{2\sqrt{a}}$ is undefined at $a = 0$.

  If we draw a careful picture of $\sqrt{x}$ around $x = 0$ we can see why this has to be the case. The figure below shows three different tangent lines to the graph of $y = f(x) = \sqrt{x}$. As the point of tangency moves closer and closer to the origin, the tangent line gets steeper and steeper. The slope of the tangent line at $(a, \sqrt{a})$ blows up as $a \to 0$.
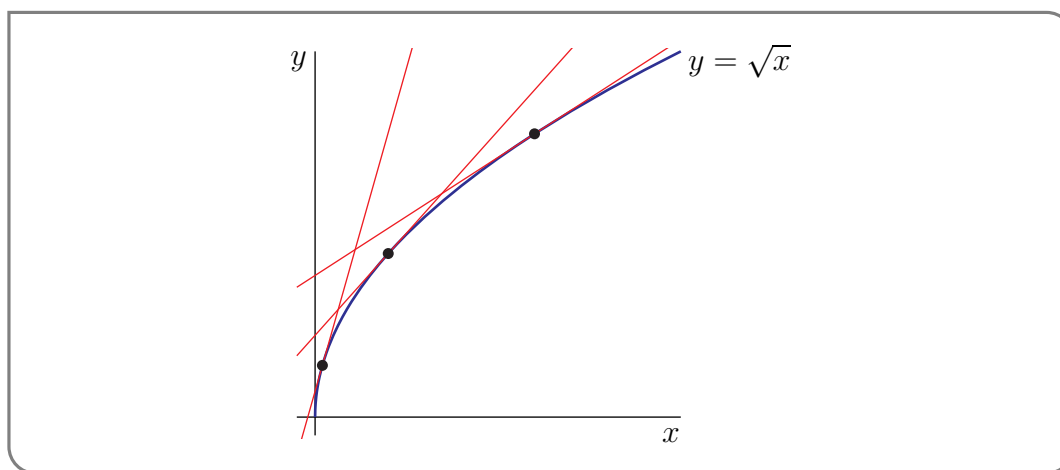
Example 2.2.9

Example 2.2.10 $\left(\frac{\mathrm{d}}{\mathrm{d}x} \{|x|\} \right)$

Compute the derivative, $f'(a)$, of the function $f(x) = |x|$ at the point $x = a$.

- We should start this example by recalling the definition of $|x|$ (we saw this back in Example 1.5.6):

$$
|x| = \begin{cases} -x & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ x & \text{if } x > 0. \end{cases}
$$

It is definitely not just "chop off the minus sign".
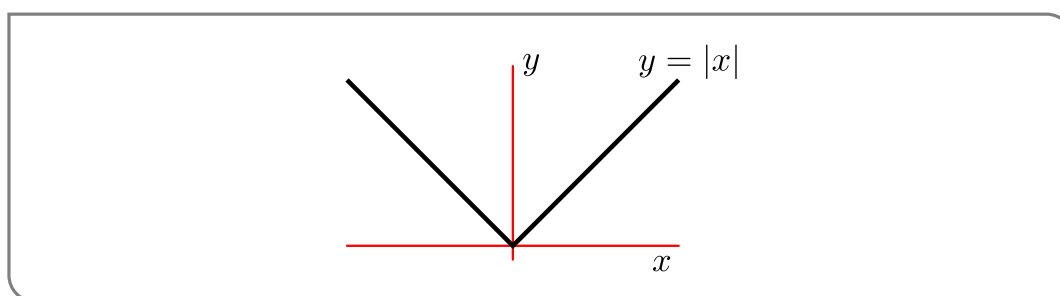
- This breaks our computation of the derivative into 3 cases depending on whether x is positive, negative or zero.

- Assume $x > 0$. Then

$$
\begin{aligned}
\frac{\mathrm{d}f}{\mathrm{d}x} &= \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \\
&= \lim_{h \to 0} \frac{|x+h| - |x|}{h}
\end{aligned}
$$

Since $x > 0$ and we are interested in the behaviour of this function as $h \to 0$ we can assume h is much smaller than x. This means $x + h > 0$ and so $|x + h| = x + h$.

$$= \lim_{h \to 0} \frac{x + h - x}{h}$$

$$= \lim_{h \to 0} \frac{h}{h} = 1 \qquad\qquad \text{as expected}$$

- Assume $x < 0$. Then

$$\frac{df}{dx} = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h}$$

$$= \lim_{h \to 0} \frac{|x + h| - |x|}{h}$$

Since $x < 0$ and we are interested in the behaviour of this function as $h \to 0$ we can assume h is much smaller than x. This means $x + h < 0$ and so $|x + h| = -(x + h)$.

$$= \lim_{h \to 0} \frac{-(x + h) - (-x)}{h}$$

$$= \lim_{h \to 0} \frac{-h}{h} = -1$$

- When $x = 0$ we have

$$f'(0) = \lim_{h \to 0} \frac{f(0 + h) - f(0)}{h}$$

$$= \lim_{h \to 0} \frac{|0 + h| - |0|}{h}$$

$$= \lim_{h \to 0} \frac{|h|}{h}$$

To proceed we need to know if $h > 0$ or $h < 0$, so we must use one-sided limits. The limit from above is:

$$\lim_{h \to 0^+} \frac{|h|}{h} = \lim_{h \to 0^+} \frac{h}{h} \qquad\qquad \text{since } h > 0, |h| = h$$

$$= 1$$

Whereas, the limit from below is:

$$\lim_{h \to 0^-} \frac{|h|}{h} = \lim_{h \to 0^-} \frac{-h}{h} \qquad\qquad \text{since } h < 0, |h| = -h$$

$$= -1$$

Since the one-sided limits differ, the limit as $h \to 0$ does not exist. And thus the derivative does not exist as $x = 0$.

In summary:

$$\frac{d}{dx}|x| = \begin{cases} -1 & \text{if } x < 0 \\ DNE & \text{if } x = 0 \\ 1 & \text{if } x > 0 \end{cases}$$
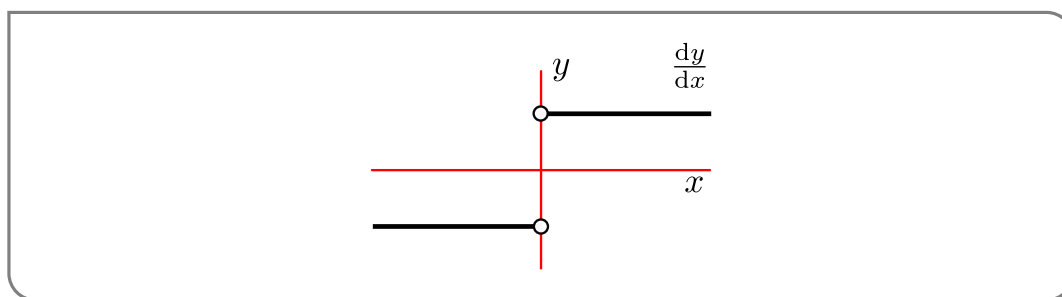
Example 2.2.10

➤ Where is the Derivative Undefined?

According to Definition 2.2.1, the derivative $f'(a)$ exists precisely when the limit $\lim\limits_{x \to a} \frac{f(x) - f(a)}{x - a}$ exists. That limit is also the slope of the tangent line to the curve $y = f(x)$ at $x = a$. That limit does not exist when the curve $y = f(x)$ does not have a tangent line at $x = a$ or when the curve does have a tangent line, but the tangent line has infinite slope. We have already seen some examples of this.

- In Example 2.2.7, we considered the function $f(x) = \frac{1}{x}$. This function "blows up" (i.e. becomes infinite) at $x = 0$. It does not have a tangent line at $x = 0$ and its derivative does not exist at $x = 0$.

- In Example 2.2.10, we considered the function $f(x) = |x|$. This function does not have a tangent line at $x = 0$, because there is a sharp corner in the graph of $y = |x|$ at $x = 0$. (Look at the graph in Example 2.2.10.) So the derivative of $f(x) = |x|$ does not exist at $x = 0$.

Here are a few more examples.

Example 2.2.11

Visually, the function

$$H(x) = \begin{cases} 0 & \text{if } x \leqslant 0 \\ 1 & \text{if } x > 0 \end{cases}$$

does not have a tangent line at $(0,0)$. Not surprisingly, when $a = 0$ and h tends to 0 with $h > 0$,

$$\frac{H(a + h) - H(a)}{h} = \frac{H(h) - H(0)}{h} = \frac{1}{h}$$

blows up. The same sort of computation shows that $f'(a)$ cannot possibly exist whenever the function f is not continuous at a. We will formalize, and prove, this statement in Theorem 2.2.14, below.

Example 2.2.11

Example 2.2.12 $\left(\frac{\mathrm{d}}{\mathrm{d}x} x^{1/3} \right)$

Visually, it looks like the function $f(x) = x^{1/3}$, sketched below, (this might be a good point to recall that cube roots of negative numbers are negative — for example, since $(-1)^3 = -1$, the cube root of -1 is -1),

$$y = x^{1/3}$$
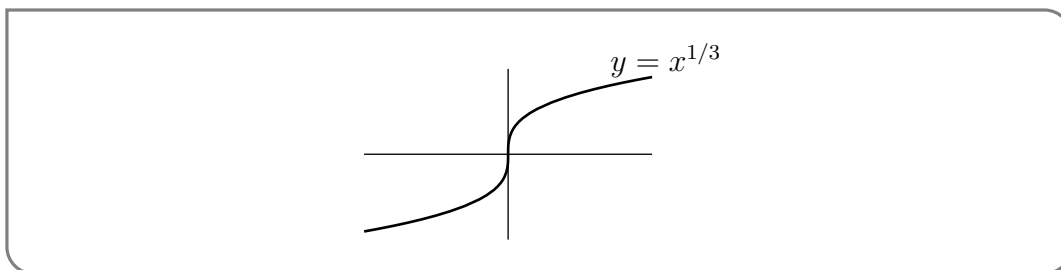
has the y–axis as its tangent line at $(0,0)$. So we would expect that $f'(0)$ does not exist. Let's check. With $a = 0$,

$$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h} = \lim_{h \to 0} \frac{f(h) - f(0)}{h} = \lim_{h \to 0} \frac{h^{1/3}}{h} = \lim_{h \to 0} \frac{1}{h^{2/3}} = DNE$$

as expected.

Example 2.2.12

Example 2.2.13 $\left(\frac{\mathrm{d}}{\mathrm{d}x} \sqrt{|x|} \right)$

We have already considered the derivative of the function $\sqrt{x}$ in Example 2.2.9. We'll now look at the function $f(x) = \sqrt{|x|}$. Recall, from Example 2.2.10, the definition of $|x|$. When $x > 0$, we have $|x| = x$ and $f(x)$ is identical to $\sqrt{x}$. When $x < 0$, we have $|x| = -x$ and $f(x) = \sqrt{-x}$. So to graph $y = \sqrt{|x|}$ when $x < 0$, you just have to graph $y = \sqrt{x}$ for $x > 0$ and then send $x \to -x$ — i.e. reflect the graph in the y–axis. Here is the graph. The pointy

$$y = \sqrt{|x|}$$
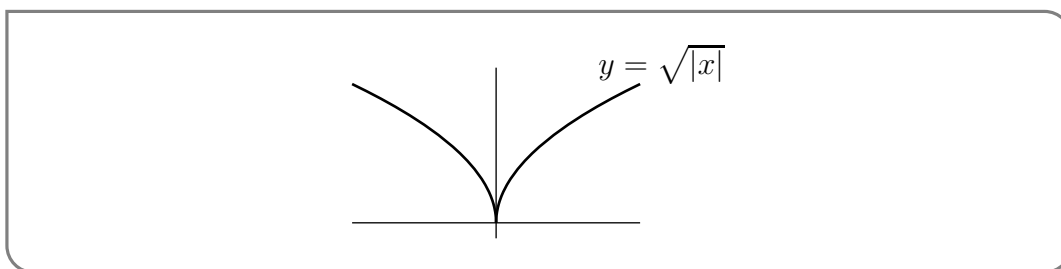
thing at the origin is called a cusp. The graph of $y = f(x)$ does not have a tangent line at $(0,0)$ and, correspondingly, $f'(0)$ does not exist because

$$\lim_{h \to 0^+} \frac{f(h) - f(0)}{h} = \lim_{h \to 0^+} \frac{\sqrt{|h|}}{h} = \lim_{h \to 0^+} \frac{1}{\sqrt{h}} = DNE$$

Example 2.2.13

Theorem 2.2.14.

If the function $f(x)$ is differentiable at $x = a$, then $f(x)$ is also continuous at $x = a$.

Proof. The function $f(x)$ is continuous at $x = a$ if and only if the limit of

$$f(a + h) - f(a) = \frac{f(a + h) - f(a)}{h} \, h$$

as $h \to 0$ exists and is zero. But if $f(x)$ is differentiable at $x = a$, then, as $h \to 0$, the first factor, $\frac{f(a+h) - f(a)}{h}$ converges to $f'(a)$ and the second factor, h, converges to zero. So the product provision of our arithmetic of limits Theorem 1.4.2 implies that the product $\frac{f(a+h) - f(a)}{h} \, h$ converges to $f'(a) \cdot 0 = 0$ too. $\qquad \square$

Notice that while this theorem is useful as stated, it is (arguably) more often applied in its contrapositive[10] form:

Theorem 2.2.15 (The contrapositive of Theorem 2.2.14).

If $f(x)$ is not continuous at $x = a$ then it is not differentiable at $x = a$.

As the above examples illustrate, this statement does not tell us what happens if f *is continuous at $x = a$ —* we have to think!

2.3 ▲ Interpretations of the Derivative

In the previous sections we defined the derivative as the slope of a tangent line, using a particular limit. This allows us to compute "the slope of a curve[11]" and provides us with one interpretation of the derivative. However, the main importance of derivatives does not come from this application. Instead, (arguably) it comes from the interpretation of the derivative as the instantaneous rate of change of a quantity.

---

10   If you have forgotten what the contrapositive is, then quickly reread Footnote 10 in Section 1.3.
11   Again — recall that we are being a little sloppy with this term — we really mean "The slope of the tangent line to the curve".

▸▸ Instantaneous Rate of Change

In fact we have already (secretly) used a derivative to compute an instantaneous rate of change in Section 1.2. For your convenience we'll review that computation here, in Example 2.3.1, and then generalise it.

Example 2.3.1

You drop a ball from a tall building. After t seconds the ball has fallen a distance of $s(t) = 4.9t^2$ metres. What is the velocity of the ball one second after it is dropped?

- In the time interval from $t = 1$ to $t = 1 + h$ the ball travels a distance

$$s(1 + h) - s(1) = 4.9(1 + h)^2 - 4.9(1)^2 = 4.9[2h + h^2]$$

- So the average velocity over this time interval is

$$\text{average velocity from } t = 1 \text{ to } t = 1 + h$$
$$= \frac{\text{distance travelled from } t = 1 \text{ to } t = 1 + h}{\text{length of time from } t = 1 \text{ to } t = 1 + h}$$
$$= \frac{s(1 + h) - s(1)}{h}$$
$$= \frac{4.9[2h + h^2]}{h}$$
$$= 4.9[2 + h]$$

- The instantaneous velocity at time $t = 1$ is then defined to be the limit

$$\text{instantaneous velocity at time } t = 1$$
$$= \lim_{h \to 0} \left[\text{average velocity from } t = 1 \text{ to } t = 1 + h\right]$$
$$= \lim_{h \to 0} \frac{s(1 + h) - s(1)}{h} = s'(1)$$
$$= \lim_{h \to 0} 4.9[2 + h]$$
$$= 9.8 \text{m/sec}$$

- We conclude that the instantaneous velocity at time $t = 1$, which is the instantaneous rate of change of distance per unit time at time $t = 1$, is the derivative $s'(1) = 9.8\text{m/sec}$.

Example 2.3.1

Now suppose, more generally, that you are taking a walk and that as you walk, you are continuously measuring some quantity, like temperature, and that the measurement

at time t is $f(t)$. Then the

$$\text{average rate of change of } f(t) \text{ from } t = a \text{ to } t = a + h$$
$$= \frac{\text{change in } f(t) \text{ from } t = a \text{ to } t = a + h}{\text{length of time from } t = a \text{ to } t = a + h}$$
$$= \frac{f(a + h) - f(a)}{h}$$

so the

$$\text{instantaneous rate of change of } f(t) \text{ at } t = a$$
$$= \lim_{h \to 0} \left[\text{average rate of change of } f(t) \text{ from } t = a \text{ to } t = a + h \right]$$
$$= \lim_{h \to 0} \frac{f(a + h) - f(a)}{h}$$
$$= f'(a)$$

In particular, if you are walking along the x–axis and your x–coordinate at time t is $x(t)$, then $x'(a)$ is the instantaneous rate of change (per unit time) of your x–coordinate at time $t = a$, which is your velocity at time a. If $v(t)$ is your velocity at time t, then $v'(a)$ is the instantaneous rate of change of your velocity at time a. This is called your acceleration at time a.

You might expect that if the instantaneous rate of change of a function at time c is strictly positive, then, in some sense, the function is increasing at $t = c$. You would be right. Indeed, if $f'(c) > 0$, then, by definition, the limit of $\frac{f(t) - f(c)}{t - c}$ as t approaches c is strictly bigger than zero. So

- for all $t > c$ that are sufficiently close[12] to c

$$\frac{f(t) - f(c)}{t - c} > 0 \implies f(t) - f(c) > 0 \qquad (\text{since } t - c > 0)$$
$$\implies f(t) > f(c)$$

- for all $t < c$ that are sufficiently close to c

$$\frac{f(t) - f(c)}{t - c} > 0 \implies f(t) - f(c) < 0 \qquad (\text{since } t - c < 0)$$
$$\implies f(t) < f(c)$$

Consequently we say that "$f(t)$ is increasing at $t = c$". If we wish to emphasise that the inequalities above are the strict inequalities $>$ and $<$, as opposed to $\geqslant$ and $\leqslant$, we will say that "$f(t)$ is strictly increasing at $t = c$".

---

12   This is typical mathematician speak — it allows us to be completely correct, without being terribly precise. In this context, "sufficiently close" means "The following need not be true for all t bigger than c, but there must exist some $b > c$ so that the following is true for all $c < t < b$". Typically we do not know what b is. And typically it does not matter what the exact value of b is. All that matters is that b exists and is strictly bigger than c.
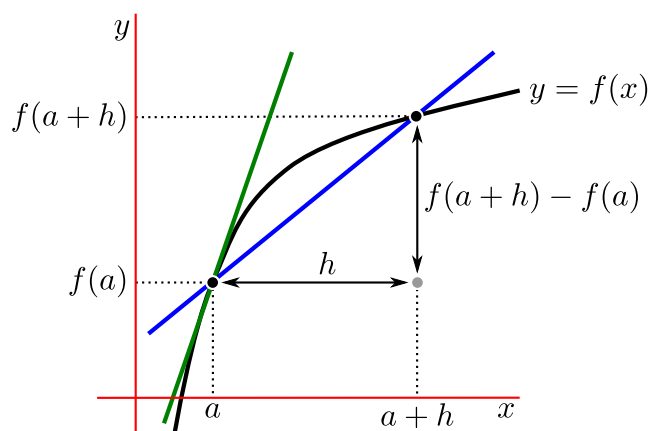
▸▸ Slope

Suppose that $y = f(x)$ is the equation of a curve in the xy–plane. That is, $f(x)$ is the y–coordinate of the point on the curve whose x–coordinate is x. Then, as we have already seen,

$$\Big[\text{the slope of the secant through } \big(a, f(a)\big) \text{ and } \big(a + h, f(a + h)\big)\Big] = \frac{f(a + h) - f(a)}{h}$$

This is shown in Figure 2.3.1 below.

Figure 2.3.1.

In order to create the tangent line (as we have done a few times now) we squeeze $h \to 0$. As we do this, the secant through $\big(a, f(a)\big)$ and $\big(a + h, f(a + h)\big)$ approaches[13] the tangent line to $y = f(x)$ at $x = a$. Since the secant becomes the tangent line in this limit, the slope of the secant becomes the slope of the tangent and

$$\Big[\text{the slope of the tangent line to } y = f(x) \text{ at } x = a\Big] = \lim_{h \to 0} \frac{f(a + h) - f(a)}{h}$$
$$= f'(a).$$

Let us go a little further and work out a general formula for the equation of the tangent line to $y = f(x)$ at $x = a$. We know that the tangent line

- has slope $f'(a)$ and

- passes through the point $\big(a, f(a)\big)$.

There are a couple of different ways to construct the equation of the tangent line from this information. One is to observe, as in Figure 2.3.2, that if (x, y) is any other point on the tangent line then the line segment from $\big(a, f(a)\big)$ to (x, y) is part of the tangent line and so also has slope $f'(a)$. That is,

$$\frac{y - f(a)}{x - a} = \Big[\text{the slope of the tangent line}\Big] = f'(a)$$

13 We are of course assuming that the curve is smooth enough to have a tangent line at a.

Cross multiplying gives us the equation of the tangent line:

$$y - f(a) = f'(a)\,(x - a) \qquad \text{or} \qquad y = f(a) + f'(a)\,(x - a)$$

Figure 2.3.2.
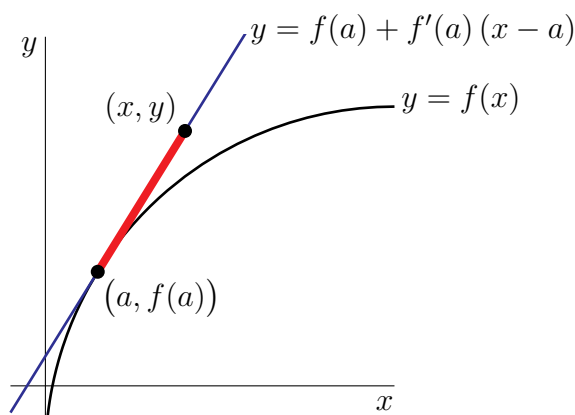
A line segment of a tangent line

A second way to derive the same equation of the same tangent line is to recall that the general equation for a line, with finite slope, is $y = mx + b$, where m is the slope and b is the y-intercept. We already know the slope — so $m = f'(a)$. To work out b we use the other piece of information — $(a, f(a))$ is on the line. So $(x, y) = (a, f(a))$ must solve $y = f'(a)\,x + b$. That is,

$$f(a) = f'(a) \cdot a + b \qquad \text{and so} \qquad b = f(a) - af'(a)$$

Hence our equation is, once again,

$$y = f'(a) \cdot x + (f(a) - af'(a)) \qquad \text{or, after rearranging a little,}$$
$$y = f(a) + f'(a)\,(x - a)$$

This is a very useful formula, so perhaps we should make it a theorem.

Theorem 2.3.2 (Tangent line).

The tangent line to the curve $y = f(x)$ at $x = a$ is given by the equation

$$y = f(a) + f'(a)\,(x - a)$$

provided the derivative $f'(a)$ exists.

The caveat at the end of the above theorem is necessary — there are certainly cases in which the derivative does not exist and so we do need to be careful.

Example 2.3.3

Find the tangent line to the curve $y = \sqrt{x}$ at $x = 4$.

Rather than redoing everything from scratch, we can, and for efficiency, should, use Theorem 2.3.2. To write this up properly, we must ensure that we tell the reader what we are doing. So something like the following:

- By Theorem 2.3.2, the tangent line to the curve $y = f(x)$ at $x = a$ is given by

$$y = f(a) + f'(a)(x - a)$$

provided $f'(a)$ exists.

- In Example 2.2.9, we found that, for any $a > 0$, the derivative of $\sqrt{x}$ at $x = a$ is

$$f'(a) = \frac{1}{2\sqrt{a}}$$

- In the current example, $a = 4$ and we have

$$f(a) = f(4) = \sqrt{x}\big|_{x=4} = \sqrt{4} = 2 \quad \text{and} \quad f'(a) = f'(4) = \frac{1}{2\sqrt{a}}\bigg|_{a=4} = \frac{1}{2\sqrt{4}} = \frac{1}{4}$$

- So the equation of the tangent line to $y = \sqrt{x}$ at $x = 4$ is

$$y = 2 + \frac{1}{4}(x - 4) \quad \text{or} \quad y = \frac{x}{4} + 1$$

We don't have to write it up using dot-points as above; we have used them here to help delineate each step in the process of computing the tangent line.

Example 2.3.3

2.4 ▲ Arithmetic of Derivatives - a Differentiation Toolbox

So far, we have evaluated derivatives only by applying Definition 2.2.1 to the function at hand and then computing the required limits directly. It is quite obvious that as the function being differentiated becomes even a little complicated, this procedure quickly becomes extremely unwieldy. It is many orders of magnitude more efficient to have access to

- a list of derivatives of some simple functions and

- a collection of rules for breaking down complicated derivative computations into sequences of simple derivative computations.

This is precisely what we did to compute limits. We started with limits of simple functions and then used "arithmetic of limits" to computed limits of complicated functions.

We have already started building our list of derivatives of simple functions. We have shown, in Examples 2.2.2, 2.2.3, 2.2.5 and 2.2.9, that

$$\frac{d}{dx}1 = 0 \qquad \frac{d}{dx}x = 1 \qquad \frac{d}{dx}x^2 = 2x \qquad \frac{d}{dx}\sqrt{x} = \frac{1}{2\sqrt{x}}$$

We'll expand this list later.

We now start building a collection of tools that help reduce the problem of computing the derivative of a complicated function to that of computing the derivatives of a number of simple functions. In this section we give three derivative "rules" as three separate theorems. We'll give the proofs of these theorems in the next section and examples of how they are used in the following section.

As was the case for limits, derivatives interact very cleanly with addition, subtraction and multiplication by a constant. The following result actually follows very directly from the first three points of Theorem 1.4.2.

Lemma 2.4.1 (Derivative of sum and difference).

Let $f(x), g(x)$ be differentiable functions and let $c \in \mathbb{R}$ be a constant. Then

$$\frac{d}{dx}\{f(x) + g(x)\} = f'(x) + g'(x)$$
$$\frac{d}{dx}\{f(x) - g(x)\} = f'(x) - g'(x)$$
$$\frac{d}{dx}\{cf(x)\} = cf'(x)$$

That is, the derivative of the sum is the sum of the derivatives, and so forth.

Following this we can combine the three statements in this lemma into a single rule which captures the "linearity of differentiation".

Theorem 2.4.2 (Linearity of differentiation).

Again, let $f(x), g(x)$ be differentiable functions, let $\alpha, \beta \in \mathbb{R}$ be constants and define the "linear combination"

$$S(x) = \alpha f(x) + \beta g(x).$$

Then the derivative of $S(x)$ at $x = a$ exists and is

$$\frac{dS}{dx} = S'(x) = \alpha f'(x) + \beta g'(x).$$

Note that we can recover the three rules in the previous lemma by setting $\alpha = \beta = 1$ or $\alpha = 1, \beta = -1$ or $\alpha = c, \beta = 0$.

Unfortunately, the derivative does not act quite as simply on products or quotients. The rules for computing derivatives of products and quotients get their own names and theorems:

> **Theorem 2.4.3** (The product rule).
>
> Let $f(x), g(x)$ be differentiable functions, then the derivative of the product $f(x)g(x)$ exists and is given by
>
> $$\frac{d}{dx}\{f(x)\,g(x)\} = f'(x)\,g(x) + f(x)\,g'(x).$$

Before we proceed to the derivative of the ratio of two functions, it is worth noting a special case of the product rule when $g(x) = f(x)$. In fact, since this is a useful special case, let us call it a corollary[14]:

> **Corollary 2.4.4** (Derivative of a square).
>
> Let $f(x)$ be a differentiable function, then the derivative of its square is:
>
> $$\frac{d}{dx}\{f(x)^2\} = 2f(x)\,f'(x)$$

With a little work this can be generalised to other powers — but that is best done once we understand how to compute the derivative of the composition of two functions. That requires the chain rule (see Theorem 2.9.2 below). But before we get to that, we need to see how to take the derivative of a quotient of two functions.

> **Theorem 2.4.5** (The quotient rule).
>
> Let $f(x), g(x)$ be differentiable functions. Then the derivative of their quotient is
>
> $$\frac{d}{dx}\left\{\frac{f(x)}{g(x)}\right\} = \frac{f'(x)\,g(x) - f(x)\,g'(x)}{g(x)^2}.$$
>
> This derivative exists except at points where $g(x) = 0$.

There is a useful special case of this theorem which we obtain by setting $f(x) = 1$. In that case, the quotient rule tells us how to compute the derivative of the reciprocal of a function.

---

14   Recall that a corollary is an important result that follows from one or more theorems — typically without too much extra work — as is the case here.

> **Corollary 2.4.6** (Derivative of a reciprocal).
>
> Let $g(x)$ be a differentiable function. Then the derivative of the reciprocal of g is given by
>
> $$\frac{d}{dx}\left\{\frac{1}{g(x)}\right\} = -\frac{g'(x)}{g(x)^2}$$
>
> and exists except at those points where $g(x) = 0$.

So we have covered, sums, differences, products and quotients. This allows us to compute derivatives of many different functions — including polynomials and rational functions. However we are still missing trigonometric functions (for example), and a rule for computing derivatives of compositions. These will follow in the near future, but there are a couple of things to do before that — understand where the above theorems come from, and practice using them.

2.5 ▲ Proofs of the Arithmetic of Derivatives

The theorems of the previous section are not too difficult to prove from the definition of the derivative (which we know) and the arithmetic of limits (which we also know). In this section we show how to construct these rules.

Throughout this section we will use our two functions $f(x)$ and $g(x)$. Since the theorems we are going to prove all express derivatives of linear combinations, products and quotients in terms of f, g and their derivatives, it is helpful to recall the definitions of the derivatives of f and g:

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \qquad \text{and} \qquad g'(x) = \lim_{h \to 0} \frac{g(x+h) - g(x)}{h}.$$

Our proofs, roughly speaking, involve doing algebraic manipulations to uncover the expressions that look like the above.

▸▸ Proof of the Linearity of Differentiation (Theorem 2.4.2)

Recall that in Theorem 2.4.2 we defined $S(x) = \alpha f(x) + \beta g(x)$, where $\alpha, \beta \in \mathbb{R}$ are constants. We wish to compute $S'(x)$, so we start with the definition:

$$S'(x) = \lim_{h \to 0} \frac{S(x+h) - S(x)}{h}$$

Let us concentrate on the numerator of the expression inside the limit and then come back to the full limit in a moment. Substitute in the definition of $S(x)$:

$$S(x+h) - S(x) = \left[\alpha f(x+h) + \beta g(x+h)\right] - \left[\alpha f(x) + \beta g(x)\right] \qquad \text{collect terms}$$
$$= \alpha\left[f(x+h) - f(x)\right] + \beta\left[g(x+h) - g(x)\right]$$

Now it is easy to see the structures we need — namely, we almost have the expressions for the derivatives $f'(x)$ and $g'(x)$. Indeed, all we need to do is divide by h and take the limit. So let's finish things off.

$$
\begin{aligned}
S'(x) &= \lim_{h \to 0} \frac{S(x+h) - S(x)}{h} & \text{from above}\\
&= \lim_{h \to 0} \frac{\alpha\left[f(x+h) - f(x)\right] + \beta\left[g(x+h) - g(x)\right]}{h}\\
&= \lim_{h \to 0} \left[\alpha\frac{f(x+h) - f(x)}{h} + \beta\frac{g(x+h) - g(x)}{h}\right] & \text{limit laws}\\
&= \alpha \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} + \beta \lim_{h \to 0} \frac{g(x+h) - g(x)}{h}\\
&= \alpha f'(x) + \beta g'(x)
\end{aligned}
$$

as required.

▸▸ **Proof of the Product Rule (Theorem 2.4.3)**

After the warm-up above, we will just jump straight in. Let $P(x) = f(x)\,g(x)$, the product of our two functions. The derivative of the product is given by

$$
P'(x) = \lim_{h \to 0} \frac{P(x+h) - P(x)}{h}
$$

Again we will focus on the numerator inside the limit and massage it into the form we need. To simplify these manipulations, define

$$
F(h) = \frac{f(x+h) - f(x)}{h} \qquad \text{and} \qquad G(h) = \frac{g(x+h) - g(x)}{h}.
$$

Then we can write

$$
f(x+h) = f(x) + hF(h) \qquad \text{and} \qquad g(x+h) = g(x) + hG(h).
$$

We can also write

$$
f'(x) = \lim_{h \to 0} F(h) \qquad \text{and} \qquad g'(x) = \lim_{h \to 0} G(h).
$$

So back to that numerator:

$$
\begin{aligned}
P(x+h) - P(x) &= f(x+h) \cdot g(x+h) - f(x) \cdot g(x) & \text{substitute}\\
&= [f(x) + hF(h)]\,[g(x) + hG(h)] - f(x) \cdot g(x) & \text{expand}\\
&= f(x)g(x) + f(x) \cdot hG(h) + hF(h) \cdot g(x) + h^2 F(h) \cdot G(h) - f(x) \cdot g(x)\\
&= f(x) \cdot hG(h) + hF(h) \cdot g(x) + h^2 F(h) \cdot G(h).
\end{aligned}
$$

Armed with this we return to the definition of the derivative:

$$
\begin{aligned}
P'(x) &= \lim_{h \to 0} \frac{P(x+h) - P(x)}{h} \\
&= \lim_{h \to 0} \frac{f(x) \cdot hG(h) + hF(h) \cdot g(x) + h^2 F(h) \cdot G(h)}{h} \\
&= \left(\lim_{h \to 0} \frac{f(x) \cdot hG(h)}{h} \right) + \left(\lim_{h \to 0} \frac{hF(h) \cdot g(x)}{h} \right) + \left(\lim_{h \to 0} \frac{h^2 F(h) \cdot G(h)}{h} \right) \\
&= \left(\lim_{h \to 0} f(x) \cdot G(h) \right) + \left(\lim_{h \to 0} F(h) \cdot g(x) \right) + \left(\lim_{h \to 0} hF(h) \cdot G(h) \right)
\end{aligned}
$$

Now since $f(x)$ and $g(x)$ do not change as we send h to zero, we can pull them outside. We can also write the third term as the product of 3 limits:

$$
\begin{aligned}
&= \left(f(x) \lim_{h \to 0} G(h) \right) + \left(g(x) \lim_{h \to 0} F(h) \right) + \left(\lim_{h \to 0} h \right) \cdot \left(\lim_{h \to 0} F(h) \right) \cdot \left(\lim_{h \to 0} G(h) \right) \\
&= f(x) \cdot g'(x) + g(x) \cdot f'(x) + 0 \cdot f'(x) \cdot g'(x) \\
&= f(x) \cdot g'(x) + g(x) \cdot f'(x).
\end{aligned}
$$

And so we recover the product rule.

⯈ (optional) — Proof of the Quotient Rule (Theorem 2.4.5)

We now give the proof of the quotient rule in two steps[15]. We assume throughout that $g(x) \neq 0$ and that $f(x)$ and $g(x)$ are differentiable, meaning that the limits defining $f'(x)$, $g'(x)$ exist.

- In the first step, we prove the quotient rule under the assumption that $f(x)/g(x)$ is differentiable.

- In the second step, we prove that $1/g(x)$ differentiable. Once we know that $1/g(x)$ is differentiable, the product rule implies that $f(x)/g(x)$ is differentiable.

Step 1: the proof of the quotient rule assumng that $\frac{f(x)}{g(x)}$ is differentiable. Write $Q(x) = \frac{f(x)}{g(x)}$. Then $f(x) = g(x) \, Q(x)$ so that $f'(x) = g'(x) \, Q(x) + g(x) \, Q'(x)$, by the product rule, and

$$
\begin{aligned}
Q'(x) &= \frac{f'(x) - g'(x) \, Q(x)}{g(x)} = \frac{f'(x) - g'(x) \frac{f(x)}{g(x)}}{g(x)} \\
&= \frac{f'(x) g(x) - f(x) g'(x)}{g(x)^2}
\end{aligned}
$$

---

15   We thank Serban Raianu for suggesting this approach.

Step 2: the proof that $1/g(x)$ is differentiable. By definition

$$\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{g(x)} = \lim_{h\to 0}\frac{1}{h}\left[\frac{1}{g(x+h)} - \frac{1}{g(x)}\right] = \lim_{h\to 0}\frac{g(x)-g(x+h)}{h\,g(x)\,g(x+h)}$$

$$= -\lim_{h\to 0}\frac{1}{g(x)}\frac{1}{g(x+h)}\frac{g(x+h)-g(x)}{h}$$

$$= -\frac{1}{g(x)}\lim_{h\to 0}\frac{1}{g(x+h)}\lim_{h\to 0}\frac{g(x+h)-g(x)}{h}$$

$$= -\frac{1}{g(x)^2}g'(x)$$

2.6 ▲ Using the Arithmetic of Derivatives – Examples

In this section we illustrate the computation of derivatives using the arithmetic of derivatives — Theorems 2.4.2, 2.4.3 and 2.4.5. To make it clear which rules we are using during the examples we will note which theorem we are using:

- LIN to stand for "linearity" $\frac{\mathrm{d}}{\mathrm{d}x}\{\alpha\,f(x)+\beta\,g(x)\} = \alpha\,f'(x)+\beta\,g'(x)$ Theorem 2.4.2

- PR to stand for "product rule" $\frac{\mathrm{d}}{\mathrm{d}x}\{f(x)\,g(x)\} = f'(x)\,g(x)+f(x)\,g'(x)$ Theorem 2.4.3

- QR to stand for "quotient rule" $\frac{\mathrm{d}}{\mathrm{d}x}\left\{\frac{f(x)}{g(x)}\right\} = \frac{f'(x)\,g(x)-f(x)\,g'(x)}{g(x)^2}$ Theorem 2.4.5

We'll start with a really easy example.

Example 2.6.1

$$\frac{\mathrm{d}}{\mathrm{d}x}\{4x+7\} = 4\cdot\frac{\mathrm{d}}{\mathrm{d}x}\{x\}+7\cdot\frac{\mathrm{d}}{\mathrm{d}x}\{1\} \qquad\qquad \text{LIN}$$

$$= 4\cdot 1+7\cdot 0 = 4$$

where we have used LIN with $f(x)=x$, $g(x)=1$, $\alpha=4$, $\beta=7$.

Example 2.6.1

Example 2.6.2

Continuing on from the previous example, we can use the product rule and the previous result to compute

$$\frac{\mathrm{d}}{\mathrm{d}x}\{x(4x+7)\} = x\cdot\frac{\mathrm{d}}{\mathrm{d}x}\{4x+7\}+(4x+7)\frac{\mathrm{d}}{\mathrm{d}x}\{x\} \qquad\qquad \text{PR}$$

$$= x\cdot 4+(4x+7)\cdot 1$$

$$= 8x+7$$

where we have used the product rule PR with $f(x)=x$ and $g(x)=4x+7$.

Example 2.6.3

In the same vein as the previous example, we can use the quotient rule to compute

$$\frac{\mathrm{d}}{\mathrm{d}x}\left\{\frac{x}{4x+7}\right\} = \frac{(4x+7)\cdot\frac{\mathrm{d}}{\mathrm{d}x}\{x\} - x\cdot\frac{\mathrm{d}}{\mathrm{d}x}\{4x+7\}}{(4x+7)^2} \qquad \text{QR}$$

$$= \frac{(4x+7)\cdot 1 - x\cdot 4}{(4x+7)^2}$$

$$= \frac{7}{(4x+7)^2}$$

where we have used the quotient rule QR with $f(x) = x$ and $g(x) = 4x + 7$.

Example 2.6.3

Now for a messier example.

Example 2.6.4

Differentiate

$$f(x) = \frac{x}{2x + \frac{1}{3x+1}}$$

This problem looks nasty. But it isn't so hard if we just build it up a bit at a time.

- First, $f(x)$ is the ratio of

$$f_1(x) = x \qquad \text{and} \qquad f_2(x) = 2x + \frac{1}{3x+1}$$

If we can find the derivatives of $f_1(x)$ and $f_2(x)$, we will be able to get the derivative of $f(x)$ just by applying the quotient rule. The derivative, $f_1'(x) = 1$, of $f_1(x)$ is easy, so let's work on $f_2(x)$.

- The function $f_2(x)$ is the linear combination

$$f_2(x) = 2f_3(x) + f_4(x) \qquad \text{with} \qquad f_3(x) = x \quad \text{and} \quad f_4(x) = \frac{1}{3x+1}$$

If we can find the derivatives of $f_3(x)$ and $f_4(x)$, we will be able to get the derivative of $f_2(x)$ just by applying linearity (Theorem 2.4.2). The derivative, $f_3'(x) = 1$, of $f_3(x)$ is easy. So let's work of $f_4(x)$.

- The function $f_4(x)$ is the ratio

$$f_4(x) = \frac{1}{f_5(x)} \qquad \text{with} \qquad f_5(x) = 3x + 1$$

If we can find the derivative of $f_5(x)$, we will be able to get the derivative of $f_4(x)$ just by applying the special case the quotient rule (Corollary 2.4.6). The derivative of $f_5(x)$ is easy.

- So we have completed breaking down $f(x)$ into easy pieces. It is now just a matter of reversing the break down steps, putting everything back together, starting with the easy pieces and working up to $f(x)$. Here goes.

$$f_5(x) = 3x + 1 \qquad \text{so } \frac{d}{dx}f_5(x) = 3\frac{d}{dx}x + \frac{d}{dx}1 = 3 \cdot 1 + 0 = 3 \qquad \text{LIN}$$

$$f_4(x) = \frac{1}{f_5(x)} \qquad \text{so } \frac{d}{dx}f_4(x) = -\frac{f_5'(x)}{f_5(x)^2} = -\frac{3}{(3x+1)^2} \qquad \text{QR}$$

$$f_2(x) = 2f_3(x) + f_4(x) \qquad \text{so } \frac{d}{dx}f_2(x) = 2f_3'(x) + f_4'(x) = 2 - \frac{3}{(3x+1)^2} \qquad \text{LIN}$$

$$f(x) = \frac{f_1(x)}{f_2(x)} \qquad \text{so } \frac{d}{dx}f(x) = \frac{f_1'(x)f_2(x) - f_1(x)f_2'(x)}{f_2(x)^2} \qquad \text{QR}$$

$$= \frac{1\left[2x + \frac{1}{3x+1}\right] - x\left[2 - \frac{3}{(3x+1)^2}\right]}{\left[2x + \frac{1}{3x+1}\right]^2}$$

Oof!

- We now have an answer. But we really should clean it up, not only to make it easier to read, but also because invariably such computations are just small steps inside much larger computations. Any future computations involving this expression will be a lot easier and less error prone if we clean it up now. Cancelling the $2x$ and the $-2x$ in

$$1\left[2x + \frac{1}{3x+1}\right] - x\left[2 - \frac{3}{(3x+1)^2}\right] = 2x + \frac{1}{3x+1} - 2x + \frac{3x}{(3x+1)^2}$$
$$= \frac{1}{3x+1} + \frac{3x}{(3x+1)^2}$$

and multiplying both the numerator and denominator by $(3x+1)^2$ gives

$$f'(x) = \frac{\frac{1}{3x+1} + \frac{3x}{(3x+1)^2}}{\left[2x + \frac{1}{3x+1}\right]^2} \frac{(3x+1)^2}{(3x+1)^2}$$
$$= \frac{(3x+1) + 3x}{\left[2x(3x+1) + 1\right]^2}$$
$$= \frac{6x+1}{\left[6x^2 + 2x + 1\right]^2}.$$

Example 2.6.4

While the linearity theorem (Theorem 2.4.2) is stated for a linear combination of two functions, it is not difficult to extend it to linear combinations of three or more functions

as the following example shows.

Example 2.6.5

We'll start by generalising linearity to three functions.

$$\frac{\mathrm{d}}{\mathrm{d}x}\{aF(x) + bG(x) + cH(x)\} = \frac{\mathrm{d}}{\mathrm{d}x}\{\, a \cdot [F(x)] \,+\, 1 \cdot [bG(x) + cH(x)] \,\}$$

$$= aF'(x) + \frac{\mathrm{d}}{\mathrm{d}x}\{bG(x) + cH(x)\}$$

by LIN with $\alpha = a$, $f(x) = F(x)$, $\beta = 1$,
and $g(x) = bG(x) + cH(x)$

$$= aF'(x) + bG'(x) + cH'(x)$$

by LIN with $\alpha = b$, $f(x) = G(x)$, $\beta = c$,
and $g(x) = H(x)$

This gives us linearity for three terms, namely (just replacing upper case names by lower case names)

$$\frac{\mathrm{d}}{\mathrm{d}x}\{af(x) + bg(x) + ch(x)\} = af'(x) + bg'(x) + ch'(x)$$

Just by repeating the above argument many times, we may generalise to linearity for n terms, for any natural number n:

$$\frac{\mathrm{d}}{\mathrm{d}x}\{a_1 f_1(x) + a_2 f_2(x) + \cdots + a_n f_n(x)\} = a_1 f'_1(x) + a_2 f'_2(x) + \cdots + a_n f'_n(x)$$

Example 2.6.5

Similarly, while the product rule is stated for the product of two functions, it is not difficult to extend it to the product of three or more functions as the following example shows.

Example 2.6.6

Once again, we'll start by generalising the product rule to three factors.

$$\frac{\mathrm{d}}{\mathrm{d}x}\{F(x)\,G(x)\,H(x)\} = F'(x)\,G(x)\,H(x) + F(x)\,\frac{\mathrm{d}}{\mathrm{d}x}\{G(x)\,H(x)\}$$

by PR with $f(x) = F(x)$ and $g(x) = G(x)H(x)$

$$= F'(x)\,G(x)\,H(x) + F(x)\,\{G'(x)\,H(x) + G(x)\,H'(x)\}$$

by PR with $f(x) = G(x)$ and $g(x) = H(x)$

This gives us a product rule for three factors, namely (just replacing upper case names by lower case names)

$$\frac{\mathrm{d}}{\mathrm{d}x}\{f(x)\,g(x)\,h(x)\} = f'(x)\,g(x)\,h(x) + f(x)\,g'(x)\,h(x) + f(x)\,g(x)\,h'(x)$$

Observe that when we differentiate a product of three factors, the answer is a sum of three terms and in each term the derivative acts on exactly one of the original factors. Just by repeating the above argument many times, we may generalise the product rule to give the derivative of a product of n factors, for any natural number n:

$$\frac{d}{dx}\{f_1(x)\,f_2(x)\cdots f_n(x)\} = \quad f_1'(x)\,f_2(x)\cdots f_n(x)$$
$$+ f_1(x)\,f_2'(x)\cdots f_n(x)$$
$$\vdots$$
$$+ f_1(x)\,f_2(x)\cdots f_n'(x)$$

We can also write the above as

$$\frac{d}{dx}\{f_1(x)\,f_2(x)\cdots f_n(x)\} = \left[\frac{f_1'(x)}{f_1(x)} + \frac{f_2'(x)}{f_2(x)} + \cdots + \frac{f_n'(x)}{f_n(x)}\right] \cdot f_1(x)\,f_2(x)\cdots f_n(x)$$

When we differentiate a product of n factors, the answer is a sum of n terms and in each term the derivative acts on exactly one of the original factors. In the first term, the derivative acts on the first of the original factors. In the second term, the derivative acts on the second of the original factors. And so on.

If we make $f_1(x) = f_2(x) = \cdots = f_n(x) = f(x)$ then each of the n terms on the right hand side of the above equation is the product of $f'(x)$ and exactly $n-1$ $f(x)$'s, and so is exactly $f(x)^{n-1}\,f'(x)$. So we get the following useful result

$$\frac{d}{dx}f(x)^n = n \cdot f(x)^{n-1} \cdot f'(x).$$

Example 2.6.6

This last result is quite useful, so let us write it as a lemma for future reference.

Lemma 2.6.7.

Let n be a natural number and f be a differentiable function. Then

$$\frac{d}{dx}f(x)^n = n \cdot f(x)^{n-1} \cdot f'(x).$$

This immediately gives us another useful result.

Example 2.6.8

We can now compute the derivative of x^n for any natural number n. Start with Lemma 2.6.7 and substitute $f(x) = x$ and $f'(x) = 1$:

$$\frac{d}{dx}x^n = n \cdot x^{n-1} \cdot 1 = n\,x^{n-1}$$

Example 2.6.8

Again — this is a result we will come back to quite a few times in the future, so we should make sure we can refer to it easily. However, at present this statement only holds when n is a positive integer. With a little more work we can extend this to compute x^q where q is any positive rational number and then any rational number at all (positive or negative). So let us hold off for a little longer. Instead we can make it a lemma, since it will be an ingredient in quite a few of the examples following below and in constructing the final corollary.

Lemma 2.6.9 (Derivative of x^n).

Let n be a positive integer then

$$\frac{d}{dx}x^n = nx^{n-1} \tag{2.6.1}$$

Back to more examples.

Example 2.6.10

$$\frac{d}{dx}\{2x^3 + 4x^5\} = 2\frac{d}{dx}\{x^3\} + 4\frac{d}{dx}\{x^5\}$$

by LIN with $\alpha = 2$, $f(x) = x^3$, $\beta = 4$, and $g(x) = x^5$

$$= 2\{3x^2\} + 4\{5x^4\}$$

by Lemma 2.6.9, once with $n = 3$, and once with $n = 5$

$$= 6x^2 + 20x^4$$

Example 2.6.10

Example 2.6.11

In this example we'll compute $\frac{d}{dx}\{(3x + 9)(x^2 + 4x^3)\}$ in two different ways. For the first, we'll start with the product rule.

$$\frac{d}{dx}\{(3x + 9)(x^2 + 4x^3)\} = \left\{\frac{d}{dx}(3x + 9)\right\}(x^2 + 4x^3) + (3x + 9)\frac{d}{dx}\{x^2 + 4x^3\}$$

$$= \{3 \times 1 + 9 \times 0\}(x^2 + 4x^3) + (3x + 9)\{2x + 4(3x^2)\}$$

$$= 3(x^2 + 4x^3) + (3x + 9)(2x + 12x^2)$$

$$= 3x^2 + 12x^3 + (6x^2 + 18x + 36x^3 + 108x^2)$$

$$= 18x + 117x^2 + 48x^3$$

For the second, we expand the product first and then differentiate.

$$\frac{d}{dx}\{(3x+9)(x^2+4x^3)\} = \frac{d}{dx}\{9x^2+39x^3+12x^4\}$$
$$= 9(2x)+39(3x^2)+12(4x^3)$$
$$= 18x+117x^2+48x^3$$

Example 2.6.11

Example 2.6.12

$$\frac{d}{dx}\left\{\frac{4x^3-7x}{4x^2+1}\right\} = \frac{(12x^2-7)(4x^2+1)-(4x^3-7x)(8x)}{(4x^2+1)^2}$$
$$\text{by QR with } f(x)=4x^3-7x,\ f'(x)=12x^2-7,$$
$$\text{and } g(x)=4x^2+1,\ g'(x)=8x$$
$$= \frac{(48x^4-16x^2-7)-(32x^4-56x^2)}{(4x^2+1)^2}$$
$$= \frac{16x^4+40x^2-7}{(4x^2+1)^2}$$

Example 2.6.12

Example 2.6.13

In this example, we'll use a little trickery to find the derivative of $\sqrt[3]{x}$. The trickery consists of observing that, by the definition of the cube root,

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

Since both sides of the expression are the same, they must have the same derivatives:

$$\frac{d}{dx}\{x\} = \frac{d}{dx}\left(\sqrt[3]{x}\right)^3.$$

We already know by Theorem 2.2.4 that

$$\frac{d}{dx}\{x\} = 1$$

and that, by Lemma 2.6.7 with $n=3$ and $f(x)=\sqrt[3]{x}$,

$$\frac{d}{dx}\left(\sqrt[3]{x}\right)^3 = 3\left(\sqrt[3]{x}\right)^2 \cdot \frac{d}{dx}\{\sqrt[3]{x}\} = 3\,x^{2/3} \cdot \frac{d}{dx}\{\sqrt[3]{x}\}.$$

Since we know that $\frac{d}{dx}\{x\} = \frac{d}{dx}(\sqrt[3]{x})^3$, we must have

$$1 = 3x^{2/3} \cdot \frac{d}{dx}\{\sqrt[3]{x}\}$$

which we can rearrange to give the result we need

$$\frac{d}{dx}\{\sqrt[3]{x}\} = \tfrac{1}{3}x^{-2/3}$$

Example 2.6.13

Example 2.6.14

In this example, we'll use the same trickery as in Example 2.6.13 to find the derivative $x^{p/q}$ for any two natural numbers p and q. By definition of the qth root,

$$x^p = \left(x^{p/q}\right)^q.$$

That is, x^p and $\left(x^{p/q}\right)^q$ are the same function, and so have the same derivative. So we differentiate both of them. We already know that, by Lemma 2.6.9 with $n = p$,

$$\frac{d}{dx}\{x^p\} = px^{p-1}$$

and that, by Lemma 2.6.7 with $n = q$ and $f(x) = x^{p/q}$,

$$\frac{d}{dx}\{(x^{p/q})^q\} = q\,(x^{p/q})^{q-1}\frac{d}{dx}\{x^{p/q}\}$$

Remember that $(x^a)^b = x^{(a \cdot b)}$. Now these two derivatives must be the same. So

$$px^{p-1} = q \cdot x^{(pq-p)/q}\frac{d}{dx}\{x^{p/q}\}$$

and, rearranging things,

$$\frac{d}{dx}\{x^{p/q}\} = \frac{p}{q}x^{p-1-(pq-p)/q}$$

$$= \frac{p}{q}x^{(pq-q-pq+p)/q}$$

$$= \frac{p}{q}x^{p/q-1}$$

So finally

$$\frac{d}{dx}\left\{x^{p/q}\right\} = \frac{p}{q}x^{p/q-1} \tag{2.6.2}$$

Notice that this has the same form as Lemma 2.6.9, above, except with $n = p/q$ allowed to be any positive rational number, not just a positive integer.

Example 2.6.14

Example 2.6.15 (Derivative of x^{-m})

In this example we'll use the quotient rule to find the derivative of x^{-m}, for any natural number m.

By the special case of the quotient rule (Corollary 2.4.6) with $g(x) = x^m$ and $g'(x) = mx^{m-1}$

$$\frac{d}{dx}\{x^{-m}\} = \frac{d}{dx}\left\{\frac{1}{x^m}\right\} = -\frac{mx^{m-1}}{(x^m)^2} = -mx^{-m-1}$$

Again, notice that this has the same form as Lemma 2.6.9, above, except with $n = -m$ being a negative integer.

Example 2.6.15

Example 2.6.16

In this example we'll use the quotient rule to find the derivative of $x^{-p/q}$, for any pair of natural numbers p and q. By the special case the quotient rule (Corollary 2.4.6) with $g(x) = x^{p/q}$ and $g'(x) = \frac{p}{q}x^{p/q-1}$,

$$\frac{d}{dx}\{x^{-p/q}\} = \frac{d}{dx}\left\{\frac{1}{x^{p/q}}\right\} = -\frac{\frac{p}{q}x^{p/q-1}}{(x^{p/q})^2} = -\frac{p}{q}x^{-p/q-1}$$

Example 2.6.16

Note that we have found, in Examples 2.2.2, 2.6.14 and 2.6.16, the derivative of x^a for any rational number a, whether 0, positive, negative, integer or fractional. In all cases, the answer is

Corollary 2.6.17 (Derivative of x^a).

Let a be a rational number, then

$$\frac{d}{dx}x^a = ax^{a-1} \qquad (2.6.3)$$

We shall show, in Example 2.10.5, that the formula $\frac{d}{dx}x^a = ax^{a-1}$ in fact applies for all real numbers a, not just rational numbers.

Back in Example 2.2.9 we computed the derivative of $\sqrt{x}$ from the definition of the derivative. The above corollary (correctly) gives

$$\frac{d}{dx}x^{1/2} = \frac{1}{2}x^{-1/2}$$

but with far less work.

Here's an (optional) messy example.

Example 2.6.18 (Optional messy example)

Find the derivative of

$$f(x) = \frac{(\sqrt{x}-1)(2-x)(1-x^2)}{\sqrt{x}(3+2x)}$$

- As we seen before, the best strategy for dealing with nasty expressions is to break them up into easy pieces. We can think of $f(x)$ as the five–fold product

$$f(x) = f_1(x) \cdot f_2(x) \cdot f_3(x) \cdot \frac{1}{f_4(x)} \cdot \frac{1}{f_5(x)}$$

with

$$f_1(x) = \sqrt{x}-1 \quad f_2(x) = 2-x \quad f_3(x) = 1-x^2 \quad f_4(x) = \sqrt{x} \quad f_5(x) = 3+2x$$

- By now, the derivatives of the f_j's should be easy to find:

$$f_1'(x) = \frac{1}{2\sqrt{x}} \quad f_2'(x) = -1 \quad f_3'(x) = -2x \quad f_4'(x) = \frac{1}{2\sqrt{x}} \quad f_5'(x) = 2$$

- Now, to get the derivative $f(x)$ we use the n–fold product rule which was developed in Example 2.6.6, together with the special case of the quotient rule (Corollary 2.4.6).

$$f'(x) = f_1' f_2 f_3 \frac{1}{f_4}\frac{1}{f_5} + f_1 f_2' f_3 \frac{1}{f_4}\frac{1}{f_5} + f_1 f_2 f_3' \frac{1}{f_4}\frac{1}{f_5} - f_1 f_2 f_3 \frac{f_4'}{f_4^2}\frac{1}{f_5} - f_1 f_2 f_3 \frac{1}{f_4}\frac{f_5'}{f_5^2}$$

$$= \left[\frac{f_1'}{f_1} + \frac{f_2'}{f_2} + \frac{f_3'}{f_3} - \frac{f_4'}{f_4} - \frac{f_5'}{f_5}\right] f_1 f_2 f_3 \frac{1}{f_4}\frac{1}{f_5}$$

$$= \left[\frac{1}{2\sqrt{x}(\sqrt{x}-1)} - \frac{1}{2-x} - \frac{2x}{1-x^2} - \frac{1}{2x} - \frac{2}{3+2x}\right] \frac{(\sqrt{x}-1)(2-x)(1-x^2)}{\sqrt{x}(3+2x)}$$

The trick that we used in going from the first line to the second line, namely multiplying term number j by $\frac{f_j(x)}{f_j(x)}$ is often useful in simplifying the derivative of a product of many factors[16].

Example 2.6.18

---

16 Also take a look at "logarithmic differentiation" in Section 2.10.

2.7 ▴ Derivatives of Exponential Functions

Now that we understand how derivatives interact with products and quotients, we are able to compute derivatives of

- polynomials,

- rational functions, and

- powers and roots of rational functions.

Notice that all of the above come from knowing[17] the derivative of x^n and applying linearity of derivatives and the product rule.

There is still one more "rule" that we need to complete our toolbox and that is the chain rule. However before we get there, we will add a few functions to our list of things we can differentiate[18]. The first of these is the exponential function.

Let $a > 0$ and set $f(x) = a^x$ — this is what is known as an exponential function. Let's see what happens when we try to compute the derivative of this function just using the definition of the derivative.

$$\frac{df}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h \to 0} \frac{a^{x+h} - a^x}{h}$$

$$= \lim_{h \to 0} a^x \cdot \frac{a^h - 1}{h} = a^x \cdot \lim_{h \to 0} \frac{a^h - 1}{h}$$

Unfortunately we cannot complete this computation because we cannot evaluate the last limit directly. For the moment, let us assume this limit exists and name it

$$C(a) = \lim_{h \to 0} \frac{a^h - 1}{h}$$

It depends only on a and is completely independent of x. Using this notation (which we will quickly improve upon below), our desired derivative is now

$$\frac{d}{dx} a^x = C(a) \cdot a^x.$$

Thus the derivative of a^x is a^x multiplied by some constant — i.e. the function a^x is nearly unchanged by differentiating. If we can tune a so that $C(a) = 1$ then the derivative would just be the original function! This turns out to be very useful.

To try finding an a that obeys $C(a) = 1$, let us investigate how $C(a)$ changes with a. Unfortunately (though this fact is not at all obvious) there is no way to write $C(a)$ as a finite combination of any of the functions we have examined so far[19]. To get started, we'll try to guess $C(a)$, for a few values of a, by plugging in some small values of h.

---

17 Differentiating powers and roots of functions is actually quite a bit easier once one knows the chain rule — which we will discuss soon.

18 One reason we add these functions is that they interact very nicely with the derivative. Another reason is that they turn up in many "real world" examples.

19 To a bit more be precise, we say that a number q is algebraic if we can write q as the zero of a polynomial with integer coefficients. When a is any positive algebraic number other than 1, $C(a)$ is not algebraic. A number that is not algebraic is called transcendental. The best known example of a transcendental number is π (which follows from the Lindemann-Weierstrass Theorem — way beyond the scope of this course).

Example 2.7.1

Let $a = 1$ then $C(1) = \lim_{h \to 0} \dfrac{1^h - 1}{h} = 0$. This is not surprising since $1^x = 1$ is constant, and

so its derivative must be zero everywhere. Let $a = 2$ then $C(2) = \lim_{h \to 0} \dfrac{2^h - 1}{h}$. Setting h to smaller and smaller numbers gives

h	0.1	0.01	0.001	0.0001	0.00001	0.000001	0.0000001
$\frac{2^h-1}{h}$	0.7177	0.6956	0.6934	0.6932	0.6931	0.6931	0.6931

Similarly when $a = 3$ we get

h	0.1	0.01	0.001	0.0001	0.00001	0.000001	0.0000001
$\frac{3^h-1}{h}$	1.1612	1.1047	1.0992	1.0987	1.0986	1.0986	1.0986

and $a = 10$

h	0.1	0.01	0.001	0.0001	0.00001	0.000001	0.0000001
$\frac{10^h-1}{h}$	2.5893	2.3293	2.3052	2.3028	2.3026	2.3026	2.3026

From this example it appears that $C(a)$ increases as we increase a, and that $C(a) = 1$ for some value of a between 2 and 3.

Example 2.7.1

We can learn a lot more about $C(a)$, and, in particular, confirm the guesses that we made in the last example, by making use of logarithms — this would be a good time for you to review them.

## ⇥ Whirlwind Review of Logarithms

Before you read much further into this little review on logarithms, you should first go back and take a look at the review of inverse functions in Section 0.6.

▸▸▸ Logarithmic Functions

We are about to define the "logarithm with base q". In principle, q is allowed to be any strictly positive real number, except $q = 1$. However we shall restrict our attention to $q > 1$, because, in practice, the only q's that are ever used are e (a number that we shall define in the next few pages), 10 and, if you are a computer scientist, 2. So, fix any $q > 1$ (if you like, pretend that $q = 10$). The function $f(x) = q^x$

- increases as x increases (for example if $x' > x$, then $10^{x'} = 10^x \cdot 10^{x'-x} > 10^x$ since $10^{x'-x} > 1$)

- obeys $\lim_{x \to -\infty} q^x = 0$ (for example 10^{-1000} is really small) and

- obeys $\lim_{x \to +\infty} q^x = +\infty$ (for example 10^{+1000} is really big).

Consequently, for any $0 < Y < \infty$, the horizontal straight line $y = Y$ crosses the graph of $y = f(x) = q^x$ at exactly one point, as illustrated in the figure below. The x–coordinate
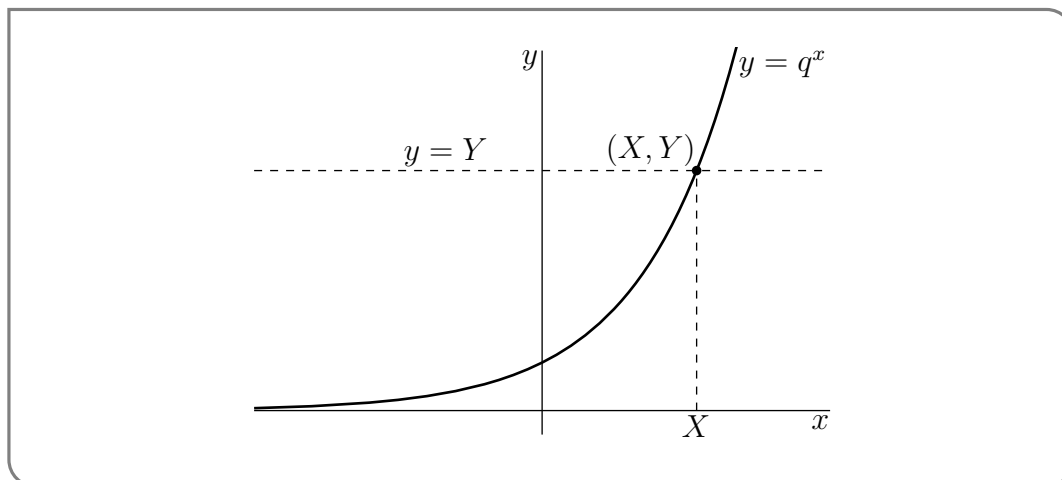
of that intersection point, denoted X in the figure, is $\log_q(Y)$. So $\log_q(Y)$ is the power to which you have to raise q to get Y. It is the inverse function of $f(x) = q^x$. Of course we are free to rename the dummy variables X and Y. If, for example, we wish to graph our logarithm function, it is natural to rename $Y \to x$ and $X \to y$, giving

Definition 2.7.2.

Let $q > 1$. Then the logarithm with base q is defined[20] by

$$y = \log_q(x) \Leftrightarrow x = q^y$$

Obviously the power to which we have to raise q to get q^x is x, so we have both

$$\log_q(q^x) = x \qquad\qquad\qquad q^{\log_q(x)} = x$$

From the exponential properties

$$q^{\log_q(xy)} = xy = q^{\log_q(x)}q^{\log_q(y)} = q^{\log_q(x)+\log_q(y)}$$
$$q^{\log_q(x/y)} = x/y = q^{\log_q(x)}/q^{\log_q(y)} = q^{\log_q(x)-\log_q(y)}$$
$$q^{\log_q(x^r)} = x^r = \left(q^{\log_q(x)}\right)^r = q^{r\log_q(x)}$$

---

20 We can also define logarithms with base $0 < r < 1$ but doing so is not necessary. To see this, set $q = 1/r > 1$. Then it is reasonable to define $\log_r(x) = -\log_q(x)$ since

$$r^{\log_r(x)} = \left(\frac{1}{q}\right)^{\log_r(x)} = \left(\frac{1}{q}\right)^{-\log_q(x)} = q^{\log_q(x)} = x$$

as required.

we have

$$\log_q(xy) = \log_q(x) + \log_q(y)$$
$$\log_q(x/y) = \log_q(x) - \log_q(y)$$
$$\log_q(x^r) = r \log_q(x)$$

Can we convert from logarithms in one base to logarithms in another? For example, if our calculator computes logarithms base 10 for us (which it very likely does), can we also use it to compute a logarithm base q? Yes, using

$$\log_q(x) = \frac{\log_{10} x}{\log_{10} q}$$

How did we get this? Well, let's start with a number x and suppose that we want to compute

$$y = \log_q x$$

We can rearrange this by exponentiating both sides

$$q^y = q^{\log_q x} = x$$

Now take log base 10 of both sides

$$\log_{10} q^y = \log_{10} x$$

But recall that $\log_q(x^r) = r \log_q(x)$, so

$$y \log_{10} q = \log_{10} x$$
$$y = \frac{\log_{10} x}{\log_{10} q}$$

▸▸ Back to that Limit

Recall that we are trying to choose a so that

$$\lim_{h \to 0} \frac{a^h - 1}{h} = C(a) = 1.$$

We can estimate the correct value of a by using our numerical estimate of $C(10)$ above. The way to do this is to first rewrite $C(a)$ in terms of logarithms.

$$a = 10^{\log_{10} a} \qquad \text{and so} \qquad a^h = 10^{h \log_{10} a}.$$
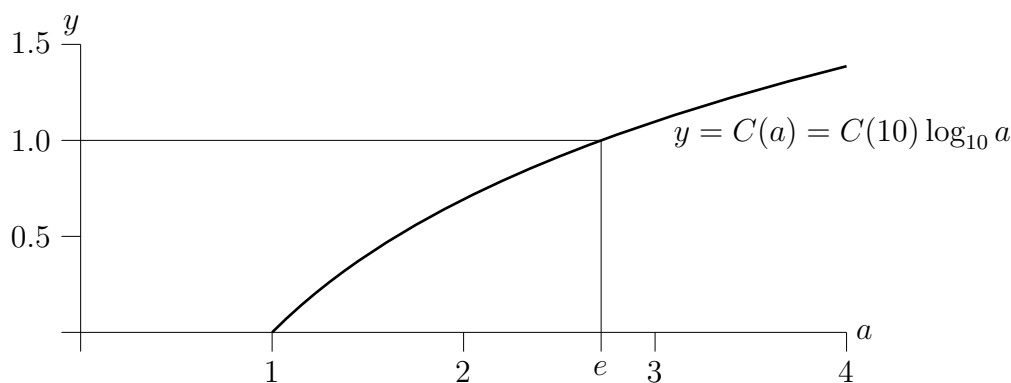
Using this we rewrite $C(a)$ as

$$C(a) = \lim_{h \to 0} \frac{1}{h} \left(10^{h \log_{10} a} - 1 \right)$$

Now set $H = h \log_{10}(a)$, and notice that as $h \to 0$ we also have $H \to 0$

$$= \lim_{H \to 0} \frac{\log_{10} a}{H} \left(10^H - 1 \right)$$

$$= \log_{10} a \cdot \lim_{H \to 0} \frac{10^H - 1}{H}$$

$$= \log_{10} a \cdot C(10).$$

Below is a sketch of $C(a)$ against a.

Figure 2.7.1.

$$y = C(a) = C(10) \log_{10} a$$

Remember that we are trying to find an a with $C(a) = 1$. We can do so by recognising that $C(a) = C(10) (\log_{10} a)$ has the following properties.

- When $a = 1$, $\log_{10}(a) = \log_{10} 1 = 0$ so that $C(a) = C(10) \log_{10}(a) = 0$. Of course, we should have expected this, because when $a = 1$ we have $a^x = 1^x = 1$ which is just the constant function and $\frac{d}{dx} 1 = 0$.

- $\log_{10} a$ increases as a increases, and hence $C(a) = C(10) \log_{10} a$ increases as a increases.

- $\log_{10} a$ tends to $+\infty$ as $a \to \infty$, and hence $C(a)$ tends to $+\infty$ as $a \to \infty$.

Hence the graph of $C(a)$ passes through $(1, 0)$, is always increasing as a increases and goes off to $+\infty$ as a goes off to $+\infty$. See Figure 2.7.1. Consequently[21] there is exactly one value of a for which $C(a) = 1$.

The value of a for which $C(a) = 1$ is given the name e. It is called Euler's constant[22].

---

21 We are applying the Intermediate Value Theorem here, but we have neglected to verify the hypothesis that $\log_{10}(a)$ is a continuous function. Please forgive us — we could do this if we really had to, but it would make a big mess without adding much understanding, if we were to do so here in the text. Better to just trust us on this.

22 Unfortunately there is another Euler's constant, γ, which is more properly called the Euler–Mascheroni constant. Anyway like many mathematical discoveries, e was first found by someone else — Napier used the constant e in order to compute logarithms but only implicitly. Bernoulli was probably the first to approximate it when examining continuous compound interest. It first appeared explicitly in work of Leibniz, though he denoted it b. It was Euler, though, who established the notation we now use and who showed how important the constant is to mathematics.

In Example 2.7.1, we estimated $C(10) \approx 2.3026$. So if we assume $C(a) = 1$ then the above equation becomes

$$2.3026 \cdot \log_{10} a \approx 1$$
$$\log_{10} a \approx \frac{1}{2.3026} \approx 0.4343$$
$$a \approx 10^{0.4343} \approx 2.7813$$

This gives us the estimate $a \approx 2.7813$ which is not too bad. In fact[23]

Equation 2.7.3 (Euler's constant).

$$e = 2.7182818284590452354\ldots$$
$$= 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots$$

We will be able to explain this last formula once we develop Taylor polynomials later in the course.

To summarise

Theorem 2.7.4.

The constant e is the unique real number that satisfies
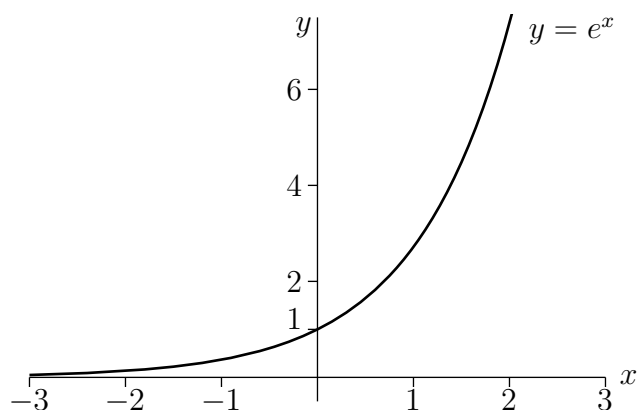
$$\lim_{h \to 0} \frac{e^h - 1}{h} = 1$$

Further,

$$\frac{d}{dx}(e^x) = e^x$$

We plot e^x in the graph below

---

23   Recall n factorial, written $n!$ is the product $n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$.

Figure 2.7.2.

And just a reminder of some of its[24] properties...

1. $e^0 = 1$

2. $e^{x+y} = e^x e^y$

3. $e^{-x} = \frac{1}{e^x}$

4. $\left(e^x\right)^y = e^{xy}$

5. $\lim\limits_{x \to \infty} e^x = \infty$, $\lim\limits_{x \to -\infty} e^x = 0$

Now consider again the problem of differentiating a^x. We saw above that

$$\frac{d}{dx} a^x = C(a) \cdot a^x \quad \text{and } C(a) = C(10) \cdot \log_{10} a \quad \text{which gives} \frac{d}{dx} a^x = C(10) \cdot \log_{10} a \cdot a^x$$

We can eliminate the $C(10)$ term with a little care. Since we know that $\frac{d}{dx} e^x = e^x$, we have $C(e) = 1$. This allows us to express

$$1 = C(e) = C(10) \cdot \log_{10} e \qquad \text{and so}$$
$$C(10) = \frac{1}{\log_{10} e}$$

Putting things back together gives

$$\frac{d}{dx} a^x = \frac{\log_{10} a}{\log_{10} e} \cdot a^x$$
$$= \log_e a \cdot a^x.$$

---

24 "The function e^x is of course the special case of the function a^x with $a = e$. So it inherits all the usual algebraic properties of a^x."

There is more than one way to get to this result. For example, let $f(x) = a^x$, then

$$\log_e f(x) = x \log_e a$$
$$f(x) = e^{x \log_e a}$$

So if we write $g(x) = e^x$ then we are really attempting to differentiate the function

$$\frac{df}{dx} = \frac{d}{dx} g(x \cdot \log_e a).$$

In order to compute this derivative we need to know how to differentiate

$$\frac{d}{dx} g(qx)$$

where q is a constant. We'll hold off on learning this for the moment until we have introduced the chain rule (see Section 2.9 and in particular Corollary 2.9.9). Similarly we'd like to know how to differentiate logarithms — again this has to wait until we have learned the chain rule.

Notice that the derivatives

$$\frac{d}{dx} x^n = nx^{n-1} \qquad \text{and} \qquad \frac{d}{dx} e^x = e^x$$

are either nearly unchanged or actually unchanged by differentiating. It turns out that some of the trigonometric functions also have this property of being "nearly unchanged" by differentiation. That brings us to the next section.

2.8 ▲ Derivatives of Trigonometric Functions

We are now going to compute the derivatives of the various trigonometric functions, $\sin x$, $\cos x$ and so on. The computations are more involved than the others that we have done so far and will take several steps. Fortunately, the final answers will be very simple.

Observe that we only need to work out the derivatives of $\sin x$ and $\cos x$, since the other trigonometric functions are really just quotients of these two functions. Recall:

$$\tan x = \frac{\sin x}{\cos x} \qquad \cot x = \frac{\cos x}{\sin x} \qquad \csc x = \frac{1}{\sin x} \qquad \sec x = \frac{1}{\cos x}.$$

The first steps towards computing the derivatives of $\sin x, \cos x$ is to find their derivatives at $x = 0$. The derivatives at general points x will follow quickly from these, using trig identities. It is important to note that we must measure angles in radians[25], rather than degrees, in what follows. Indeed — unless explicitly stated otherwise, any number that is put into a trigonometric function is measured in radians.

---

25 In science, radians is the standard unit for measuring angles. While you may be more familiar with degrees, radians should be used in any computation involving calculus. Using degrees will cause errors. Thankfully it is easy to translate between these two measures since $360° = 2\pi$ radians. See Appendix B.2.1.

▸▸ These Proofs are Optional, the Results are Not.

While we expect you to read and follow these proofs, we do not expect you to be able to reproduce them. You will be required to know the results, in particular Theorem 2.8.5 below.

▸▸ Step 1: $\frac{d}{dx}\{\sin x\}\big|_{x=0}$

By definition, the derivative of $\sin x$ evaluated at $x = 0$ is

$$\frac{d}{dx}\{\sin x\}\bigg|_{x=0} = \lim_{h\to 0}\frac{\sin h - \sin 0}{h} = \lim_{h\to 0}\frac{\sin h}{h}$$

We will prove this limit by use of the squeeze theorem (Theorem 1.4.17). To get there we will first need to do some geometry. But first we will build some intuition.

The figure below contains part of a circle of radius 1. Recall that an arc of length h on such a circle subtends an angle of h **radians** at the centre of the circle. So the darkened arc in the figure has length h and the darkened vertical line in the figure has length $\sin h$. We must determine what happens to the ratio of the lengths of the darkened vertical line and darkened arc as h tends to zero.
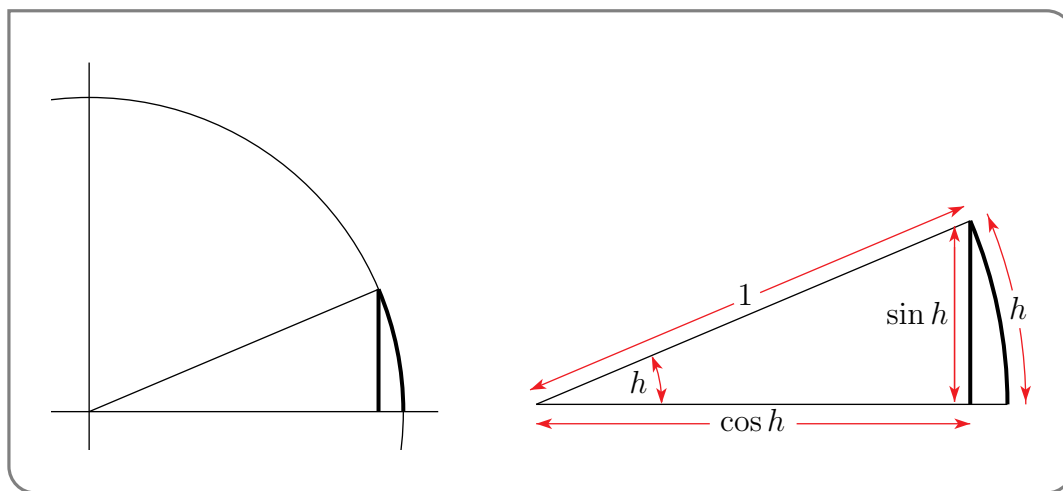
Here is a magnified version of the part of the above figure that contains the darkened arc and vertical line.
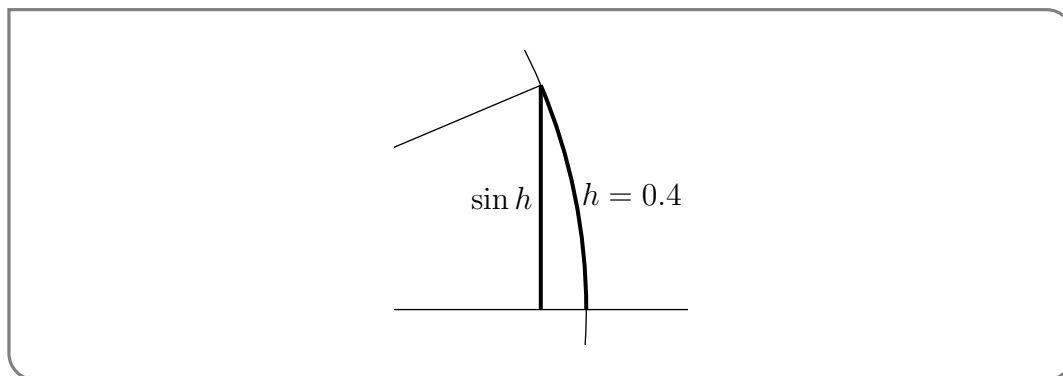
This particular figure has been drawn with $h = .4$ radians. Here are three more such blow ups. In each successive figure, the value of h is smaller. To make the figures clearer, the degree of magnification was increased each time h was decreased.
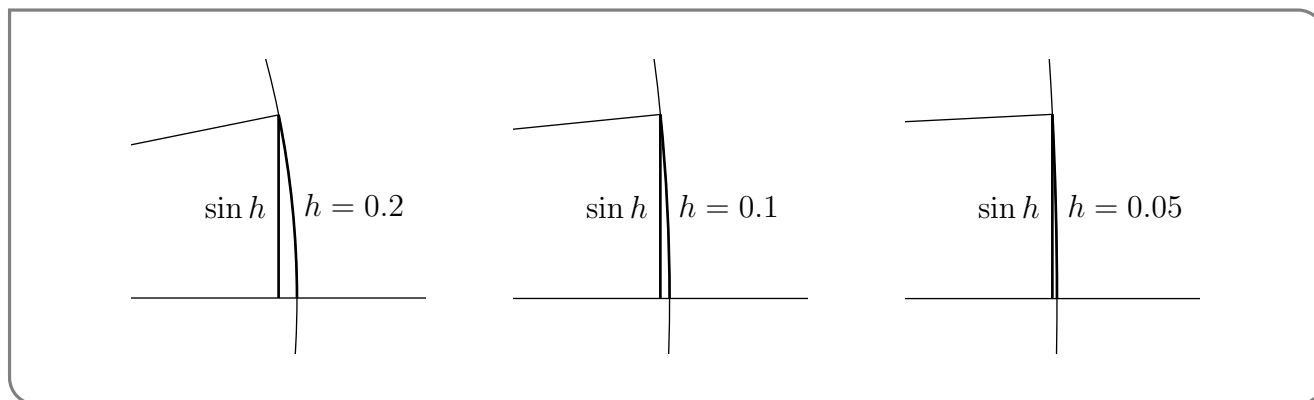
As we make h smaller and smaller and look at the figure with ever increasing magnification, the arc of length h and vertical line of length $\sin h$ look more and more alike. We would guess from this that
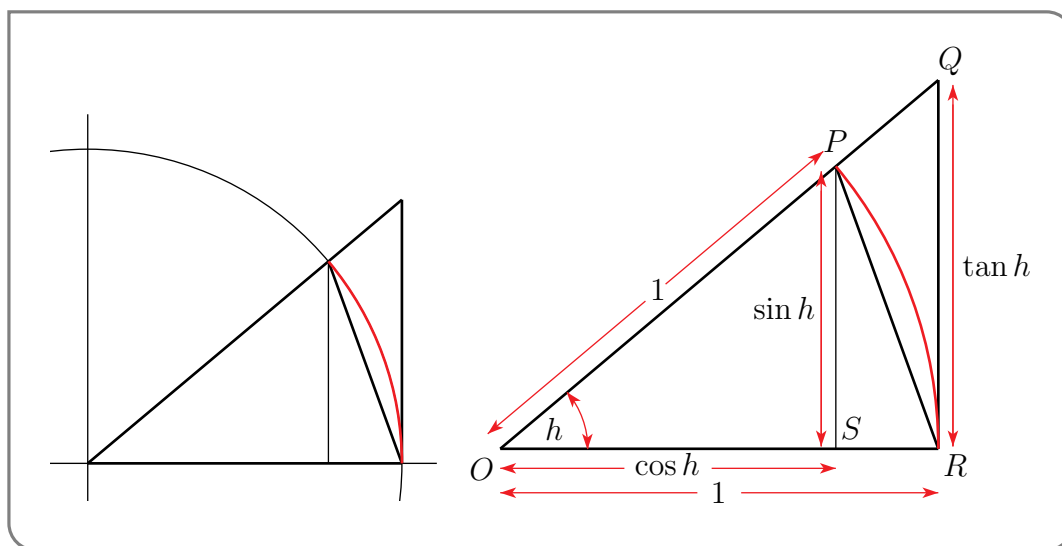
$$\lim_{h \to 0} \frac{\sin h}{h} = 1$$

The following tables of values

h	$\sin h$	$\frac{\sin h}{h}$
0.4	.3894	.9735
0.2	.1987	.9934
0.1	.09983	.9983
0.05	.049979	.99958
0.01	.00999983	.999983
0.001	.0099999983	.9999983

h	$\sin h$	$\frac{\sin h}{h}$
−0.4	−.3894	.9735
−0.2	−.1987	.9934
−0.1	−.09983	.9983
−0.05	−.049979	.99958
−0.01	−.00999983	.999983
−0.001	−.0099999983	.9999983

suggests the same guess. Here is an argument that shows that the guess really is correct.

▸▸▸ **Proof that** $\lim\limits_{h\to 0} \frac{\sin h}{h} = 1$:

The circle in the figure above has radius 1. Hence

$$|OP| = |OR| = 1 \qquad\qquad |PS| = \sin h$$
$$|OS| = \cos h \qquad\qquad |QR| = \tan h$$

Now we can use a few geometric facts about this figure to establish both an upper bound and a lower bound on $\frac{\sin h}{h}$ with both the upper and lower bounds tending to 1 as h tends to 0. So the squeeze theorem will tell us that $\frac{\sin h}{h}$ also tends to 1 as h tends to 0.

- The triangle OPR has base 1 and height $\sin h$, and hence

$$\text{area of } \triangle OPR = \tfrac{1}{2} \times 1 \times \sin h = \frac{\sin h}{2}.$$

- The triangle OQR has base 1 and height $\tan h$, and hence

$$\text{area of } \triangle OQR = \tfrac{1}{2} \times 1 \times \tan h = \frac{\tan h}{2}.$$

- The "piece of pie" OPR cut out of the circle is the fraction $\frac{h}{2\pi}$ of the whole circle (since the angle at the corner of the piece of pie is h radians and the angle for the whole circle is 2π radians). Since the circle has radius 1 we have

$$\text{area of pie } OPR = \frac{h}{2\pi} \cdot (\text{area of circle}) = \frac{h}{2\pi}\pi \cdot 1^2 = \frac{h}{2}$$

Now the triangle OPR is contained inside the piece of pie OPR. and so the area of the triangle is smaller than the area of the piece of pie. Similarly, the piece of pie OPR is contained inside the triangle OQR. Thus we have

$$\text{area of triangle } OPR \leqslant \text{area of pie } OPR \leqslant \text{area of triangle } OQR$$

Substituting in the areas we worked out gives

$$\frac{\sin h}{2} \leqslant \frac{h}{2} \leqslant \frac{\tan h}{2}$$

which cleans up to give

$$\sin h \leqslant h \leqslant \frac{\sin h}{\cos h}$$

We rewrite these two inequalities so that $\frac{\sin h}{h}$ appears in both.

- Since $\sin h \leqslant h$, we have that $\dfrac{\sin h}{h} \leqslant 1$.

- Since $h \leqslant \dfrac{\sin h}{\cos h}$ we have that $\cos h \leqslant \dfrac{\sin h}{h}$.

Thus we arrive at the "squeezable" inequality

$$\cos h \leqslant \frac{\sin h}{h} \leqslant 1$$

We know[26] that

$$\lim_{h \to 0} \cos h = 1.$$

Since $\frac{\sin h}{h}$ is sandwiched between $\cos h$ and 1, we can apply the squeeze theorem for limits (Theorem 1.4.17) to deduce the following lemma:

> **Lemma 2.8.1.**
>
> $$\lim_{h \to 0} \frac{\sin h}{h} = 1.$$

Since this argument took a bit of work, perhaps we should remind ourselves why we needed it in the first place. We were computing

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}x}\{\sin x\}\Big|_{x=0} &= \lim_{h \to 0} \frac{\sin h - \sin 0}{h} \\
&= \lim_{h \to 0} \frac{\sin h}{h} \qquad\qquad \text{(This is why!)} \\
&= 1
\end{aligned}$$

This concludes Step 1. We now know that $\frac{\mathrm{d}}{\mathrm{d}x} \sin x\big|_{x=0} = 1$. The remaining steps are easier.

---

26 Again, refresh your memory by looking up Appendix A.5.

▸ **Step 2:** $\frac{d}{dx}\{\cos x\}\big|_{x=0}$

By definition, the derivative of $\cos x$ evaluated at $x = 0$ is

$$\lim_{h\to 0}\frac{\cos h - \cos 0}{h} = \lim_{h\to 0}\frac{\cos h - 1}{h}$$

Fortunately we don't have to wade through geometry like we did for the previous step. Instead we can recycle our work and massage the above limit to rewrite it in terms of expressions involving $\frac{\sin h}{h}$. Thanks to Lemma 2.8.1 the work is then easy.

We'll show you two ways to proceed — one uses a method similar to "multiplying by the conjugate" that we have already used a few times (see Example 1.4.16 and 2.2.9), while the other uses a nice trick involving the double–angle formula[27].

▸▸ **Method 1 — Multiply by the "Conjugate"**

Start by multiplying the expression inside the limit by 1, written as $\dfrac{\cos h + 1}{\cos h + 1}$:

$$
\begin{aligned}
\frac{\cos h - 1}{h} &= \frac{\cos h - 1}{h} \cdot \frac{\cos h + 1}{\cos h + 1} \\
&= \frac{\cos^2 h - 1}{h(1 + \cos h)} && \left(\text{since } (a-b)(a+b) = a^2 - b^2\right) \\
&= -\frac{\sin^2 h}{h(1 + \cos h)} && \left(\text{since } \sin^2 h + \cos^2 h = 1\right) \\
&= -\frac{\sin h}{h} \cdot \frac{\sin h}{1 + \cos h}
\end{aligned}
$$

Now we can take the limit as $h \to 0$ via Lemma 2.8.1.

$$
\begin{aligned}
\lim_{h\to 0}\frac{\cos h - 1}{h} &= \lim_{h\to 0}\left(\frac{-\sin h}{h} \cdot \frac{\sin h}{1 + \cos h}\right) \\
&= -\lim_{h\to 0}\left(\frac{\sin h}{h}\right) \cdot \lim_{h\to 0}\left(\frac{\sin h}{1 + \cos h}\right) \\
&= -1 \cdot \frac{0}{2} \\
&= 0
\end{aligned}
$$

▸▸ **Method 2 — via the Double Angle Formula**

The other way involves the double angle formula[28],

$$\cos 2\theta = 1 - 2\sin^2(\theta) \qquad \text{or} \qquad \cos 2\theta - 1 = -2\sin^2(\theta)$$

27 See Appendix A.14 if you have forgotten. You should also recall that $\sin^2 \theta + \cos^2 \theta = 1$. Sorry for nagging.
28 We hope you looked this up in in Appendix A.14. Nag.

Setting $\theta = h/2$, we have

$$\frac{\cos h - 1}{h} = \frac{-2\left(\sin \frac{h}{2}\right)^2}{h}$$

Now this begins to look like $\frac{\sin h}{h}$, except that inside the $\sin(\cdot)$ we have $h/2$. So, setting $\theta = h/2$,

$$\frac{\cos h - 1}{h} = -\frac{\sin^2 \theta}{\theta} = -\theta \cdot \frac{\sin^2 \theta}{\theta^2}$$
$$= -\theta \cdot \frac{\sin \theta}{\theta} \cdot \frac{\sin \theta}{\theta}$$

When we take the limit as $h \to 0$, we are also taking the limit as $\theta = h/2 \to 0$, and so

$$\lim_{h \to 0} \frac{\cos h - 1}{h} = \lim_{\theta \to 0} \left[-\theta \cdot \frac{\sin \theta}{\theta} \cdot \frac{\sin \theta}{\theta} \right]$$
$$= \lim_{\theta \to 0} [-\theta] \cdot \lim_{\theta \to 0} \left[\frac{\sin \theta}{\theta} \right] \cdot \lim_{\theta \to 0} \left[\frac{\sin \theta}{\theta} \right]$$
$$= 0 \cdot 1 \cdot 1$$
$$= 0$$

where we have used the fact that $\lim_{h \to 0} \dfrac{\sin h}{h} = 1$ and that the limit of a product is the product of limits (i.e. Lemma 2.8.1 and Theorem 1.4.2).

Thus we have now produced two proofs of the following lemma:

Lemma 2.8.2.

$$\lim_{h \to 0} \frac{\cos h - 1}{h} = 0$$

Again, there has been a bit of work to get to here, so we should remind ourselves why we needed it. We were computing

$$\frac{d}{dx} \{\cos x\} \Big|_{x=0} = \lim_{h \to 0} \frac{\cos h - \cos 0}{h}$$
$$= \lim_{h \to 0} \frac{\cos h - 1}{h}$$
$$= 0$$

Armed with these results we can now build up the derivatives of sine and cosine.

▶▶ **Step 3: $\frac{\mathrm{d}}{\mathrm{d}x}\{\sin x\}$ and $\frac{\mathrm{d}}{\mathrm{d}x}\{\cos x\}$ for General x**

To proceed to the general derivatives of $\sin x$ and $\cos x$ we are going to use the above two results and a couple of trig identities. Remember the addition formulae[29]

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

To compute the derivative of $\sin(x)$ we just start from the definition of the derivative:

$$\frac{\mathrm{d}}{\mathrm{d}x}\sin x = \lim_{h\to 0}\frac{\sin(x+h) - \sin x}{h}$$

$$= \lim_{h\to 0}\frac{\sin x \cos h + \cos x \sin h - \sin x}{h}$$

$$= \lim_{h\to 0}\left[\sin x \frac{\cos h - 1}{h} + \cos x \frac{\sin h - 0}{h}\right]$$

$$= \sin x \lim_{h\to 0}\frac{\cos h - 1}{h} + \cos x \lim_{h\to 0}\frac{\sin h - 0}{h}$$

$$= \sin x \underbrace{\left[\frac{\mathrm{d}}{\mathrm{d}x}\cos x\right]_{x=0}}_{=0} + \cos x \underbrace{\left[\frac{\mathrm{d}}{\mathrm{d}x}\sin x\right]_{x=0}}_{=1}$$

$$= \cos x$$

The computation of the derivative of $\cos x$ is very similar.

$$\frac{\mathrm{d}}{\mathrm{d}x}\cos x = \lim_{h\to 0}\frac{\cos(x+h) - \cos x}{h}$$

$$= \lim_{h\to 0}\frac{\cos x \cos h - \sin x \sin h - \cos x}{h}$$

$$= \lim_{h\to 0}\left[\cos x \frac{\cos h - 1}{h} - \sin x \frac{\sin h - 0}{h}\right]$$

$$= \cos x \lim_{h\to 0}\frac{\cos h - 1}{h} - \sin x \lim_{h\to 0}\frac{\sin h - 0}{h}$$

$$= \cos x \underbrace{\left[\frac{\mathrm{d}}{\mathrm{d}x}\cos x\right]_{x=0}}_{=0} - \sin x \underbrace{\left[\frac{\mathrm{d}}{\mathrm{d}x}\sin x\right]_{x=0}}_{=1}$$

$$= -\sin x$$

We have now found the derivatives of both $\sin x$ and $\cos x$, *provided x is measured in radians.*

---

29   You really should. Look this up in Appendix A.8 if you have forgotten.

> **Lemma 2.8.3.**
>
> $$\frac{\mathrm{d}}{\mathrm{d}x}\sin x = \cos x \qquad\qquad \frac{\mathrm{d}}{\mathrm{d}x}\cos x = -\sin x$$
>
> The above formulas hold provided x is measured in radians.

These formulae are pretty easy to remember — applying $\frac{\mathrm{d}}{\mathrm{d}x}$ to $\sin x$ and $\cos x$ just exchanges $\sin x$ and $\cos x$, except for the minus sign[30] in the derivative of $\cos x$.

Remark 2.8.4 (Optional — Another derivation of $\frac{\mathrm{d}}{\mathrm{d}x}\cos x = -\sin x$). We remark that, once one knows that $\frac{\mathrm{d}}{\mathrm{d}x}\sin x = \cos x$, it is easy to use it and the trig identity $\cos(x) = \sin\left(\frac{\pi}{2} - x\right)$ to derive $\frac{\mathrm{d}}{\mathrm{d}x}\cos x = -\sin x$. Here is how[31].

$$\frac{\mathrm{d}}{\mathrm{d}x}\cos x = \lim_{h\to 0}\frac{\cos(x+h)-\cos x}{h} = \lim_{h\to 0}\frac{\sin\left(\frac{\pi}{2}-x-h\right)-\sin\left(\frac{\pi}{2}-x\right)}{h}$$

$$= -\lim_{h'\to 0}\frac{\sin\left(x'+h'\right)-\sin(x')}{h'} \qquad \text{with } x' = \frac{\pi}{2}-x,\ h'=-h$$

$$= -\frac{\mathrm{d}}{\mathrm{d}x'}\sin x'\bigg|_{x'=\frac{\pi}{2}-x} = -\cos\left(\frac{\pi}{2}-x\right)$$

$$= -\sin x$$

Note that, if x is measured in degrees, then the formulas of Lemma 2.8.3 are wrong. There are similar formulas, but we need the chain rule to build them — that is the subject of the next section. But first we should find the derivatives of the other trig functions.

▸ Step 4: the Remaining Trigonometric Functions

It is now an easy matter to get the derivatives of the remaining trigonometric functions using basic trig identities and the quotient rule. Remember[32] that

$$\tan x = \frac{\sin x}{\cos x} \qquad\qquad \cot x = \frac{\cos x}{\sin x} = \frac{1}{\tan x}$$

$$\csc x = \frac{1}{\sin x} \qquad\qquad \sec x = \frac{1}{\cos x}$$

---

30 There is a bad pun somewhere in here about sine errors and sign errors.
31 We thank Serban Raianu for suggesting that we include this.
32 You really should. If you do not then take a quick look at the appropriate appendix.

So, by the quotient rule,

$$\frac{d}{dx}\tan x = \frac{d}{dx}\frac{\sin x}{\cos x} = \frac{\left(\overbrace{\frac{d}{dx}\sin x}^{\cos x}\right)\cos x - \sin x\left(\overbrace{\frac{d}{dx}\cos x}^{-\sin x}\right)}{\cos^2 x} = \sec^2 x$$

$$\frac{d}{dx}\csc x = \frac{d}{dx}\frac{1}{\sin x} = -\frac{\left(\overbrace{\frac{d}{dx}\sin x}^{\cos x}\right)}{\sin^2 x} = -\csc x \cot x$$

$$\frac{d}{dx}\sec x = \frac{d}{dx}\frac{1}{\cos x} = -\frac{\left(\overbrace{\frac{d}{dx}\cos x}^{-\sin x}\right)}{\cos^2 x} = \sec x \tan x$$

$$\frac{d}{dx}\cot x = \frac{d}{dx}\frac{\cos x}{\sin x} = \frac{\left(\overbrace{\frac{d}{dx}\cos x}^{-\sin x}\right)\sin x - \cos x\left(\overbrace{\frac{d}{dx}\sin x}^{\cos x}\right)}{\sin^2 x} = -\csc^2 x$$

⇥ **Summary**

To summarise all this work, we can write this up as a theorem:

> **Theorem 2.8.5** (Derivatives of trigonometric functions).
>
> The derivatives of $\sin x$ and $\cos x$ are
>
> $$\frac{d}{dx}\sin x = \cos x \qquad\qquad \frac{d}{dx}\cos x = -\sin x$$
>
> Consequently the derivatives of the other trigonometric functions are
>
> $$\frac{d}{dx}\tan x = \sec^2 x \qquad\qquad \frac{d}{dx}\cot x = -\csc^2 x$$
> $$\frac{d}{dx}\csc x = -\csc x \cot x \qquad\qquad \frac{d}{dx}\sec x = \sec x \tan x$$

Of these 6 derivatives you should really memorise those of sine, cosine and tangent. We certainly expect you to be able to work out those of cotangent, cosecant and secant.

2.9 ▴ One More Tool – the Chain Rule

We have built up most of the tools that we need to express derivatives of complicated functions in terms of derivatives of simpler known functions. We started by learning how to evaluate

- derivatives of sums, products and quotients

- derivatives of constants and monomials

These tools allow us to compute derivatives of polynomials and rational functions. In the previous sections, we added exponential and trigonometric functions to our list. The final tool we add is called the chain rule. It tells us how to take the derivative of a composition of two functions. That is if we know $f(x)$ and $g(x)$ and their derivatives, then the chain rule tells us the derivative of $f(g(x))$.

Before we get to the statement of the rule, let us look at an example showing how such a composition might arise (in the "real-world").

Example 2.9.1

You are out in the woods after a long day of mathematics and are walking towards your camp fire on a beautiful still night. The heat from the fire means that the air temperature depends on your position. Let your position at time t be $x(t)$. The temperature of the air at position x is $f(x)$. What instantaneous rate of change of temperature do you feel at time t?
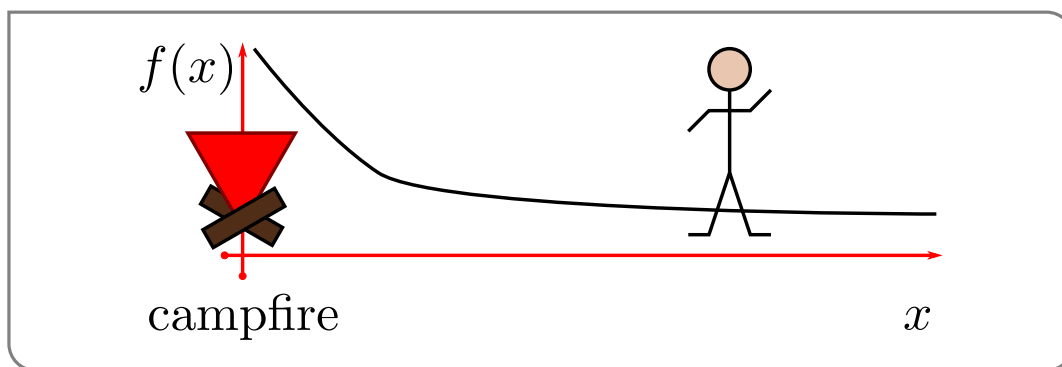
- Because your position at time t is $x = x(t)$, the temperature you feel at time t is $F(t) = f(x(t))$.

- The instantaneous rate of change of temperature that you feel is $F'(t)$. We have a complicated function, $F(t)$, constructed by composing two simpler functions, $x(t)$ and $f(x)$.

- We wish to compute the derivative, $F'(t) = \frac{d}{dt} f(x(t))$, of the complicated function $F(t)$ in terms of the derivatives, $x'(t)$ and $f'(x)$, of the two simple functions. This is exactly what the chain rule does.

Example 2.9.1

▸ Statement of the Chain Rule

> **Theorem 2.9.2** (The chain rule — version 1).
>
> Let $a \in \mathbb{R}$ and let $g(x)$ be a function that is differentiable at $x = a$. Now let $f(u)$ be a function that is differentiable at $u = g(a)$. Then the function $F(x) = f(g(x))$ is differentiable at $x = a$ and
>
> $$F'(a) = f'\big(g(a)\big) g'(a)$$

Here, as was the case earlier in this chapter, we have been very careful to give the point at which the derivative is evaluated a special name (i.e. a). But of course this evaluation point can really be any point (where the derivative is defined). So it is very common to just call the evaluation point "x" rather than give it a special name like "a", like this:

> **Theorem 2.9.3** (The chain rule — version 2).
>
> Let f and g be differentiable functions then
>
> $$\frac{\mathrm{d}}{\mathrm{d}x} f(g(x)) = f'\big(g(x)\big) \cdot g'(x)$$

Notice that when we form the composition $f(g(x))$ there is an "outside" function (namely $f(x)$) and an "inside" function (namely $g(x)$). The chain rule tells us that when we differentiate a composition that we have to differentiate the outside and then multiply by the derivative of the inside.

$$\frac{\mathrm{d}}{\mathrm{d}x} f(g(x)) = \underbrace{f'\big(g(x)\big)}_{\text{diff outside}} \cdot \underbrace{g'(x)}_{\text{diff inside}}$$

Here is another statement of the chain rule which makes this idea more explicit.

> **Theorem 2.9.4** (The chain rule — version 3).
>
> Let $y = f(u)$ and $u = g(x)$ be differentiable functions, then
>
> $$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}y}{\mathrm{d}u} \cdot \frac{\mathrm{d}u}{\mathrm{d}x}$$

This particular form is easy to remember because it looks like we can just "cancel" the $\mathrm{d}u$ between the two terms.

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}y}{\cancel{\mathrm{d}u}} \cdot \frac{\cancel{\mathrm{d}u}}{\mathrm{d}x}$$

Of course, $\mathrm{d}u$ is not, by itself, a number or variable[33] that can be cancelled. But this is still a good memory aid.

The hardest part about applying the chain rule is recognising when the function you are trying to differentiate is really the composition of two simpler functions. This takes a little practice. We can warm up with a couple of simple examples.

Example 2.9.5

Let $f(u) = u^5$ and $g(x) = \sin(x)$. Then set $F(x) = f(g(x)) = (\sin(x))^5$. To find the derivative of $F(x)$ we can simply apply the chain rule — the pieces of the composition have been laid out for us. Here they are.

$$f(u) = u^5 \qquad\qquad f'(u) = 5u^4$$
$$g(x) = \sin(x) \qquad\qquad g'(x) = \cos x$$

We now just put them together as the chain rule tells us

$$\frac{\mathrm{d}F}{\mathrm{d}x} = f'(g(x)) \cdot g'(x)$$
$$= 5(g(x))^4 \cdot \cos(x) \qquad\qquad \text{since } f'(u) = 5u^4$$
$$= 5(\sin(x))^4 \cdot \cos(x)$$

Notice that it is quite easy to extend this to any power. Set $f(u) = u^n$. Then follow the same steps and we arrive at

$$F(x) = (\sin(x))^n \qquad\qquad F'(x) = n(\sin(x))^{n-1} \cos(x)$$

Example 2.9.5

This example shows one of the ways that the chain rule appears very frequently — when we need to differentiate the power of some simpler function. More generally we have the following.

Example 2.9.6

Let $f(u) = u^n$ and let $g(x)$ be any differentiable function. Set $F(x) = f(g(x)) = g(x)^n$. Then

$$\frac{\mathrm{d}F}{\mathrm{d}x} = \frac{\mathrm{d}}{\mathrm{d}x}(g(x)^n) = ng(x)^{n-1} \cdot g'(x)$$

This is precisely the result in Example 2.6.6 and Lemma 2.6.7.

---

33 In this context $\mathrm{d}u$ is called a differential. There are ways to understand and manipulate these in calculus but they are beyond the scope of this course.

Example 2.9.7

Let $f(u) = \cos(u)$ and $g(x) = 3x - 2$. Find the derivative of

$$F(x) = f(g(x)) = \cos(3x - 2).$$

Again we should approach this by first writing down f and g and their derivatives and then putting everything together as the chain rule tells us.

$$f(u) = \cos(u) \qquad\qquad f'(u) = -\sin(u)$$
$$g(x) = 3x - 2 \qquad\qquad g'(x) = 3$$

So the chain rule says

$$\begin{aligned}
F'(x) &= f'(g(x)) \cdot g'(x) \\
&= -\sin(g(x)) \cdot 3 \\
&= -3\sin(3x - 2)
\end{aligned}$$

This example shows a second way that the chain rule appears very frequently — when we need to differentiate some function of $ax + b$. More generally we have the following.

Example 2.9.8

Let $a, b \in \mathbb{R}$ and let $f(x)$ be a differentiable function. Set $g(x) = ax + b$. Then

$$\begin{aligned}
\frac{d}{dx} f(ax + b) &= \frac{d}{dx} f(g(x)) \\
&= f'(g(x)) \cdot g'(x) \\
&= f'(ax + b) \cdot a
\end{aligned}$$

So the derivative of $f(ax + b)$ with respect to x is just $af'(ax + b)$.

The above is a very useful result that follows from the chain rule, so let's make it a corollary to highlight it.

> **Corollary 2.9.9.**
>
> Let $a, b \in \mathbb{R}$ and let $f(x)$ be a differentiable function, then
>
> $$\frac{d}{dx} f(ax + b) = af'(ax + b).$$

Example 2.9.10 (Example 2.9.1, continued)

Let us now go back to our motivating campfire example. There we had

$$
\begin{aligned}
f(x) &= \text{ temperature at position } x \\
x(t) &= \text{ position at time } t \\
F(t) &= f(x(t)) = \text{ temperature at time } t
\end{aligned}
$$

The chain rule gave

$$
F'(t) = f'(x(t)) \cdot x'(t)
$$

Notice that the units of measurement on both sides of the equation agree — as indeed they must. To see this, let us assume that t is measured in seconds, that $x(t)$ is measured in metres and that $f(x)$ is measured in degrees. Because of this $F(x(t))$ must also be measured in degrees (since it is a temperature).

What about the derivatives? These are rates of change. So

- $F'(t)$ has units $\frac{\text{degrees}}{\text{second}}$,

- $f'(x)$ has units $\frac{\text{degrees}}{\text{metre}}$, and

- $x'(t)$ has units $\frac{\text{metre}}{\text{second}}$.

Hence the product

$$
f'(x(t)) \cdot x'(t) \text{ has units } = \frac{\text{degrees}}{\text{metre}} \cdot \frac{\text{metre}}{\text{second}} = \frac{\text{degrees}}{\text{second}}.
$$

has the same units as $F'(t)$. So the units on both sides of the equation agree. Checking that the units on both sides of an equation agree is a good check of consistency, but of course it does not prove that both sides are in fact the same.

Example 2.9.10

▸▸ (optional) — Derivation of the Chain Rule

First, let's review what our goal is. We have been given a function $g(x)$, that is differentiable at some point $x = a$, and another function $f(u)$, that is differentiable at the point $u = b = g(a)$. We have defined the composite function $F(x) = f(g(x))$ and we wish to show that

$$
F'(a) = f'(g(a)) \cdot g'(a)
$$

Before we can compute $F'(a)$, we need to set up some ground work, and in particular the definitions of our given derivatives:

$$
f'(b) = \lim_{H \to 0} \frac{f(b+H) - f(b)}{H} \qquad \text{and} \qquad g'(a) = \lim_{h \to 0} \frac{g(a+h) - g(a)}{h}.
$$

We are going to use similar manipulation tricks as we did back in the proofs of the arithmetic of derivatives in Section 2.5. Unfortunately, we have already used up the symbols "F" and "H", so we are going to make use of the Greek letters γ, φ.

As was the case in our derivation of the product rule it is convenient to introduce a couple of new functions. Set

$$\varphi(H) = \frac{f(b + H) - f(b)}{H}$$

Then we have

$$\lim_{H \to 0} \varphi(H) = f'(b) = f'(g(a)) \qquad \text{since } b = g(a), \qquad (2.9.1)$$

and we can also write (with a little juggling)

$$f(b + H) = f(b) + H\varphi(H)$$

Similarly set

$$\gamma(h) = \frac{g(a + h) - g(a)}{h}$$

which gives us

$$\lim_{h \to 0} \gamma(h) = g'(a) \qquad \text{and} \qquad g(a + h) = g(a) + h\gamma(h).$$

Now we can start computing

$$F'(a) = \lim_{h \to 0} \frac{F(a + h) - F(a)}{h}$$
$$= \lim_{h \to 0} \frac{f(g(a + h)) - f(g(a))}{h}$$

We know that $g(a) = b$ and $g(a + h) = g(a) + h\,\gamma(h)$, so

$$F'(a) = \lim_{h \to 0} \frac{f(g(a) + h\gamma(h)) - f(g(a))}{h}$$
$$= \lim_{h \to 0} \frac{f(b + h\gamma(h)) - f(b)}{h}$$

Now for the sneaky bit. We can turn $f(b + h\gamma(h))$ into $f(b + H)$ by setting

$$H = h\gamma(h)$$

Now notice that as $h \to 0$ we have

$$\lim_{h \to 0} H = \lim_{h \to 0} h \cdot \gamma(h)$$
$$= \lim_{h \to 0} h \cdot \lim_{h \to 0} \gamma(h)$$
$$= 0 \cdot g'(a) = 0$$

So as $h \to 0$ we also have $H \to 0$.

We now have

$$
\begin{aligned}
F'(a) &= \lim_{h \to 0} \frac{f(b+H) - f(b)}{h} \\
&= \lim_{h \to 0} \underbrace{\frac{f(b+H) - f(b)}{H}}_{= \varphi(H)} \cdot \underbrace{\frac{H}{h}}_{= \gamma(h)} && \text{if } H = h\gamma(h) \neq 0 \\
&= \lim_{h \to 0} \big(\varphi(H) \cdot \gamma(h)\big) \\
&= \lim_{h \to 0} \varphi(H) \cdot \lim_{h \to 0} \gamma(h) && \text{since } H \to 0 \text{ as } h \to 0 \\
&= \lim_{H \to 0} \varphi(H) \cdot \lim_{h \to 0} \gamma(h) && = f'(b) \cdot g'(a)
\end{aligned}
$$

This is exactly the RHS of the chain rule. It is possible to have $H = 0$ in the second line above. But that possibility is easy to deal with:

- If $g'(a) \neq 0$, then, since $\lim_{h \to 0} \gamma(h) = g'(a)$, $H = h\gamma(h)$ cannot be 0 for small nonzero h. Technically, there is an $h_0 > 0$ such that $H = h\gamma(h) \neq 0$ for all $0 < |h| < h_0$. In taking the limit $h \to 0$, above, we need only consider $0 < |h| < h_0$ and so, in this case, the above computation is completely correct.
- If $g'(a) = 0$, the above computation is still fine provided we exclude all h's for which $H = h\gamma(h) \neq 0$. When $g'(a) = 0$, the right hand side, $f'(g(a)) \cdot g'(a)$, of the chain rule is 0. So the above computation gives

$$
\lim_{\substack{h \to 0 \\ \gamma(h) \neq 0}} \frac{f(b+H) - f(b)}{h} = f'(g(a)) \cdot g'(a) = 0
$$

On the other hand, when $H = 0$, we have $f(b+H) - f(b) = 0$. So

$$
\lim_{\substack{h \to 0 \\ \gamma(h) = 0}} \frac{f(b+H) - f(b)}{h} = 0
$$

too. That's all we need.

▸ Chain Rule Examples

We'll now use the chain rule to compute some more derivatives.

Example 2.9.11

Find $\frac{\mathrm{d}}{\mathrm{d}x}(1+3x)^{75}$.

This is a concrete version of Example 2.9.8. We are to find the derivative of a function that is built up by first computing $1 + 3x$ and then taking the 75th power of the result. So we set

$$
\begin{aligned}
f(u) &= u^{75} && f'(u) = 75u^{74} \\
g(x) &= 1 + 3x && g'(x) = 3 \\
F(x) &= f(g(x)) = g(x)^{75} = (1+3x)^{75}
\end{aligned}
$$

By the chain rule

$$F'(x) = f'(g(x)) \, g'(x) = 75 \, g(x)^{74} \, g'(x) = 75 \, (1+3x)^{74} \cdot 3$$
$$= 225 \, (1+3x)^{74}$$

Example 2.9.11

Example 2.9.12

Find $\frac{d}{dx} \sin(x^2)$.

In this example we are to compute the derivative of sin with a (slightly) complicated argument. So we apply the chain rule with f being sin and $g(x)$ being the complicated argument. That is, we set

$$f(u) = \sin u \qquad\qquad f'(u) = \cos u$$
$$g(x) = x^2 \qquad\qquad g'(x) = 2x$$
$$F(x) = f(g(x)) = \sin(g(x)) = \sin(x^2)$$

By the chain rule

$$F'(x) = f'(g(x)) \, g'(x) = \cos(g(x)) \, g'(x) = \cos(x^2) \cdot 2x$$
$$= 2x \cos(x^2)$$

Example 2.9.12

Example 2.9.13

Find $\frac{d}{dx} \sqrt[3]{\sin(x^2)}$.

In this example we are to compute the derivative of the cube root of a (moderately) complicated argument, namely $\sin(x^2)$. So we apply the chain rule with f being "cube root" and $g(x)$ being the complicated argument. That is, we set

$$f(u) = \sqrt[3]{u} = u^{\frac{1}{3}} \qquad\qquad f'(u) = \tfrac{1}{3} u^{-\frac{2}{3}}$$
$$g(x) = \sin(x^2) \qquad\qquad g'(x) = 2x \cos(x^2)$$
$$F(x) = f(g(x)) = \sqrt[3]{g(x)} = \sqrt[3]{\sin(x^2)}$$

In computing $g'(x)$ here, we have already used the chain rule once (in Example 2.9.12). By the chain rule

$$F'(x) = f'(g(x)) \, y'(x) = \tfrac{1}{3} g(x)^{-\frac{2}{3}} \cdot 2x \cos(x^2)$$
$$= \frac{2x}{3} \frac{\cos(x^2)}{[\sin(x^2)]^{\frac{2}{3}}}$$

Example 2.9.14

Find the derivative of $\frac{d}{dx} f(g(h(x)))$.

This is very similar to the previous example. Let us set $F(x) = f(g(h(x)))$ with $u = g(h(x))$ then the chain rule tells us

$$\frac{dF}{dx} = \frac{df}{du} \cdot \frac{du}{dx}$$
$$= f'(g(h(x))) \cdot \frac{d}{dx} g(h(x))$$

We now just apply the chain rule again

$$= f'(g(h(x))) \cdot g'(h(x)) \cdot h'(x).$$

Indeed it is not too hard to generalise further (in the manner of Example 2.6.6 to find the derivative of the composition of 4 or more functions (though things start to become tedious to write down):

$$\frac{d}{dx} f_1(f_2(f_3(f_4(x)))) = f_1'(f_2(f_3(f_4(x)))) \cdot \frac{d}{dx} f_2(f_3(f_4(x)))$$
$$= f_1'(f_2(f_3(f_4(x)))) \cdot f_2'(f_3(f_4(x))) \cdot \frac{d}{dx} f_3(f_4(x))$$
$$= f_1'(f_2(f_3(f_4(x)))) \cdot f_2'(f_3(f_4(x))) \cdot f_3'(f_4(x)) \cdot f_4'(x)$$

Example 2.9.15

We can also use the chain rule to recover Corollary 2.4.6 and from there we can use the product rule to recover the quotient rule.

We want to differentiate $F(x) = \frac{1}{g(x)}$ so set $f(u) = \frac{1}{u}$ and $u = g(x)$. Then the chain rule tells us

$$\frac{d}{dx} \left\{ \frac{1}{g(x)} \right\} = \frac{dF}{dx} = \frac{df}{du} \cdot \frac{du}{dx}$$
$$= \frac{-1}{u^2} \cdot g'(x)$$
$$= -\frac{g'(x)}{g(x)^2}.$$

Once we know this, a quick application of the product rule will give us the quotient rule.

$$\frac{d}{dx}\left\{\frac{f(x)}{g(x)}\right\} = \frac{d}{dx}\left\{f(x)\cdot\frac{1}{g(x)}\right\} \qquad\qquad \text{use PR}$$

$$= f'(x)\cdot\frac{1}{g(x)} + f(x)\cdot\frac{d}{dx}\left\{\frac{1}{g(x)}\right\} \qquad\qquad \text{use the result from above}$$

$$= f'(x)\cdot\frac{1}{g(x)} - f(x)\cdot\frac{g'(x)}{g(x)^2} \qquad\qquad \text{place over a common denominator}$$

$$= \frac{f'(x)\cdot g(x) - f(x)\cdot g'(x)}{g(x)^2}$$

which is exactly the quotient rule.

Example 2.9.15

Example 2.9.16

Compute the following derivative:

$$\frac{d}{dx}\cos\left(\frac{x^5\sqrt{3+x^6}}{(4+x^2)^3}\right)$$

This time we are to compute the derivative of cos with a really complicated argument.

- So, to start, we apply the chain rule with $g(x) = \frac{x^5\sqrt{3+x^6}}{(4+x^2)^3}$ being the really complicated argument and f being cos. That is, $f(u) = \cos(u)$. Since $f'(u) = -\sin(u)$, the chain rule gives

$$\frac{d}{dx}\cos\left(\frac{x^5\sqrt{3+x^6}}{(4+x^2)^3}\right) = -\sin\left(\frac{x^5\sqrt{3+x^6}}{(4+x^2)^3}\right)\frac{d}{dx}\left\{\frac{x^5\sqrt{3+x^6}}{(4+x^2)^3}\right\}$$

- This reduced our problem to that of computing the derivative of the really complicated argument $\frac{x^5\sqrt{3+x^6}}{(4+x^2)^3}$. We can think of the argument as being built up out of three pieces, namely x^5, multiplied by $\sqrt{3+x^6}$, divided by $(4+x^2)^3$, or, equivalently, multiplied by $(4+x^2)^{-3}$. So we may rewrite $\frac{x^5\sqrt{3+x^6}}{(4+x^2)^3}$ as $x^5\left(3+x^6\right)^{1/2}\left(4+x^2\right)^{-3}$, and then apply the product rule to reduce the problem to that of computing the derivatives of the three pieces.

- Here goes (recall Example 2.6.6):

$$\frac{d}{dx}\left[x^5\left(3+x^6\right)^{1/2}\left(4+x^2\right)^{-3}\right] = \frac{d}{dx}\left[x^5\right]\cdot\left(3+x^6\right)^{1/2}\cdot\left(4+x^2\right)^{-3}$$

$$+ x^5\cdot\frac{d}{dx}\left[\left(3+x^6\right)^{1/2}\right]\cdot\left(4+x^2\right)^{-3}$$

$$+ x^5\cdot\left(3+x^6\right)^{1/2}\cdot\frac{d}{dx}\left[\left(4+x^2\right)^{-3}\right]$$

This has reduced our problem to computing the derivatives of x^5, which is easy, and of $(3 + x^6)^{1/2}$ and $(4 + x^2)^{-3}$, both of which can be done by the chain rule. Doing so,

$$\frac{d}{dx}\left[x^5\,(3 + x^6)^{1/2}\,(4 + x^2)^{-3}\right] = \overbrace{\frac{d}{dx}\left[x^5\right]}^{5x^4} \cdot (3 + x^6)^{1/2} \cdot (4 + x^2)^{-3}$$

$$+ x^5 \cdot \overbrace{\frac{d}{dx}\left[(3 + x^6)^{1/2}\right]}^{\frac{1}{2}(3+x^6)^{-1/2}\cdot 6x^5} \cdot (4 + x^2)^{-3}$$

$$+ x^5 \cdot (3 + x^6)^{1/2} \cdot \overbrace{\frac{d}{dx}\left[(4 + x^2)^{-3}\right]}^{-3(4+x^2)^{-4}\cdot 2x}$$

- Now we can clean things up in a sneaky way by observing
  - differentiating x^5, to get $5x^4$, is the same as multiplying x^5 by $\frac{5}{x}$, and
  - differentiating $(3 + x^6)^{\frac{1}{2}}$ to get $\frac{1}{2}(3 + x^6)^{-1/2} \cdot 6x^5$ is the same as multiplying $(3 + x^6)^{\frac{1}{2}}$ by $\frac{3x^5}{3+x^6}$, and
  - differentiating $(4 + x^2)^{-3}$ to get $-3(4 + x^2)^{-4} \cdot 2x$ is the same as multiplying $(4 + x^2)^{-3}$ by $-\frac{6x}{4+x^2}$.

Using these sneaky tricks we can write our solution quite neatly:

$$\frac{d}{dx}\cos\left(\frac{x^5\sqrt{3 + x^6}}{(4 + x^2)^3}\right) = -\sin\left(\frac{x^5\sqrt{3 + x^6}}{(4 + x^2)^3}\right)\frac{x^5\sqrt{3 + x^6}}{(4 + x^2)^3}\left\{\frac{5}{x} + \frac{3x^5}{3 + x^6} - \frac{6x}{4 + x^2}\right\}$$

- This method of cleaning up the derivative of a messy product is actually something more systematic in disguise — namely logarithmic differentiation. We will come to this later.

Example 2.9.16

2.10 ▴ The Natural Logarithm

The chain rule opens the way to understanding derivatives of more complicated function. Not only compositions of known functions as we have seen the examples of the previous section, but also functions which are defined implicitly.

Consider the logarithm base e — $\log_e(x)$ is the power that e must be raised to to give x. That is, $\log_e(x)$ is defined by

$$e^{\log_e x} = x$$

i.e. — it is the inverse of the exponential function with base e. Since this choice of base works so cleanly and easily with respect to differentiation, this base turns out to be (arguably) the most natural choice for the base of the logarithm. And as we saw in our whirlwind review of logarithms in Section 2.7, it is easy to use logarithms of one base to compute logarithms with another base:

$$\log_q x = \frac{\log_e x}{\log_e q}$$

So we are (relatively) free to choose a base which is convenient for our purposes.

The logarithm with base e, is called the "natural logarithm". The "naturalness" of logarithms base e is exactly that this choice of base works very nicely in calculus (and so wider mathematics) in ways that other bases do not[34]. There are several different "standard" notations for the logarithm base e;

$$\log_e x = \log x = \ln x.$$

We recommend that you be able to recognise all of these.

In this text we will write the natural logarithm as "log" with no base. The reason for this choice is that base e is the standard choice of base for logarithms in mathematics[35] The natural logarithm inherits many properties of general logarithms[36]. So, for all $x, y > 0$ the following hold:

- $e^{\log x} = x$,

- for any real number X, $\log\left(e^X\right) = X$,

- for any $a > 1$, $\log_a x = \frac{\log x}{\log a}$ and $\log x = \frac{\log_a x}{\log_a e}$

- $\log 1 = 0$, $\log e = 1$

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

- $\log\left(\frac{x}{y}\right) = \log x - \log y$, $\log\left(\frac{1}{y}\right) = -\log y$

- $\log(x^X) = X \log x$

- $\lim_{x \to \infty} \log x = \infty$, $\lim_{x \to 0} \log x = -\infty$

And finally we should remember that $\log x$ has domain (i.e. is defined for) $x > 0$ and range (i.e. takes all of the values in) the set of all real numbers.

---

34 The interested reader should head to Wikipedia and look up the natural logarithm.
35 In other disciplines other bases are natural; in computer science, since numbers are stored in binary it makes sense to use the binary logarithm — i.e. base 2. While in some sciences and finance, it makes sense to use the decimal logarithm — i.e. base 10.
36 Again take a quick look at the whirlwind review of logarithms in Section 2.7.

Figure 2.10.1.
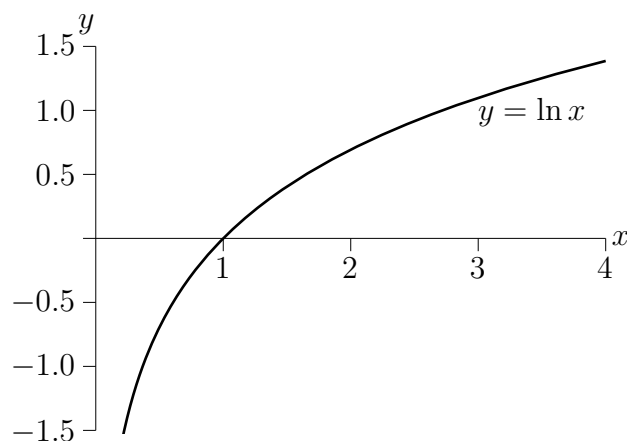
To compute the derivative of $\log x$ we could attempt to start with the limit definition of the derivative

$$\frac{d}{dx}\log x = \lim_{h \to 0} \frac{\log(x+h) - \log(x)}{h}$$

$$= \lim_{h \to 0} \frac{\log((x+h)/x)}{h}$$

$$= \text{um}\dots$$

This doesn't look good. But all is not lost — we have the chain rule, and we know that the logarithm satisfies the equation:

$$x = e^{\log x}$$

Since both sides of the equation are the same function, both sides of the equation have the same derivative. i.e. we are using[37]

$$\text{if } f(x) = g(x) \text{ for all } x, \text{ then } f'(x) = g'(x)$$

So now differentiate both sides:

$$\frac{d}{dx}x = \frac{d}{dx}e^{\log x}$$

The left-hand side is easy, and the right-hand side we can process using the chain rule with $f(u) = e^u$ and $u = \log x$.

$$1 = \frac{df}{du} \cdot \frac{du}{dx}$$

$$= e^u \cdot \underbrace{\frac{d}{dx}\log x}_{\text{what we want to compute}}$$

---

37 Notice that just because the derivatives are the same, doesn't mean the original functions are the same. Both $f(x) = x^2$ and $g(x) = x^2 + 3$ have derivative $f'(x) = g'(x) = 2x$, but $f(x) \neq g(x)$.

Recall that $e^u = e^{\log x} = x$, so

$$1 = x \cdot \underbrace{\frac{\mathrm{d}}{\mathrm{d}x} \log x}_{\text{now what?}}$$

We can now just rearrange this equation to make the thing we want the subject:

$$\frac{\mathrm{d}}{\mathrm{d}x} \log x = \frac{1}{x}$$

Thus we have proved:

Theorem 2.10.1.

$$\frac{\mathrm{d}}{\mathrm{d}x} \log x = \frac{1}{x}$$

where $\log x$ is the logarithm base e.

Example 2.10.2

Let $f(x) = \log 3x$. Find $f'(x)$.

There are two ways to approach this — we can simplify then differentiate, or differentiate and then simplify. Neither is difficult.

- Simplify and then differentiate:

$$
\begin{aligned}
f(x) &= \log 3x &&\text{log of a product} \\
&= \log 3 + \log x \\
f'(x) &= \frac{\mathrm{d}}{\mathrm{d}x} \log 3 + \frac{\mathrm{d}}{\mathrm{d}x} \log x \\
&= \frac{1}{x}.
\end{aligned}
$$

- Differentiation and then simplify:

$$
\begin{aligned}
f'(x) &= \frac{\mathrm{d}}{\mathrm{d}x} \log(3x) &&\text{chain rule} \\
&= \frac{1}{3x} \cdot 3 \\
&= \frac{1}{x}
\end{aligned}
$$

Example 2.10.2

Example 2.10.3 (The derivative of $\log cx$)

Notice that we can extend the previous example for any positive constant — not just 3. Let $c > 0$ be a constant, then

$$\frac{\mathrm{d}}{\mathrm{d}x}\log cx = \frac{\mathrm{d}}{\mathrm{d}x}\left(\log c + \log x\right)$$
$$= \frac{1}{x}$$

Example 2.10.3

Example 2.10.4 (The derivative of $\log |x|$)

We can push this further still. Let $g(x) = \log |x|$, then[38]

- If $x > 0$, $|x| = x$ and so

$$g'(x) = \frac{\mathrm{d}}{\mathrm{d}x}\log x = \frac{1}{x}$$

- If $x < 0$ then $|x| = -x$. If $|h|$ is strictly smaller than $|x|$, then we also have that $x + h < 0$ and $|x + h| = -(x + h) = |x| - h$. Write $X = |x|$ and $H = -h$. Then, by the definition of the derivative,

$$g'(x) = \lim_{h \to 0} \frac{\log |x + h| - \log |x|}{h} = \lim_{h \to 0} \frac{\log(|x| - h) - \log |x|}{h}$$
$$= \lim_{H \to 0} \frac{\log(X + H) - \log X}{-H} = -\lim_{H \to 0} \frac{\log(X + H) - \log X}{H}$$
$$= -\frac{\mathrm{d}}{\mathrm{d}X}\log X = -\frac{1}{X} = -\frac{1}{|x|}$$
$$= \frac{1}{x}$$

- Since $\log 0$ is undefined, $g'(0)$ does not exist.

Putting this together gives:

$$\frac{\mathrm{d}}{\mathrm{d}x}\log |x| = \frac{1}{x}$$

Example 2.10.4

Example 2.10.5 (The derivative of x^a)

Just after Corollary 2.6.17, we said that we would, in the future, find the derivative of x^a

---

38   It's probably a good moment to go back and look at Example 2.2.10.

for all real numbers. The future is here. Let $x > 0$ and a be any real number. Exponentiating both sides of $\log(x^a) = a \log x$ gives us $x^a = e^{a \log x}$ and then

$$\begin{aligned}
\frac{d}{dx} x^a &= \frac{d}{dx} e^{a \log x} = e^{a \log x} \frac{d}{dx} (a \log x) && \text{by the chain rule} \\
&= \frac{a}{x} e^{a \log x} = \frac{a}{x} x^a \\
&= a x^{a-1}
\end{aligned}$$

as expected.

Example 2.10.5

We can extend Theorem 2.10.1 to compute the derivative of logarithms of other bases in a straightforward way. Since for any positive $a \neq 1$:

$$\begin{aligned}
\log_a x &= \frac{\log x}{\log a} = \frac{1}{\log a} \cdot \log x && \text{since } a \text{ is a constant} \\
\frac{d}{dx} \log_a x &= \frac{1}{\log a} \cdot \frac{1}{x}
\end{aligned}$$

▸▸ Back to $\frac{d}{dx} a^x$

We can also now finally get around to computing the derivative of a^x (which we started to do back in Section 2.7).

$$\begin{aligned}
f(x) &= a^x && \text{take log of both sides} \\
\log f(x) &= x \log a && \text{exponentiate both sides base } e \\
f(x) &= e^{x \log a} && \text{chain rule} \\
f'(x) &= e^{x \log a} \cdot \log a \\
&= a^x \cdot \log a
\end{aligned}$$

Notice that we could have also done the following:

$$\begin{aligned}
f(x) &= a^x && \text{take log of both sides} \\
\log f(x) &= x \log a && \text{differentiate both sides} \\
\frac{d}{dx} (\log f(x)) &= \log a
\end{aligned}$$

We then process the left-hand side using the chain rule

$$\begin{aligned}
f'(x) \cdot \frac{1}{f(x)} &= \log a \\
f'(x) &= f(x) \cdot \log a = a^x \cdot \log a
\end{aligned}$$

We will see $\frac{d}{dx} \log f(x)$ more below in the subsection on "logarithmic differentiation".
   To summarise the results above:

> **Corollary 2.10.6.**
>
> $$\frac{d}{dx}a^x = \log a \cdot a^x \qquad\qquad\qquad \text{for any } a > 0$$
>
> $$\frac{d}{dx}\log_a x = \frac{1}{x \cdot \log a} \qquad\qquad \text{for any } a > 0, a \neq 1$$
>
> where $\log x$ is the natural logarithm.

Recall that we need the caveat $a \neq 1$ because the logarithm base 1 is not well defined. This is because $1^x = 1$ for any x. We do not need a similar caveat for the derivative of the exponential because we know (recall Example 2.7.1)

$$\frac{d}{dx}1^x = \frac{d}{dx}1 = 0 \qquad\qquad\qquad \text{while the above corollary tells us}$$
$$= \log 1 \cdot 1^x = 0 \cdot 1 = 0.$$

▸▸ Logarithmic Differentiation

We want to go back to some previous slightly messy examples (Examples 2.6.6 and 2.6.18) and now show you how they can be done more easily.

Example 2.10.7

Consider again the derivative of the product of 3 functions:

$$P(x) = F(x) \cdot G(x) \cdot H(x)$$

Start by taking the logarithm of both sides:

$$\log P(x) = \log (F(x) \cdot G(x) \cdot H(x))$$
$$= \log F(x) + \log G(x) + \log H(x)$$

Notice that the product of functions on the right-hand side has become a sum of functions. Differentiating sums is much easier than differentiating products. So when we differentiate we have

$$\frac{d}{dx}\log P(x) = \frac{d}{dx}\log F(x) + \frac{d}{dx}\log G(x) + \frac{d}{dx}\log H(x)$$

A quick application of the chain rule shows that $\frac{d}{dx}\log f(x) = f'(x)/f(x)$:

$$\frac{P'(x)}{P(x)} = \frac{F'(x)}{F(x)} + \frac{G'(x)}{G(x)} + \frac{H'(x)}{H(x)}$$

Multiply through by $P(x) = F(x)G(x)H(x)$:

$$P'(x) = \left(\frac{F'(x)}{F(x)} + \frac{G'(x)}{G(x)} + \frac{H'(x)}{H(x)}\right) \cdot F(x)G(x)H(x)$$
$$= F'(x)G(x)H(x) + F(x)G'(x)H(x) + F(x)G(x)H'(x)$$

which is what found in Example 2.6.6 by repeated application of the product rule. The above generalises quite easily to more than 3 functions.

Example 2.10.7

This same trick of "take a logarithm and then differentiate" — or logarithmic differentiation — will work any time you have a product (or ratio) of functions.

Example 2.10.8

Let's use logarithmic differentiation on the function from Example 2.6.18:

$$f(x) = \frac{(\sqrt{x}-1)(2-x)(1-x^2)}{\sqrt{x}(3+2x)}$$

Beware however, that we may only take the logarithm of positive numbers, and this $f(x)$ is often negative. For example, if $1 < x < 2$, the factor $(1-x^2)$ in the definition of $f(x)$ is negative while all of the other factors are positive, so that $f(x) < 0$. None–the–less, we can use logarithmic differentiation to find $f'(x)$, by exploiting the observation that $\frac{d}{dx}\log|f(x)| = \frac{f'(x)}{f(x)}$. (To see this, use the chain rule and Example 2.10.4.) So we take the logarithm of $|f(x)|$ and expand.

$$\begin{aligned}
\log|f(x)| &= \log\frac{|\sqrt{x}-1|\,|2-x|\,|1-x^2|}{\sqrt{x}|3+2x|} \\
&= \log|\sqrt{x}-1| + \log|2-x| + \log|1-x^2| - \underbrace{\log(\sqrt{x})}_{=\frac{1}{2}\log x} - \log|3+2x|
\end{aligned}$$

Now we can essentially just differentiate term-by-term:

$$\begin{aligned}
\frac{d}{dx}\log|f(x)| &= \frac{d}{dx}\left(\log|\sqrt{x}-1| + \log|2-x| + \log|1-x^2| - \frac{1}{2}\log(x) - \log|3+2x|\right) \\
\frac{f'(x)}{f(x)} &= \frac{1/(2\sqrt{x})}{\sqrt{x}-1} + \frac{-1}{2-x} + \frac{-2x}{1-x^2} - \frac{1}{2x} - \frac{2}{3+2x} \\
f'(x) &= f(x)\cdot\left(\frac{1}{2\sqrt{x}(\sqrt{x}-1)} - \frac{1}{2-x} - \frac{2x}{1-x^2} - \frac{1}{2x} - \frac{2}{3+2x}\right) \\
&= \frac{(\sqrt{x}-1)(2-x)(1-x^2)}{\sqrt{x}(3+2x)}\cdot\left(\frac{1}{2\sqrt{x}(\sqrt{x}-1)} - \frac{1}{2-x} - \frac{2x}{1-x^2} - \frac{1}{2x} - \frac{2}{3+2x}\right)
\end{aligned}$$

just as we found previously.

Example 2.10.8

2.11 ▴ Implicit Differentiation

Implicit differentiation is a simple trick that is used to compute derivatives of functions either

- when you don't know an explicit formula for the function, but you know an equation that the function obeys or

- even when you have an explicit, but complicated, formula for the function, and the function obeys a simple equation.

The trick is just to differentiate both sides of the equation and then solve for the derivative we are seeking. In fact we have already done this, without using the name "implicit differentiation", when we found the derivative of $\log x$ in the previous section. There we knew that the function $f(x) = \log x$ satisfied the equation $e^{f(x)} = x$ for all x. That is, the functions $e^{f(x)}$ and x are in fact the same function and so have the same derivative. So we had

$$\frac{d}{dx}e^{f(x)} = \frac{d}{dx}x = 1$$

We then used the chain rule to get $\frac{d}{dx}e^{f(x)} = e^{f(x)}f'(x)$, which told us that $f'(x)$ obeys the equation

$$e^{f(x)}f'(x) = 1 \qquad\qquad \text{and we can now solve for } f'(x)$$

$$f'(x) = e^{-f(x)} = e^{-\log x} = \frac{1}{x}.$$

The typical way to get used to implicit differentiation is to play with problems involving tangent lines to curves. So here are a few examples finding the equations of tangent lines to curves. Recall, from Theorem 2.3.2, that, in general, the tangent line to the curve $y = f(x)$ at (x_0, y_0) is $y = f(x_0) + f'(x_0)(x - x_0) = y_0 + f'(x_0)(x - x_0)$.

Example 2.11.1

Find the equation of the tangent line to $y = y^3 + xy + x^3$ at $x = 1$.
    This is a very standard sounding example, but made a little complicated by the fact that the curve is given by a cubic equation — which means we cannot solve directly for y in terms of x or vice versa. So we really do need implicit differentiation.

- First notice that when $x = 1$ the equation, $y = y^3 + xy + x^3$, of the curve simplifies to $y = y^3 + y + 1$ or $y^3 = -1$, which we can solve[39]: $y = -1$. So we know that the curve passes through $(1, -1)$ when $x = 1$.

- Now, to find the slope of the tangent line at $(1, -1)$, pretend that our curve is $y = f(x)$ so that $f(x)$ obeys

$$f(x) = f(x)^3 + xf(x) + x^3$$

for all x. Differentiating both sides gives

$$f'(x) = 3f(x)^2f'(x) + f(x) + xf'(x) + 3x^2$$

---

[39] This type of luck rarely happens in the "real world". But it happens remarkably frequently in textbooks, problem sets and tests.

- At this point we could isolate for $f'(x)$ and write it in terms of $f(x)$ and x, but since we only want answers when $x = 1$, let us substitute in $x = 1$ and $f(1) = -1$ (since the curve passes through $(1, -1)$) and clean things up before doing anything else.

- Subbing in $x = 1$, $f(1) = -1$ gives

$$f'(1) = 3f'(1) - 1 + f'(1) + 3 \qquad \text{and so } f'(1) = -\frac{2}{3}$$

- The equation of the tangent line is

$$y = y_0 + f'(x_0)(x - x_0) = -1 - \frac{2}{3}(x - 1) = -\frac{2}{3}x - \frac{1}{3}$$

We can further clean up the equation of the line to write it as $2x + 3y = -1$.

Example 2.11.1

In the previous example we replace y by $f(x)$ in the middle of the computation. We don't actually have to do this. When we are writing out our solution we can remember that y is a function of x. So we can start with

$$y = y^3 + xy + x^3$$

and differentiate remembering that $y \equiv y(x)$

$$y' = 3y^2 y' + xy' + y + 3x^2$$

And now substitute $x = 1, y = -1$ to get

$$y'(1) = 3 \cdot y'(1) + y'(1) - 1 + 3 \qquad \text{and so}$$
$$y'(1) = -\frac{2}{3}$$

The next one is at the same time a bit easier (because it is a quadratic) and a bit harder (because we are asked for the tangent at a general point on the curve, not a specific one).

Example 2.11.2

Let (x_0, y_0) be a point on the ellipse $3x^2 + 5y^2 = 7$. Find the equation for the tangent lines when $x = 1$ and y is positive. Then find an equation for the tangent line to the ellipse at a general point (x_0, y_0).

Since we are not given a specific point x_0 we are going to have to be careful with the second half of this question.

- When $x = 1$ the equation simplifies to

$$3 + 5y^2 = 7$$
$$5y^2 = 4$$
$$y = \pm\frac{2}{\sqrt{5}}.$$

We are only interested in positive y, so our point on the curve is $(1, 2/\sqrt{5})$.

- Now we use implicit differentiation to find $\frac{dy}{dx}$ at this point. First we pretend that we have solved the curve explicitly, for some interval of x's, as $y = f(x)$. The equation becomes

$$3x^2 + 5f(x)^2 = 7 \qquad \text{now differentiate}$$
$$6x + 10f(x)f'(x) = 0$$
$$f'(x) = -\frac{3x}{5f(x)}$$

- When $x = 1, y = 2/\sqrt{5}$ this becomes

$$f'(1) = -\frac{3}{5 \cdot 2/\sqrt{5}} = -\frac{3}{2\sqrt{5}}$$

So the tangent line passes through $(1, 2/\sqrt{5})$ and has slope $-\frac{3}{2\sqrt{5}}$. Hence the tangent line has equation

$$y = y_0 + f'(x_0)(x - x_0)$$
$$= \frac{2}{\sqrt{5}} - \frac{3}{2\sqrt{5}}(x - 1)$$
$$= \frac{7 - 3x}{2\sqrt{5}} \qquad \text{or equivalently}$$
$$3x + 2\sqrt{5}y = 7$$

Now we should go back and do the same but for a general point on the curve (x_0, y_0):

- A good first step here is to sketch the curve. Since this is an ellipse, it is pretty straight-forward.
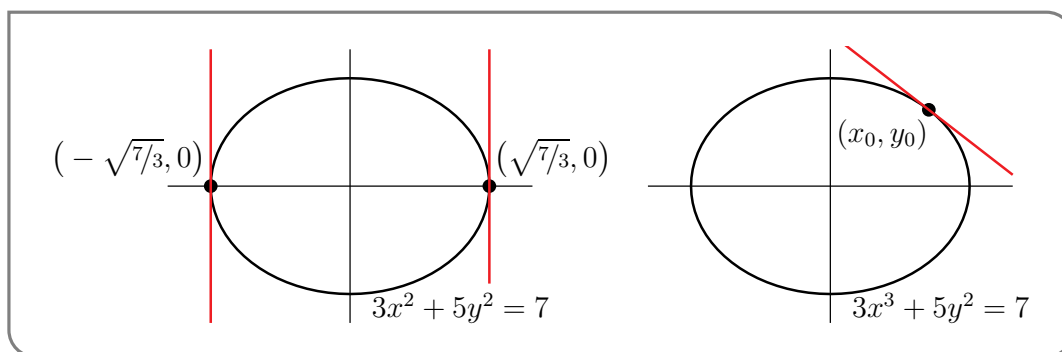
- Notice that there are two points on the ellipse — the extreme right and left points $(x_0, y_0) = \pm\left(\sqrt{7/3}, 0\right)$ — at which the tangent line is vertical. In those two cases, the tangent line is just $x = x_0$.

- Since this is a quadratic for y, we could solve it explicitly to get

$$y = \pm\sqrt{\frac{7 - 3x^2}{5}}$$

and choose the positive or negative branch as appropriate. Then we could differentiate to find the slope and put things together to get the tangent line.

But even in this relatively easy case, it is computationally cleaner, and hence less vulnerable to mechanical errors, to use implicit differentiation. So that's what we'll do.

- Now we could again "pretend" that we have solved the equation for the ellipse for $y = f(x)$ near (x_0, y_0), but let's not do that. Instead (as we did just before this example) just remember that when we differentiate y is really a function of x. So starting from

$$3x^2 + 5y^2 = 7 \qquad\qquad \text{differentiating gives}$$
$$6x + 5 \cdot 2y \cdot y' = 0$$

We can then solve this for y':

$$y' = -\frac{3x}{5y}$$

where y' and y are both functions of x.

- Hence at the point (x_0, y_0) we have

$$y'\big|_{(x_0, y_0)} = -\frac{3x_0}{5y_0}$$

This is the slope of the tangent line at (x_0, y_0) and so its equation is

$$y = y_0 + y' \cdot (x - x_0)$$
$$= y_0 - \frac{3x_0}{5y_0}(x - x_0)$$

We can simplify this by multiplying through by $5y_0$ to get

$$5y_0 y = 5y_0^2 - 3x_0 x + 3x_0^2$$

We can clean this up more by moving all the terms that contain x or y to the left-hand side and everything else to the right:

$$3x_0 x + 5y_0 y = 3x_0^2 + 5y_0^2$$

But there is one more thing we can do, our original equation is $3x^2 + 5y^2 = 7$ for all points on the curve, so we know that $3x_0^2 + 5y_0^2 = 7$. This cleans up the right-hand side.

$$3x_0 x + 5y_0 y = 7$$

- In deriving this formula for the tangent line at (x_0, y_0) we have assumed that $y_0 \neq 0$. But in fact the final answer happens to also work when $y_0 = 0$ (which means $x_0 = \pm\sqrt{7/3}$), so that the tangent line is $x = x_0$.

We can also check that our answer for general (x_0, y_0) reduces to our answer for $x_0 = 1$.

- When $x_0 = 1$ we worked out that $y_0 = 2/\sqrt{5}$.

- Plugging this into our answer above gives

$$3x_0 x + 5y_0 y = 7 \qquad\qquad \text{sub in } (x_0, y_0) = (1, 2/\sqrt{5}):$$

$$3x + 5\frac{2}{\sqrt{5}}y = 7 \qquad\qquad \text{clean up a little}$$

$$3x + 2\sqrt{5}y = 7$$

as required.

Example 2.11.2

Example 2.11.3

At which points does the curve $x^2 - xy + y^2 = 3$ cross the x–axis? Are the tangent lines to the curve at those points parallel?

This is a 2 part question — first the x-intercepts and then we need to examine tangent lines.

- Finding where the curve crosses the x-axis is straight forward. It does so when $y = 0$. This means x satisfies

$$x^2 - x \cdot 0 + 0^2 = 3 \qquad\qquad \text{so } x = \pm\sqrt{3}.$$

So the curve crosses the x–axis at two points $(\pm\sqrt{3}, 0)$.

- Now we need to find the tangent lines at those points. But we don't actually need the lines, just their slopes. Again we can pretend that near one of those points the curve is $y = f(x)$. Applying $\frac{d}{dx}$ to both sides of $x^2 - xf(x) + f(x)^2 = 3$ gives

$$2x - f(x) - xf'(x) + 2f(x)f'(x) = 0$$

etc etc.

- But let us stop "pretending". Just make sure we remember that y is a function of x when we differentiate:

$$x^2 - xy + y^2 = 3 \qquad \text{start with the curve, and differentiate}$$
$$2x - xy' - y + 2yy' = 0$$

Now substitute in the first point, $x = +\sqrt{3}, y = 0$:

$$2\sqrt{3} - \sqrt{3}y' + 0 = 0$$
$$y' = 2$$

And now do the second point $x = -\sqrt{3}, y = 0$:

$$-2\sqrt{3} + \sqrt{3}y' + 0 = 0$$
$$y' = 2$$

Thus the slope is the same at $x = \sqrt{3}$ and $x = -\sqrt{3}$ and the tangent lines are parallel.
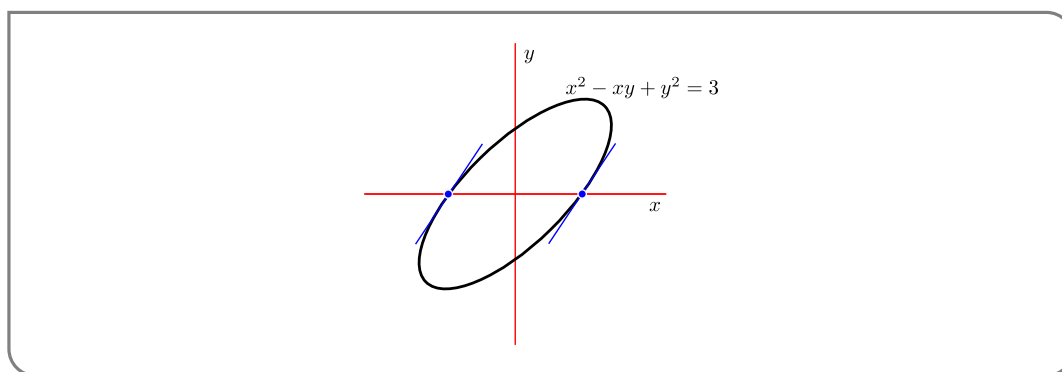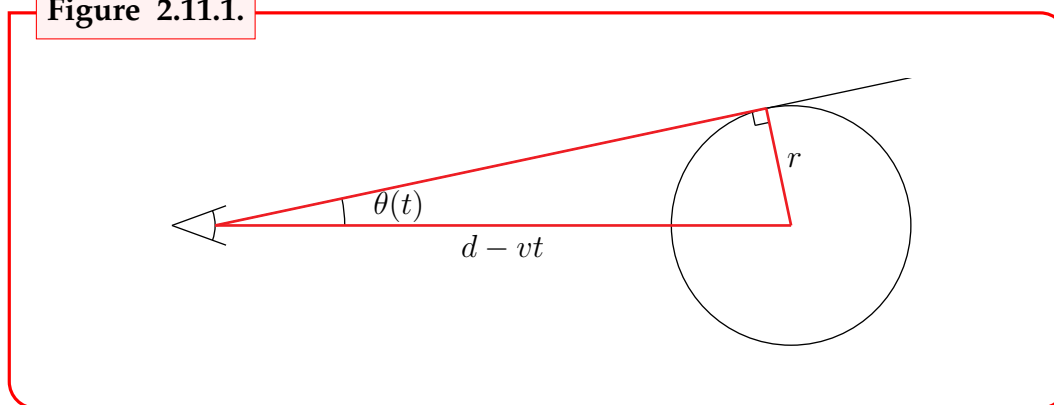
Example 2.11.3

Okay — let's get away from curves and do something a little different.

Example 2.11.4

You are standing at the origin. At time zero a pitcher throws a ball at your head[40].

Figure 2.11.1.

The position of the (centre of the) ball at time t is $x(t) = d - vt$, where d is the distance from your head to the pitcher's mound and v is the ball's velocity. Your eye sees the ball filling [41] an angle $2\theta(t)$ with

$$\sin\big(\theta(t)\big) = \frac{r}{d - vt}$$

where r is the radius of the baseball. The question is "How fast is θ growing at time t?" That is, what is $\frac{d\theta}{dt}$?

- We don't know (yet) how to solve this equation to find $\theta(t)$ explicitly. So we use implicit differentiation.

- To do so we apply $\frac{d}{dt}$ to both sides of our equation. This gives

$$\cos\big(\theta(t)\big) \cdot \theta'(t) = \frac{rv}{(d - vt)^2}$$

---

40   It seems that it is not a friendly game today.
41   This is the "visual angle" or "angular size".

- Then we solve for $\theta'(t)$:

$$\theta'(t) = \frac{rv}{(d - vt)^2 \cos\left(\theta(t)\right)}$$

- As is often the case, when using implicit differentiation, this answer is not very sat-isfying because it contains $\theta(t)$, for which we still do not have an explicit formula. However in this case we can get an explicit formula for $\cos\left(\theta(t)\right)$, without having an explicit formula for $\theta(t)$, just by looking at the right–angled triangle in Figure 2.11.1, above.

- The hypotenuse of that triangle has length $d - vt$. By Pythagoras, the length of the side of the triangle adjacent of the angle $\theta(t)$ is $\sqrt{(d - vt)^2 - r^2}$. So

$$\cos\left(\theta(t)\right) = \frac{\sqrt{(d - vt)^2 - r^2}}{d - vt}$$

and

$$\theta'(t) = \frac{rv}{(d - vt)\sqrt{(d - vt)^2 - r^2}}$$

Example 2.11.4

Okay — just one more tangent-to-the-curve example and then we'll go on to something different.

Example 2.11.5

Let (x_0, y_0) be a point on the astroid[42]

$$x^{2/3} + y^{2/3} = 1.$$

Find an equation for the tangent line to the astroid at (x_0, y_0).

- As was the case in examples above we can rewrite the equation of the astroid near (x_0, y_0) in the form $y = f(x)$, with an explicit $f(x)$, by solving the equation $x^{2/3} + y^{2/3} = 1$. But again, it is computationally cleaner, and hence less vulnerable to me-chanical errors, to use implicit differentiation. So that's what we'll do.

- First up, since (x_0, y_0) lies on the curve, it satisfies

$$x_0^{2/3} + y_0^{2/3} = 1.$$

---

42 Here is where is the astroid comes from. Imagine two circles, one of radius $1/4$ and one of radius 1. Paint a red dot on the smaller circle. Then imagine the smaller circle rolling around the inside of the larger circle. The curve traced by the red dot is our astroid. Google "astroid" (be careful about the spelling) to find animations showing this.
The astroid was first discussed by Johann Bernoulli in 1691–92. It also appears in the work of Leibniz.

- Now, no pretending that $y = f(x)$, this time — just make sure we remember when we differentiate that y changes with x.

$$x^{2/3} + y^{2/3} = 1 \qquad \text{start with the curve, and differentiate}$$

$$\frac{2}{3}x^{-1/3} + \frac{2}{3}y^{-1/3}y' = 0$$

- Note the derivative of $x^{2/3}$, namely $\frac{2}{3}x^{-1/3}$, and the derivative of $y^{2/3}$, namely $\frac{2}{3}y^{-1/3}y'$, are defined only when $x \neq 0$ and $y \neq 0$. We are interested in the case that $x = x_0$ and $y = y_0$. So we better assume that $x_0 \neq 0$ and $y_0 \neq 0$. Probably something weird happens when $x_0 = 0$ or $y_0 = 0$. We'll come back to this shortly.

- To continue on, we set $x = x_0, y = y_0$ in the equation above, and then solve for y':

$$\frac{2}{3}x_0^{-1/3} + \frac{2}{3}y_0^{-1/3}y'(x) = 0 \implies y'(x_0) = -\left(\frac{y_0}{x_0}\right)^{1/3}$$

This is the slope of the tangent line and its equation is

$$y = y_0 + f'(x_0)(x - x_0) = y_0 - \left(\frac{y_0}{x_0}\right)^{1/3}(x - x_0)$$

Now let's think a little bit about what the tangent line slope of $-\sqrt[3]{y_0/x_0}$ tells us about the astroid.

- First, as a preliminary observation, note that since $x_0^{2/3} \geq 0$ and $y_0^{2/3} \geq 0$ the equation $x_0^{2/3} + y_0^{2/3} = 1$ of the astroid forces $0 \leq x_0^{2/3}, y_0^{2/3} \leq 1$ and hence $-1 \leq x_0, y_0 \leq 1$.

- For all $x_0, y_0 > 0$ the slope $-\sqrt[3]{y_0/x_0} < 0$. So at all points on the astroid that are in the first quadrant, the tangent line has negative slope, i.e. is "leaning backwards".

- As x_0 tends to zero, y_0 tends to ± 1 and the tangent line slope tends to infinity. So at points on the astroid near $(0, \pm 1)$, the tangent line is almost vertical.

- As y_0 tends to zero, x_0 tends to ± 1 and the tangent line slope tends to zero. So at points on the astroid near $(\pm 1, 0)$, the tangent line is almost horizontal.

Here is a figure illustrating all this.
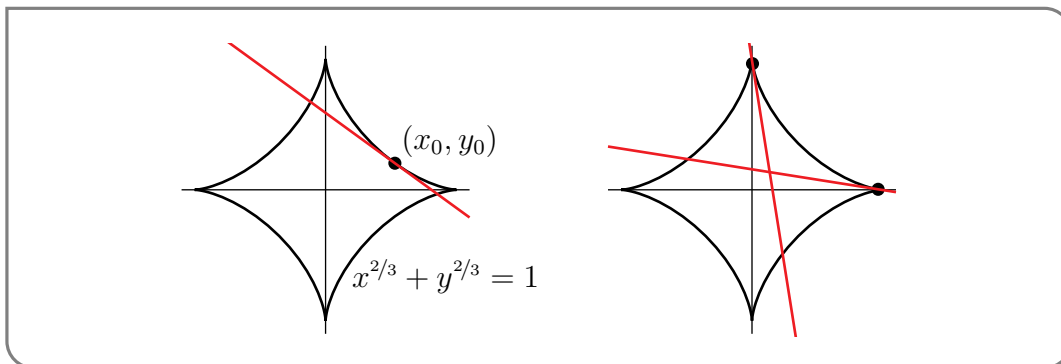
Sure enough, as we speculated earlier, something weird does happen to the astroid when x_0 or y_0 is zero. The astroid is pointy, and does not have a tangent there.

Example 2.11.5

2.12 ▲ Inverse Trigonometric Functions

One very useful application of implicit differentiation is to find the derivatives of inverse functions. We have already used this approach to find the derivative of the inverse of the exponential function — the logarithm.

We are now going to consider the problem of finding the derivatives of the inverses of trigonometric functions. Now is a very good time to go back and reread Section 0.6 on inverse functions — especially Definition 0.6.3. Most importantly, given a function $f(x)$, its inverse function $f^{-1}(x)$ only exists, with domain D, when $f(x)$ passes the "horizontal line test", which says that for each Y in D the horizontal line $y = Y$ intersects the graph $y = f(x)$ exactly once. (That is, $f(x)$ is a one-to-one function.)

Let us start by playing with the sine function and determine how to restrict the domain of $\sin x$ so that its inverse function exists.

Example 2.12.1

Let $y = f(x) = \sin(x)$. We would like to find the inverse function which takes y and returns to us a unique x-value so that $\sin(x) = y$.
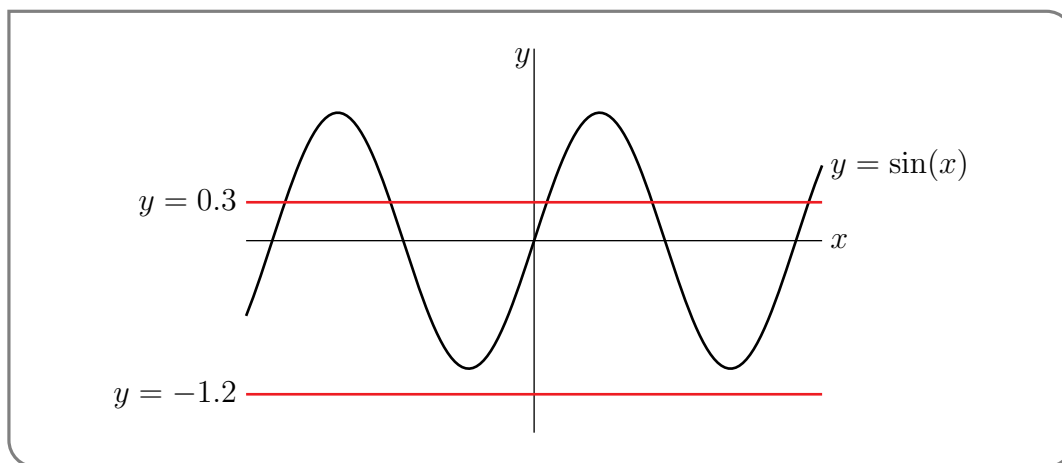
- For each real number Y, the number of x-values that obey $\sin(x) = Y$, is exactly the number of times the horizontal straight line $y = Y$ intersects the graph of $\sin(x)$.

- When $-1 \leqslant Y \leqslant 1$, the horizontal line intersects the graph infinitely many times. This is illustrated in the figure above by the line $y = 0.3$.

- On the other hand, when $Y < -1$ or $Y > 1$, the line $y = Y$ never intersects the graph of $\sin(x)$. This is illustrated in the figure above by the line $y = -1.2$.

This is exactly the horizontal line test and it shows that the sine function is not one-to-one.

Now consider the function

$$y = \sin(x) \qquad\qquad \text{with domain } -\frac{\pi}{2} \leqslant x \leqslant \frac{\pi}{2}$$

This function has the same formula but the domain has been restricted so that, as we'll now show, the horizontal line test is satisfied.
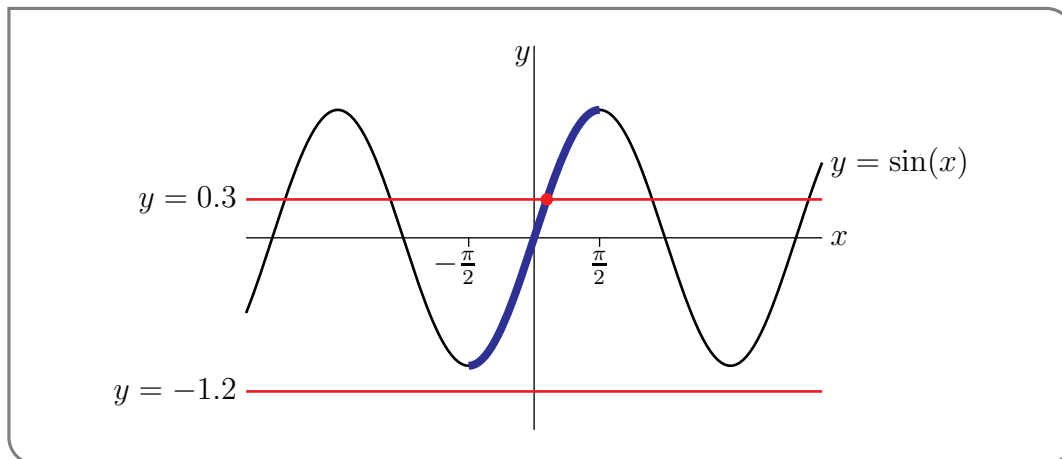
As we saw above when $|Y| > 1$ no x obeys $\sin(x) = Y$ and, for each $-1 \leqslant Y \leqslant 1$, the line $y = Y$ (illustrated in the figure above with $y = 0.3$) crosses the curve $y = \sin(x)$ infinitely many times, so that there are infinitely many x's that obey $f(x) = \sin x = Y$. However exactly one of those crossings (the dot in the figure) has $-\pi/2 \leqslant x \leqslant \pi/2$.

That is, for each $-1 \leqslant Y \leqslant 1$, there is exactly one x, call it X, that obeys both

$$\sin X = Y \qquad\qquad \text{and} \qquad\qquad -\frac{\pi}{2} \leqslant X \leqslant \frac{\pi}{2}$$

That unique value, X, is typically denoted $\arcsin(Y)$. That is

$$\sin(\arcsin(Y)) = Y \qquad\qquad \text{and} \qquad\qquad -\frac{\pi}{2} \leqslant \arcsin(Y) \leqslant \frac{\pi}{2}$$

Renaming $Y \to x$, the inverse function $\arcsin(x)$ is defined for all $-1 \leqslant x \leqslant 1$ and is determined by the equation

$$\sin\big(\arcsin(x)\big) = x \qquad \text{and} \qquad -\frac{\pi}{2} \leqslant \arcsin(x) \leqslant \frac{\pi}{2}. \qquad\qquad (2.12.1)$$

Note that many texts will use $\sin^{-1}(x)$ to denote arcsine, however we will use $\arcsin(x)$ since we feel that it is clearer[43]; the reader should recognise both.

Example 2.12.1

Example 2.12.2

Since

$$\sin\frac{\pi}{2} = 1 \qquad \sin\frac{\pi}{6} = \frac{1}{2}$$

---

43   The main reason being that people frequently confuse $\sin^{-1}(x)$ with $(\sin(x))^{-1} = \frac{1}{\sin x}$. We feel that prepending the prefix "arc" less likely to lead to such confusion. The notations $\mathrm{asin}(x)$ and $\mathrm{Arcsin}(x)$ are also used.

and $-\pi/2 \leqslant \pi/6, \pi/2 \leqslant \pi/2$, we have

$$\arcsin 1 = \frac{\pi}{2} \qquad \arcsin \frac{1}{2} = \frac{\pi}{6}$$

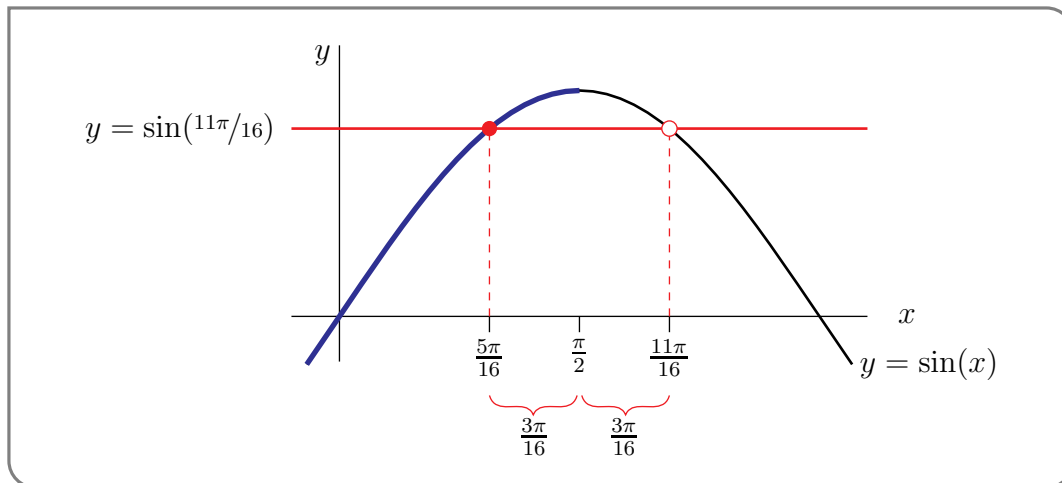Even though

$$\sin(2\pi) = 0$$

it is **not** true that $\arcsin 0 = 2\pi$, and it is **not** true that $\arcsin\big(\sin(2\pi)\big) = 2\pi$, because 2π is not between $-\pi/2$ and $\pi/2$. More generally

$$\arcsin\big(\sin(x)\big) = \text{ the unique angle } \theta \text{ between } -\pi/2 \text{ and } \pi/2 \text{ obeying } \sin\theta = \sin x$$
$$= x \quad \text{if and only if } -\pi/2 \leqslant x \leqslant \pi/2$$

So, for example, $\arcsin\big(\sin\left(11\pi/16\right)\big)$ cannot be $11\pi/16$ because $11\pi/16$ is bigger than $\pi/2$. So how do we find the correct answer? Start by sketching the graph of $\sin(x)$.

It looks like the graph of $\sin x$ is symmetric about $x = \pi/2$. The mathematical way to say that "the graph of $\sin x$ is symmetric about $x = \pi/2$" is "$\sin(\pi/2 - \theta) = \sin(\pi/2 + \theta)$" for all θ. That is indeed true[44].

Now $11\pi/16 = \pi/2 + 3\pi/16$ so

$$\sin\left(\frac{11\pi}{16}\right) = \sin\left(\frac{\pi}{2} + \frac{3\pi}{16}\right) = \sin\left(\frac{\pi}{2} - \frac{3\pi}{16}\right) = \sin\left(\frac{5\pi}{16}\right)$$

and, since $5\pi/16$ is indeed between $-\pi/2$ and $\pi/2$,

$$\arcsin\left(\sin\left(\frac{11\pi}{16}\right)\right) = \frac{5\pi}{16} \qquad \left(\text{and } \textbf{not } \frac{11\pi}{16}\right).$$

Example 2.12.2

---

44   Indeed both are equal to $\cos\theta$. You can see this by playing with the trig identities in Appendix A.8.

▸▸ Derivatives of Inverse Trig Functions

Now that we have explored the arcsine function we are ready to find its derivative. Lets call

$$\arcsin(x) = \theta(x),$$

so that the derivative we are seeking is $\frac{d\theta}{dx}$. The above equation is (after taking sine of both sides) equivalent to

$$\sin(\theta) = x$$

Now differentiate this using implicit differentiation (we just have to remember that θ varies with x and use the chain rule carefully):

$$\cos(\theta) \cdot \frac{d\theta}{dx} = 1$$
$$\frac{d\theta}{dx} = \frac{1}{\cos(\theta)} \qquad \text{substitute } \theta = \arcsin x$$
$$\frac{d}{dx} \arcsin x = \frac{1}{\cos(\arcsin x)}$$

This doesn't look too bad, but it's not really very satisfying because the right hand side is expressed in terms of $\arcsin(x)$ and we do not have an explicit formula for $\arcsin(x)$.

However even without an explicit formula for $\arcsin(x)$, it is a simple matter to get an explicit formula for $\cos\left(\arcsin(x)\right)$, which is all we need. Just draw a right–angled triangle with one angle being $\arcsin(x)$. This is done in the figure below[45].
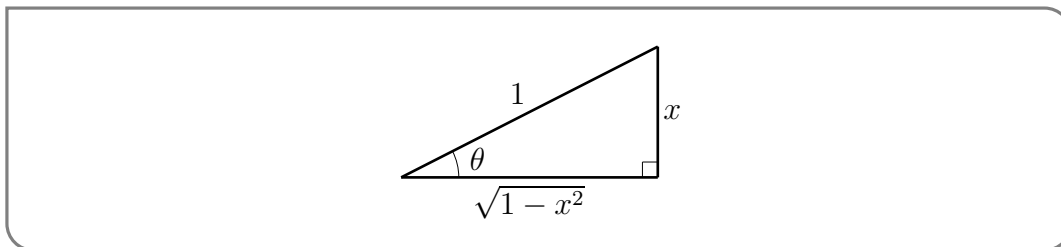
Since $\sin(\theta) = x$ (see (2.12.1)), we have made the side opposite the angle θ of length x and the hypotenuse of length 1. Then, by Pythagoras, the side adjacent to θ has length $\sqrt{1 - x^2}$ and so

$$\cos\left(\arcsin(x)\right) = \cos(\theta) = \sqrt{1 - x^2}$$

which in turn gives us the answer we need:

$$\frac{d}{dx} \arcsin(x) = \frac{1}{\sqrt{1 - x^2}}$$

The definitions for arccos, arctan and arccot are developed in the same way. Here are the graphs that are used.

---

45   The figure is drawn for the case that $0 \leqslant \arcsin(x) \leqslant \pi/2$. Virtually the same argument works for the case $-\pi/2 \leqslant \arcsin(x) \leqslant 0$
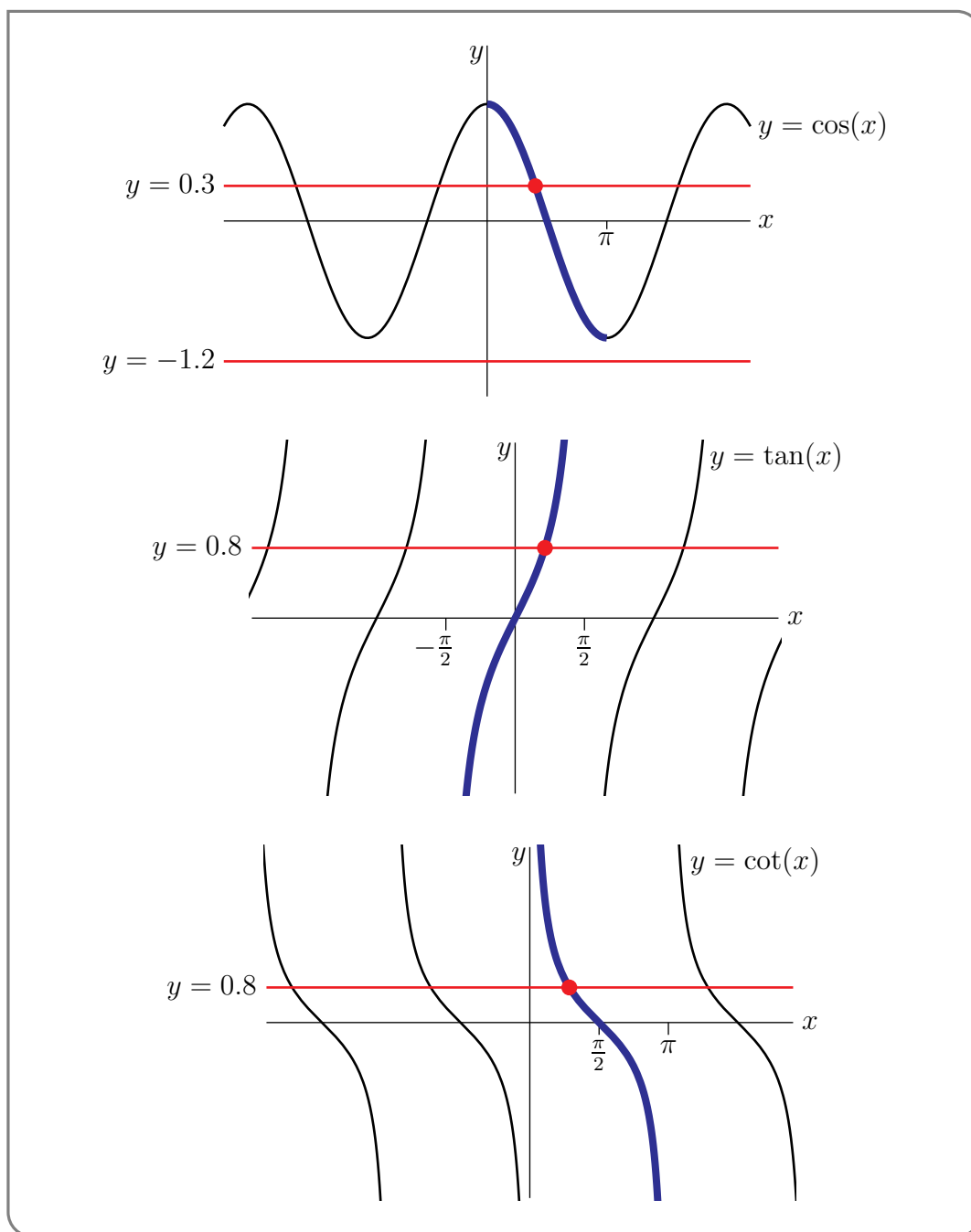
The definitions for the remaining two inverse trigonometric functions may also be developed in the same way[46][47]. But it's a little easier to use

$$\csc x = \frac{1}{\sin x} \qquad \sec x = \frac{1}{\cos x}$$

46 In fact, there are two different widely used definitions of arcsec x. Under our definition, below, $\theta = \text{arcsec}\, x$ takes values in $0 \leqslant \theta \leqslant \pi$. Some people, perfectly legitimately, define $\theta = \text{arcsec}\, x$ to take values in the union of $0 \leqslant \theta < \frac{\pi}{2}$ and $\pi \leqslant \theta < \frac{3\pi}{2}$. Our definition is sometimes called the "trigonometry friendly" definition. The definition itself has the advantage of simplicity. The other definition is sometimes called the "calculus friendly" definition. It eliminates some absolute values and hence simplifies some computations. Similarly, there are two different widely used definitions of arccsc x.

47 One could also define $\text{arccot}(x) = \text{arctan}(1/x)$ with $\text{arccot}(0) = \frac{\pi}{2}$. We have chosen not to do so, because the definition we have chosen is both continuous and standard.

Definition 2.12.3.

arcsin x is defined for $|x| \leqslant 1$. It is the unique number obeying

$$\sin\big(\arcsin(x)\big) = x \qquad \text{and} \qquad -\frac{\pi}{2} \leqslant \arcsin(x) \leqslant \frac{\pi}{2}$$

arccos x is defined for $|x| \leqslant 1$. It is the unique number obeying

$$\cos\big(\arccos(x)\big) = x \qquad \text{and} \qquad 0 \leqslant \arccos(x) \leqslant \pi$$

arctan x is defined for all $x \in \mathbb{R}$. It is the unique number obeying

$$\tan\big(\arctan(x)\big) = x \qquad \text{and} \qquad -\frac{\pi}{2} < \arctan(x) < \frac{\pi}{2}$$

$\operatorname{arccsc} x = \arcsin\frac{1}{x}$ is defined for $|x| \geqslant 1$. It is the unique number obeying

$$\csc\big(\operatorname{arccsc}(x)\big) = x \qquad \text{and} \qquad -\frac{\pi}{2} \leqslant \operatorname{arccsc}(x) \leqslant \frac{\pi}{2}$$

Because $\csc(0)$ is undefined, $\operatorname{arccsc}(x)$ never takes the value 0.

$\operatorname{arcsec} x = \arccos\frac{1}{x}$ is defined for $|x| \geqslant 1$. It is the unique number obeying

$$\sec\big(\operatorname{arcsec}(x)\big) = x \qquad \text{and} \qquad 0 \leqslant \operatorname{arcsec}(x) \leqslant \pi$$

Because $\sec(\pi/2)$ is undefined, $\operatorname{arcsec}(x)$ never takes the value $\pi/2$.

arccot x is defined for all $x \in \mathbb{R}$. It is the unique number obeying

$$\cot\big(\operatorname{arccot}(x)\big) = x \qquad \text{and} \qquad 0 < \operatorname{arccot}(x) < \pi$$

Example 2.12.4

To find the derivative of arccos we can follow the same steps:

- Write $\arccos(x) = \theta(x)$ so that $\cos\theta = x$ and the desired derivative is $\frac{d\theta}{dx}$.

- Differentiate implicitly, remembering that θ is a function of x:

$$-\sin\theta\frac{d\theta}{dx} = 1$$
$$\frac{d\theta}{dx} = -\frac{1}{\sin\theta}$$
$$\frac{d}{dx}\arccos x = -\frac{1}{\sin(\arccos x)}.$$

- To simplify this expression, again draw the relevant triangle
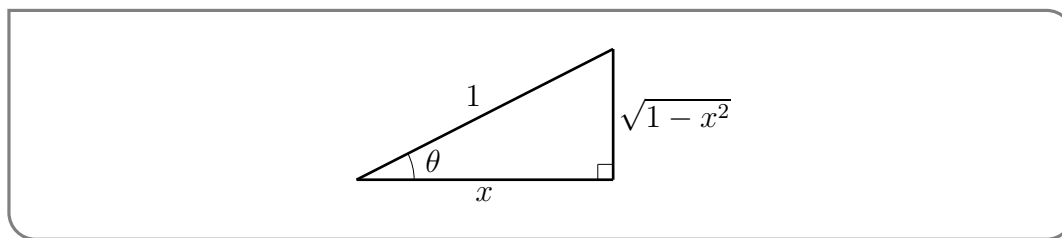
from which we see

$$\sin(\arccos x) = \sin\theta = \sqrt{1-x^2}.$$

- Thus

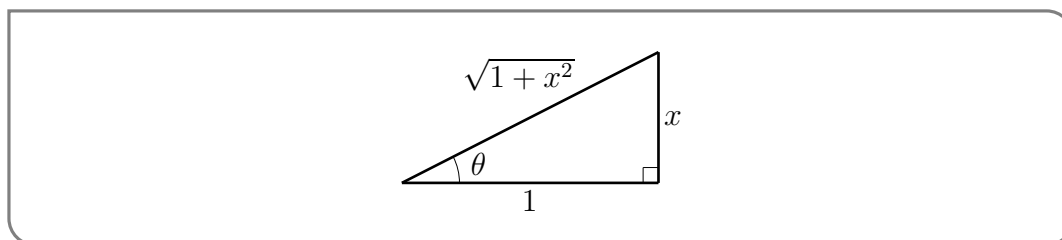$$\frac{d}{dx}\arccos x = -\frac{1}{\sqrt{1-x^2}}.$$

Example 2.12.4

Example 2.12.5

Very similar steps give the derivative of $\arctan x$:

- Start with $\theta = \arctan x$, so $\tan\theta = x$.

- Differentiate implicitly:

$$\sec^2\theta\frac{d\theta}{dx} = 1$$
$$\frac{d\theta}{dx} = \frac{1}{\sec^2\theta} = \cos^2\theta$$
$$\frac{d}{dx}\arctan x = \cos^2(\arctan x).$$

- To simplify this expression, we draw the relevant triangle

from which we see

$$\cos^2(\arctan x) = \cos^2\theta = \frac{1}{1+x^2}$$

- Thus

$$\frac{d}{dx} \arctan x = \frac{1}{1+x^2}.$$

An almost identical computation gives the derivative of arccot x:

- Start with $\theta = \text{arccot } x$, so $\cot \theta = x$.

- Differentiate implicitly:

$$-\csc^2 \theta \frac{d\theta}{dx} = 1$$

$$\frac{d}{dx} \text{arccot } x = \frac{d\theta}{dx} = -\frac{1}{\csc^2 \theta} = -\sin^2 \theta = -\frac{1}{1+x^2}$$

from the triangle
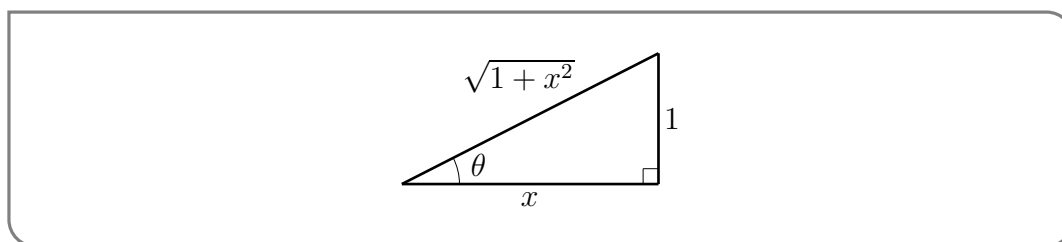
Example 2.12.5

Example 2.12.6

To find the derivative of arccsc we can use its definition and the chain rule.

$$\theta = \text{arccsc } x \qquad \qquad \text{take cosecant of both sides}$$

$$\csc \theta = x \qquad \qquad \text{but } \csc \theta = \frac{1}{\sin \theta}, \text{ so flip both sides}$$

$$\sin \theta = \frac{1}{x} \qquad \qquad \text{now take arcsine of both sides}$$

$$\theta = \arcsin \left(\frac{1}{x} \right)$$

Now just differentiate:

$$\frac{d\theta}{dx} = \frac{d}{dx} \arcsin \left(\frac{1}{x} \right) \qquad \qquad \text{chain rule carefully}$$

$$= \frac{1}{\sqrt{1-x^{-2}}} \cdot \frac{-1}{x^2}$$

To simplify further we will factor x^{-2} out of the square root. We need to be a little careful doing that. Take another look at examples 1.5.6 and 1.5.7 and the discussion between them before proceeding.

$$= \frac{1}{\sqrt{x^{-2}(x^2-1)}} \cdot \frac{-1}{x^2}$$

$$= \frac{1}{|x^{-1}| \cdot \sqrt{x^2-1}} \cdot \frac{-1}{x^2} \qquad \text{note that } x^2 \cdot |x^{-1}| = |x|.$$

$$= -\frac{1}{|x|\sqrt{x^2-1}}$$

In the same way, we can find the derivative of the remaining inverse trig function. We just use its definition, a derivative we already know and the chain rule.

$$\frac{\mathrm{d}}{\mathrm{d}x} \operatorname{arcsec}(x) = \frac{\mathrm{d}}{\mathrm{d}x} \arccos\left(\frac{1}{x}\right) = -\frac{1}{\sqrt{1-1/x^2}} \cdot \left(-\frac{1}{x^2}\right) = \frac{1}{|x|\sqrt{x^2-1}}$$

Example 2.12.6

By way of summary, we have

Theorem 2.12.7.

The derivatives of the inverse trigonometric functions are

$$\frac{\mathrm{d}}{\mathrm{d}x} \arcsin(x) = \frac{1}{\sqrt{1-x^2}} \qquad\qquad \frac{\mathrm{d}}{\mathrm{d}x} \operatorname{arccsc}(x) = -\frac{1}{|x|\sqrt{x^2-1}}$$

$$\frac{\mathrm{d}}{\mathrm{d}x} \arccos(x) = -\frac{1}{\sqrt{1-x^2}} \qquad\qquad \frac{\mathrm{d}}{\mathrm{d}x} \operatorname{arcsec}(x) = \frac{1}{|x|\sqrt{x^2-1}}$$

$$\frac{\mathrm{d}}{\mathrm{d}x} \arctan(x) = \frac{1}{1+x^2} \qquad\qquad \frac{\mathrm{d}}{\mathrm{d}x} \operatorname{arccot}(x) = -\frac{1}{1+x^2}$$

2.13 ▴ The Mean Value Theorem

Consider the following situation. Two towns are separated by a 120km long stretch of road. The police in town A observe a car leaving at 1pm. Their colleagues in town B see the car arriving at 2pm. After a quick phone call between the two police stations, the driver is issued a fine for going $120km/h$ at some time between 1pm and 2pm. It is intuitively obvious[48] that, because his average velocity was $120km/h$, the driver must have been going at least $120km/h$ at some point. From a knowledge of the average velocity of the car, we are able to deduce something about an instantaneous velocity[49].

---

48 Unfortunately there are many obvious things that are decidedly false — for example "There are more rational numbers than integers." or "Viking helmets had horns on them".
49 Recall that speed and velocity are not the same.

Let us turn this around a little bit. Consider the premise of a 90s action film[50] — a bus must travel at a velocity of no less than $80km/h$. Being a bus, it is unable to go faster than, say, $120km/h$. The film runs for about 2 hours, and let's assume that there is about thirty minutes of non-action — so the bus' velocity is constrained between 80 and $120km/h$ for a total of 1.5 hours.

It is again obvious that the bus must have travelled between $80 \times 1.5 = 120$ and $120 \times 1.5 = 180km$ during the film. This time, from a knowledge of the instantaneous rate of change of position — the derivative — throughout a 90 minute time interval, we are able to say something about the net change of position during the 90 minutes.

In both of these scenarios we are making use of a piece of mathematics called the Mean Value Theorem. It says that, under appropriate hypotheses, the average rate of change $\frac{f(b)-f(a)}{b-a}$ of a function over an interval is achieved exactly by the instantaneous rate of change $f'(c)$ of the function at some[51] (unknown) point $a \leqslant c \leqslant b$. We shall get to a precise statement in Theorem 2.13.4. We start working up to it by first considering the special case in which $f(a) = f(b)$.

▸▸ Rolle's Theorem

Theorem 2.13.1 (Rolle's theorem).

Let a and b be real numbers with $a < b$. And let f be a function so that

- $f(x)$ is continuous on the closed interval $a \leqslant x \leqslant b$,

- $f(x)$ is differentiable on the open interval $a < x < b$, and

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

then there is a c strictly between a and b, i.e. obeying $a < c < b$, such that

$$f'(c) = 0.$$

Again, like the two scenarios above, this theorem says something intuitively obvious. Consider — if you throw a ball straight up into the air and then catch it, at some time in between the throw and the catch it must be stationary. Translating this into mathematical statements, let $s(t)$ be the height of the ball above the ground in metres, and let t be time

---

- Velocity specifies the direction of motion as well as the rate of change. Objects moving along a straight line have velocities that are positive or negative numbers indicating which direction the object is moving along the line.

- Speed, on the other hand, is the distance travelled per unit time and is always a non-negative number — it is the absolute value of velocity.

50   The sequel won a Raspberry award for "Worst remake or sequel".
51   There must be at least one such point — there could be more than one — but there cannot be zero.

from the moment the ball is thrown in seconds. Then we have

$$s(0) = 1 \qquad \text{we release the ball at about hip-height}$$
$$s(4) = 1 \qquad \text{we catch the ball } 4s \text{ later at hip-height}$$

Then we know there is some time in between — say at $t = c$ — when the ball is stationary (in this case when the ball is at the top of its trajectory). I.e.

$$v(c) = s'(c) = 0.$$

Rolle's theorem guarantees that for any differentiable function that starts and ends at the same value, there will always be at least one point between the start and finish where the derivative is zero.

Figure 2.13.1.
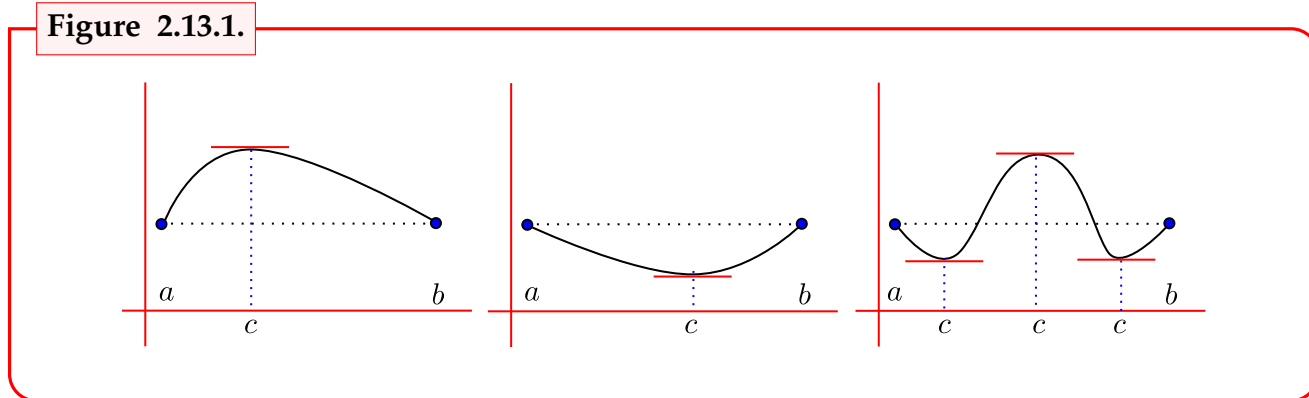
There can, of course, also be multiple points at which the derivative is zero — but there must always be at least one. Notice, however, the theorem[52] does not tell us the value of c, just that such a c must exist.

Example 2.13.2

We can use Rolle's theorem to show that the function

$$f(x) = \sin(x) - \cos(x)$$

has a point c between 0 and $\frac{3\pi}{2}$ so that $f'(c) = 0$.

To apply Rolle's theorem we first have to show the function satisfies the conditions of the theorem on the interval $[0, \frac{3\pi}{2}]$.

- Since f is the sum of sine and cosine it is continuous on the interval and also differentiable on the interval.

- Further, since

$$f(0) = \sin 0 - \cos 0 = 0 - 1 = -1$$
$$f\left(\frac{3\pi}{2}\right) = \sin \frac{3\pi}{2} - \cos \frac{3\pi}{2} = -1 - 0 = -1$$

we can now apply Rolle's theorem.

---

52   Notice this is very similar to the intermediate value theorem (see Theorem 1.6.12)

- Rolle's theorem implies that there must be a point $c \in (0, 3\pi/2)$ so that $f'(c) = 0$.

While Rolle's theorem doesn't tell us the value of c, this example is sufficiently simple that we can find it directly.
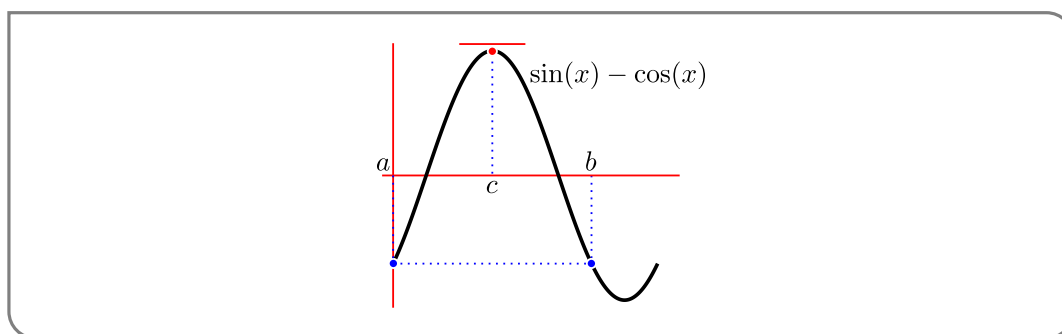
$$
\begin{aligned}
f'(x) &= \cos x + \sin x \\
f'(c) &= \cos c + \sin c = 0 &&\text{rearrange} \\
\sin c &= -\cos c &&\text{and divide by } \cos c \\
\tan c &= -1
\end{aligned}
$$

Hence $c = \frac{3\pi}{4}$. We have sketched the function and the relevant points below.

$\sin(x) - \cos(x)$

A more substantial application of Rolle's theorem (in conjunction with the intermediate value theorem — Theorem 1.6.12) is to show that a function does not have multiple zeros in an interval:

Example 2.13.3

Show that the equation $2x - 1 = \sin(x)$ has exactly 1 solution.

- Start with a rough sketch of each side of the equation

$2x - 1$

$\sin x$
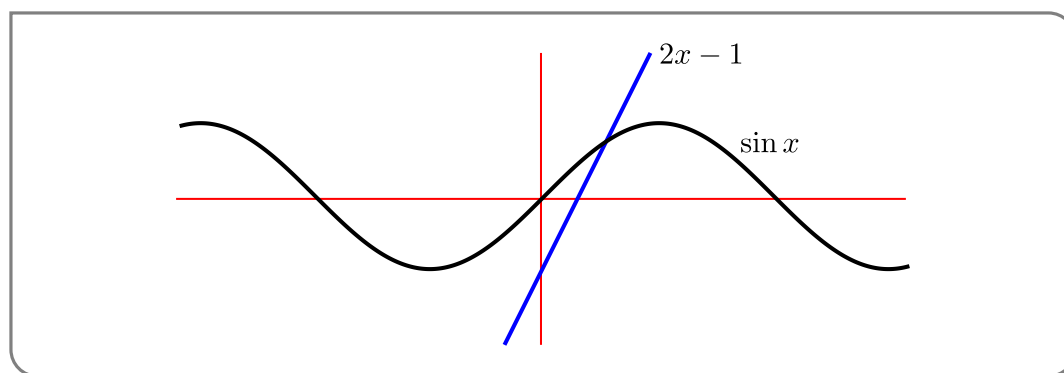
This seems like it should be true.

- Notice that the problem we are trying to solve is equivalent to showing that the function

$$
f(x) = 2x - 1 - \sin(x)
$$

has only a single zero.

- Since $f(x)$ is the sum of a polynomial and a sine function, it is continuous and differentiable everywhere. Thus we can apply both the IVT and Rolle's theorem.

- Notice that $f(0) = -1$ and $f(2) = 4 - 1 - \sin(2) = 3 - \sin(2) \geqslant 2$, since $-1 \leqslant \sin(2) \leqslant 1$. Thus by the IVT we know there is at least one number c between 0 and 2 so that $f(c) = 0$.

- But our job is only half done — this shows that there is at least one zero, but it does not tell us there is no more than one. We have more work to do, and Rolle's theorem is the tool we need.

- Consider what would happen if $f(x)$ is zero in 2 places — that is, there are numbers a, b so that $f(a) = f(b) = 0$.

  - Since $f(x)$ is differentiable everywhere and $f(a) = f(b) = 0$, we can apply Rolle's theorem.

  - Hence we know there is a point c between a and b so that $f'(c) = 0$.

  - But let us examine $f'(x)$:

  $$f'(x) = 2 - \cos x$$

  Since $-1 \leqslant \cos x \leqslant 1$, we must have that $f'(x) \geqslant 1$.

  - But this contradicts Rolle's theorem which tells us there must be a point at which the derivative is zero.

  Thus the function cannot be zero at two different places — otherwise we'd have a contradiction.
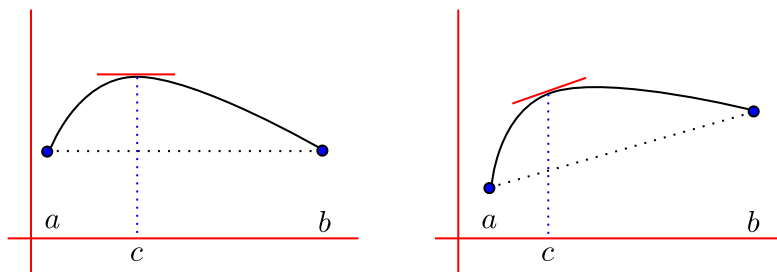
We can actually nail down the value of c using the bisection approach we used in example 1.6.15. If we do this carefully we find that $c \approx 0.887862\ldots$

Example 2.13.3

⇥ Back to the MVT

Rolle's theorem can be generalised in a straight-forward way; given a differentiable function $f(x)$ we can still say something about $\frac{df}{dx}$, even if $f(a) \neq f(b)$. Consider the following sketch:

Figure 2.13.2.

All we have done is tilt the picture so that $f(a) < f(b)$. Now we can no longer guarantee that there will be a point on the graph where the tangent line is horizontal, but there will be a point where the tangent line is parallel to the secant joining $(a, f(a))$ to $(b, f(b))$.

To state this in terms of our first scenario back at the beginning of this section, suppose that you are driving along the x–axis. At time $t = a$ you are at $x = f(a)$ and at time $t = b$ you are at $x = f(b)$. For simplicity, let's suppose that $b > a$ and $f(b) \geqslant f(a)$, just like in the above sketch. Then during the time interval in question you travelled a net distance of $f(b) - f(a)$. It took you $b - a$ units of time to travel that distance, so your average velocity was $\frac{f(b)-f(a)}{b-a}$. You may very well have been going faster than $\frac{f(b)-f(a)}{b-a}$ part of the time and slower than $\frac{f(b)-f(a)}{b-a}$ part of the time. But it is reasonable to guess that at some time between $t = a$ and $t = b$ your instantaneous velocity was exactly $\frac{f(b)-f(a)}{b-a}$. The mean value theorem says that, under reasonable assumptions about f, this is indeed the case.

Theorem 2.13.4 (The mean value theorem).

Let a and b be real numbers with $a < b$. And let $f(x)$ be a function so that

- $f(x)$ is continuous on the closed interval $a \leqslant x \leqslant b$, and

- $f(x)$ is differentiable on the open interval $a < x < b$

then there is a $c \in (a, b)$, such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$

which we can also express as

$$f(b) = f(a) + f'(c)(b - a).$$

Let us start to explore the mean value theorem — which is very frequently known as the MVT. A simple example to start:

Example 2.13.5

Consider the polynomial $f(x) = 3x^2 - 4x + 2$ on $[-1, 1]$.

- Since f is a polynomial it is continuous on the interval and also differentiable on the interval. Hence we can apply the MVT.

- The MVT tells us that there is a point $c \in (-1, 1)$ so that
$$f'(c) = \frac{f(1) - f(-1)}{1 - (-1)} = \frac{1 - 9}{2} = -4$$

This example is sufficiently simple that we can find the point c and the corresponding tangent line:

- The derivative is
$$f'(x) = 6x - 4$$

- So we need to solve $f'(c) = -4$:
$$6c - 4 = -4$$
which tells us that $c = 0$.

- The tangent line has slope -4 and passes through $(0, f(0)) = (0, 2)$, and so is given by
$$y = -4x + 2$$

- The secant line joining $(-1, f(-1)) = (-1, 9)$ to $(1, f(1)) = (1, 1)$ is just
$$y = 5 - 4x$$

- Here is a sketch of the curve and the two lines:
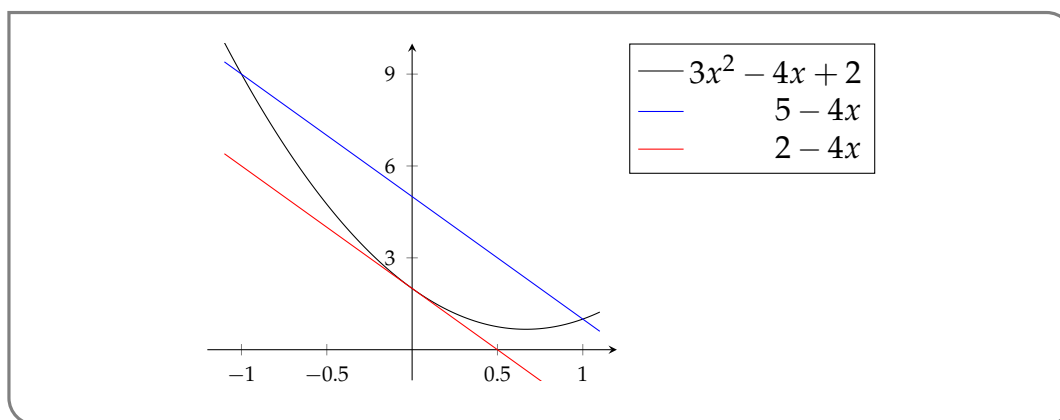
Example 2.13.5

Example 2.13.6

We can return to our initial car-motivated examples. Say you are driving along a straight road in a car that can go at most $80km/h$. How far can you go in 2 hours? — the answer is easy, but we can also solve this using MVT.

- Let $s(t)$ be the position of the car in km at time t measured in hours.

- Then $s(0) = 0$ and $s(2) = q$, where q is the quantity that we need to bound.

- We are told that $|s'(t)| \leqslant 80$, or equivalently

$$-80 \leqslant s'(t) \leqslant 80$$

- By the MVT there is some c between 0 and 2 so that

$$s'(c) = \frac{q-0}{2} = \frac{q}{2}$$

- Now since $-80 \leqslant s'(c) \leqslant 80$ we must have $-80 \leqslant q/2 \leqslant 80$ and hence $-160 \leqslant q = s(2) \leqslant 160$.

Example 2.13.6

More generally if we have some information about the derivative, then we can use the MVT to leverage this information to tell us something about the function.

Example 2.13.7

Let $f(x)$ be a differentiable function so that

$$f(1) = 10 \qquad \text{and} \qquad -1 \leqslant f'(x) \leqslant 2 \text{ everywhere}$$

Obtain upper and lower bounds on $f(5)$.
    Okay — what do we do?

- Since $f(x)$ is differentiable we can use the MVT.

- Say $f(5) = q$, then the MVT tells us that there is some c between 1 and 5 such that

$$f'(c) = \frac{q-10}{5-1} = \frac{q-10}{4}$$

- But we know that $-1 \leqslant f'(c) \leqslant 2$, so

$$-1 \leqslant f'(c) \leqslant 2$$
$$-1 \leqslant \frac{q-10}{4} \leqslant 2$$
$$-4 \leqslant q - 10 \leqslant 8$$
$$6 \leqslant q \leqslant 18$$

- Thus we must have $6 \leqslant f(5) \leqslant 18$.

Example 2.13.7

▸▸▸ (Optional) — Why is the MVT True?

We won't give a real proof for this theorem, but we'll look at a picture which shows why it is true. Here is the picture. It contains a sketch of the graph of $f(x)$, with x running from a to b, as well as a line segment which is the secant of the graph from the point $(a, f(a))$ to the point $(b, f(b))$. The slope of the secant is exactly $\frac{f(b)-f(a)}{b-a}$. Remember that we are
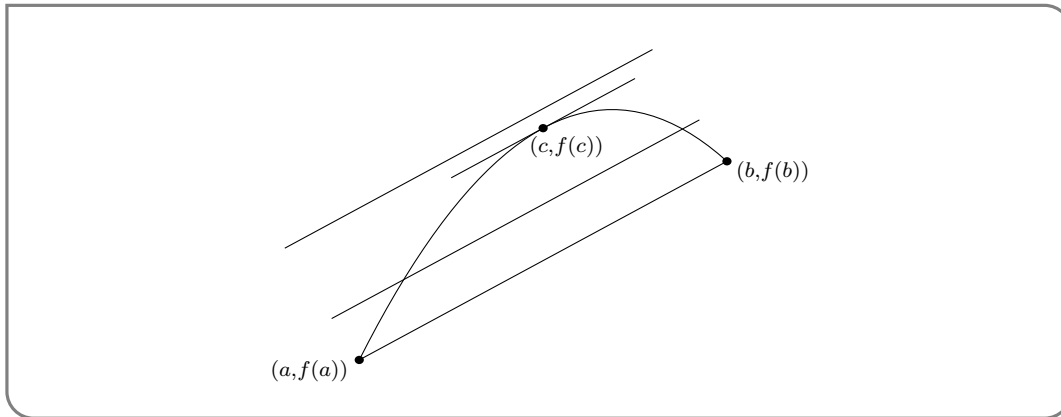
looking for a point, $(c, f(c))$, on the graph of $f(x)$ with the property that $f'(c) = \frac{f(b)-f(a)}{b-a}$, i.e. with the property that the slope of the tangent line at $(c, f(c))$ is the same as the slope of the secant. So imagine that you start moving the secant upward, carefully keeping the moved line segment parallel to the secant. So the slope of the moved line segment is always exactly $\frac{f(b)-f(a)}{b-a}$. When we first start moving the line segment it is not tangent to the curve — it crosses the curve. This is illustrated in the figure by the second line segment from the bottom. If we move the line segment too far it does not touch the curve at all. This is illustrated in the figure by the top segment. But if we stop moving the line segment just before it stops intersecting the curve at all, we get exactly the tangent line to the curve at the point on the curve that is farthest from the secant. This tangent line has exactly the desired slope. This is illustrated in the figure by the third line segment from the bottom.

▸▸ Be Careful with Hypotheses

The mean value theorem has hypotheses — $f(x)$ has to be continuous for $a \leqslant x \leqslant b$ and has to be differentiable for $a < x < b$. If either hypothesis is violated, the conclusion of the mean value theorem can fail. That is, the curve $y = f(x)$ need not have a tangent line at some $x = c$ between a and b whose slope, $f'(c)$, is the same as the slope, $\frac{f(b)-f(a)}{b-a}$, of the secant joining the points $(a, f(a))$ and $(b, f(b))$ on the curve. If $f'(x)$ fails to exist for even a single value of x between a and b, all bets are off. The following two examples illustrate this.

Example 2.13.8

For the first "bad" example, $a = 0$, $b = 2$ and

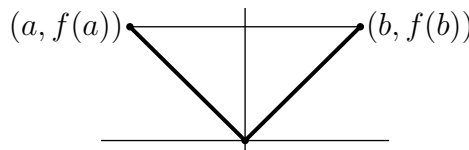$$f(x) = \begin{cases} 0 & \text{if } x \leqslant 1 \\ 1 & \text{if } x > 1 \end{cases}$$
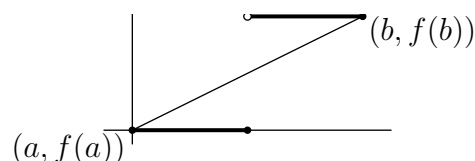
For this example, $f'(x) = 0$ at every x where it is defined. That is, at every $x \neq 1$. But the slope of the secant joining $(a, f(a)) = (0, 0)$ and $(b, f(b)) = (2, 1)$ is $\frac{1}{2}$.
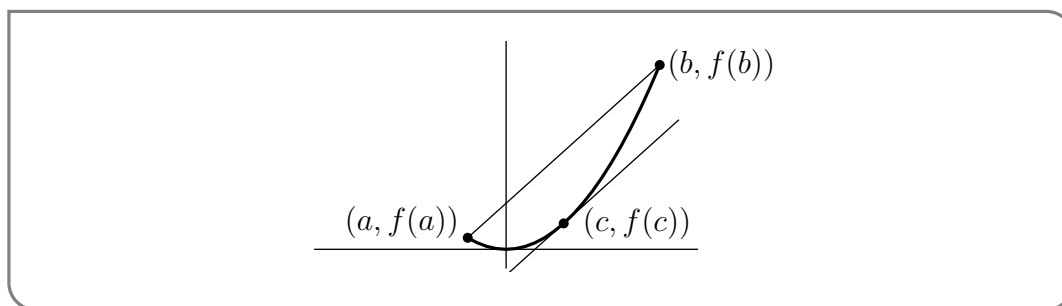
Example 2.13.8

Example 2.13.9

For the second "bad" example, $a = -1$, $b = 1$ and $f(x) = |x|$. For this function

$$f'(x) = \begin{cases} -1 & \text{if } x < 0 \\ \text{undefined} & \text{if } x = 0 \\ 1 & \text{if } x > 0 \end{cases}$$

For this example, $f'(x) = \pm 1$ at every x where it is defined. That is, at every $x \neq 0$. But the slope of the secant joining $(a, f(a)) = (-1, 1)$ and $(b, f(b)) = (1, 1)$ is 0.

Example 2.13.9

Example 2.13.10

Here is one "good" example, where the hypotheses of the mean value theorem are satisfied. Let $f(x) = x^2$. Then $f'(x) = 2x$. For any $a < b$,

$$\frac{f(b) - f(a)}{b - a} = \frac{b^2 - a^2}{b - a} = b + a$$

So $f'(c) = 2c$ is exactly $\frac{f(b) - f(a)}{b - a}$ when $c = \frac{a+b}{2}$, which, in this example, happens to be exactly half way between $x = a$ and $x = b$.

Example 2.13.10

    Recall from Section 2.3 that if $f'(c) > 0$, then $f(x)$ is increasing at $x = c$. A simple consequence of the mean value theorem is that if you know the sign of $f'(c)$ for all c's between a and b, with $b > a$, then $f(b) - f(a) = f'(c)(b - a)$ must have the same sign.

> **Corollary 2.13.11** (Consequences of the mean value theorem).
>
> Let A and B be real numbers with $A < B$. Let the function $f(x)$ be defined and continuous on the closed interval $A \leqslant x \leqslant B$ and be differentiable on the open interval $A < x < B$.
>
> (a) If $f'(c) = 0$ for all $A < c < B$, then $f(b) = f(a)$ for all $A \leqslant a < b \leqslant B$.
>   — That is, $f(x)$ is constant on $A \leqslant x \leqslant B$.
>
> (b) If $f'(c) \geqslant 0$ for all $A < c < B$, then $f(b) \geqslant f(a)$ for all $A \leqslant a \leqslant b \leqslant B$.
>   — That is, $f(x)$ is increasing on $A \leqslant x \leqslant B$.
>
> (c) If $f'(c) > 0$ for all $A < c < B$, then $f(b) < f(a)$ for all $A \leqslant a < b \leqslant B$.
>   — That is, $f(x)$ is strictly increasing on $A \leqslant x \leqslant B$.
>
> (d) If $f'(c) \leqslant 0$ for all $A < c < B$, then $f(b) \leqslant f(a)$ for all $A \leqslant a \leqslant b \leqslant B$.
>   — That is, $f(x)$ is decreasing on $A \leqslant x \leqslant B$.
>
> (e) If $f'(c) < 0$ for all $A < c < B$, then $f(b) < f(a)$ for all $A \leqslant a < b \leqslant B$.
>   — That is, $f(x)$ is strictly decreasing on $A \leqslant x \leqslant B$.

It is not hard to see why the above is true:

- Say $f'(x) = 0$ at every point in the interval $[A, B]$. Now pick any $a, b \in [A, B]$ with $a < b$. Then the MVT tells us that there is $c \in (a, b)$ so that

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$

  If $f(b) \neq f(a)$ then we must have that $f'(c) \neq 0$ — contradicting what we are told about $f'(x)$. Thus we must have that $f(b) = f(a)$.

- Similarly, say $f'(x) \geqslant 0$ at every point in the interval $[A, B]$. Now pick any $a, b \in [A, B]$ with $a < b$. Then the MVT tells us that there is $c \in (a, b)$ so that

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$

  Since $b > a$, the denominator is positive. Now if $f(b) < f(a)$ the numerator would be negative, making the right-hand side negative, and contradicting what we are told about $f'(x)$. Hence we must have $f(b) \geqslant f(a)$.

- The other parts are similar.

A nice corollary of the above corollary is the following:

> **Corollary 2.13.12.**
>
> If $f'(x) = g'(x)$ for all x in the open interval (a, b), then $f - g$ is a constant on (a, b). That is $f(x) = g(x) + c$, where c is some constant.

We can prove this by setting $h(x) = f(x) - g(x)$. Then $h'(x) = 0$ and so the previous corollary tells us that $h(x)$ is constant.

Example 2.13.13

Using this corollary we can prove results like the following:

$$\arcsin x + \arccos x = \frac{\pi}{2} \qquad \text{for all } -1 < x < 1$$

How does this work? Let $f(x) = \arcsin x + \arccos x$. Then

$$f'(x) = \frac{1}{\sqrt{1-x^2}} + \frac{-1}{\sqrt{1-x^2}} = 0$$

Thus f must be a constant. To find out which constant, we can just check its value at a convenient point, like $x = 0$.

$$\arcsin(0) + \arccos(0) = \pi/2 + 0 = \pi/2$$

Since the function is constant, this must be the value.

Example 2.13.13

2.14 ▲ Higher Order Derivatives

The operation of differentiation takes as input one function, $f(x)$, and produces as output another function, $f'(x)$. Now $f'(x)$ is once again a function. So we can differentiate it again, assuming that it is differentiable, to create a third function, called the second derivative of f. And we can differentiate the second derivative again to create a fourth function, called the third derivative of f. And so on.

Notation 2.14.1.

- $f''(x)$ and $f^{(2)}(x)$ and $\frac{d^2 f}{dx^2}(x)$ all mean $\frac{d}{dx}\left(\frac{d}{dx}f(x)\right)$

- $f'''(x)$ and $f^{(3)}(x)$ and $\frac{d^3 f}{dx^3}(x)$ all mean $\frac{d}{dx}\left(\frac{d}{dx}\left(\frac{d}{dx}f(x)\right)\right)$

- $f^{(4)}(x)$ and $\frac{d^4 f}{dx^4}(x)$ both mean $\frac{d}{dx}\left(\frac{d}{dx}\left(\frac{d}{dx}\left(\frac{d}{dx}f(x)\right)\right)\right)$

- and so on.

Here is a simple example. Then we'll think a little about the significance of second order derivatives. Then we'll do a more a computationally complex example.

Example 2.14.2

Let n be a natural number and let $f(x) = x^n$. Then

$$\frac{\mathrm{d}}{\mathrm{d}x}x^n = nx^{n-1}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}x^n = \frac{\mathrm{d}}{\mathrm{d}x}\left(nx^{n-1}\right) = n(n-1)x^{n-2}$$

$$\frac{\mathrm{d}^3}{\mathrm{d}x^3}x^n = \frac{\mathrm{d}}{\mathrm{d}x}\left(n(n-1)x^{n-2}\right) = n(n-1)(n-2)x^{n-3}$$

Each time we differentiate, we bring down the exponent, which is exactly one smaller than the previous exponent brought down, and we reduce the exponent by one. By the time we have differentiated $n - 1$ times, the exponent has decreased to $n - (n - 1) = 1$ and we have brought down the factors $n(n - 1)(n - 2) \cdots 2$. So

$$\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}x^n = n(n-1)(n-2)\cdots 2x$$

and

$$\frac{\mathrm{d}^n}{\mathrm{d}x^n}x^n = n(n-1)(n-2)\cdots 1$$

The product of the first n natural numbers, $1 \cdot 2 \cdot 3 \cdot \cdots \cdot n$, is called "$n$ factorial" and is denoted $n!$. So we can also write

$$\frac{\mathrm{d}^n}{\mathrm{d}x^n}x^n = n!$$

If $m > n$, then

$$\frac{\mathrm{d}^m}{\mathrm{d}x^m}x^n = 0$$

Example 2.14.2

Example 2.14.3

Recall that the derivative $v'(a)$ is the (instantaneous) rate of change of the function $v(t)$ at $t = a$. Suppose that you are walking on the x–axis and that $x(t)$ is your x–coordinate at time t. Also suppose, for simplicity, that you are moving from left to right. Then $v(t) = x'(t)$ is your velocity at time t and $v'(a) = x''(a)$ is the rate at which your velocity is changing at time $t = a$. It is called your acceleration. In particular, if $x''(a) > 0$, then your velocity is increasing, i.e. you are speeding up, at time a. If $x''(a) < 0$, then your velocity is decreasing, i.e. you are slowing down, at time a. That's one interpretation of the second derivative.

Example 2.14.3

Example 2.14.4 (Example 2.11.1, continued)

Find y'' if $y = y^3 + xy + x^3$.

Solution. This problem concerns some function $y(x)$ that is not given to us explicitly. All that we are told is that $y(x)$ satisfies

$$y(x) = y(x)^3 + xy(x) + x^3 \tag{E1}$$

for all x. We are asked to find $y''(x)$. We cannot solve this equation to get an explicit formula for $y(x)$. So we use implicit differentiation, as we did in Example 2.11.1. That is, we apply $\frac{d}{dx}$ to both sides of (E1). This gives

$$y'(x) = 3y(x)^2 y'(x) + y(x) + x y'(x) + 3x^2 \tag{E2}$$

which we can solve for $y'(x)$, by moving all $y'(x)$'s to the left hand side, giving

$$[1 - x - 3y(x)^2] y'(x) = y(x) + 3x^2$$

and then dividing across.

$$y'(x) = \frac{y(x) + 3x^2}{1 - x - 3y(x)^2} \tag{E3}$$

To get $y''(x)$, we have two options.
Method 1. Apply $\frac{d}{dx}$ to both sides of (E2). This gives

$$y''(x) = 3y(x)^2 y''(x) + 6y(x) y'(x)^2 + 2y'(x) + x y''(x) + 6x$$

We can now solve for $y''(x)$, giving

$$y''(x) = \frac{6x + 2y'(x) + 6y(x)y'(x)^2}{1 - x - 3y(x)^2} \tag{E4}$$

Then we can substitute in (E3), giving

$$y''(x) = 2 \frac{3x + \frac{y(x)+3x^2}{1-x-3y(x)^2} + 3y(x)\left(\frac{y(x)+3x^2}{1-x-3y(x)^2}\right)^2}{1 - x - 3y(x)^2}$$

$$= 2 \frac{3x[1 - x - 3y(x)^2]^2 + [y(x) + 3x^2][1 - x - 3y(x)^2] + 3y(x)[y(x) + 3x^2]^2}{[1 - x - 3y(x)^2]^3}$$

Method 2. Alternatively, we can also differentiate (E3).

$$y''(x) = \frac{[y'(x) + 6x][1 - x - 3y(x)^2] - [y(x) + 3x^2][-1 - 6y(x)y'(x)]}{[1 - x - 3y(x)^2]^2}$$

$$= \frac{[\frac{y(x)+3x^2}{1-x-3y(x)^2} + 6x][1 - x - 3y(x)^2] - [y(x) + 3x^2][-1 - 6y(x)\frac{y(x)+3x^2}{1-x-3y(x)^2}]}{[1 - x - 3y(x)^2]^2}$$

$$= \frac{2[y(x) + 3x^2][1 - x - 3y(x)^2] + 6x[1 - x - 3y(x)^2]^2 + 6y(x)[y(x) + 3x^2]^2}{[1 - x - 3y(x)^2]^3}$$

Remark 1. We have now computed $y''(x)$ — sort of. The answer is in terms of $y(x)$, which we don't know. Since we cannot get an explicit formula for $y(x)$, there's not a great deal that we can do, in general.

Remark 2. Even though we cannot solve $y = y^3 + xy + x^3$ explicitly for $y(x)$, for general x, it is sometimes possible to solve equations like this for some special values of x. In fact, we saw in Example 2.11.1 that when $x = 1$, the given equation reduces to $y(1) = y(1)^3 + 1 \cdot y(1) + 1^3$, or $y(1)^3 = -1$, which we can solve to get $y(1) = -1$. Substituting into (E2), as we did in Example 2.11.1 gives

$$y'(1) = \frac{-1+3}{1-1-3(-1)^2} = -\frac{2}{3}$$

and substituting into (E4) gives

$$y''(1) = \frac{6 + 2\left(-\frac{2}{3}\right) + 6(-1)\left(-\frac{2}{3}\right)^2}{1 - 1 - 3(-1)^2} = \frac{6 - \frac{4}{3} - \frac{8}{3}}{-3} = -\frac{2}{3}$$

(It's a fluke that, in this example, $y'(1)$ and $y''(1)$ happen to be equal.) So we now know that, even though we can't solve $y = y^3 + xy + x^3$ explicitly for $y(x)$, the graph of the solution passes through $(1, -1)$ and has slope $-\frac{2}{3}$ (i.e. is sloping downwards by between $30°$ and $45°$) there and, furthermore, the slope of the graph decreases as x increases through $x = 1$.
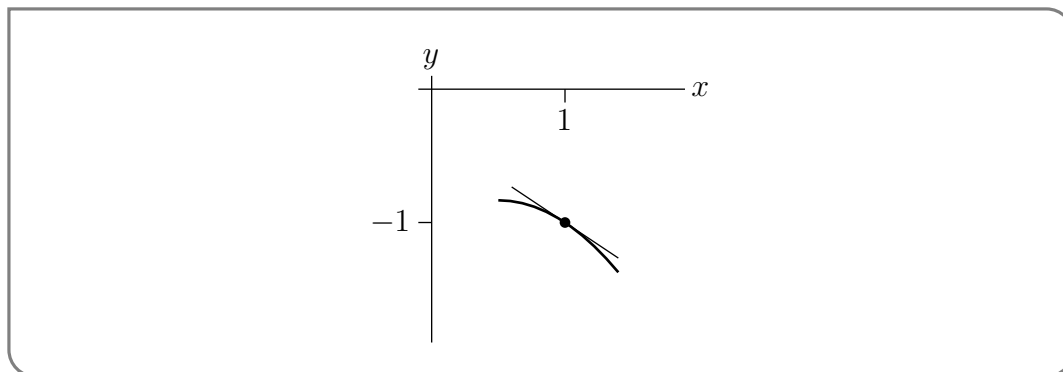
Here is a sketch of the part of the graph very near $(1, -1)$. The tangent line to the graph at $(1, -1)$ is also shown. Note that the tangent line is sloping down to the right, as we expect, and that the graph lies below the tangent line near $(1, -1)$. That's because the slope $f'(x)$ is decreasing (becoming more negative) as x passes through 1.

Example 2.14.4

Warning 2.14.5.

Many people will suppress the (x) in $y(x)$ when doing computations like those in Example 2.14.4. This gives shorter, easier to read formulae, like $y' = \frac{y+3x^2}{1-x-3y^2}$. If you do this, you must never forget that y is a function of x and is *not* a constant. If you do forget, you'll make the very serious error of saying that $\frac{dy}{dx} = 0$, which is false.

2.15 ▲ (Optional) — Is $\lim\limits_{x\to c} f'(x)$ Equal to $f'(c)$?

Consider the function

$$f(x) = \begin{cases} \frac{\sin x^2}{x} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

For any $x \neq 0$ we can easily use our differentiation rules to find

$$f'(x) = \frac{2x^2 \cos x^2 - \sin x^2}{x^2}$$

But for $x = 0$ none of our usual differentation rules apply. So how do we find $f'(0)$? One obviously legitimate strategy is to directly apply the Definition 2.2.1 of the derivative. As an alternative, we will now consider the question "Can one find $f'(0)$ by taking the limit of $f'(x)$ as x tends to zero?". There is bad news and there is good news.

- The bad news is that, even for functions $f(x)$ that are differentiable for all x, $f'(x)$ need not be continuous. That is, it is *not* always true that $\lim_{x\to 0} f'(x) = f'(0)$. We will see a function for which $\lim_{x\to 0} f'(x) \neq f'(0)$ in Example 2.15.1, below.

- The good news is that Theorem 2.15.2, below provides conditions which are sufficient to guarantee that $f(x)$ is differentiable at $x = 0$ and that $\lim_{x\to 0} f'(x) = f'(0)$.

Example 2.15.1

Consider the function

$$f(x) = \begin{cases} x^2 \sin \frac{1}{x} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

For $x \neq 0$ we have, by the product and chain rules,

$$f'(x) = 2x \sin \frac{1}{x} + x^2 \left(\cos \frac{1}{x} \right) \left(-\frac{1}{x^2} \right)$$
$$= 2x \sin \frac{1}{x} - \cos \frac{1}{x}$$

As $\left| \sin \frac{1}{x} \right| \leq 1$, we have

$$\lim_{x\to 0} 2x \sin \frac{1}{x} = 0$$

On the other hand, as x tends to zero, $\frac{1}{x}$ goes to $\pm\infty$. So

$$\lim_{x\to 0} \cos \frac{1}{x} = DNE \implies \lim_{x\to 0} f'(x) = DNE$$

206

We will now see that, despite this, $f'(0)$ is perfectly well defined. By definition

$$f'(0) = \lim_{h \to 0} \frac{f(h) - f(0)}{h}$$

$$= \lim_{h \to 0} \frac{h^2 \sin \frac{1}{h} - 0}{h}$$

$$= \lim_{h \to 0} h \sin \frac{1}{h}$$

$$= 0 \qquad \text{since } \left| \sin \frac{1}{h} \right| \leqslant 1$$

So $f'(0)$ exists, but is not equal to $\lim_{x \to 0} f'(x)$, which does not exist.

Example 2.15.1

Now for the good news.

Theorem 2.15.2.

Let $a < c < b$. Assume that

- the function $f(x)$ is continous on the interval $a < x < b$ and

- is differentiable at every x in the intervals $a < x < c$ and $c < x < b$ and

- the limit $\lim_{x \to c} f'(x)$ exists.

Then f is differentiable at $x = c$ and

$$f'(c) = \lim_{x \to c} f'(x)$$

Proof. By hypothesis, there is a number L such that

$$\lim_{x \to c} f'(x) = L$$

By definition

$$f'(c) = \lim_{h \to 0} \frac{f(c + h) - f(c)}{h}$$

By the Mean Value Theorem (Theorem 2.13.4) there is, for each h, an (unknown) number x_h between c and $c + h$ such that $f'(x_h) = \frac{f(c+h) - f(c)}{h}$. So

$$f'(c) = \lim_{h \to 0} f'(x_h)$$

As h tends to zero, $c + h$ tends to c, and so x_h is forced to tend to c, and $f'(x_h)$ is forced to tend to L so that

$$f'(c) = \lim_{h \to 0} f'(x_h) = L$$

$\square$

In the next example we evaluate $f'(0)$ by applying Theorem 2.15.2.

Example 2.15.3

Let
$$f(x) = \begin{cases} \frac{\sin x^2}{x} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

We have already observed above that, for $x \neq 0$,

$$f'(x) = \frac{2x^2 \cos x^2 - \sin x^2}{x^2} = 2 \cos x^2 - \frac{\sin x^2}{x^2}$$

We use Theorem 2.15.2 with $c = 0$ to show that $f(x)$ is differentiable at $x = 0$ and to evaluate $f'(0)$. That theorem has two hypotheses that we have not yet verified, namely the continuity of $f(x)$ at $x = 0$, and the existence of the limit $\lim_{x \to 0} f'(x)$. We verify them now.

- We already know, by Lemma 2.8.1, that $\lim_{h \to 0} \frac{\sin h}{h} = 1$. So

$$\lim_{x \to 0} \frac{\sin x^2}{x^2} = \lim_{h \to 0^+} \frac{\sin h}{h} \qquad \text{with } h = x^2$$
$$= 1$$

  and

$$\lim_{x \to 0} f(x) = \lim_{x \to 0} \frac{\sin x^2}{x} = \lim_{x \to 0} x \frac{\sin x^2}{x^2} = \lim_{x \to 0} x \lim_{x \to 0} \frac{\sin x^2}{x^2} = 0 \times 1 = 0$$

  and $f(x)$ is continuous at $x = 0$.

- The limit of the derivative is

$$\lim_{x \to 0} f'(x) = \lim_{x \to 0} \left[2 \cos x^2 - \frac{\sin x^2}{x^2} \right] = 2 \times 1 - 1 = 1$$

So, by Theorem 2.15.2, $f(x)$ is differentiable at $x = 0$ and $f'(0) = 1$.

Example 2.15.3

APPLICATIONS OF DERIVATIVES

In Section 2.2 we defined the derivative at $x = a$, $f'(a)$, of an abstract function $f(x)$, to be its instantaneous rate of change at $x = a$:

$$f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a}$$

This abstract definition, and the whole theory that we have developed to deal with it, turns out be extremely useful simply because "instantaneous rate of change" appears in a huge number of settings. Here are a few examples.

- If you are moving along a line and $x(t)$ is your position on the line at time t, then your rate of change of position, $x'(t)$, is your velocity. If, instead, $v(t)$ is your velocity at time t, then your rate of change of velocity, $v'(t)$, is your acceleration. We shall explore this further in Section 3.1.

- If $P(t)$ is the size of some population (say the number of humans on the earth) at time t, then $P'(t)$ is the rate at which the size of that population is changing. It is called the net birth rate. We shall explore it further in Section 3.3.3.

- Radiocarbon dating, a procedure used to determine the age of, for example, archaeological materials, is based on an understanding of the rate at which an unstable isotope of carbon decays. We shall look at this procedure in Section 3.3.1

- A capacitor is an electrical component that is used to repeatedly store and release electrical charge (say electrons) in an electronic circuit. If $Q(t)$ is the charge on a capacitor at time t, then $Q'(t)$ is the instantaneous rate at which charge is flowing into the capacitor. That's called the current. The standard unit of charge is the coulomb. One coulomb is the magnitude of the charge of approximately 6.241×10^{18} electrons. The standard unit for current is the amp. One amp represents one coulomb per second.

3.1 ⬧ Velocity and Acceleration

If you are moving along the x–axis and your position at time t is $x(t)$, then your velocity at time t is $v(t) = x'(t)$ and your acceleration at time t is $a(t) = v'(t) = x''(t)$. Your speed, i.e. the distance that you are travelling per unit time, is $|v(t)| = |x'(t)|$.

Example 3.1.1

Suppose that you are moving along the x–axis and that at time t your position is given by

$$x(t) = t^3 - 3t + 2.$$

We're going to try and get a good picture of what your motion is like. We can learn quite a bit just by looking at the sign of the velocity $v(t) = x'(t)$ at each time t.

- If $x'(t) > 0$, then at that instant x is increasing, i.e. you are moving to the right.

- If $x'(t) = 0$, then at that instant you are not moving at all.

- If $x'(t) < 0$, then at that instant x is decreasing, i.e. you are moving to the left.

From the given formula for $x(t)$ it is straight forward to work out the velocity

$$v(t) = x'(t) = 3t^2 - 3 = 3(t^2 - 1) = 3(t+1)(t-1)$$

This is zero only when $t = -1$ and when $t = +1$; at no other value[1] of t can this polynomial be equal zero. Consequently in any time interval that does *not* include either $t = -1$ or $t = +1$, $v(t)$ takes only a single sign[2]. So

- For all $t < -1$, both $(t+1)$ and $(t-1)$ are negative (sub in, for example, $t = -10$) so the product $v(t) = x'(t) = 3(t+1)(t-1) > 0$.

- For all $-1 < t < 1$, the factor $(t+1) > 0$ and the factor $(t-1) < 0$ (sub in, for example, $t = 0$) so the product $v(t) = x'(t) = 3(t+1)(t-1) < 0$.

- For all $t > 1$, both $(t+1)$ and $(t-1)$ are positive (sub in, for example, $t = +10$) so the product $v(t) = x'(t) = 3(t+1)(t-1) > 0$.

The figure below gives a summary of the sign information we have about $t - 1$, $t + 1$ and $x'(t)$.

1   This is because the equation $ab = 0$ is only satisfied for real numbers a and b when either $a = 0$ or $b = 0$ or both $a = b = 0$. Hence if a polynomial is the product of two (or more) factors, then it is only zero when at least one of those factors is zero. There are more complicated mathematical environments in which you have what are called "zero divisors" but they are beyond the scope of this course.

2   This is because if $v(t_a) < 0$ and $v(t_b) > 0$ then, by the intermediate value theorem, the continuous function $v(t) = x'(t)$ must take the value 0 for some t between t_a and t_b.
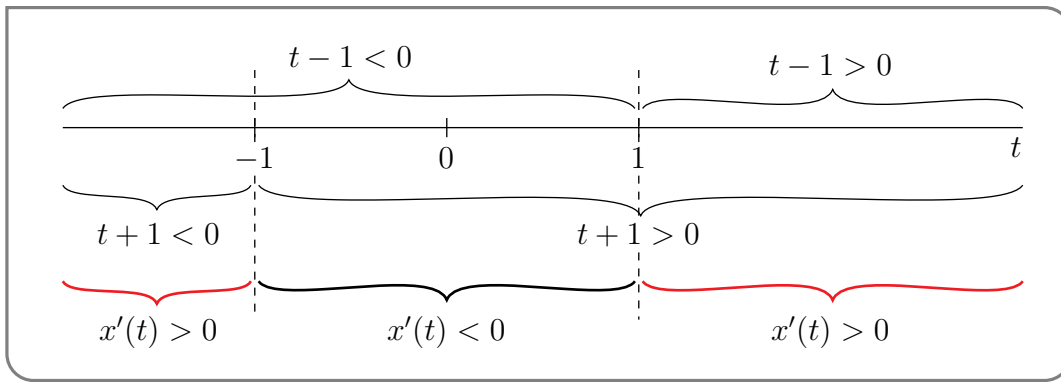
It is now easy to put together a mental image of your trajectory.

- For t large and negative (i.e. far in the past), $x(t)$ is large and negative and $v(t)$ is large and positive. For example[3], when $t = -10^6$, $x(t) \approx t^3 = -10^{18}$ and $v(t) \approx 3t^2 = 3 \cdot 10^{12}$. So you are moving quickly to the right.

- For $t < -1$, $v(t) = x'(t) > 0$ so that $x(t)$ is increasing and you are moving to the right.

- At $t = -1$, $v(-1) = 0$ and you have come to a halt at position $x(-1) = (-1)^3 - 3(-1) + 2 = 4$.

- For $-1 < t < 1$, $v(t) = x'(t) < 0$ so that $x(t)$ is decreasing and you are moving to the left.

- At $t = +1$, $v(1) = 0$ and you have again come to a halt, but now at position $x(1) = 1^3 - 3 + 2 = 0$.

- For $t > 1$, $v(t) = x'(t) > 0$ so that $x(t)$ is increasing and you are again moving to the right.

- For t large and positive (i.e. in the far future), $x(t)$ is large and positive and $v(t)$ is large and positive. For example[4], when $t = 10^6$, $x(t) \approx t^3 = 10^{18}$ and $v(t) \approx 3t^2 = 3 \cdot 10^{12}$. So you are moving quickly to the right.

Here is a sketch of the graphs of $x(t)$ and $v(t)$. The heavy lines in the graphs indicate when you are moving to the right — that is where $v(t) = x'(t)$ is positive.

---

3   Notice here we are using the fact that when t is very large t^3 is much bigger than t^2 and t^1. So we can approximate the value of the polynomial $x(t)$ by the largest term — in this case t^3. We can do similarly with $v(t)$ — the largest term is $3t^2$.

4   We are making a similar rough approximation here.
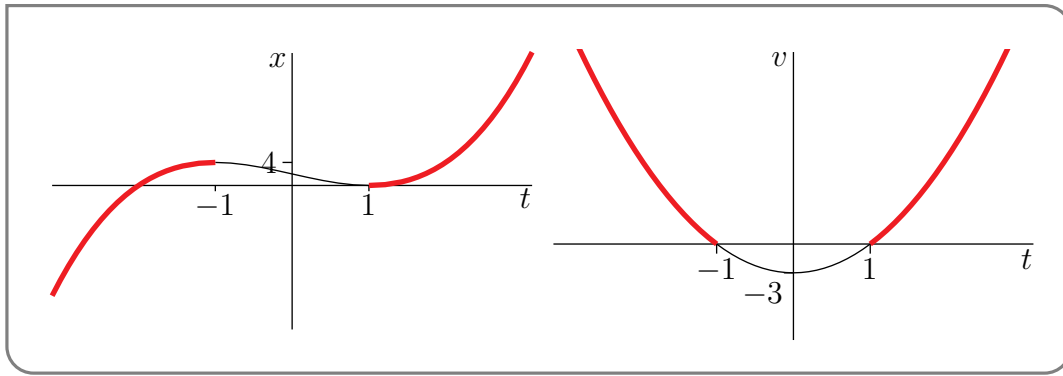
And here is a schematic picture of the whole trajectory.
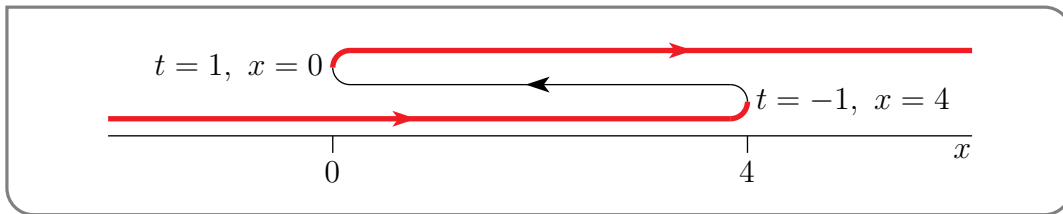
Example 3.1.1

Example 3.1.2

In this example we are going to figure out how far a body falling from rest will fall in a given time period.

- We should start by defining some variables and their units. Denote

    – time in seconds by t,

    – mass in kilograms by m,

    – distance fallen (in metres) at time t by $s(t)$, velocity (in m/sec) by $v(t) = s'(t)$ and acceleration (in m/sec^2) by $a(t) = v'(t) = s''(t)$.

  It makes sense to choose a coordinate system so that the body starts to fall at $t = 0$.

- We will use Newton's second law of motion

$$\text{the force applied to the body at time } t = m \cdot a(t).$$

  together with the assumption that the only force acting on the body is gravity (in particular, no air resistance). Note that near the surface of the Earth,

$$\text{the force due to gravity acting on a body of mass } m = m \cdot g.$$

  The constant g, called the acceleration of gravity[5], is about 9.8m/sec^2.

---

5   It is also called the standard acceleration due to gravity or standard gravity. For those of you who prefer imperial units (or US customary units), it is about 32 ft/sec^2, 77165 cubits/minute^2, or 631353 furlongs/hour^2.

- Since the body is falling from rest, we know that its initial velocity is zero. That is

$$v(0) = 0.$$

Newton's second law then implies that

$$
\begin{aligned}
m \cdot a(t) &= \text{force due to gravity} \\
m \cdot v'(t) &= m \cdot g \qquad\qquad\qquad \text{cancel the } m \\
v'(t) &= g
\end{aligned}
$$

- In order to find the velocity, we need to find a function of t whose derivative is constant. We are simply going to guess such a function and then we will verify that our guess has all of the desired properties. It's easy to guess a function whose derivative is the constant g. Certainly gt has the correct derivative. So does

$$v(t) = gt + c$$

for any constant c. One can then verify[6] that $v'(t) = g$. Using the fact that $v(0) = 0$ we must then have $c = 0$ and so

$$v(t) = gt.$$

- Since velocity is the derivative of position, we know that

$$s'(t) = v(t) = g \cdot t.$$

To find $s(t)$ we are again going to guess and check. It's not hard to see that we can use

$$s(t) = \frac{g}{2}t^2 + c$$

where again c is some constant. Again we can verify that this works simply by differentiating[7]. Since we have defined $s(t)$ to be the distance fallen, it follows that $s(0) = 0$ which in turn tells us that $c = 0$. Hence

$$
\begin{aligned}
s(t) &= \frac{g}{2}t^2 \qquad\qquad\qquad \text{but } g = 9.8, \text{ so} \\
&= 4.9t^2,
\end{aligned}
$$

which is exactly the $s(t)$ used way back in Section 1.2.

---

6   While it is clear that this satisfies the equation we want, it is less clear that it is the only function that works. To see this, assume that there are two functions $f(t)$ and $h(t)$ which both satisfy $v'(t) = g$. Then $f'(t) = h'(t) = g$ and so $f'(t) - h'(t) = 0$. Equivalently

$$\frac{\mathrm{d}}{\mathrm{d}t}(f(t) - h(t)) = 0.$$

The only function whose derivative is zero everywhere is the constant function (see Section 2.13 and Corollary 2.13.11). Thus $f(t) - h(t) = \text{constant}$. So all the functions that satisfy $v'(t) = g$ must be of the form $gt + \text{constant}$.

7   To show that any solution of $s'(t) = gv$ must be of this form we can use the same reasoning we used to get $v(t) = gt + \text{constant}$.

Example 3.1.2

Let's now do a similar but more complicated example.

Example 3.1.3

A car's brakes can decelerate the car at 64000km/hr^2. How fast can the car be driven if it must be able to stop within a distance of 50m?

Solution.   Before getting started, notice that there is a small "trick" in this problem — several quantities are stated but their units are different. The acceleration is stated in kilometres per hour^2, but the distance is stated in metres. Whenever we come across a "real world" problem[8] we should be careful of the units used.

- We should first define some variables and their units. Denote

  - time (in hours) by t,
  - the position of the car (in kilometres) at time t by $x(t)$, and
  - the velocity (in kilometres per hour) by is $v(t)$.

  We can also choose a coordinate system such that $x(0) = 0$ and the car starts braking at time $t = 0$.

- Now let us rewrite the information in the problem in terms of these variables.

  - We are told that, at maximum braking, the acceleration $v'(t) = x''(t)$ of the car is -64000.
  - We need to determine the maximum initial velocity $v(0)$ so that the stopping distance is at most $50m = 0.05km$ (being careful with our units). Let us call the stopping distance x_{stop} which is really $x(t_{stop})$ where t_{stop} is the stopping time.

- In order to determine x_{stop} we first need to determine t_{stop}, which we will do by assuming maximum braking from a, yet to be determined, initial velocity of $v(0) = q$ m/sec.

- Assuming that the car undergoes a constant acceleration at this maximum braking power, we have

$$v'(t) = -64000$$

This equation is very similar to the ones we had to solve in Example 3.1.2 just above.

As we did there[9], we are going to just guess $v(t)$. First, we just guess one function whose derivative is -64000, namely $-64000t$. Next we observe that, since the derivative of a constant is zero, any function of the form

$$v(t) = -64000\,t + c$$

8   Well — "realer world" would perhaps be a betterer term.
9   Now is a good time to go back and have a read of that example.

with constant c, has the correct derivative. Finally, the requirement that the initial velocity $v(0) = q$ forces $c = q$, so

$$v(t) = q - 64000\,t$$

- From this we can easily determine the stopping time t_{stop}, when the initial velocity is q, since this is just when $v(t) = 0$:

$$0 = v(t_{stop}) = q - 64000 \cdot t_{stop} \qquad \text{and so}$$
$$t_{stop} = \frac{q}{64000}.$$

- Armed with the stopping time, how do we get at the stopping distance? We need to find the formula satisfied by $x(t)$. Again (as per Example 3.1.2) we make use of the fact that

$$x'(t) = v(t) = q - 64000t.$$

So we need to guess a function $x(t)$ so that $x'(t) = q - 64000t$. It is not hard to see that

$$x(t) = qt - 32000t^2 + \text{constant}$$

works. Since we know that $x(0) = 0$, this constant is just zero and

$$x(t) = qt - 32000t^2.$$

- We are now ready to compute the stopping distance (in terms of the, still yet to be determined, initial velocity q):

$$
\begin{aligned}
x_{stop} = x(t_{stop}) &= qt_{stop} - 32000t_{stop}^2 \\
&= \frac{q^2}{64000} - \frac{32000q^2}{64000^2} \\
&= \frac{q^2}{64000}\left(1 - \frac{1}{2}\right) \\
&= \frac{q^2}{2 \times 64000}
\end{aligned}
$$

Notice that the stopping distance is a quadratic function of the initial velocity — if you go twice as fast, you need four times the distance to stop.

- But we are told that the stopping distance must be less than $50m = 0.05km$. This means that

$$x_{stop} = \frac{q^2}{2 \times 64000} \leqslant \frac{5}{100}$$
$$q^2 \leqslant \frac{2 \times 64000 \times 5}{100} = \frac{64000 \times 10}{100} = 6400$$

Thus we must have $q \leqslant 80$. Hence the initial velocity can be no greater than $80km/h$.

Example 3.1.3

3.2 ▲ Related Rates

Consider the following problem

> A spherical balloon is being inflated at a rate of $13cm^3/sec$. How fast is the radius changing when the balloon has radius $15cm$?

There are several pieces of information in the statement:

- The balloon is spherical

- The volume is changing at a rate of $13cm^3/sec$ — so we need variables for volume (in cm^3) and time (in sec). Good choices are V and t.

- We are asked for the rate at which the radius is changing — so we need a variable for radius and units. A good choice is r, measured in cm — since volume is measured in cm^3.

Since the balloon is a sphere we know[10] that

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

Since both the volume and radius are changing with time, both V and r are implicitly functions of time; we could really write

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

We are told the rate at which the volume is changing and we need to find the rate at which the radius is changing. That is, from a knowledge of $\frac{dV}{dt}$, find the related rate[11] $\frac{dr}{dt}$.

In this case, we can just differentiate our equation by t to get

$$\frac{dV}{dt} = 4\pi r^2 \frac{dr}{dt}$$

This can then be rearranged to give

$$\frac{dr}{dt} = \frac{1}{4\pi r^2}\frac{dV}{dt}.$$

Now we were told that $\frac{dV}{dt} = 13$, so

$$\frac{dr}{dt} = \frac{13}{4\pi r^2}.$$

We were also told that the radius is $15cm$, so at that moment in time

$$\frac{dr}{dt} = \frac{13}{\pi 4 \times 15^2}.$$

This is a very typical example of a related rate problem. This section is really just a collection of problems, but all will follow a similar pattern.

---

10 If you don't know the formula for the volume of a sphere, now is a good time to revise by looking at Appendix A.11.

11 Related rate problems are problems in which you are given the rate of change of one quantity and are to determine the rate of change of another, related, quantity.

- The statement of the problem will tell you quantities that must be related (above it was volume, radius and, implicitly, time).

- Typically a little geometry (or some physics or...) will allow you to relate these quantities (above it was the formula that links the volume of a sphere to its radius).

- Implicit differentiation will then allow you to link the rate of change of one quantity to another.

Another balloon example

Example 3.2.1

Consider a helium balloon rising vertically from a fixed point 200m away from you. You are trying to work out how fast it is rising. Now — computing the velocity directly is difficult, but you can measure angles. You observe that when it is at an angle of $\pi/4$ its angle is changing by 0.05 radians per second.

- Start by drawing a picture with the relevant variables
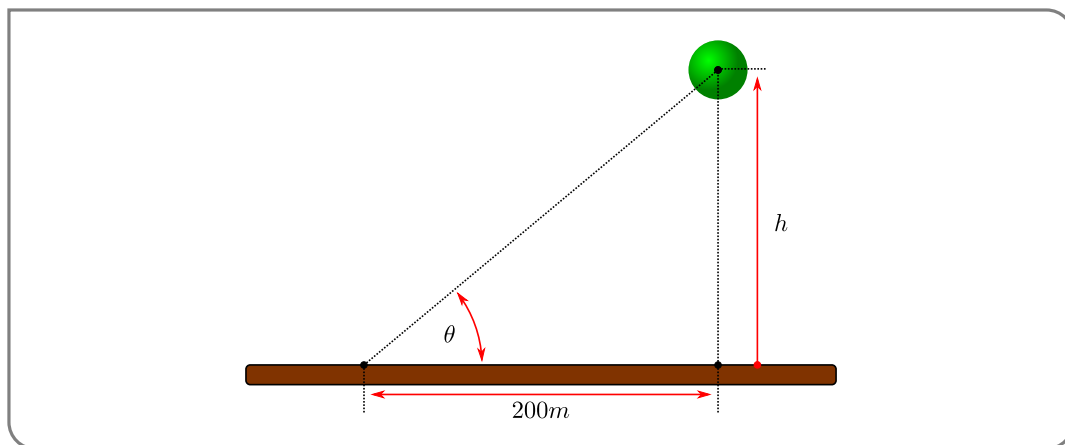
- So denote the angle to be θ (in radians), the height of the balloon (in m) by h and time (in seconds) by t. Then trigonometry tells us

$$h = 200 \cdot \tan \theta$$

- Differentiating allows us to relate the rates of change

$$\frac{dh}{dt} = 200 \sec^2 \theta \cdot \frac{d\theta}{dt}$$

- We are told that when $\theta = \pi/4$ we observe $\frac{d\theta}{dt} = 0.05$, so

$$\frac{dh}{dt} = 200 \cdot \sec^2(\pi/4) \cdot 0.05$$
$$= 200 \cdot 0.05 \cdot \left(\sqrt{2}\right)^2$$
$$= 200 \cdot \frac{5}{100} \cdot 2 \qquad\qquad = 20m/s$$

- So the balloon is rising at a rate of 20m/s.

The following problem is perhaps *the* classic related rate problem.

Example 3.2.2

A 5m ladder is leaning against a wall. The floor is quite slippery and the base of the ladder slides out from the wall at a rate of $1m/s$. How fast is the top of the ladder sliding down the wall when the base of the ladder is 3m from the wall?

- A good first step is to draw a picture stating all relevant quantities. This will also help us define variables and units.
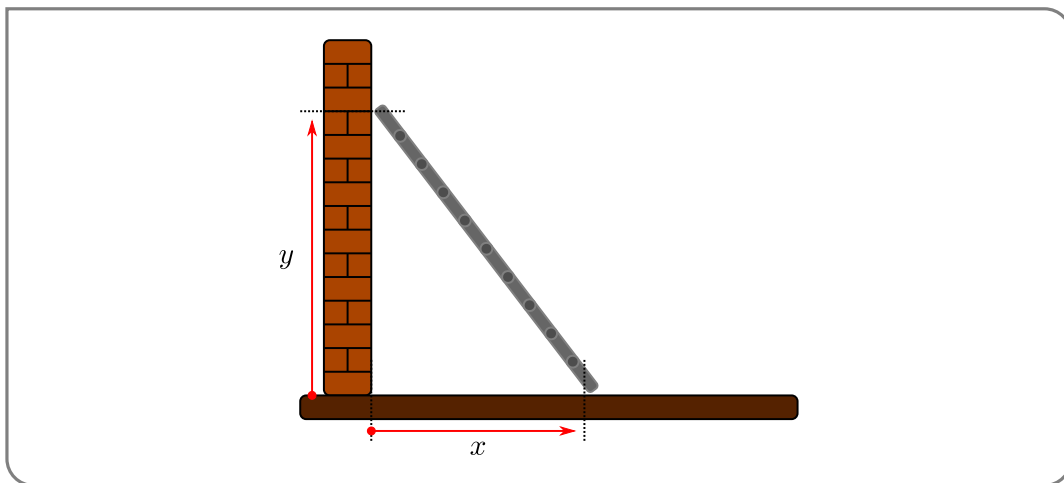
- So now define $x(t)$ to be the distance between the bottom of the ladder and the wall, at time t, and let $y(t)$ be the distance between the top of the ladder and the ground at time t. Measure time in seconds, but both distances in meters.

- We can relate the quantities using Pythagoras:

$$x^2 + y^2 = 5^2$$

- Differentiating with respect to time then gives

$$2x\frac{\mathrm{d}x}{\mathrm{d}t} + 2y\frac{\mathrm{d}y}{\mathrm{d}t} = 0$$

- We know that $\frac{\mathrm{d}x}{\mathrm{d}t} = 1$ and $x = 3$, so

$$6 \cdot 1 + 2y\frac{\mathrm{d}y}{\mathrm{d}t} = 0$$

but we need to determine y before we can go further. Thankfully we know that $x^2 + y^2 = 25$ and $x = 3$, so $y^2 = 25 - 9 = 16$ and[12] so $y = 4$.

---

12   Since the ladder isn't buried in the ground, we can discard the solution $y = -4$.

- So finally putting everything together

$$6 \cdot 1 + 8\frac{dy}{dt} = 0$$
$$\frac{dy}{dt} = -\frac{3}{4}m/s.$$

Thus the top of the ladder is sliding towards the floor at a rate of $3/4m/s$.

Example 3.2.2

The next example is complicated by the rates of change being stated not just as "the rate of change per unit time" but instead being stated as "the percentage rate of change per unit time". If a quantity f is changing with rate $\frac{df}{dt}$, then we can say that

f is changing at a rate of $100 \cdot \dfrac{\frac{df}{dt}}{f}$ percent.

Thus if, at time t, f has rate of change $r\%$, then

$$100\frac{f'(t)}{f(t)} = r \implies f'(t) = \frac{r}{100}f(t)$$

so that if h is a very small time increment

$$\frac{f(t+h) - f(t)}{h} \approx \frac{r}{100}f(t) \implies f(t+h) \approx f(t) + \frac{rh}{100}f(t)$$

That is, over a very small time interval h, f increases by the fraction $\frac{rh}{100}$ of its value at time t.

So armed with this, let's look at the problem.

Example 3.2.3

The quantities P, Q and R are functions of time and are related by the equation $R = PQ$. Assume that P is increasing instantaneously at the rate of 8% per year (meaning that $100\frac{P'}{P} = 8$) and that Q is decreasing instantaneously at the rate of 2% per year (meaning that $100\frac{Q'}{Q} = -2$). Determine the percentage rate of change for R.

Solution. This one is a little different — we are given the variables and the formula, so no picture drawing or defining required. Though we do need to define a time variable — let t denote time in years.

- Since $R(t) = P(t) \cdot Q(t)$ we can differentiate with respect to t to get

$$\frac{dR}{dt} = PQ' + QP'$$

- But we need the percentage change in R, namely

$$100\frac{R'}{R} = 100\frac{PQ' + QP'}{R}$$

but $R = PQ$, so rewrite it as

$$= 100\frac{PQ' + QP'}{PQ}$$
$$= 100\frac{PQ'}{PQ} + 100\frac{QP'}{PQ}$$
$$= 100\frac{Q'}{Q} + 100\frac{P'}{P}$$

so we have stated the instantaneous percentage rate of change in R as the sum of the percentage rate of change in P and Q.

- We know the percentage rate of change of P and Q, so

$$100\frac{R'}{R} = -2 + 8 = 6$$

That is, the instantaneous percentage rate of change of R is 6% per year.

Example 3.2.3

Yet another falling object example.

Example 3.2.4

A ball is dropped from a height of 49m above level ground. The height of the ball at time t is $h(t) = 49 - 4.9t^2$ m. A light, which is also 49m above the ground, is 10m to the left of the ball's original position. As the ball descends, the shadow of the ball caused by the light moves across the ground. How fast is the shadow moving one second after the ball is dropped?

Solution. There is quite a bit going on in this example, so read carefully.

- First a diagram; the one below is perhaps a bit over the top.
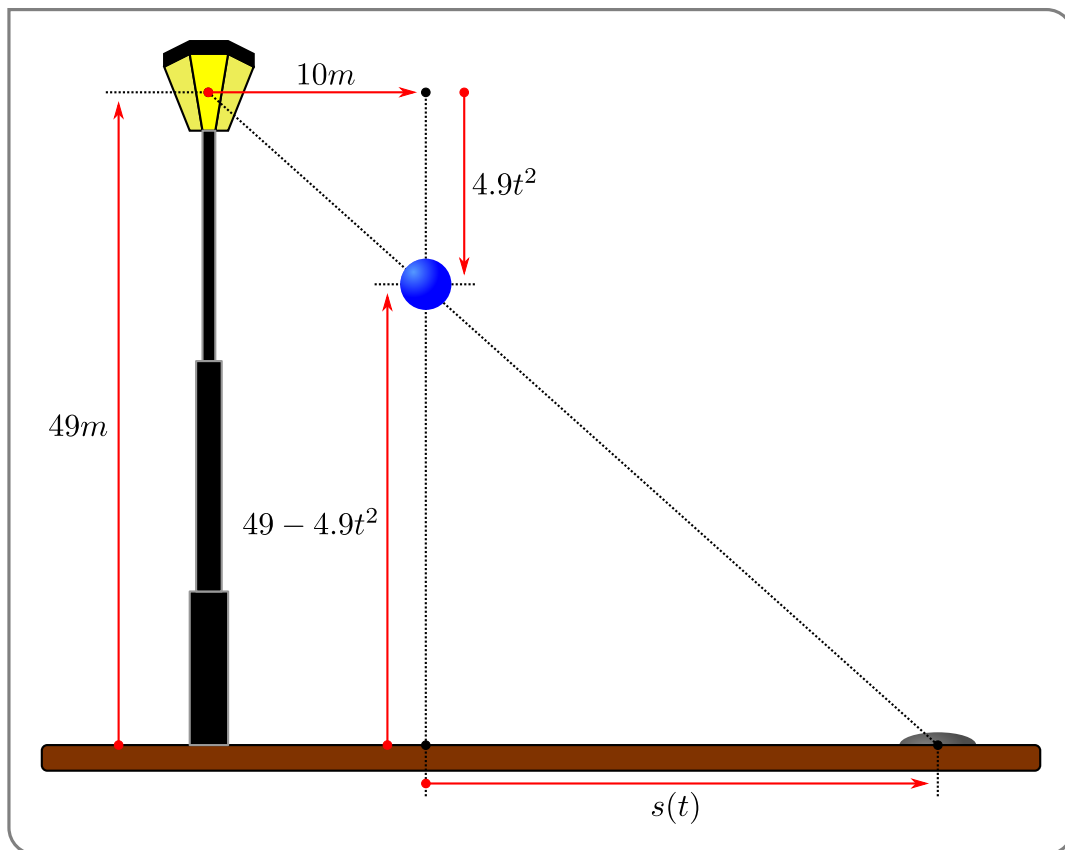
- Let's call $s(t)$ the distance from the shadow to the point on the ground directly underneath the ball.

- By similar triangles we see that

$$\frac{4.9t^2}{10} = \frac{49 - 4.9t^2}{s(t)}$$

We can then solve for $s(t)$ by just multiplying both sides by $\frac{10}{4.9t^2}s(t)$. This gives

$$s(t) = 10\frac{49 - 4.9t^2}{4.9t^2} = \frac{100}{t^2} - 10$$

- Differentiating with respect to t will then give us the rates,

$$s'(t) = -2\frac{100}{t^3}$$

- So, at $t = 1$, $s'(1) = -200$m/sec. That is, the shadow is moving to the left at 200m/sec.

Example 3.2.4

A more nautical example.

Example 3.2.5

Two boats spot each other in the ocean at midday — boat A is 15km west of boat B. Boat A is travelling east at 3km/h and boat B is travelling north at 4km/h. How fast is the distance between the boats changing at 3pm?

- First we draw a picture.
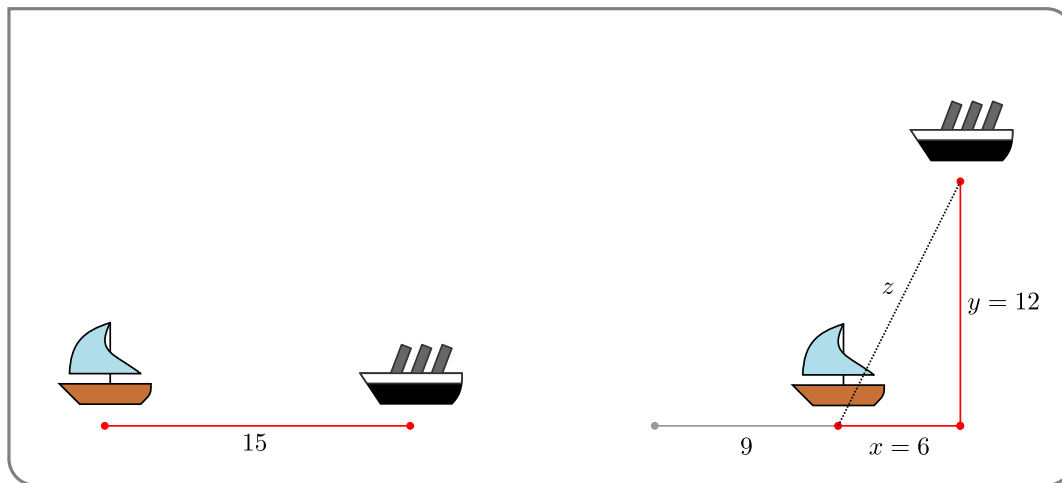
- Let $x(t)$ be the distance at time t, in km, from boat A to the original position of boat B (i.e. to the position of boat B at noon). And let $y(t)$ be the distance at time t, in km, of boat B from its original position. And let $z(t)$ be the distance between the two boats at time t.

- Additionally we are told that $x' = -3$ and $y' = 4$ — notice that $x' < 0$ since that distance is getting smaller with time, while $y' > 0$ since that distance is increasing with time.

- Further at $3pm$ boat A has travelled 9km towards the original position of boat B, so $x = 15 - 9 = 6$, while boat B has travelled 12km away from its original position, so $y = 12$.

- The distances x, y and z form a right-angled triangle, and Pythagoras tells us that

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

At 3pm we know $x = 6, y = 12$ so

$$z^2 = 36 + 144 = 180$$
$$z = \sqrt{180} = 6\sqrt{5}.$$

- Differentiating then gives

$$2z\frac{\mathrm{d}z}{\mathrm{d}t} = 2x\frac{\mathrm{d}x}{\mathrm{d}t} + 2y\frac{\mathrm{d}y}{\mathrm{d}t}$$
$$= 12 \cdot (-3) + 24 \cdot (4)$$
$$= 60.$$

Dividing through by $2z = 12\sqrt{5}$ then gives

$$\frac{dz}{dt} = \frac{60}{12\sqrt{5}} = \frac{5}{\sqrt{5}} = \sqrt{5}$$

So the distance between the boats is increasing at $\sqrt{5}km/h$.

Example 3.2.5

One last one before we move on to another topic.

Example 3.2.6

Consider a cylindrical fuel tank of radius r and length L (in some appropriate units) that is lying on its side. Suppose that fuel is being pumped into the tank at a rate q. At what rate is the fuel level rising?
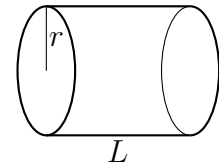
Solution. If the tank were vertical everything would be much easier. Unfortunately the tank is on its side, so we are going to have to work a bit harder to establish the relation between the depth and volume. Also notice that we have not been supplied with units for this problem — so we do not need to state the units of our variables.

- Again — draw a picture. Here is an end view of the tank; the shaded part of the circle is filled with fuel.
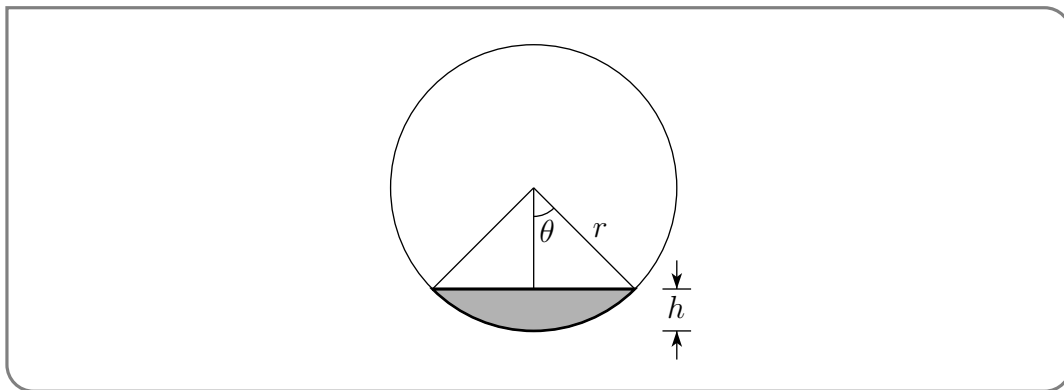
- Let us denote by $V(t)$ the volume of fuel in the tank at time t and by $h(t)$ the fuel level at time t.

- We have been told that $V'(t) = q$ and have been asked to determine $h'(t)$. While it is possible to do so by finding a formula relating $V(t)$ and $h(t)$, it turns out to be quite a bit easier to first find a formula relating V and the angle θ shown in the end view. We can then translate this back into a formula in terms of h using the relation
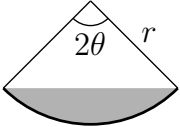
$$h(t) = r - r\cos\theta(t).$$

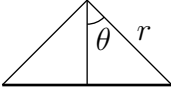Once we know $\theta'(t)$, we can easily obtain $h'(t)$ by differentiating the above equation.

- The computation that follows below gets a little involved in places, so we will drop the "(t)" on the variables V, h and θ. The reader must never forget that these three quantities are really functions of time, while r and L are constants that do not depend on time.

- The volume of fuel is L times the cross–sectional area filled by the fuel. That is,

$$V = L \times \text{Area}\left(\rightharpoondown \right)$$

While we do not have a canned formula for the area of a chord of a circle like this, it is easy to express the area of the chord in terms of two areas that we can compute.

$$V = L \times \text{Area}\left(\rightharpoondown \right) = L \times \left[\text{Area}\left(\overset{2\theta}{\diagup} \, r \right) - \text{Area}\left(\overset{\theta}{\diagup} \, r \right) \right]$$

  - The piece of pie $\overset{2\theta}{\diagup} \, r$ is the fraction $\frac{2\theta}{2\pi}$ of the full circle, so its area is

$\frac{2\theta}{2\pi} \pi r^2 = \theta r^2$.

  - The triangle $\overset{\theta}{\diagup} \, r$ as height $r \cos \theta$ and base $2r \sin \theta$ and hence has area

$\frac{1}{2}(r \cos \theta)(2r \sin \theta) = r^2 \sin \theta \cos \theta = \frac{r^2}{2} \sin(2\theta)$, where we have used a double-angle formula (see Appendix A.14).

Subbing these two areas into the above expression for V gives

$$V = L \times \left[\theta r^2 - \frac{r^2}{2} \sin 2\theta \right] = \frac{L r^2}{2} \left[2\theta - \sin 2\theta \right]$$

Oof!

- Now we can differentiate to find the rate of change. Recalling that $V = V(t)$ and $\theta = \theta(t)$, while r and L are constants,

$$V' = \frac{L r^2}{2} \left[2\theta' - 2 \cos 2\theta \cdot \theta' \right]$$
$$= L r^2 \cdot \theta' \cdot \left[1 - \cos 2\theta \right]$$

Solving this for θ' and using $V' = q$ gives

$$\theta' = \frac{q}{L r^2 (1 - \cos 2\theta)}$$

This is the rate at which θ is changing, but we need the rate at which h is changing. We get this from

$$h = r - r \cos \theta \qquad\qquad \text{differentiating this gives}$$
$$h' = r \sin \theta \cdot \theta'$$

Substituting our expression for θ' into the expression for h' gives

$$h' = r\sin\theta \cdot \frac{q}{Lr^2(1 - \cos 2\theta)}$$

- We can clean this up a bit more — recall more double-angle formulas[13]

$$
\begin{aligned}
h' &= r\sin\theta \cdot \frac{q}{Lr^2(1 - \cos 2\theta)} && \text{substitute } \cos 2\theta = 1 - 2\sin^2\theta \\
&= r\sin\theta \cdot \frac{q}{Lr^2 \cdot 2\sin^2\theta} && \text{now cancel } r\text{'s and a } \sin\theta \\
&= \frac{q}{2Lr\sin\theta}
\end{aligned}
$$

- But we can clean this up even more — instead of writing this rate in terms of θ it is more natural to write it in terms of h (since the initial problem is stated in terms of h). From the triangle
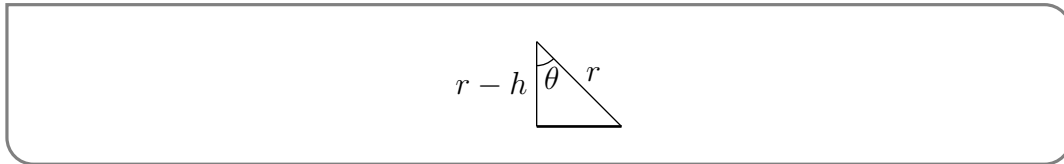
and by Pythagoras we have

$$\sin\theta = \frac{\sqrt{r^2 - (r - h)^2}}{r} = \frac{\sqrt{2rh - h^2}}{r}$$

and hence

$$h' = \frac{q}{2L\sqrt{2rh - h^2}}.$$

- As a check, notice that h' becomes undefined when $h < 0$ and also when $h > 2r$, because then the argument of the square root in the denominator is negative. Both make sense — the fuel level in the tank must obey $0 \leqslant h \leqslant 2r$.

Example 3.2.6

3.3 ▲ Exponential Growth and Decay — a First Look at Differential Equations

A differential equation is an equation for an unknown function that involves the derivative of the unknown function. For example, Newton's law of cooling says:

---

13   Take another look at Appendix A.14.

The rate of change of temperature of an object is proportional to the difference in temperature between the object and its surroundings.

We can write this more mathematically using a differential equation — an equation for the unknown function $T(t)$ that also involves its derivative $\frac{dT}{dt}(t)$. If we denote by $T(t)$ the temperature of the object at time t and by A the temperature of its surroundings, Newton's law of cooling says that there is some constant of proportionality, K, such that

$$\frac{dT}{dt}(t) = K\big[T(t) - A\big]$$

Differential equations play a central role in modelling a huge number of different phenomena, including the motion of particles, electromagnetic radiation, financial options, ecosystem populations and nerve action potentials. Most universities offer half a dozen different undergraduate courses on various aspects of differential equations. We are barely going to scratch the surface of the subject. At this point we are going to restrict ourselves to a few very simple differential equations for which we can just guess the solution. In particular, we shall learn how to solve systems obeying Newton's law of cooling in Section 3.3.2, below. But first, here is another slightly simpler example.

3.3.1 ▸▸ Carbon Dating

Scientists can determine the age of objects containing organic material by a method called *carbon dating* or *radiocarbon dating*[14]. Cosmic rays hitting the atmosphere convert nitrogen into a radioactive isotope of carbon, ^{14}C, with a half–life of about 5730 years[15]. Vegetation absorbs carbon dioxide from the atmosphere through photosynthesis and animals acquire ^{14}C by eating plants. When a plant or animal dies, it stops replacing its carbon and the amount of ^{14}C begins to decrease through radioactive decay. More precisely, let $Q(t)$ denote the amount of ^{14}C in the plant or animal t years after it dies. The number of radioactive decays per unit time, at time t, is proportional to the amount of ^{14}C present at time t, which is $Q(t)$. Thus

Equation 3.3.1 (Radioactive decay).

$$\frac{dQ}{dt}(t) = -kQ(t)$$

Here k is a constant of proportionality that is determined by the half–life. We shall explain what half-life is and also determine the value of k in Example 3.3.3, below. Before we do so, let's think about the sign in equation (3.3.1).

- Recall that $Q(t)$ denotes a quantity, namely the amount of ^{14}C present at time t. There cannot be a negative amount of ^{14}C, nor can this quantity be zero (otherwise we wouldn't use carbon dating, so we must have $Q(t) > 0$.

---

14   Willard Libby, of Chicago University was awarded the Nobel Prize in Chemistry in 1960, for developing radiocarbon dating.

15   A good question to ask yourself is "How can a scientist (who presumably doesn't live 60 centuries) measure this quantity?" One way exploits the little piece of calculus we are about to discuss.

- As the time t increases, $Q(t)$ decreases, because ^{14}C is being continuously converted into ^{14}N by radioactive decay[16]. Thus $\frac{dQ}{dt}(t) < 0$.

- The signs $Q(t) > 0$ and $\frac{dQ}{dt}(t) < 0$ are consistent with equation (3.3.1) provided the constant of proportionality $k > 0$.

- In equation (3.3.1), we chose to call the constant of proportionality "$-k$". We did so in order to make $k > 0$. We could just as well have chosen to call the constant of proportionality "K". That is, we could have replaced equation (3.3.1) by $\frac{dQ}{dt}(t) = KQ(t)$. The constant of proportionality K would have to be negative, (and K and k would be related by $K = -k$).

Now, let's guess some solutions to equation (3.3.1). We wish to guess a function $Q(t)$ whose derivative is just a constant times itself. Here is a short table of derivatives. It is certainly not complete, but it contains the most important derivatives that we know.

$F(t)$	1	t^a	$\sin t$	$\cos t$	$\tan t$	e^t	$\log t$	$\arcsin t$	$\arctan t$
$\frac{d}{dt}F(t)$	0	at^{a-1}	$\cos t$	$-\sin t$	$\sec^2 t$	e^t	$\frac{1}{t}$	$\frac{1}{\sqrt{1-t^2}}$	$\frac{1}{1+t^2}$

There is exactly one function in this table whose derivative is just a (nonzero) constant times itself. Namely, the derivative of e^t is exactly $e^t = 1 \times e^t$. This is almost, but not quite what we want. We want the derivative of $Q(t)$ to be the constant $-k$ (rather than the constant 1) times $Q(t)$. We want the derivative to "pull a constant" out of our guess. That is exactly what happens when we differentiate e^{at}, where a is a constant. Differentiating gives

$$\frac{d}{dt}e^{at} = ae^{at}$$

i.e. "pulls the constant a out of e^{at}".

We have succeeded in guessing a single function, namely e^{-kt}, that obeys equation (3.3.1). Can we guess any other solutions? Yes. If C is any constant, Ce^{-kt} also obeys equation (3.3.1):

$$\frac{d}{dt}(Ce^{-kt}) = C\frac{d}{dt}e^{-kt} = Ce^{-kt}(-k) = -k(Ce^{-kt})$$

You can try guessing some more solutions, but you won't find any, because with a little trickery we can prove that a function $Q(t)$ obeys equation (3.3.1) if and only if $Q(t)$ is of the form Ce^{-kt}, where C is some constant.

The trick[17] is to imagine that $Q(t)$ is any (at this stage, unknown) solution to (3.3.1) and to compare $Q(t)$ and our known solution e^{-kt} by studying the ratio $Q(t)/e^{-kt}$. We will show that $Q(t)$ obeys (3.3.1) if and only if the ratio $Q(t)/e^{-kt}$ is a constant, i.e. if and only if the derivative of the ratio is zero. By the product rule

$$\frac{d}{dt}\left[Q(t)/e^{-kt}\right] = \frac{d}{dt}\left[e^{kt}Q(t)\right] = ke^{kt}Q(t) + e^{kt}Q'(t)$$

---

16 The precise transition is $^{14}C \rightarrow {}^{14}N + e^- + \bar{\nu}_e$ where e^- is an electron and $\bar{\nu}_e$ is an electron neutrino.
17 Notice that is very similar to what we needed in Example 3.1.2, except that here the constant is multiplicative rather than additive. That is $const \times f(t)$ rather than $const + f(t)$.

Since e^{kt} is never 0, the right hand side is zero if and only if $kQ(t) + Q'(t) = 0$; that is $Q'(t) = -kQ(t)$. Thus

$$\frac{\mathrm{d}}{\mathrm{d}t}Q(t) = -kQ(t) \iff \frac{\mathrm{d}}{\mathrm{d}t}\left[Q(t)/e^{-kt}\right] = 0$$

as required.

We have succeed in finding all functions that obey (3.3.1). That is, we have found the general solution to (3.3.1). This is worth stating as a theorem.

> **Theorem 3.3.2.**
>
> A differentiable function $Q(t)$ obeys the differential equation
>
> $$\frac{\mathrm{d}Q}{\mathrm{d}t}(t) = -kQ(t)$$
>
> if and only if there is a constant C such that
>
> $$Q(t) = Ce^{-kt}$$

Before we start to apply the above theorem, we take this opportunity to remind the reader that in this text we will use $\log x$ with no base to indicate the natural logarithm. That is

$$\log x = \log_e x = \ln x$$

Both of the notations $\log(x)$ and $\ln(x)$ are used widely and the reader should be comfortable with both.

Example 3.3.3

In this example, we determine the value of the constant of proportionality k in equation (3.3.1) that corresponds to the half–life of ^{14}C, which is 5730 years.

- Imagine that some plant or animal contains a quantity Q_0 of ^{14}C at its time of death. Let's choose the zero point of time $t = 0$ to be the instant that the plant or animal died.

- Denote by $Q(t)$ the amount of ^{14}C in the plant or animal t years after it died. Then $Q(t)$ must obey both equation (3.3.1) and $Q(0) = Q_0$.

- Since $Q(t)$ must obey equation (3.3.1), Theorem 3.3.2 tells us that there must be a constant C such that $Q(t) = Ce^{-kt}$. To also have $Q_0 = Q(0) = Ce^{-k\times 0}$, the constant C must be Q_0. That is, $Q(t) = Q_0 e^{-kt}$ for all $t \geqslant 0$.

- By definition, the half–life of ^{14}C is the length of time that it takes for half of the ^{14}C to decay. That is, the half–life $t_{1/2}$ is determined by

$$Q(t_{1/2}) = \tfrac{1}{2}Q(0) = \tfrac{1}{2}Q_0 \qquad\qquad \text{but we know } Q(t) = Q_0 e^{-kt}$$
$$Q_0 e^{-kt_{1/2}} = \tfrac{1}{2}Q_0 \qquad\qquad\qquad \text{now cancel } Q_0$$
$$e^{-kt_{1/2}} = \tfrac{1}{2}$$

Taking the logarithm of both sides gives

$$-kt_{1/2} = \log \frac{1}{2} = -\log 2 \qquad \text{and so}$$

$$k = \frac{\log 2}{t_{1/2}}.$$

We are told that, for ^{14}C, the half–life $t_{1/2} = 5730$, so

$$k = \frac{\log 2}{5730} = 0.000121 \qquad \text{to 6 digits}$$

Example 3.3.3

From the work in the above example we have accumulated enough new facts to make a corollary to Theorem 3.3.2.

Corollary 3.3.4.

The function $Q(t)$ satisfies the equation

$$\frac{dQ}{dt} = -kQ(t)$$

if and only if

$$Q(t) = Q(0) \cdot e^{-kt}.$$

The half-life is defined to be the time $t_{1/2}$ which obeys

$$Q(t_{1/2}) = \frac{1}{2} \cdot Q(0).$$

The half-life is related to the constant k by

$$t_{1/2} = \frac{\log 2}{k}$$

Now here is a typical problem that is solved using Corollary 3.3.4.

Example 3.3.5

A particular piece of parchment contains about 64% as much ^{14}C as plants do today. Estimate the age of the parchment.

Solution. Let $Q(t)$ denote the amount of ^{14}C in the parchment t years after it was first created. By equation (3.3.1) and Example 3.3.3,

$$\frac{dQ}{dt} = -kQ(t) \qquad \text{with } k = \frac{\log 2}{5730} = 0.000121.$$

By Corollary 3.3.4

$$Q(t) = Q(0) \cdot e^{-kt}$$

The time at which $Q(t)$ reaches $0.64Q(0)$ is determined by

$$Q(t) = 0.64Q(0) \qquad \text{but } Q(t) = Q(0)e^{-kt}$$
$$Q(0)e^{-kt} = 0.64Q(0) \qquad \text{cancel } Q(0)$$
$$e^{-kt} = 0.64 \qquad \text{take logarithms}$$
$$-kt = \log 0.64$$
$$t = \frac{\log 0.64}{-k} = \frac{\log 0.64}{-0.000121} = 3700 \qquad \text{to 2 significant digits.}$$

That is, the parchment[18] is about 37 centuries old.

Example 3.3.5

We have stated that the half-life of ^{14}C is 5730 years. How can this be determined? We can explain this using the following example.

Example 3.3.6

A scientist in a B-grade science fiction film is studying a sample of the rare and fictitious element, implausium[19]. With great effort he has produced a sample of pure implausium. The next day — 17 hours later — he comes back to his lab and discovers that his sample is now only 37% pure. What is the half-life of the element?

Solution. We can again set up our problem using Corollary 3.3.4. Let $Q(t)$ denote the quantity of implausium at time t, measured in hours. Then we know

$$Q(t) = Q(0) \cdot e^{-kt}$$

We also know that

$$Q(17) = 0.37Q(0).$$

That enables us to determine k via

$$Q(17) = 0.37Q(0) = Q(0)e^{-17k} \qquad \text{divide both sides by } Q(0)$$
$$0.37 = e^{-17k}$$

and so

$$k = -\frac{\log 0.37}{17} = 0.05849$$

We can then convert this to the half life using Corollary 3.3.4:

$$t_{1/2} = \frac{\log 2}{k} \approx 11.85 \text{ hours}$$

18 The British Museum has an Egyptian mathematical text from the seventeenth century B.C.
19 Implausium leads to even weaker plots than unobtainium.

While this example is entirely fictitious, one really can use this approach to measure the half-life of materials.

Example 3.3.6

3.3.2 ⇥ Newton's Law of Cooling

Recall Newton's law of cooling from the start of this section:

> The rate of change of temperature of an object is proportional to the difference in temperature between the object and its surroundings. The temperature of the surroundings is sometimes called the ambient temperature.

We translated this statement into the following differential equation

Equation 3.3.7 (Newton's law of cooling).

$$\frac{dT}{dt}(t) = K\big[T(t) - A\big]$$

where $T(t)$ is the temperature of the object at time t, A is the temperature of its surroundings, and K is a constant of proportionality. This mathematical model of temperature change works well when studying a small object in a large, fixed temperature, environment. For example, a hot cup of coffee in a large room[20].

Before we worry about solving this equation, let's think a little about the sign of the constant of proportionality. At any time t, there are three possibilities.

- If $T(t) > A$, that is, if the body is warmer than its surroundings, we would expect heat to flow from the body into its surroundings and so we would expect the body to cool off so that $\frac{dT}{dt}(t) < 0$. For this expectation to be consistent with equation (3.3.7), we need $K < 0$.

- If $T(t) < A$, that is the body is cooler than its surroundings, we would expect heat to flow from the surroundings into the body and so we would expect the body to warm up so that $\frac{dT}{dt}(t) > 0$. For this expectation to be consistent with equation (3.3.7), we again need $K < 0$.

- Finally if $T(t) = A$, that is the body and its environment have the same temperature, we would not expect any heat to flow between the two and so we would expect that $\frac{dT}{dt}(t) = 0$. This does not impose any condition on K.

In conclusion, we would expect $K < 0$. Of course, we could have chosen to call the constant of proportionality $-k$, rather than K. Then the differential equation would be $\frac{dT}{dt} = -k(T - A)$ and we would expect $k > 0$.

---

20   It does not work so well when the object is of a similar size to its surroundings since the temperature of the surroundings will rise as the object cools. It also fails when there are phase transitions involved — for example, an ice-cube melting in a warm room does not obey Newton's law of cooling.

Now to find the general solution to equation (3.3.7). Since this equation is so similar in form to equation (3.3.1), we might expect a similar solution. Start by trying $T(t) = Ce^{Kt}$ and let's see what goes wrong. Substitute it into the equation:

$$\frac{\mathrm{d}T}{\mathrm{d}t} = K(T(t) - A)$$
$$KCe^{Kt} = KCe^{KT} - KA$$
$$?0 = -KA? \qquad\qquad \text{the constant } A \text{ causes problems!}$$

Let's try something a little different — recall that the derivative of a constant is zero. So we can add or subtract a constant from $T(t)$ without changing its derivative. Set $Q(t) = T(t) + B$, then

$$\frac{\mathrm{d}Q}{\mathrm{d}t}(t) = \frac{\mathrm{d}T}{\mathrm{d}t}(t) \qquad\qquad \text{by Newton's law of cooling}$$
$$= K(T(t) - A) = K(Q(t) - B - A)$$

So if we choose $B = -A$ then we will have

$$\frac{\mathrm{d}Q}{\mathrm{d}t}(t) = KQ(t)$$

which is exactly the same form as equation (3.3.1), but with $K = -k$. So by Theorem 3.3.2

$$Q(t) = Q(0)e^{Kt}$$

We can translate back to $T(t)$, since $Q(t) = T(t) - A$ and $Q(0) = T(0) - A$. This gives us the solution.

> **Corollary 3.3.8.**
>
> A differentiable function $T(t)$ obeys the differential equation
>
> $$\frac{\mathrm{d}T}{\mathrm{d}t}(t) = K[T(t) - A]$$
>
> if and only if
>
> $$T(t) = [T(0) - A]\,e^{Kt} + A$$

Just before we put this into action, we remind the reader that $\log x = \log_e x = \ln x$.

> **Example 3.3.9**
>
> The temperature of a glass of iced tea is initially $5°$. After 5 minutes, the tea has heated to $10°$ in a room where the air temperature is $30°$.
>
> (a) Determine the temperature as a function of time.
>
> (b) What is the temperature after 10 minutes?

(c) Determine when the tea will reach a temperature of $20°$.

Solution. Part (a)

- Denote by $T(t)$ the temperature of the tea t minutes after it was removed from the fridge, and let $A = 30$ be the ambient temperature.

- By Newton's law of cooling,

$$\frac{dT}{dt} = K(T - A) = K(T - 30)$$

for some, as yet unknown, constant of proportionality K.

- By Corollary 3.3.8,

$$T(t) = [T(0) - 30]\, e^{Kt} + 30 = 30 - 25e^{Kt}$$

since the initial temperature $T(0) = 5$.

- This solution is not complete because it still contains an unknown constant, namely K. We have not yet used the given data that $T(5) = 10$. We can use it to determine K. At $t = 5$,

$$T(5) = 30 - 25e^{5K} = 10 \qquad\qquad \text{rearrange}$$

$$e^{5K} = \frac{20}{25}$$

$$5K = \log\frac{20}{25} \qquad\qquad \text{and so}$$

$$K = \frac{1}{5}\log\frac{4}{5} = -0.044629 \qquad\qquad \text{to 6 digits}$$

Part (b)

- To find the temperature at 10 minutes we can just use the solution we have determined above.

$$T(10) = 30 - 25e^{10K}$$

$$= 30 - 25e^{10\times\frac{1}{5}\log\frac{4}{5}}$$

$$= 30 - 25e^{2\log\frac{4}{5}} = 30 - 25e^{\log\frac{16}{25}}$$

$$= 30 - 16 = 14°$$

Part (c)

- We can find when the temperature is $20°$ by solving $T(t) = 20$:

$$20 = 30 - 25e^{Kt} \qquad\qquad \text{rearrange}$$

$$e^{Kt} = \frac{10}{25} = \frac{2}{5}$$

$$Kt = \log\frac{2}{5}$$

$$t = \frac{\log\frac{2}{5}}{K}$$

$$= 20.5 \text{ minutes} \qquad\qquad \text{to 1 decimal place}$$

Example 3.3.9

A slightly more gruesome example.

Example 3.3.10

A dead body is discovered at 3:45pm in a room where the temperature is 20°C. At that time the temperature of the body is 27°C. Two hours later, at 5:45pm, the temperature of the body is 25.3 °C. What was the time of death? Note that the normal (adult human) body temperature is 37°.

Solution. We will assume[21] that the body's temperature obeys Newton's law of cooling.

- Denote by $T(t)$ the temperature of the body at time t, with $t = 0$ corresponding to 3:45pm. We wish to find the time of death — call it t_d.

- There is a lot of data in the statement of the problem; we are told that

  - the ambient temperature: $A = 20$
  - the temperature of the body when discovered: $T(0) = 27$
  - the temperature of the body 2 hours later: $T(2) = 25.3$
  - assuming the person was a healthy adult right up until he died, the temperature at the time of death: $T(t_d) = 37$.

- Since we assume the temperature of the body obeys Newton's law of cooling, we use Corollary 3.3.8 to find,

$$T(t) = [T(0) - A] e^{Kt} + A = 20 + 7e^{Kt}$$

Two unknowns remain, K and t_d.

- We can find the constant K by using $T(2) = 25.3$:

$$25.3 = T(2) = 20 + 7e^{2K} \qquad\qquad \text{rearrange}$$
$$7e^{2K} = 5.3 \qquad\qquad \text{rearrange a bit more}$$
$$2K = \log\left(\tfrac{5.3}{7}\right)$$
$$K = \tfrac{1}{2}\log\left(\tfrac{5.3}{7}\right) = -0.139 \qquad\qquad \text{to 3 decimal places}$$

- Since we know[22] that t_d is determined by $T(t_d) = 37$, we have

$$37 = T(t_d) = 20 + 7e^{-0.139t_d} \qquad\qquad \text{rearrange}$$
$$e^{-0.139t_d} = \tfrac{17}{7}$$
$$-0.139t_d = \log\left(\tfrac{17}{7}\right)$$
$$t_d = -\tfrac{1}{0.139}\log\left(\tfrac{17}{7}\right)$$
$$= -6.38 \qquad\qquad \text{to 2 decimal places}$$

---

21 We don't know any other method!
22 Actually, we are assuming again.

Now 6.38 hours is 6 hours and $0.38 \times 60 = 23$ minutes. So the time of death was 6 hours and 23 minutes before 3:45pm, which is 9:22am.

Example 3.3.10

A slightly tricky example — we need to determine the ambient temperature from three measurements at different times.

Example 3.3.11

A glass of room-temperature water is carried out onto a balcony from an apartment where the temperature is 22°C. After one minute the water has temperature 26°C and after two minutes it has temperature 28°C. What is the outdoor temperature?

Solution. We will assume that the temperature of the thermometer obeys Newton's law of cooling.

- Let A be the outdoor temperature and $T(t)$ be the temperature of the water t minutes after it is taken outside.

- By Newton's law of cooling,

$$T(t) = A + \big(T(0) - A\big)e^{Kt}$$

by Corollary 3.3.8. Notice there are 3 unknowns here — A, $T(0)$ and K — so we need three pieces of information to find them all.

- We are told $T(0) = 22$, so

$$T(t) = A + \big(22 - A\big)e^{Kt}.$$

- We are also told $T(1) = 26$, which gives

$$26 = A + (22 - A)e^K \qquad\qquad \text{rearrange things}$$
$$e^K = \frac{26 - A}{22 - A}$$

- Finally, $T(2) = 28$, so

$$28 = A + (22 - A)e^{2K} \qquad\qquad\qquad \text{rearrange}$$
$$e^{2K} = \frac{28 - A}{22 - A} \qquad\qquad \text{but } e^K = \frac{26 - A}{22 - A}, \text{ so}$$
$$\left(\frac{26 - A}{22 - A}\right)^2 = \frac{28 - A}{22 - A} \qquad\qquad \text{multiply through by } (22 - A)^2$$
$$(26 - A)^2 = (28 - A)(22 - A)$$

We can expand out both sides and collect up terms to get

$$\underbrace{26^2}_{=676} - 52A + A^2 = \underbrace{28 \times 22}_{=616} - 50A + A^2$$
$$60 = 2A$$
$$30 = A$$

So the temperature outside is 30°.

3.3.3 ⤑ Population Growth

Suppose that we wish to predict the size $P(t)$ of a population as a function of the time t. In the most naive model of population growth, each couple produces β offspring (for some constant β) and then dies. Thus over the course of one generation $\beta\frac{P(t)}{2}$ children are produced and $P(t)$ parents die so that the size of the population grows from $P(t)$ to

$$P(t + t_g) = \underbrace{P(t) + \beta\frac{P(t)}{2}}_{\text{parents+offspring}} - \underbrace{P(t)}_{\text{parents die}} = \frac{\beta}{2}P(t)$$

where t_g denotes the lifespan of one generation. The rate of change of the size of the population per unit time is

$$\frac{P(t + t_g) - P(t)}{t_g} = \frac{1}{t_g}\left[\frac{\beta}{2}P(t) - P(t)\right] = bP(t)$$

where $b = \frac{\beta - 2}{2t_g}$ is the net birthrate per member of the population per unit time. If we approximate

$$\frac{P(t+t_g) - P(t)}{t_g} \approx \frac{\mathrm{d}P}{\mathrm{d}t}(t)$$

we get the differential equation

> **Equation 3.3.12** (Simple population model).
>
> $$\frac{\mathrm{d}P}{\mathrm{d}t} = bP(t)$$

By Corollary 3.3.4, with $-k$ replaced by b,

$$P(t) = P(0) \cdot e^{bt}$$

This is called the Malthusian[23] growth model. It is, of course, very simplistic. One of its main characteristics is that, since $P(t + T) = P(0) \cdot e^{b(t+T)} = P(t) \cdot e^{bT}$, every time you *add* T to the time, the population size is *multiplied* by e^{bT}. In particular, the population size doubles every $\frac{\log 2}{b}$ units of time. The Malthusian growth model can be a reasonably good model only when the population size is very small compared to its environment[24].

---

23   This is named after Rev. Thomas Robert Malthus. He described this model in a 1798 paper called "An essay on the principle of population".
24   That is, the population has plenty of food and space to grow.

A more sophisticated model of population growth, that takes into account the "carrying capacity of the environment" is considered in the optional subsection below.

Example 3.3.13

In 1927 the population of the world was about 2 billion. In 1974 it was about 4 billion. Estimate when it reached 6 billion. What will the population of the world be in 2100, assuming the Malthusian growth model?

Solution. We follow our usual pattern for dealing with such problems.

- Let $P(t)$ be the world's population t years after 1927. Note that 1974 corresponds to $t = 1974 - 1927 = 47$.

- We are assuming that $P(t)$ obeys equation (3.3.12). So, by Corollary 3.3.4 with $-k$ replaced by b,

$$P(t) = P(0) \cdot e^{bt}$$

Notice that there are 2 unknowns here — b and $P(0)$ — so we need two pieces of information to find them.

- We are told $P(0) = 2$, so

$$P(t) = 2 \cdot e^{bt}$$

- We are also told $P(47) = 4$, which gives

$$
\begin{aligned}
4 &= 2 \cdot e^{47b} && \text{clean up} \\
e^{47b} &= 2 && \text{take the log and clean up} \\
b &= \frac{\log 2}{47} = 0.0147 && \text{to 3 significant digits}
\end{aligned}
$$

- We now know $P(t)$ completely, so we can easily determine the predicted population[25] in 2100, i.e. at $t = 2100 - 1927 = 173$.

$$P(173) = 2e^{173b} = 2e^{173 \times 0.0147} = 25.4 \text{ billion}$$

- Finally, our crude model predicts that the population is 6 billion at the time t that obeys

$$
\begin{aligned}
P(t) = 2e^{bt} &= 6 && \text{clean up} \\
e^{bt} &= 3 && \text{take the log and clean up} \\
t = \frac{\log 3}{b} &= 47\frac{\log 3}{\log 2} = 74.5
\end{aligned}
$$

which corresponds[26] to the middle of 2001.

Example 3.3.13

---

25 The *2015 Revision of World Population*, a publication of the United Nations, predicts that the world's population in 2100 will be about 11 billion. They are predicting a reduction in the world population growth rate due to lower fertility rates, which the Malthusian growth model does not take into account.
26 The world population really reached 6 billion in about 1999.

▸▸▸ (Optional) — Logistic Population Growth

Logistic growth adds one more wrinkle to the simple population model. It assumes that the population only has access to limited resources. As the size of the population grows the amount of food available to each member decreases. This in turn causes the net birth rate b to decrease. In the logistic growth model $b = b_0 \left(1 - \frac{P}{K}\right)$, where K is called the carrying capacity of the environment, so that

$$P'(t) = b_0 \left(1 - \frac{P(t)}{K}\right) P(t)$$

We can learn quite a bit about the behaviour of solutions to differential equations like this, without ever finding formulae for the solutions, just by watching the sign of $P'(t)$. For concreteness, we'll look at solutions of the differential equation

$$\frac{dP}{dt}(t) = \left(6000 - 3P(t)\right) P(t)$$

We'll sketch the graphs of four functions $P(t)$ that obey this equation.

- For the first function, $P(0) = 0$.
- For the second function, $P(0) = 1000$.
- For the third function, $P(0) = 2000$.
- For the fourth function, $P(0) = 3000$.

The sketches will be based on the observation that $(6000 - 3P) P = 3(2000 - P) P$

- is zero for $P = 0,\ 2000$,
- is strictly positive for $0 < P < 2000$ and
- is strictly negative for $P > 2000$.

Consequently

$$\frac{dP}{dt}(t) \begin{cases} = 0 & \text{if } P(t) = 0 \\ > 0 & \text{if } 0 < P(t) < 2000 \\ = 0 & \text{if } P(t) = 2000 \\ < 0 & \text{if } P(t) > 2000 \end{cases}$$

Thus if $P(t)$ is some function that obeys $\frac{dP}{dt}(t) = \left(6000 - 3P(t)\right)P(t)$, then as the graph of $P(t)$ passes through $\left(t, P(t)\right)$

$$\text{the graph has} \begin{cases} \text{slope zero,} & \text{i.e. is horizontal,} & \text{if } P(t) = 0 \\ \text{positive slope,} & \text{i.e. is increasing,} & \text{if } 0 < P(t) < 2000 \\ \text{slope zero,} & \text{i.e. is horizontal,} & \text{if } P(t) = 2000 \\ \text{negative slope,} & \text{i.e. is decreasing,} & \text{if } P(t) > 2000 \end{cases}$$

as illustrated in the figure
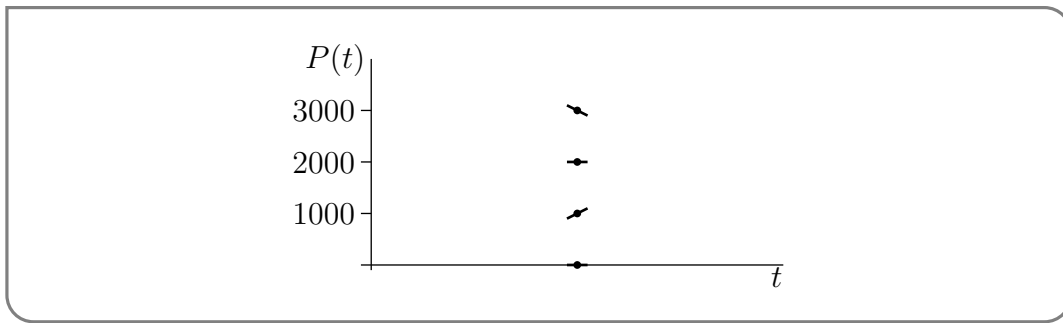
As a result,

- if $P(0) = 0$, the graph starts out horizontally. In other words, as t starts to increase, $P(t)$ remains at zero, so the slope of the graph remains at zero. The population size remains zero for all time. As a check, observe that the function $P(t) = 0$ obeys $\frac{dP}{dt}(t) = (6000 - 3P(t))P(t)$ for all t.

- Similarly, if $P(0) = 2000$, the graph again starts out horizontally. So $P(t)$ remains at 2000 and the slope remains at zero. The population size remains 2000 for all time. Again, the function $P(t) = 2000$ obeys $\frac{dP}{dt}(t) = (6000 - 3P(t))P(t)$ for all t.

- If $P(0) = 1000$, the graph starts out with positive slope. So $P(t)$ increases with t. As $P(t)$ increases towards 2000, the slope $(6000 - 3P(t))P(t)$, while remaining positive, gets closer and closer to zero. As the graph approaches height 2000, it becomes more and more horizontal. The graph cannot actually cross from below 2000 to above 2000, because to do so, it would have to have strictly positive slope for some value of P above 2000, which is not allowed.

- If $P(0) = 3000$, the graph starts out with negative slope. So $P(t)$ decreases with t. As $P(t)$ decreases towards 2000, the slope $(6000 - 3P(t))P(t)$, while remaining negative, gets closer and closer to zero. As the graph approaches height 2000, it becomes more and more horizontal. The graph cannot actually cross from above 2000 to below 2000, because to do so, it would have to have negative slope for some value of P below 2000. which is not allowed.

These curves are sketched in the figure below. We conclude that for any initial population size $P(0)$, except $P(0) = 0$, the population size approaches 2000 as $t \to \infty$.

Figure 3.3.1.
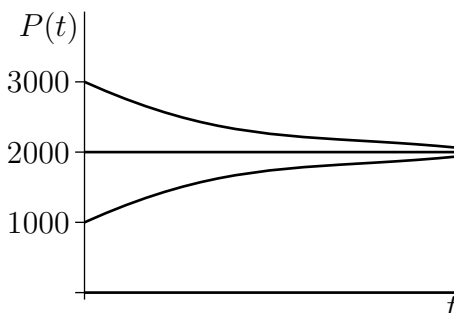
3.4 ▴ Approximating Functions Near a Specified Point — Taylor Polynomials

Suppose that you are interested in the values of some function $f(x)$ for x near some fixed point a. When the function is a polynomial or a rational function we can use some arithmetic (and maybe some hard work) to write down the answer. For example:

$$f(x) = \frac{x^2 - 3}{x^2 - 2x + 4}$$

$$f(1/5) = \frac{\frac{1}{25} - 3}{\frac{1}{25} - \frac{2}{5} + 4} = \frac{\frac{1-75}{25}}{\frac{1-10+100}{25}}$$

$$= \frac{-74}{91}$$

Tedious, but we can do it. On the other hand if you are asked to compute $\sin(1/10)$ then what can we do? We know that a calculator can work it out

$$\sin(1/10) = 0.09983341\ldots$$

but how does the calculator do this? How did people compute this before calculators[27]? A hint comes from the following sketch of $\sin(x)$ for x around 0.

Figure 3.4.1.
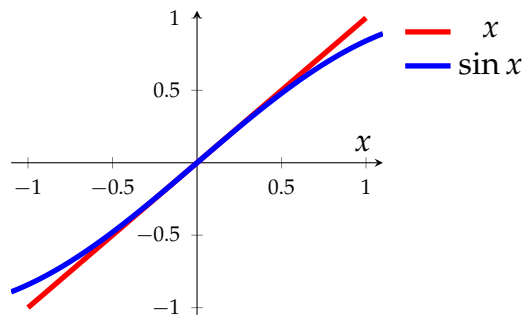
The above figure shows that the curves $y = x$ and $y = \sin x$ are almost the same when x is close to 0. Hence if we want the value of $\sin(1/10)$ we could just use this approximation $y = x$ to get

$$\sin(1/10) \approx 1/10.$$

Of course, in this case we simply observed that one function was a good approximation of the other. We need to know how to find such approximations more systematically.

More precisely, say we are given a function $f(x)$ that we wish to approximate close to some point $x = a$, and we need to find another function $F(x)$ that

- is simple and easy to compute[28]

---

27 Originally the word "calculator" referred not to the software or electronic (or even mechanical) device we think of today, but rather to a person who performed calculations.
28 It is no good approximating a function with something that is even more difficult to work with.

- is a good approximation to $f(x)$ for x values close to a.

Further, we would like to understand how good our approximation actually is. Namely we need to be able to estimate the error $|f(x) - F(x)|$.

There are many different ways to approximate a function and we will discuss one family of approximations: Taylor polynomials. This is an infinite family of ever improving approximations, and our starting point is the very simplest.

3.4.1 ▸ Zeroth Approximation — the Constant Approximation

The simplest functions are those that are constants. And our zeroth[29] approximation will be by a constant function. That is, the approximating function will have the form $F(x) = A$, for some constant A. Notice that this function is a polynomial of degree zero.

To ensure that $F(x)$ is a good approximation for x close to a, we choose A so that $f(x)$ and $F(x)$ take exactly the same value when $x = a$.

$$F(x) = A \quad \text{so} \quad F(a) = A = f(a) \implies A = f(a)$$

Our first, and crudest, approximation rule is

Equation 3.4.1 (Constant approximation).

$$f(x) \approx f(a)$$

An important point to note is that we need to know $f(a)$ — if we cannot compute that easily then we are not going to be able to proceed. We will often have to choose a (the point around which we are approximating $f(x)$) with some care to ensure that we can compute $f(a)$.

Here is a figure showing the graphs of a typical $f(x)$ and approximating function $F(x)$. At $x = a$, $f(x)$ and $F(x)$ take the same value. For x very near a, the values of $f(x)$ and $F(x)$
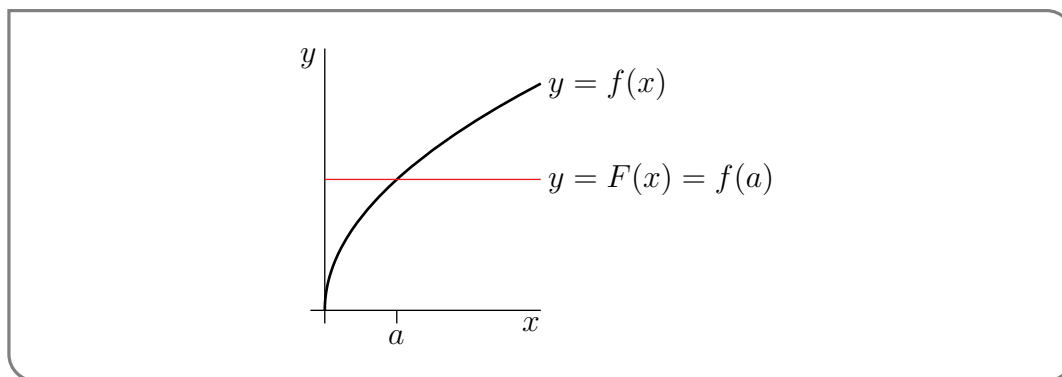
remain close together. But the quality of the approximation deteriorates fairly quickly as x moves away from a. Clearly we could do better with a straight line that follows the slope of the curve. That is our next approximation.

---

29   It barely counts as an approximation at all, but it will help build intuition. Because of this, and the fact that a constant is a polynomial of degree 0, we'll start counting our approximations from zero rather than 1.

But before then, an example:

Example 3.4.2

Use the constant approximation to estimate $e^{0.1}$.

Solution. First set $f(x) = e^x$.

- Now we first need to pick a point $x = a$ to approximate the function. This point needs to be close to 0.1 and we need to be able to evaluate $f(a)$ easily. The obvious choice is $a = 0$.

- Then our constant approximation is just

$$F(x) = f(0) = e^0 = 1$$
$$F(0.1) = 1$$

Note that $e^{0.1} = 1.105170918\ldots$, so even this approximation isn't too bad..

Example 3.4.2

3.4.2 ▸▸ First Approximation — the Linear approximation

Our first[30] approximation improves on our zeroth approximation by allowing the approximating function to be a linear function of x rather than just a constant function. That is, we allow $F(x)$ to be of the form $A + Bx$, for some constants A and B.

To ensure that $F(x)$ is a good approximation for x close to a, we still require that $f(x)$ and $F(x)$ have the same value at $x = a$ (that was our zeroth approximation). Our additional requirement is that their tangent lines at $x = a$ have the same slope — that the derivatives of $f(x)$ and $F(x)$ are the same at $x = a$. Hence

$$F(x) = A + Bx \qquad \Longrightarrow \qquad F(a) = A + Ba = f(a)$$
$$F'(x) = B \qquad \Longrightarrow \qquad F'(a) = \qquad B = f'(a)$$

So we must have $B = f'(a)$. Substituting this into $A + Ba = f(a)$ we get $A = f(a) - af'(a)$. So we can write

$$F(x) = A + Bx = \overbrace{f(a) - af'(a)}^{A} + f'(a) \cdot x$$
$$= f(a) + f'(a) \cdot (x - a)$$

We write it in this form because we can now clearly see that our first approximation is just an extension of our zeroth approximation. This first approximation is also often called the linear approximation of $f(x)$ about $x = a$.

Equation 3.4.3 (Linear approximation).

$$f(x) \approx f(a) + f'(a)(x - a)$$

---

30   Recall that we started counting from zero.

We should again stress that in order to form this approximation we need to know $f(a)$ and $f'(a)$ — if we cannot compute them easily then we are not going to be able to proceed.

Recall, from Theorem 2.3.2, that $y = f(a) + f'(a)(x - a)$ is exactly the equation of the tangent line to the curve $y = f(x)$ at a. Here is a figure showing the graphs of a typical $f(x)$ and the approximating function $F(x)$. Observe that the graph of $f(a) + f'(a)(x - a)$
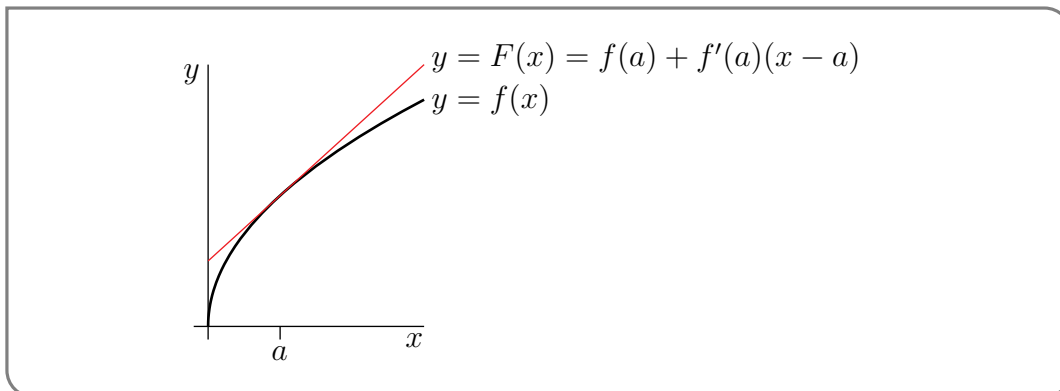
remains close to the graph of $f(x)$ for a much larger range of x than did the graph of our constant approximation, $f(a)$. One can also see that we can improve this approximation if we can use a function that curves down rather than being perfectly straight. That is our next approximation.

But before then, back to our example:

Example 3.4.4

Use the linear approximation to estimate $e^{0.1}$.

Solution. First set $f(x) = e^x$ and $a = 0$ as before.

- To form the linear approximation we need $f(a)$ and $f'(a)$:

$$f(x) = e^x \qquad\qquad f(0) = 1$$
$$f'(x) = e^x \qquad\qquad f'(0) = 1$$

- Then our linear approximation is

$$F(x) = f(0) + x f'(0) = 1 + x$$
$$F(0.1) = 1.1$$

Recall that $e^{0.1} = 1.105170918\ldots$, so the linear approximation is almost correct to 3 digits.

Example 3.4.4

It is worth doing another simple example here.

Example 3.4.5

Use a linear approximation to estimate $\sqrt{4.1}$.

Solution. First set $f(x) = \sqrt{x}$. Hence $f'(x) = \frac{1}{2\sqrt{x}}$. Then we are trying to approximate $f(4.1)$. Now we need to choose a sensible a value.

- We need to choose a so that $f(a)$ and $f'(a)$ are easy to compute.

    - We could try $a = 4.1$ — but then we need to compute $f(4.1)$ and $f'(4.1)$ — which is our original problem and more!

    - We could try $a = 0$ — then $f(0) = 0$ and $f'(0) = DNE$.

    - Setting $a = 1$ gives us $f(1) = 1$ and $f'(1) = \frac{1}{2}$. This would work, but we can get a better approximation by choosing a is closer to 4.1.

    - Indeed we can set a to be the square of any rational number and we'll get a result that is easy to compute.

    - Setting $a = 4$ gives $f(4) = 2$ and $f'(4) = \frac{1}{4}$. This seems good enough.

- Substitute this into equation (3.4.3) to get

$$f(4.1) \approx f(4) + f'(4) \cdot (4.1 - 4)$$
$$= 2 + \frac{0.1}{4} = 2 + 0.025 = 2.025$$

Notice that the true value is $\sqrt{4.1} = 2.024845673\ldots$.

Example 3.4.5

3.4.3 ▸▸ Second Approximation — the Quadratic Approximation

We next develop a still better approximation by now allowing the approximating function be to a quadratic function of x. That is, we allow $F(x)$ to be of the form $A + Bx + Cx^2$, for some constants A, B and C. To ensure that $F(x)$ is a good approximation for x close to a, we choose A, B and C so that

- $f(a) = F(a)$ (just as in our zeroth approximation),

- $f'(a) = F'(a)$ (just as in our first approximation), and

- $f''(a) = F''(a)$ — this is a new condition.

These conditions give us the following equations

$$
\begin{aligned}
F(x) &= A + Bx + Cx^2 &\implies& & F(a) &= A + Ba + Ca^2 = f(a) \\
F'(x) &= B + 2Cx &\implies& & F'(a) &= \quad\quad B + 2Ca = f'(a) \\
F''(x) &= 2C &\implies& & F''(a) &= \quad\quad\quad\quad 2C = f''(a)
\end{aligned}
$$

Solve these for C first, then B and finally A.

$$
\begin{aligned}
C &= \tfrac{1}{2}f''(a) & \text{substitute} \\
B &= f'(a) - 2Ca = f'(a) - af''(a) & \text{substitute again} \\
A &= f(a) - Ba - Ca^2 = f(a) - a[f'(a) - af''(a)] - \tfrac{1}{2}f''(a)a^2
\end{aligned}
$$

Then put things back together to build up $F(x)$:

$$F(x) = f(a) - f'(a)a + \tfrac{1}{2}f''(a)a^2 \qquad \text{(this line is } A\text{)}$$
$$+ f'(a)\, x - f''(a)ax \qquad \text{(this line is } Bx\text{)}$$
$$+ \tfrac{1}{2}f''(a)x^2 \qquad \text{(this line is } Cx^2\text{)}$$
$$= f(a) + f'(a)(x - a) + \tfrac{1}{2}f''(a)(x - a)^2$$

Oof! We again write it in this form because we can now clearly see that our second approximation is just an extension of our first approximation.

Our second approximation is called the quadratic approximation:

Equation 3.4.6 (Quadratic approximation).

$$f(x) \approx f(a) + f'(a)(x - a) + \tfrac{1}{2}f''(a)(x - a)^2$$

Here is a figure showing the graphs of a typical $f(x)$ and approximating function $F(x)$.
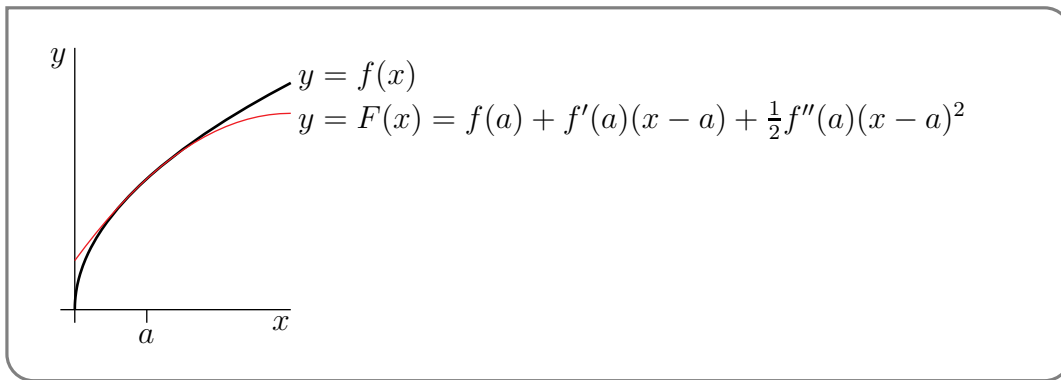
This new approximation looks better than both the first and second.

Now there is actually an easier way to derive this approximation, which we show you now. Let us rewrite[31] $F(x)$ so that it is easy to evaluate it and its derivatives at $x = a$:

$$F(x) = \alpha + \beta \cdot (x - a) + \gamma \cdot (x - a)^2$$

Then

$$
\begin{aligned}
F(x) &= \alpha + \beta \cdot (x - a) + \gamma \cdot (x - a)^2 & F(a) &= \alpha = f(a) \\
F'(x) &= \beta + 2\gamma \cdot (x - a) & F'(a) &= \beta = f'(a) \\
F''(x) &= 2\gamma & F''(a) &= 2\gamma = f''(a)
\end{aligned}
$$

And from these we can clearly read off the values of α, β and γ and so recover our function $F(x)$. Additionally if we write things this way, then it is quite clear how to extend this to a cubic approximation and a quartic approximation and so on.

Return to our example:

31 Any polynomial of degree two can be written in this form. For example, when $a = 1$, $3 + 2x + x^2 = 6 + 4(x - 1) + (x - 1)^2$.

Example 3.4.7

Use the quadratic approximation to estimate $e^{0.1}$.

Solution. Set $f(x) = e^x$ and $a = 0$ as before.

- To form the quadratic approximation we need $f(a), f'(a)$ and $f''(a)$:

$$
\begin{aligned}
f(x) &= e^x & f(0) &= 1 \\
f'(x) &= e^x & f'(0) &= 1 \\
f''(x) &= e^x & f''(0) &= 1
\end{aligned}
$$

- Then our quadratic approximation is

$$
F(x) = f(0) + xf'(0) + \frac{1}{2}x^2 f''(0) = 1 + x + \frac{x^2}{2}
$$
$$
F(0.1) = 1.105
$$

Recall that $e^{0.1} = 1.105170918\ldots$, so the quadratic approximation is quite accurate with very little effort.

Example 3.4.7

Before we go on, let us first introduce (or revise) some notation that will make our discussion easier.

▸ Whirlwind Tour of Summation Notation

In the remainder of this section we will frequently need to write sums involving a large number of terms. Writing out the summands explicitly can become quite impractical — for example, say we need the sum of the first 11 squares:

$$
1 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2 + 7^2 + 8^2 + 9^2 + 10^2 + 11^2
$$

This becomes tedious. Where the pattern is clear, we will often skip the middle few terms and instead write

$$
1 + 2^2 + \cdots + 11^2.
$$

A far more precise way to write this is using Σ (capital-sigma) notation. For example, we can write the above sum as

$$
\sum_{k=1}^{11} k^2
$$

This is read as

The sum from k equals 1 to 11 of k^2.

More generally

Notation 3.4.8.

Let $m \leqslant n$ be integers and let $f(x)$ be a function defined on the integers. Then we write

$$\sum_{k=m}^{n} f(k)$$

to mean the sum of $f(k)$ for k from m to n:

$$f(m) + f(m+1) + f(m+2) + \cdots + f(n-1) + f(n).$$

Similarly we write

$$\sum_{i=m}^{n} a_i$$

to mean

$$a_m + a_{m+1} + a_{m+2} + \cdots + a_{n-1} + a_n$$

for some set of coefficients $\{a_m, \ldots, a_n\}$.

Consider the example

$$\sum_{k=3}^{7} \frac{1}{k^2} = \frac{1}{3^2} + \frac{1}{4^2} + \frac{1}{5^2} + \frac{1}{6^2} + \frac{1}{7^2}$$

It is important to note that the right hand side of this expression evaluates to a number[32]; it does not contain "k". The summation index k is just a "dummy" variable and it does not have to be called k. For example

$$\sum_{k=3}^{7} \frac{1}{k^2} = \sum_{i=3}^{7} \frac{1}{i^2} = \sum_{j=3}^{7} \frac{1}{j^2} = \sum_{\ell=3}^{7} \frac{1}{\ell^2}$$

Also the summation index has no meaning outside the sum. For example

$$k \sum_{k=3}^{7} \frac{1}{k^2}$$

has no mathematical meaning; It is gibberish[33].

---

32 Some careful addition shows it is $\frac{46181}{176400}$.

33 Or possibly gobbledygook. For a discussion of statements without meaning and why one should avoid them we recommend the book "Bendable learnings: the wisdom of modern management" by Don Watson.

3.4.4 ⤇ Still Better Approximations — Taylor Polynomials

We can use the same strategy to generate still better approximations by polynomials[34] of any degree we like. As was the case with the approximations above, we determine the coefficients of the polynomial by requiring, that at the point $x = a$, the approximation and its first n derivatives agree with those of the original function.

Rather than simply moving to a cubic polynomial, let us try to write things in a more general way. We will consider approximating the function $f(x)$ using a polynomial, $T_n(x)$, of degree n — where n is a non-negative integer. As we discussed above, the algebra is easier if we write

$$T_n(x) = c_0 + c_1(x - a) + c_2(x - a)^2 + \cdots + c_n(x - a)^n$$
$$= \sum_{k=0}^{n} c_k(x - a)^k \qquad\qquad \text{using } \Sigma \text{ notation}$$

The above form[35] [36] makes it very easy to evaluate this polynomial and its derivatives at $x = a$. Before we proceed, we remind the reader of some notation (see Notation 2.2.8):

- Let $f(x)$ be a function and k be a positive integer. We can denote its kth derivative with respect to x by

$$\frac{d^k f}{dx^k} \qquad\qquad \left(\frac{d}{dx}\right)^k f(x) \qquad\qquad f^{(k)}(x)$$

  Additionally we will need

---

34 Polynomials are generally a good choice for an approximating function since they are so easy to work with. Depending on the situation other families of functions may be more appropriate. For example if you are approximating a periodic function, then sums of sines and cosines might be a better choice; this leads to Fourier series.

35 Any polynomial in x of degree n can also be expressed as a polynomial in $(x - a)$ of the same degree n and vice versa. So $T_n(x)$ really still is a polynomial of degree n.

36 Furthermore when x is close to a, $(x - a)^k$ decreases very quickly as k increases, which often makes the "high k" terms in $T_n(x)$ very small. This can be a considerable advantage when building up approximations by adding more and more terms. If we were to rewrite $T_n(x)$ in the form $\sum_{k=0}^{n} b_k x^k$ the "high k" terms would typically not be very small when x is close to a.

Definition 3.4.9 (Factorial).

Let n be a positive integer[37], then n-factorial, denoted $n!$, is the product

$$n! = n \times (n-1) \times \cdots \times 3 \times 2 \times 1$$

Further, we use the convention that

$$0! = 1$$

The first few factorials are

$$1! = 1 \qquad\qquad 2! = 2 \qquad\qquad 3! = 6$$
$$4! = 24 \qquad\qquad 5! = 120 \qquad\qquad 6! = 720$$

Now consider $T_n(x)$ and its derivatives:

$$
\begin{aligned}
T_n(x) &= c_0 \;+c_1(x-a) \;+c_2(x-a)^2 \;+c_3(x-a)^3 \;+\cdots+\; c_n(x-a)^n \\
T_n'(x) &= \qquad\quad c_1 \;+2c_2(x-a) \;+3c_3(x-a)^2 \;+\cdots+\; nc_n(x-a)^{n-1} \\
T_n''(x) &= \qquad\qquad\qquad\quad 2c_2 \;+6c_3(x-a) \;+\cdots+\; n(n-1)c_n(x-a)^{n-2} \\
T_n'''(x) &= \qquad\qquad\qquad\qquad\qquad\quad 6c_3 \;+\cdots+\; n(n-1)(n-2)c_n(x-a)^{n-3} \\
&\;\;\vdots \\
T_n^{(n)}(x) &= \qquad\qquad\qquad\qquad\qquad\qquad\qquad n! \cdot c_n
\end{aligned}
$$

Now notice that when we substitute $x = a$ into the above expressions only the constant terms survive and we get

$$
\begin{aligned}
T_n(a) &= c_0 \\
T_n'(a) &= c_1 \\
T_n''(a) &= 2 \cdot c_2 \\
T_n'''(a) &= 6 \cdot c_3 \\
&\;\;\vdots \\
T_n^{(n)}(a) &= n! \cdot c_n
\end{aligned}
$$

So now if we want to set the coefficients of $T_n(x)$ so that it agrees with $f(x)$ at $x = a$ then we need

$$T_n(a) = c_0 = f(a) \qquad\qquad\qquad c_0 = f(a) = \frac{1}{0!}f(a)$$

37   It is actually possible to define the factorial of positive real numbers and even negative numbers but it requires more advanced calculus and is outside the scope of this course. The interested reader should look up the Gamma function.

We also want the first n derivatives of $T_n(x)$ to agree with the derivatives of $f(x)$ at $x = a$, so

$$T_n'(a) = c_1 = f'(a) \qquad\qquad c_1 = f'(a) = \frac{1}{1!}f'(a)$$

$$T_n''(a) = 2 \cdot c_2 = f''(a) \qquad\qquad c_2 = \frac{1}{2}f''(a) = \frac{1}{2!}f''(a)$$

$$T_n'''(a) = 6 \cdot c_3 = f'''(a) \qquad\qquad c_3 = \frac{1}{6}f'''(a) = \frac{1}{3!}f'''(a)$$

More generally, making the kth derivatives agree at $x = a$ requires :

$$T_n^{(k)}(a) = k! \cdot c_k = f^{(k)}(a) \qquad\qquad c_k = \frac{1}{k!}f^{(k)}(a)$$

And finally the nth derivative:

$$T_n^{(n)}(a) = n! \cdot c_n = f^{(n)}(a) \qquad\qquad c_n = \frac{1}{n!}f^{(n)}(a)$$

Putting this all together we have

Equation 3.4.10 (Taylor polynomial).

$$f(x) \approx T_n(x) = f(a) + f'(a)(x-a) + \frac{1}{2}f''(a) \cdot (x-a)^2 + \cdots + \frac{1}{n!}f^{(n)}(a) \cdot (x-a)^n$$

$$= \sum_{k=0}^{n} \frac{1}{k!}f^{(k)}(a) \cdot (x-a)^k$$

Let us formalise this definition.

Definition 3.4.11 (Taylor polynomial).

Let a be a constant and let n be a non-negative integer. The nth order[38] Taylor polynomial for $f(x)$ about $x = a$ is

$$T_n(x) = \sum_{k=0}^{n} \frac{1}{k!}f^{(k)}(a) \cdot (x-a)^k.$$

The special case $a = 0$ is called a Maclaurin[39] polynomial.

Before we proceed with some examples, a couple of remarks are in order.

---

38   It is sometimes called the nth degree Taylor polynomial, but its degree will actually be less than n if $f^{(n)}(a) = 0$.

39   The polynomials are named after Brook Taylor who devised a general method for constructing them in 1715. Slightly later, Colin Maclaurin made extensive use of the special case $a = 0$ (with attribution of the general case to Taylor) and it is now named after him. The special case of $a = 0$ was worked on previously by James Gregory and Isaac Newton, and some specific cases were known to the 14th century Indian mathematician Madhava of Sangamagrama.

- While we can compute a Taylor polynomial about any a-value (providing the derivatives exist), in order to be a *useful* approximation, we must be able to compute $f(a), f'(a), \ldots, f^{(n)}(a)$ easily. This means we must choose the point a with care. Indeed for many functions the choice $a = 0$ is very natural — hence the prominence of Maclaurin polynomials.

- If we have computed the approximation $T_n(x)$, then we can readily extend this to the next Taylor polynomial $T_{n+1}(x)$ since

$$T_{n+1}(x) = T_n(x) + \frac{1}{(n+1)!} f^{(n+1)}(a) \cdot (x-a)^{n+1}$$

This is very useful if we discover that $T_n(x)$ is an insufficient approximation, because then we can produce $T_{n+1}(x)$ without having to start again from scratch.

3.4.5 ▸▸ Some Examples

Let us return to our running example of e^x:

Example 3.4.12

The constant, linear and quadratic approximations we used above were the first few Maclaurin polynomial approximations of e^x. That is

$$T_0(x) = 1 \qquad\qquad T_1(x) = 1 + x \qquad\qquad T_2(x) = 1 + x + \frac{x^2}{2}$$

Since $\frac{d}{dx} e^x = e^x$, the Maclaurin polynomials are very easy to compute. Indeed this invariance under differentiation means that

$$f^{(n)}(x) = e^x \qquad\qquad n = 0, 1, 2, \ldots \qquad\qquad \text{so}$$
$$f^{(n)}(0) = 1$$

Substituting this into equation (3.4.10) we get

$$T_n(x) = \sum_{k=0}^{n} \frac{1}{k!} x^k$$

Thus we can write down the seventh Maclaurin polynomial very easily:

$$T_7(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \frac{x^5}{120} + \frac{x^6}{720} + \frac{x^7}{5040}$$

The following figure contains sketches of the graphs of e^x and its Taylor polynomials $T_n(x)$ for $n = 0, 1, 2, 3, 4$.
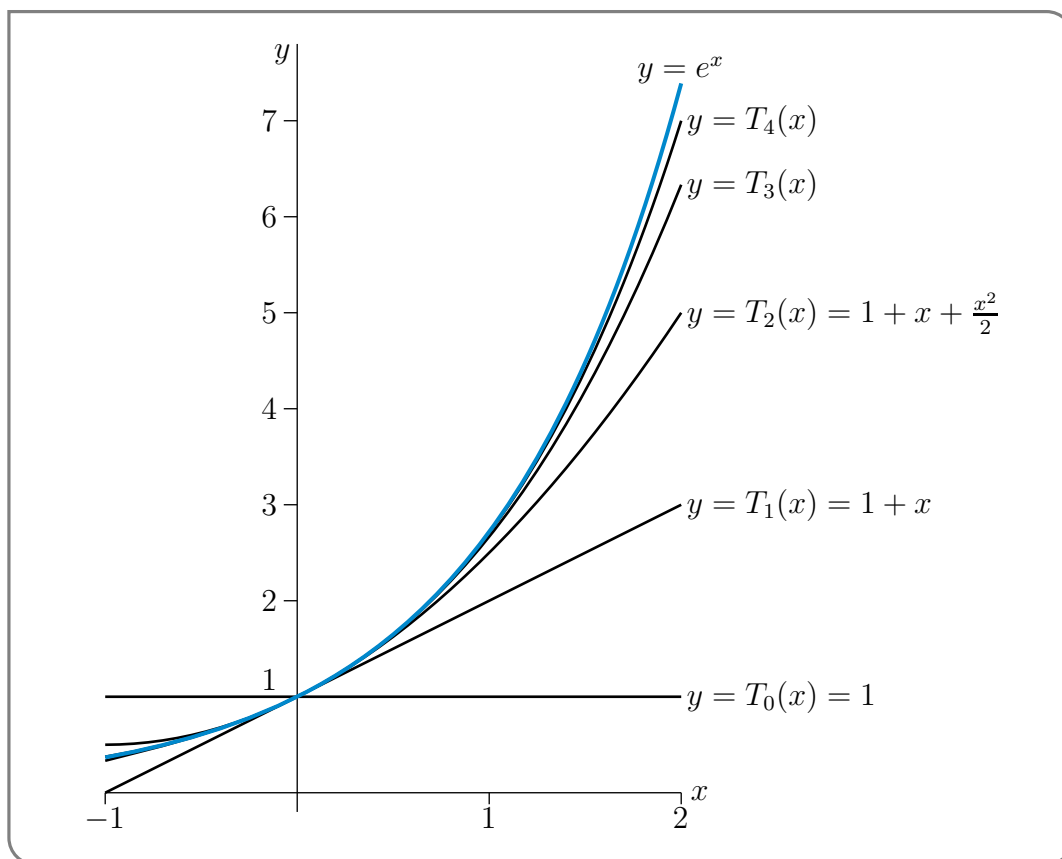
Also notice that if we use $T_7(1)$ to approximate the value of e^1 we obtain:

$$e^1 \approx T_7(1) = 1 + 1 + \frac{1}{2} + \frac{1}{6} + \frac{1}{24} + \frac{1}{120} + \frac{1}{720} + \frac{1}{5040}$$
$$= \frac{685}{252} = 2.718253968\ldots$$

The true value of e is $2.718281828\ldots$, so the approximation has an error of about 3×10^{-5}.

   Under the assumption that the accuracy of the approximation improves with n (an assumption we examine in Subsection 3.4.8 below) we can see that the approximation of e above can be improved by adding more and more terms. Indeed this is how the expression for e in equation (2.7.3) in Section 2.7 comes about.

Example 3.4.12

Now that we have examined Maclaurin polynomials for e^x we should take a look at $\log x$. Notice that we cannot compute a Maclaurin polynomial for $\log x$ since it is not defined at $x = 0$.

Example 3.4.13

Compute the 5th order Taylor polynomial for $\log x$ about $x = 1$.

Solution. We have been told $a = 1$ and fifth order, so we should start by writing down the

function and its first five derivatives:

$$f(x) = \log x \qquad\qquad f(1) = \log 1 = 0$$
$$f'(x) = \frac{1}{x} \qquad\qquad f'(1) = 1$$
$$f''(x) = \frac{-1}{x^2} \qquad\qquad f''(1) = -1$$
$$f'''(x) = \frac{2}{x^3} \qquad\qquad f'''(1) = 2$$
$$f^{(4)}(x) = \frac{-6}{x^4} \qquad\qquad f^{(4)}(1) = -6$$
$$f^{(5)}(x) = \frac{24}{x^5} \qquad\qquad f^{(5)}(1) = 24$$

Substituting this into equation (3.4.10) gives

$$T_5(x) = 0 + 1 \cdot (x-1) + \frac{1}{2} \cdot (-1) \cdot (x-1)^2 + \frac{1}{6} \cdot 2 \cdot (x-1)^3 + \frac{1}{24} \cdot (-6) \cdot (x-1)^4 + \frac{1}{120} \cdot 24 \cdot (x-1)^5$$
$$= (x-1) - \frac{1}{2}(x-1)^2 + \frac{1}{3}(x-1)^3 - \frac{1}{4}(x-1)^4 + \frac{1}{5}(x-1)^5$$

Again, it is not too hard to generalise the above work to find the Taylor polynomial of order n: With a little work one can show that

$$T_n(x) = \sum_{k=1}^{n} \frac{(-1)^{k+1}}{k}(x-1)^k.$$

Example 3.4.13

For cosine:

Example 3.4.14

Find the 4th order Maclaurin polynomial for $\cos x$.

Solution. We have $a = 0$ and we need to find the first 4 derivatives of $\cos x$.

$$f(x) = \cos x \qquad\qquad f(0) = 1$$
$$f'(x) = -\sin x \qquad\qquad f'(0) = 0$$
$$f''(x) = -\cos x \qquad\qquad f''(0) = -1$$
$$f'''(x) = \sin x \qquad\qquad f'''(0) = 0$$
$$f^{(4)}(x) = \cos x \qquad\qquad f^{(4)}(0) = 1$$

Substituting this into equation (3.4.10) gives

$$T_4(x) = 1 + 1 \cdot (0) \cdot x + \frac{1}{2} \cdot (-1) \cdot x^2 + \frac{1}{6} \cdot 0 \cdot x^3 + \frac{1}{24} \cdot (1) \cdot x^4$$
$$= 1 - \frac{x^2}{2} + \frac{x^4}{24}$$

Notice that since the 4th derivative of $\cos x$ is $\cos x$ again, we also have that the fifth derivative is the same as the first derivative, and the sixth derivative is the same as the second derivative and so on. Hence the next four derivatives are

$$f^{(4)}(x) = \cos x \qquad\qquad f^{(4)}(0) = 1$$
$$f^{(5)}(x) = -\sin x \qquad\qquad f^{(5)}(0) = 0$$
$$f^{(6)}(x) = -\cos x \qquad\qquad f^{(6)}(0) = -1$$
$$f^{(7)}(x) = \sin x \qquad\qquad f^{(7)}(0) = 0$$
$$f^{(8)}(x) = \cos x \qquad\qquad f^{(8)}(0) = 1$$

Using this we can find the 8th order Maclaurin polynomial:

$$T_8(x) = 1 - \frac{x^2}{2} + \frac{x^4}{24} - \frac{x^6}{6!} + \frac{x^8}{8!}$$

Continuing this process gives us the $2n^{\text{th}}$ Maclaurin polynomial

$$T_{2n}(x) = \sum_{k=0}^{n} \frac{(-1)^k}{(2k)!} \cdot x^{2k}$$

Warning 3.4.15.

The above formula only works when x is measured in radians, because all of our derivative formulae for trig functions were developed under the assumption that angles are measured in radians.

Below we plot $\cos x$ against its first few Maclaurin polynomial approximations:

$\cos x \approx 1$

$\cos x \approx 1 - \frac{1}{2!}x^2$

$$\cos x \approx 1 - \tfrac{1}{2!}x^2 + \tfrac{1}{4!}x^4 \qquad\qquad \cos x \approx 1 - \tfrac{1}{2!}x^2 + \tfrac{1}{4!}x^4 - \tfrac{1}{6!}x^6$$
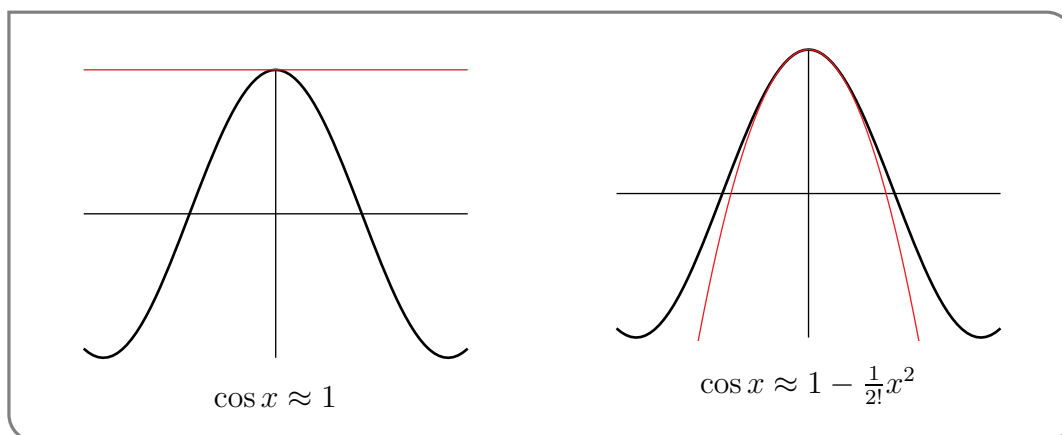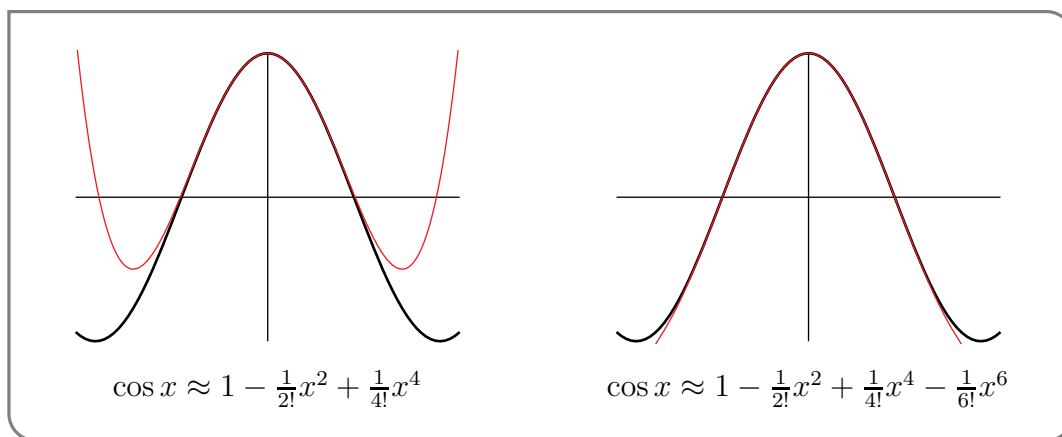
Example 3.4.15

The above work is quite easily recycled to get the Maclaurin polynomial for sine:

Example 3.4.16

Find the 5th order Maclaurin polynomial for $\sin x$.

Solution. We could simply work as before and compute the first five derivatives of $\sin x$. But set $g(x) = \sin x$ and notice that $g(x) = -f'(x)$, where $f(x) = \cos x$. Then we have

$$g(0) = -f'(0) = 0$$
$$g'(0) = -f''(0) = 1$$
$$g''(0) = -f'''(0) = 0$$
$$g'''(0) = -f^{(4)}(0) = -1$$
$$g^{(4)}(0) = -f^{(5)}(0) = 0$$
$$g^{(5)}(0) = -f^{(6)}(0) = 1$$

Hence the required Maclaurin polynomial is

$$T_5(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!}$$

Just as we extended to the $2n^{\text{th}}$ Maclaurin polynomial for cosine, we can also extend our work to compute the $(2n+1)^{\text{th}}$ Maclaurin polynomial for sine:

$$T_{2n+1}(x) = \sum_{k=0}^{n} \frac{(-1)^k}{(2k+1)!} \cdot x^{2k+1}$$

Warning 3.4.17.

The above formula only works when x is measured in radians, because all of our derivative formulae for trig functions were developed under the assumption that angles are measured in radians.

Below we plot $\sin x$ against its first few Maclaurin polynomial approximations.

$\sin x \approx x$

$\sin x \approx x - \frac{1}{3!}x^3$

$\sin x \approx x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5$

$\sin x \approx x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \frac{1}{7!}x^7$
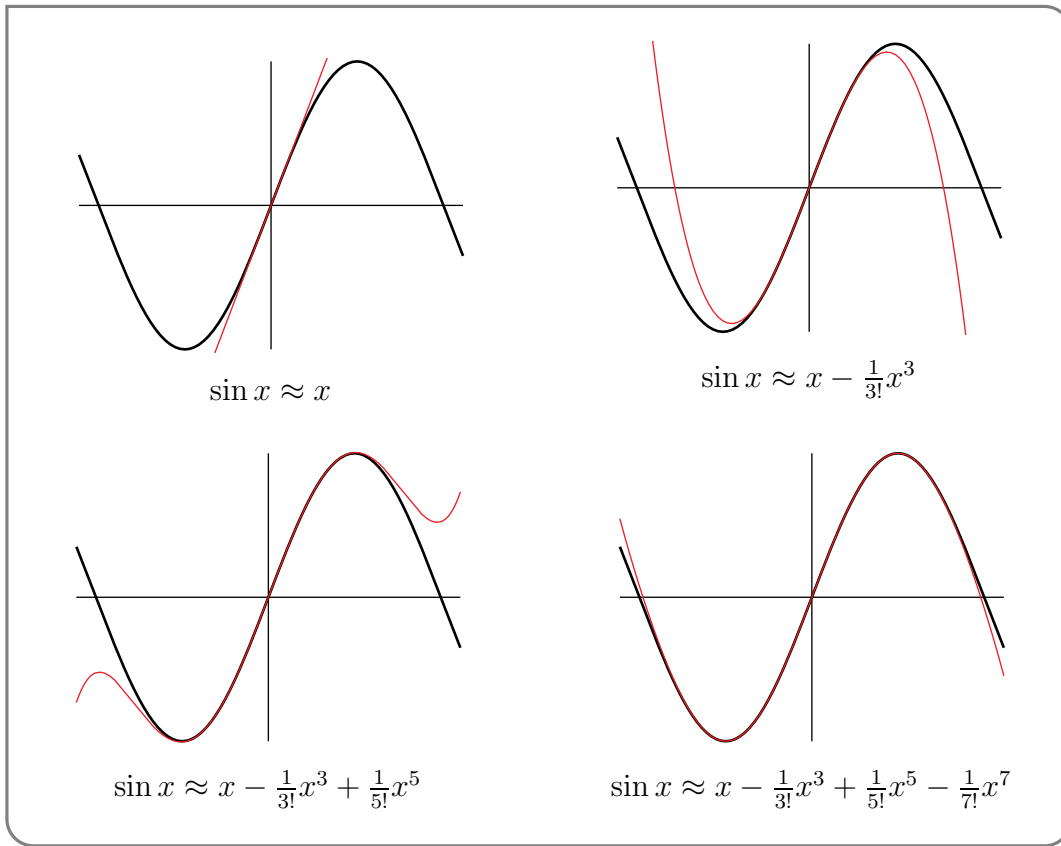
Example 3.4.17

To get an idea of how good these Taylor polynomials are at approximating sin and cos, let's concentrate on $\sin x$ and consider x's whose magnitude $|x| \leqslant 1$. There are tricks that you can employ[40] to evaluate sine and cosine at values of x outside this range.

If $|x| \leqslant 1$ radians[41], then the magnitudes of the successive terms in the Taylor polynomials for $\sin x$ are bounded by

$$|x| \leqslant 1 \qquad \frac{1}{3!}|x|^3 \leqslant \frac{1}{6} \qquad \frac{1}{5!}|x|^5 \leqslant \frac{1}{120} \approx 0.0083$$
$$\frac{1}{7!}|x|^7 \leqslant \frac{1}{7!} \approx 0.0002 \qquad \frac{1}{9!}|x|^9 \leqslant \frac{1}{9!} \approx 0.000003 \qquad \frac{1}{11!}|x|^{11} \leqslant \frac{1}{11!} \approx 0.000000025$$

From these inequalities, and the graphs on the previous pages, it certainly looks like, for x not too large, even relatively low order Taylor polynomials give very good approximations. In Section 3.4.8 we'll see how to get rigorous error bounds on our Taylor polynomial approximations.

---

40 If you are writing software to evaluate $\sin x$, you can always use the trig identity $\sin(x) = \sin(x - 2n\pi)$, to easily restrict to $|x| \leqslant \pi$. You can then use the trig identity $\sin(x) = -\sin(x \pm \pi)$ to reduce to $|x| \leqslant \frac{\pi}{2}$. Finally you can use the trig identity $\sin(x) = \mp \cos(\frac{\pi}{2} \pm x))$ to reduce to $|x| \leqslant \frac{\pi}{4} < 1$.

41 Recall that the derivative formulae that we used to derive the Taylor polynomials are valid only when x is in radians. The restriction $-1 \leqslant x \leqslant 1$ radians translates to angles bounded by $\frac{180}{\pi} \approx 57°$.

3.4.6 ‣ Estimating Change and Δx, Δy Notation

Suppose that we have two variables x and y that are related by $y = f(x)$, for some function f. One of the most important applications of calculus is to help us understand what happens to y when we make a small change in x.

> **Notation 3.4.18.**
>
> Let x, y be variables related by a function f. That is $y = f(x)$. Then we denote a small change in the variable x by Δx (read as "delta x"). The corresponding small change in the variable y is denoted Δy (read as "delta y").
>
> $$\Delta y = f(x + \Delta x) - f(x)$$

In many situations we do not need to compute Δy exactly and are instead happy with an approximation. Consider the following example.

Example 3.4.19

Let x be the number of cars manufactured per week in some factory and let y the cost of manufacturing those x cars. Given that the factory currently produces a cars per week, we would like to estimate the increase in cost if we make a small change in the number of cars produced.

Solution. We are told that a is the number of cars currently produced per week; the cost of production is then $f(a)$.

- Say the number of cars produced is changed from a to $a + \Delta x$ (where Δx is some small number.

- As x undergoes this change, the costs change from $y = f(a)$ to $f(a + \Delta x)$. Hence

$$\Delta y = f(a + \Delta x) - f(a)$$

- We can estimate this change using a linear approximation. Substituting $x = a + \Delta x$ into the equation (3.4.3) yields the approximation

$$f(a + \Delta x) \approx f(a) + f'(a)(a + \Delta x - a)$$

and consequently the approximation

$$\Delta y = f(a + \Delta x) - f(a) \approx f(a) + f'(a)\Delta x - f(a)$$

simplifies to the following neat estimate of Δy:

> **Equation 3.4.20** (Linear approximation of Δy).
>
> $$\Delta y \approx f'(a)\Delta x$$

- In the automobile manufacturing example, when the production level is a cars per week, increasing the production level by Δx will cost approximately $f'(a)\Delta x$. The additional cost per additional car, $f'(a)$, is called the "marginal cost" of a car.

- If we instead use the quadratic approximation (given by equation (3.4.6)) then we estimate

$$f(a + \Delta x) \approx f(a) + f'(a)\Delta x + \tfrac{1}{2}f''(a)\Delta x^2$$

and so

$$\Delta y = f(a + \Delta x) - f(a) \approx f(a) + f'(a)\Delta x + \tfrac{1}{2}f''(a)\Delta x^2 - f(a)$$

which simplifies to

Equation 3.4.21 (Quadratic approximation of Δy).

$$\Delta y \approx f'(a)\Delta x + \tfrac{1}{2}f''(a)\Delta x^2$$

Example 3.4.21

3.4.7 ▸▸ Further Examples

In this subsection we give further examples of computation and use of Taylor approximations.

Example 3.4.22

Estimate $\tan 46°$, using the constant-, linear- and quadratic-approximations (equations (3.4.1), (3.4.3) and (3.4.6)).

Solution. Note that we need to be careful to translate angles measured in degrees to radians.

- Set $f(x) = \tan x$, $x = 46\frac{\pi}{180}$ radians and $a = 45\frac{\pi}{180} = \frac{\pi}{4}$ radians. This is a good choice for a because

  - $a = 45°$ is close to $x = 46°$. As noted above, it is generally the case that the closer x is to a, the better various approximations will be.
  - We know the values of all trig functions at $45°$.

- Now we need to compute f and its first two derivatives at $x = a$. It is a good time to recall the special $1 : 1 : \sqrt{2}$ triangle

So
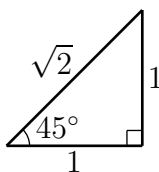
$$f(x) = \tan x \qquad\qquad f(\pi/4) = 1$$

$$f'(x) = \sec^2 x = \frac{1}{\cos^2 x} \qquad\qquad f'(\pi/4) = \frac{1}{1/\sqrt{2}^2} = 2$$

$$f''(x) = \frac{2\sin x}{\cos^3 x} \qquad\qquad f''(\pi/4) = \frac{2/\sqrt{2}}{1/\sqrt{2}^3} = 4$$

- As $x - a = 46\frac{\pi}{180} - 45\frac{\pi}{180} = \frac{\pi}{180}$ radians, the three approximations are

$$f(x) \approx f(a) \qquad\qquad\qquad\qquad\qquad = 1$$
$$f(x) \approx f(a) + f'(a)(x-a) \qquad\qquad = 1 + 2\frac{\pi}{180} \qquad = 1.034907$$
$$f(x) \approx f(a) + f'(a)(x-a) + \tfrac{1}{2}f''(a)(x-a)^2 = 1 + 2\frac{\pi}{180} + \tfrac{1}{2}4\left(\frac{\pi}{180}\right)^2 = 1.035516$$

For comparison purposes, $\tan 46°$ really is 1.035530 to 6 decimal places.

Example 3.4.22

Warning 3.4.23.

All of our derivative formulae for trig functions were developed under the assumption that angles are measured in radians. Those derivatives appeared in the approximation formulae that we used in Example 3.4.22, so we were obliged to express $x - a$ in radians.

Example 3.4.24

Suppose that you are ten meters from a vertical pole. You were contracted to measure the height of the pole. You can't take it down or climb it. So you measure the angle subtended by the top of the pole. You measure $\theta = 30°$, which gives

$$h = 10\tan 30° = \tfrac{10}{\sqrt{3}} \approx 5.77\text{m}$$

This is just standard trigonometry — if we know the angle exactly then we know the height exactly.

However, in the "real world" angles are hard to measure with such precision. If the contract requires you the measurement of the pole to be accurate within 10 cm, how accurate does your measurement of the angle θ need to be?

Solution. For simplicity[42], we are going to assume that the pole is perfectly straight and perfectly vertical and that your distance from the pole was exactly 10 m.

- Write $\theta = \theta_0 + \Delta\theta$ where θ is the exact angle, θ_0 is the measured angle and $\Delta\theta$ is the error.

---

42   Mathematicians love assumptions that let us tame the real world.

- Similarly write $h = h_0 + \Delta h$, where h is the exact height and $h_0 = \frac{10}{\sqrt{3}}$ is the computed height. Their difference, Δh, is the error.

- Then

$$h_0 = 10 \tan \theta_0 \qquad\qquad\qquad h_0 + \Delta h = 10 \tan(\theta_0 + \Delta \theta)$$
$$\Delta h = 10 \tan(\theta_0 + \Delta \theta) - 10 \tan \theta_0$$

We could attempt to solve this equation for $\Delta \theta$ in terms of Δh — but it is far simpler to approximate Δh using the linear approximation in equation 3.4.20.

- To use equation 3.4.20, replace y with h, x with θ and a with θ_0. Our function $f(\theta) = 10 \tan \theta$ and $\theta_0 = 30° = \pi/6$ radians. Then

$$\Delta y \approx f'(a)\Delta x \qquad\qquad \text{becomes} \qquad\qquad \Delta h \approx f'(\theta_0)\Delta\theta$$

Since $f(\theta) = 10 \tan \theta$, $f'(\theta) = 10 \sec^2 \theta$ and

$$f'(\theta_0) = 10 \sec^2(\pi/6) = 10 \cdot \left(\frac{2}{\sqrt{3}}\right)^2 = \frac{40}{3}$$

- Putting things together gives

$$\Delta h \approx f'(\theta_0)\Delta\theta \qquad\qquad \text{becomes} \qquad\qquad \Delta h \approx \frac{40}{3}\Delta\theta$$

We can then solve this equation for $\Delta\theta$ in terms of Δh:

$$\Delta\theta \approx \frac{3}{40}\Delta h$$

- We are told that we must have $|\Delta h| < 0.1$, so we must have

$$|\Delta\theta| \leqslant \frac{3}{400}$$

This is measured in radians, so converting back to degrees

$$\frac{3}{400} \cdot \frac{180}{\pi} = 0.43°$$

Example 3.4.24

Definition 3.4.25.

Suppose that you measure, approximately, some quantity. Suppose that the exact value of that quantity is Q_0 and that your measurement yielded $Q_0 + \Delta Q$. Then $|\Delta Q|$ is called the absolute error of the measurement and $100\frac{|\Delta Q|}{Q_0}$ is called the percentage error of the measurement. As an example, if the exact value is 4 and the measured value is 5, then the absolute error is $|5 - 4| = 1$ and the percentage error is $100\frac{|5-4|}{4} = 25$. That is, the error, 1, was 25% of the exact value, 4.

Example 3.4.26

Suppose that the radius of a sphere has been measured with a percentage error of at most $\varepsilon\%$. Find the corresponding approximate percentage errors in the surface area and volume of the sphere.

Solution. We need to be careful in this problem to convert between absolute and percentage errors correctly.

- Suppose that the exact radius is r_0 and that the measured radius is $r_0 + \Delta r$.

- Then the absolute error in the measurement is $|\Delta r|$ and, by definition, the percentage error is $100\frac{|\Delta r|}{r_0}$. We are told that $100\frac{|\Delta r|}{r_0} \leqslant \varepsilon$.

- The surface area[43] of a sphere of radius r is $A(r) = 4\pi r^2$. The error in the surface area computed with the measured radius is

$$\Delta A = A(r_0 + \Delta r) - A(r_0) \approx A'(r_0)\Delta r$$
$$= 8\pi r_0 \Delta r$$

where we have made use of the linear approximation, equation (3.4.20).

- The corresponding percentage error is then

$$100\frac{|\Delta A|}{A(r_0)} \approx 100\frac{|A'(r_0)\Delta r|}{A(r_0)} = 100\frac{8\pi r_0|\Delta r|}{4\pi r_0^2} = 2 \times 100\frac{|\Delta r|}{r_0} \leqslant 2\varepsilon$$

- The volume of a sphere[44] of radius r is $V(r) = \frac{4}{3}\pi r^3$. The error in the volume computed with the measured radius is

$$\Delta V = V(r_0 + \Delta r) - V(r_0) \approx V'(r_0)\Delta r$$
$$= 4\pi r_0^2 \Delta r$$

where we have again made use of the linear approximation, equation (3.4.20).

---

43 We do not expect you to remember the surface areas of solids for this course.
44 We do expect you to remember the formula for the volume of a sphere.

- The corresponding percentage error is

$$100\frac{|\Delta V|}{V(r_0)} \approx 100\frac{|V'(r_0)\Delta r|}{V(r_0)} = 100\frac{4\pi r_0^2|\Delta r|}{4\pi r_0^3/3} = 3 \times 100\frac{|\Delta r|}{r_0} \leqslant 3\varepsilon$$

We have just computed an approximation to ΔV. This problem is actually sufficiently simple that we can compute ΔV exactly:

$$\Delta V = V(r_0 + \Delta r) - V(r_0) = \tfrac{4}{3}\pi(r_0 + \Delta r)^3 - \tfrac{4}{3}\pi r_0^3$$

- Applying $(a + b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$ with $a = r_0$ and $b = \Delta r$, gives

$$V(r_0 + \Delta r) - V(r_0) = \tfrac{4}{3}\pi \left[r_0^3 + 3r_0^2\Delta r + 3r_0\,(\Delta r)^2 + (\Delta r)^3 \right] - \tfrac{4}{3}\pi r_0^3$$
$$= \tfrac{4}{3}\pi[3r_0^2\Delta r + 3r_0\,(\Delta r)^2 + (\Delta r)^3]$$

- Thus the difference between the exact error and the linear approximation to the error is obtained by retaining only the last two terms in the square brackets. This has magnitude

$$\tfrac{4}{3}\pi\left|3r_0\,(\Delta r)^2 + (\Delta r)^3\right| = \tfrac{4}{3}\pi\left|3r_0 + \Delta r\right|(\Delta r)^2$$

or in percentage terms

$$100 \cdot \frac{1}{\tfrac{4}{3}\pi r_0^3} \cdot \tfrac{4}{3}\pi\left|3r_0\,(\Delta r)^2 + (\Delta r)^3\right| = 100 \left|3\frac{\Delta r^2}{r_0^2} + \frac{\Delta r^3}{r_0^3}\right|$$
$$= \left(100\frac{3\Delta r}{r_0}\right) \cdot \left(\frac{\Delta r}{r_0}\right)\left|1 + \frac{\Delta r}{3r_0}\right|$$
$$\leqslant 3\varepsilon\left(\frac{\varepsilon}{100}\right) \cdot \left(1 + \frac{\varepsilon}{300}\right)$$

Since ε is small, we can assume that $1 + \frac{\varepsilon}{300} \approx 1$. Hence the difference between the exact error and the linear approximation of the error is roughly a factor of $\frac{\varepsilon}{100}$ smaller than the linear approximation 3ε.

- As an aside, notice that if we argue that Δr is very small and so we can ignore terms involving $(\Delta r)^2$ and $(\Delta r)^3$ as being really really small, then we obtain

$$V(r_0 + \Delta r) - V(r_0) = \tfrac{4}{3}\pi[3r_0^2\Delta r \underbrace{+3r_0\,(\Delta r)^2 + (\Delta r)^3}_{\text{really really small}}]$$
$$\approx \tfrac{4}{3}\pi \cdot 3r_0^2\Delta r = 4\pi r_0^2\Delta r$$

which is precisely the result of our linear approximation above.

Example 3.4.26

Example 3.4.27

To compute the height h of a lamp post, the length s of the shadow of a two meter pole is measured. The pole is 6 m from the lamp post. If the length of the shadow was measured to be 4 m, with an error of at most one cm, find the height of the lamp post and estimate the percentage error in the height.

Solution. We should first draw a picture[45]
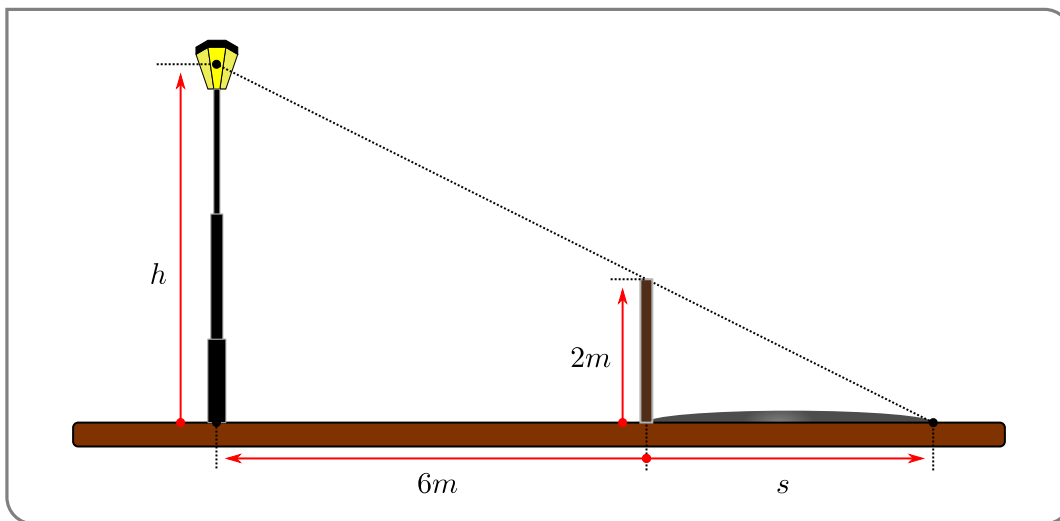
- By similar triangles we see that

$$\frac{2}{s} = \frac{h}{6+s}$$

from which we can isolate h as a function of s:

$$h = \frac{2(6+s)}{s} = \frac{12}{s} + 2$$

- The length of the shadow was measured to be $s_0 = 4$ m. The corresponding height of the lamp post is

$$h_0 = \frac{12}{4} + 2 = 5m$$

- If the error in the measurement of the length of the shadow was Δs, then the exact shadow length was $s = s_0 + \Delta s$ and the exact lamp post height is $h = f(s_0 + \Delta s)$, where $f(s) = \frac{12}{s} + 2$. The error in the computed lamp post height is

$$\Delta h = h - h_0 = f(s_0 + \Delta s) - f(s_0)$$

- We can then make a linear approximation of this error using equation (3.4.20):

$$\Delta h \approx f'(s_0)\Delta s = -\frac{12}{s_0^2}\Delta s = -\frac{12}{4^2}\Delta s$$

45   We get to reuse that nice lamp post picture from Example 3.2.4.

- We are told that $|\Delta s| \leqslant \frac{1}{100}$ m. Consequently, approximately,

$$|\Delta h| \leqslant \frac{12}{4^2} \frac{1}{100} = \frac{3}{400}$$

The percentage error is then approximately

$$100 \frac{|\Delta h|}{h_0} \leqslant 100 \frac{3}{400 \times 5} = 0.15\%$$

Example 3.4.27

3.4.8 ▸▸ The Error in the Taylor Polynomial Approximations

Any time you make an approximation, it is desirable to have some idea of the size of the error you introduced. That is, we would like to know the difference $R(x)$ between the original function $f(x)$ and our approximation $F(x)$:

$$R(x) = f(x) - F(x).$$

Of course if we know $R(x)$ exactly, then we could recover $f(x) = F(x) + R(x)$ — so this is an unrealistic hope. In practice we would simply like to bound $R(x)$:

$$|R(x)| = |f(x) - F(x)| \leqslant M$$

where (hopefully) M is some small number. It is worth stressing that we do not need the tightest possible value of M, we just need a relatively easily computed M that isn't too far off the true value of $|f(x) - F(x)|$.

We will now develop a formula for the error introduced by the constant approximation, equation (3.4.1) (developed back in Section 3.4.1)

$$f(x) \approx f(a) = T_0(x) \qquad\qquad 0^{\text{th}} \text{ Taylor polynomial}$$

The resulting formula can be used to get an upper bound on the size of the error $|R(x)|$.

The main ingredient we will need is the Mean-Value Theorem (Theorem 2.13.4) — so we suggest you quickly revise it. Consider the following obvious statement:

$$
\begin{aligned}
f(x) &= f(x) & \text{now some sneaky manipulations}\\
&= f(a) + (f(x) - f(a))\\
&= \underbrace{f(a)}_{=T_0(x)} + (f(x) - f(a)) \cdot \underbrace{\frac{x-a}{x-a}}_{=1}\\
&= T_0(x) + \underbrace{\frac{f(x) - f(a)}{x-a}}_{\text{looks familiar}} \cdot (x-a)
\end{aligned}
$$

Indeed, this equation is important in the discussion that follows, so we'll highlight it

Equation 3.4.28 (We will need it again soon).

$$f(x) = T_0(x) + \left[\frac{f(x) - f(a)}{x - a} \right] (x - a)$$

The coefficient $\dfrac{f(x) - f(a)}{x - a}$ of $(x - a)$ is the average slope of $f(t)$ as t moves from $t = a$ to $t = x$. We can picture this as the slope of the secant joining the points $(a, f(a))$ and $(x, f(x))$ in the sketch below.
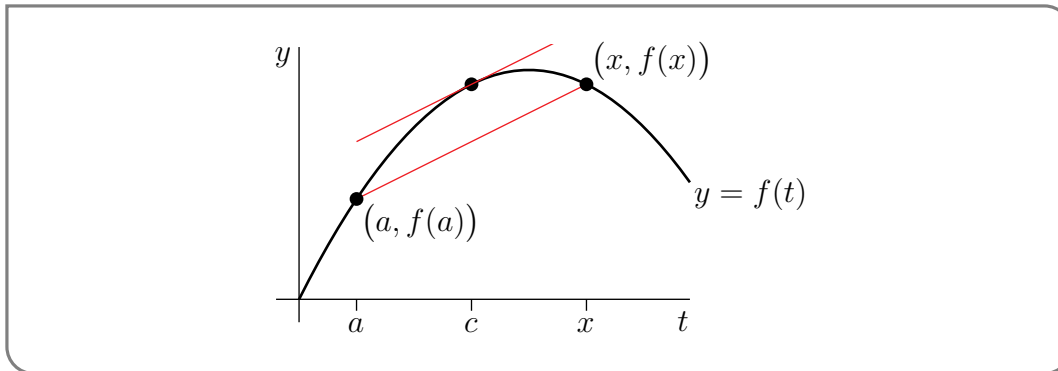
As t moves from a to x, the instantaneous slope $f'(t)$ keeps changing. Sometimes $f'(t)$ might be larger than the average slope $\frac{f(x)-f(a)}{x-a}$, and sometimes $f'(t)$ might be smaller than the average slope $\frac{f(x)-f(a)}{x-a}$. However, by the Mean-Value Theorem (Theorem 2.13.4), there must be some number c, strictly between a and x, for which $f'(c) = \dfrac{f(x) - f(a)}{x - a}$ exactly.

Substituting this into formula (3.4.28) gives

Equation 3.4.29 (Towards the error).

$$f(x) = T_0(x) + f'(c)(x - a) \qquad \text{for some } c \text{ strictly between } a \text{ and } x$$

Notice that this expression as it stands is not quite what we want. Let us massage this around a little more into a more useful form

Equation 3.4.30 (The error in constant approximation).

$$f(x) - T_0(x) = f'(c) \cdot (x - a) \qquad \text{for some } c \text{ strictly between } a \text{ and } x$$

Notice that the MVT doesn't tell us the value of c, however we do know that it lies strictly between x and a. So if we can get a good bound on $f'(c)$ on this interval then we can get a good bound on the error.

Example 3.4.31

Let us return to Example 3.4.2, and we'll try to bound the error in our approximation of $e^{0.1}$.

- Recall that $f(x) = e^x$, $a = 0$ and $T_0(x) = e^0 = 1$.

- Then by equation (3.4.30)

$$e^{0.1} - T_0(0.1) = f'(c) \cdot (0.1 - 0) \qquad \text{with } 0 < c < 0.1$$

- Now $f'(c) = e^c$, so we need to bound e^c on $(0, 0.1)$. Since e^c is an increasing function, we know that

$$e^0 < f'(c) < e^{0.1} \qquad \text{when } 0 < c < 0.1$$

So one is tempted to write that

$$|e^{0.1} - T_0(0.1)| = |R(x)| = |f'(c)| \cdot (0.1 - 0)$$
$$< e^{0.1} \cdot 0.1$$

And while this is true, it is rather circular. We have just bounded the error in our approximation of $e^{0.1}$ by $\frac{1}{10}e^{0.1}$ — if we actually knew $e^{0.1}$ then we wouldn't need to estimate it!

- While we don't know $e^{0.1}$ exactly, we do know[46] that $1 = e^0 < e^{0.1} < e^1 < 3$. This gives us

$$|R(0.1)| < 3 \times 0.1 = 0.3$$

That is — the error in our approximation of $e^{0.1}$ is no greater than 0.3. Recall that we don't need the error exactly, we just need a good idea of how large it actually is.

- In fact the real error here is

$$|e^{0.1} - T_0(0.1)| = |e^{0.1} - 1| = 0.1051709\ldots$$

so we have over-estimated the error by a factor of 3.

But we can actually go a little further here — we can bound the error above and below. If we do not take absolute values, then since

$$e^{0.1} - T_0(0.1) = f'(c) \cdot 0.1 \qquad \text{and } 1 < f'(c) < 3$$

we can write

$$1 \times 0.1 \leqslant (e^{0.1} - T_0(0.1)) \leqslant 3 \times 0.1$$

so

$$T_0(0.1) + 0.1 \leqslant e^{0.1} \leqslant T_0(0.1) + 0.3$$
$$1.1 \leqslant e^{0.1} \leqslant 1.3$$

So while the upper bound is weak, the lower bound is quite tight.

---

46   Oops! Do we really know that $e < 3$? We haven't proved it. We will do so soon.

Example 3.4.31

There are formulae similar to equation (3.4.29), that can be used to bound the error in our other approximations; all are based on generalisations of the MVT. The next one — for linear approximations — is

$$f(x) = \underbrace{f(a) + f'(a)(x - a)}_{=T_1(x)} + \tfrac{1}{2}f''(c)(x - a)^2 \qquad \text{for some } c \text{ strictly between } a \text{ and } x$$

which we can rewrite in terms of $T_1(x)$:

Equation 3.4.32 (The error in linear approximation).

$$f(x) - T_1(x) = \tfrac{1}{2}f''(c)(x - a)^2 \qquad \text{for some } c \text{ strictly between } a \text{ and } x$$

It implies that the error that we make when we approximate $f(x)$ by $T_1(x) = f(a) + f'(a)(x - a)$ is exactly $\tfrac{1}{2}f''(c)(x - a)^2$ for some c strictly between a and x.

More generally

$$f(x) = \underbrace{f(a) + f'(a) \cdot (x - a) + \cdots + \frac{1}{n!}f^{(n)}(a) \cdot (x - a)^n}_{=T_n(x)} + \frac{1}{(n + 1)!}f^{(n+1)}(c) \cdot (x - a)^{n+1}$$

for some c strictly between a and x. Again, rewriting this in terms of $T_n(x)$ gives

Equation 3.4.33.

$$f(x) - T_n(x) = \frac{1}{(n + 1)!}f^{(n+1)}(c) \cdot (x - a)^{n+1} \quad \text{for some } c \text{ strictly between } a \text{ and } x$$

That is, the error introduced when $f(x)$ is approximated by its Taylor polynomial of order n, is precisely the last term of the Taylor polynomial of order $n + 1$, but with the derivative evaluated at some point between a and x, rather than exactly at a. These error formulae are proven in the optional Section 3.4.9 later in this chapter.

Example 3.4.34

Approximate $\sin 46°$ using Taylor polynomials about $a = 45°$, and estimate the resulting error.

Solution.

- Start by defining $f(x) = \sin x$ and

$$a = 45° = 45\tfrac{\pi}{180}\text{radians} \qquad x = 46° = 46\tfrac{\pi}{180}\text{radians} \qquad x - a = \tfrac{\pi}{180}\text{radians}$$

- The first few derivatives of f at a are

$$f(x) = \sin x \qquad f(a) = \frac{1}{\sqrt{2}}$$

$$f'(x) = \cos x \qquad f'(a) = \frac{1}{\sqrt{2}}$$

$$f''(x) = -\sin x \quad f''(a) = -\frac{1}{\sqrt{2}}$$

$$f^{(3)}(x) = -\cos x \, f^{(3)}(a) = -\frac{1}{\sqrt{2}}$$

- The constant, linear and quadratic Taylor approximations for $\sin(x)$ about $\frac{\pi}{4}$ are

$$T_0(x) = f(a) \qquad\qquad = \frac{1}{\sqrt{2}}$$

$$T_1(x) = T_0(x) + f'(a) \cdot (x - a) \quad = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}\left(x - \frac{\pi}{4}\right)$$

$$T_2(x) = T_1(x) + \tfrac{1}{2}f''(a) \cdot (x - a)^2 = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}\left(x - \frac{\pi}{4}\right) - \frac{1}{2\sqrt{2}}\left(x - \frac{\pi}{4}\right)^2$$

- So the approximations for $\sin 46°$ are

$$\sin 46° \approx T_0\left(\frac{46\pi}{180}\right) = \frac{1}{\sqrt{2}} \qquad\qquad\qquad = 0.70710678$$

$$\sin 46° \approx T_1\left(\frac{46\pi}{180}\right) = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}\left(\frac{\pi}{180}\right) \qquad = 0.71944812$$

$$\sin 46° \approx T_2\left(\frac{46\pi}{180}\right) = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}\left(\frac{\pi}{180}\right) - \frac{1}{2\sqrt{2}}\left(\frac{\pi}{180}\right)^2 \quad = 0.71934042$$

- The errors in those approximations are (respectively)

$$\text{error in } 0.70710678 = f'(c)(x - a) \qquad = \cos c \cdot \left(\frac{\pi}{180}\right)$$

$$\text{error in } 0.71944812 = \frac{1}{2}f''(c)(x - a)^2 \quad = -\frac{1}{2} \cdot \sin c \cdot \left(\frac{\pi}{180}\right)^2$$

$$\text{error in } 0.71923272 = \frac{1}{3!}f^{(3)}(c)(x - a)^3 = -\frac{1}{3!} \cdot \cos c \cdot \left(\frac{\pi}{180}\right)^3$$

In each of these three cases c must lie somewhere between $45°$ and $46°$.

- Rather than carefully estimating $\sin c$ and $\cos c$ for c in that range, we make use of a simpler (but much easier bound). No matter what c is, we know that $|\sin c| \leqslant 1$ and $|\cos c| \leqslant 1$. Hence

$$\left|\text{error in } 0.70710678\right| \leqslant \left(\frac{\pi}{180}\right) \qquad < 0.018$$

$$\left|\text{error in } 0.71944812\right| \leqslant \frac{1}{2}\left(\frac{\pi}{180}\right)^2 < 0.00015$$

$$\left|\text{error in } 0.71934042\right| \leqslant \frac{1}{3!}\left(\frac{\pi}{180}\right)^3 < 0.0000009$$

Example 3.4.34

Example 3.4.35 (Showing $e < 3$)

In Example 3.4.31 above we used the fact that $e < 3$ without actually proving it. Let's do so now.

- Consider the linear approximation of e^x about $a = 0$.

$$T_1(x) = f(0) + f'(0) \cdot x = 1 + x$$

So at $x = 1$ we have

$$e \approx T_1(1) = 2$$

- The error in this approximation is

$$e^x - T_1(x) = \frac{1}{2}f''(c) \cdot x^2 = \frac{e^c}{2} \cdot x^2$$

So at $x = 1$ we have

$$e - T_1(1) = \frac{e^c}{2}$$

where $0 < c < 1$.

- Now since e^x is an increasing[47] function, it follows that $e^c < e$. Hence

$$e - T_1(1) = \frac{e^c}{2} < \frac{e}{2}$$

Moving the $\frac{e}{2}$ to the left hand side and the $T_1(1)$ to the right hand side gives

$$\frac{e}{2} < T_1(1) = 2$$

So $e < 4$.

- This isn't as tight as we would like — so now do the same with the quadratic approximation with $a = 0$:

$$e^x \approx T_2(x) = 1 + x + \frac{x^2}{2}$$

So when $x = 1$ we have

$$e \approx T_2(1) = 1 + 1 + \frac{1}{2} = \frac{5}{2}$$

---

47 Since the derivative of e^x is e^x which is positive everywhere, the function is increasing everywhere.

- The error in this approximation is

$$e^x - T_2(x) = \frac{1}{3!}f'''(c) \cdot x^3 = \frac{e^c}{6} \cdot x^3$$

So at $x = 1$ we have

$$e - T_2(1) = \frac{e^c}{6}$$

where $0 < c < 1$.

- Again since e^x is an increasing function we have $e^c < e$. Hence

$$e - T_2(1) = \frac{e^c}{6} < \frac{e}{6}$$

That is

$$\frac{5e}{6} < T_2(1) = \frac{5}{2}$$

So $e < 3$ as required.

Example 3.4.35

Example 3.4.36 (More on e^x)

We wrote down the general n^{th} order Maclaurin polynomial approximation of e^x in Example 3.4.12 above.

- Recall that

$$T_n(x) = \sum_{k=0}^{n} \frac{1}{k!}x^k$$

- The error in this approximation is (by equation (3.4.33))

$$e^x - T_n(x) = \frac{1}{(n+1)!}e^c$$

where c is some number between 0 and x.

- So setting $x = 1$ in this gives

$$e - T_n(1) = \frac{1}{(n+1)!}e^c$$

where $0 < c < 1$.

- Since e^x is an increasing function we know that $1 = e^0 < e^c < e^1 < 3$, so the above expression becomes

$$\frac{1}{(n+1)!} \leqslant e - T_n(1) = \frac{1}{(n+1)!}e^c \leqslant \frac{3}{(n+1)!}$$

- So when $n = 9$ we have

$$\frac{1}{10!} \leqslant e - \left(1 + 1 + \frac{1}{2} + \cdots + \frac{1}{9!}\right) \leqslant \frac{3}{10!}$$

- Now $1/10! < 3/10! < 10^{-6}$, so the approximation of e by

$$e \approx 1 + 1 + \frac{1}{2} + \cdots + \frac{1}{9!} = \frac{98641}{36288} = 2.718281\ldots$$

is correct to 6 decimal places.

- More generally we know that using $T_n(1)$ to approximate e will have an error of at most $\frac{3}{(n+1)!}$ — so it converges very quickly.

Example 3.4.36

Example 3.4.37 (Example 3.4.24 Revisited)

Recall[48] that in Example 3.4.24 (measuring the height of the pole), we used the linear approximation

$$f(\theta_0 + \Delta\theta) \approx f(\theta_0) + f'(\theta_0)\Delta\theta$$

with $f(\theta) = 10 \tan\theta$ and $\theta_0 = 30\frac{\pi}{180}$ to get

$$\Delta h = f(\theta_0 + \Delta\theta) - f(\theta_0) \approx f'(\theta_0)\Delta\theta \qquad \text{which implies that} \qquad \Delta\theta \approx \frac{\Delta h}{f'(\theta_0)}$$

- While this procedure is fairly reliable, it did involve an approximation. So that you could not 100% guarantee to your client's lawyer that an accuracy of 10 cm was achieved.

- On the other hand, if we use the *exact* formula (3.4.29), with the replacements $x \to \theta_0 + \Delta\theta$ and $a \to \theta_0$

$$f(\theta_0 + \Delta\theta) = f(\theta_0) + f'(c)\Delta\theta \qquad \text{for some } c \text{ between } \theta_0 \text{ and } \theta_0 + \Delta\theta$$

in place of the approximate formula (3.4.3), this legality is taken care of:

$$\Delta h = f(\theta_0 + \Delta\theta) - f(\theta_0) = f'(c)\Delta\theta \qquad \text{for some } c \text{ between } \theta_0 \text{ and } \theta_0 + \Delta\theta$$

We can clean this up a little more since in our example $f'(\theta) = 10 \sec^2\theta$. Thus for some c between θ_0 and $\theta_0 + \Delta\theta$:

$$|\Delta h| = 10 \sec^2(c)|\Delta\theta|$$

---

48   Now is a good time to go back and re-read it.

- Of course we do not know exactly what c is. But suppose that we know that the angle was somewhere between $25°$ and $35°$. In other words suppose that, even though we don't know precisely what our measurement error was, it was certainly no more than $5°$.

- Now on the range $25° < c < 35°$, $\sec(c)$ is an increasing and positive function. Hence on this range

$$1.217\cdots = \sec^2 25° \leqslant \sec^2 c \leqslant \sec^2 35° = 1.490\cdots < 1.491$$

So

$$12.17 \cdot |\Delta\theta| \leqslant |\Delta h| = 10\sec^2(c) \cdot |\Delta\theta| \leqslant 14.91 \cdot |\Delta\theta|$$

- Since we require $|\Delta h| < 0.1$, we need $14.91|\Delta\theta| < 0.1$, that is

$$|\Delta\theta| < \frac{0.1}{14.91} = 0.0067\ldots$$

So we must measure angles with an accuracy of no less than 0.0067 radians — which is

$$\frac{180}{\pi} \cdot 0.0067 = 0.38°.$$

Hence a measurement error of $0.38°$ or less is acceptable.

Example 3.4.37

3.4.9 ➤ (Optional) — Derivation of the Error Formulae

In this section we will derive the formula for the error that we gave in equation (3.4.33) — namely

$$R_n(x) = f(x) - T_n(x) = \frac{1}{(n+1)!}f^{(n+1)}(c) \cdot (x-a)^{n+1}$$

for some c strictly between a and x, and where $T_n(x)$ is the nth order Taylor polynomial approximation of $f(x)$ about $x = a$:

$$T_n(x) = \sum_{k=0}^{n} \frac{1}{k!}f^{(k)}(a).$$

Recall that we have already proved a special case of this formula for the constant approximation using the Mean-Value Theorem (Theorem 2.13.4). To prove the general case we need the following generalisation[49] of that theorem:

---

[49] It is not a terribly creative name for the generalisation, but it is an accurate one.

> **Theorem 3.4.38** (Generalised Mean-Value Theorem).
>
> Let the functions $F(x)$ and $G(x)$ both be defined and continuous on $a \leqslant x \leqslant b$ and both be differentiable on $a < x < b$. Furthermore, suppose that $G'(x) \neq 0$ for all $a < x < b$. Then, there is a number c obeying $a < c < b$ such that
>
> $$\frac{F(b) - F(a)}{G(b) - G(a)} = \frac{F'(c)}{G'(c)}$$

Notice that setting $G(x) = x$ recovers the original Mean-Value Theorem. It turns out that this theorem is not too difficult to prove from the MVT using some sneaky algebraic manipulations:

Proof.   • First we construct a new function $h(x)$ as a linear combination of $F(x)$ and $G(x)$ so that $h(a) = h(b) = 0$. Some experimentation yields

$$h(x) = \big[F(b) - F(a)\big] \cdot \big[G(x) - G(a)\big] - \big[G(b) - G(a)\big] \cdot \big[F(x) - F(a)\big]$$

• Since $h(a) = h(b) = 0$, the Mean-Value theorem (actually Rolle's theorem) tells us that there is a number c obeying $a < c < b$ such that $h'(c) = 0$:

$$h'(x) = \big[F(b) - F(a)\big] \cdot G'(x) - \big[G(b) - G(a)\big] \cdot F'(x) \qquad \text{so}$$
$$0 = \big[F(b) - F(a)\big] \cdot G'(c) - \big[G(b) - G(a)\big] \cdot F'(c)$$

Now move the $G'(c)$ terms to one side and the $F'(c)$ terms to the other:

$$\big[F(b) - F(a)\big] \cdot G'(c) = \big[G(b) - G(a)\big] \cdot F'(c).$$

• Since we have $G'(x) \neq 0$, we know that $G'(c) \neq 0$. Further the Mean-Value theorem ensures[50] that $G(a) \neq G(b)$. Hence we can move terms about to get

$$\big[F(b) - F(a)\big] = \big[G(b) - G(a)\big] \cdot \frac{F'(c)}{G'(c)}$$
$$\frac{F(b) - F(a)}{G(b) - G(a)} = \frac{F'(c)}{G'(c)}$$

as required.

$\square$

Armed with the above theorem we can now move on to the proof of the Taylor remainder formula.

Proof of equation (3.4.33). We begin by proving the remainder formula for $n = 1$. That is

$$f(x) - T_1(x) = \frac{1}{2}f''(c) \cdot (x - a)^2$$

---

50   Otherwise if $G(a) = G(b)$ the MVT tells us that there is some point c between a and b so that $G'(c) = 0$.

- Start by setting

$$F(x) = f(x) - T_1(x) \qquad\qquad G(x) = (x - a)^2$$

Notice that, since $T_1(a) = f(a)$ and $T_1'(x) = f'(a)$,

$$F(a) = 0 \qquad\qquad G(a) = 0$$
$$F'(x) = f'(x) - f'(a) \qquad\qquad G'(x) = 2(x - a)$$

- Now apply the generalised MVT with $b = x$: there exists a point q between a and x such that

$$\frac{F(x) - F(a)}{G(x) - G(a)} = \frac{F'(q)}{G'(q)}$$
$$\frac{F(x) - 0}{G(x) - 0} = \frac{f'(q) - f'(a)}{2(q - a)}$$
$$2 \cdot \frac{F(x)}{G(x)} = \frac{f'(q) - f'(a)}{q - a}$$

- Consider the right-hand side of the above equation and set $g(x) = f'(x)$. Then we have the term $\frac{g(q) - g(a)}{q - a}$ — this is exactly the form needed to apply the MVT. So now apply the standard MVT to the right-hand side of the above equation — there is some c between q and a so that

$$\frac{f'(q) - f'(a)}{q - a} = \frac{g(q) - g(a)}{q - a} = g'(c) = f''(c)$$

Notice that here we have assumed that $f''(x)$ exists.

- Putting this together we have that

$$2 \cdot \frac{F(x)}{G(x)} = \frac{f'(q) - f'(a)}{q - a} = f''(c)$$
$$2\frac{f(x) - T_1(x)}{(x - a)^2} = f''(c)$$
$$f(x) - T_1(x) = \frac{1}{2!} f''(c) \cdot (x - a)^2$$

as required.

Oof! We have now proved the cases $n = 1$ (and we did $n = 0$ earlier).

To proceed — assume we have proved our result for $n = 1, 2, \ldots, k$. We realise that we haven't done this yet, but bear with us. Using that assumption we will prove the result is true for $n = k + 1$. Once we have done that, then

- we have proved the result is true for $n = 1$, and

- we have shown if the result is true for $n = k$ then it is true for $n = k + 1$

Hence it must be true for all $n \geqslant 1$. This style of proof is called mathematical induction. You can think of the process as something like climbing a ladder:

- prove that you can get onto the ladder (the result is true for $n = 1$), and

- if I can stand on the current rung, then I can step up to the next rung (if the result is true for $n = k$ then it is also true for $n = k + 1$)

Hence I can climb as high as like.

- Let $k > 0$ and assume we have proved

$$f(x) - T_k(x) = \frac{1}{(k+1)!} f^{(k+1)}(c) \cdot (x - a)^{k+1}$$

  for some c between a and x.

- Now set

$$F(x) = f(x) - T_{k+1}(x) \qquad\qquad G(x) = (x - a)^{k+1}$$

  and notice that, since $T_{k+1}(a) = f(a)$,

$$F(a) = f(a) - T_{k+1}(a) = 0 \qquad G(a) = 0 \qquad\qquad G'(x) = (k+1)(x-a)^k$$

  and apply the generalised MVT with $b = x$: hence there exists a q between a and x so that

$$\frac{F(x) - F(a)}{G(x) - G(a)} = \frac{F'(q)}{G'(q)} \qquad\qquad \text{which becomes}$$

$$\frac{F(x)}{(x-a)^{k+1}} = \frac{F'(q)}{(k+1)(q-a)^k} \qquad\qquad \text{rearrange}$$

$$F(x) = \frac{(x-a)^{k+1}}{(k+1)(q-a)^k} \cdot F'(q)$$

- We now examine $F'(q)$. First carefully differentiate $F(x)$:

$$F'(x) = \frac{d}{dx}\left[f(x) - \left(f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2 + \cdots + \frac{1}{k!}f^{(k)}(x-a)^k \right) \right]$$

$$= f'(x) - \left(f'(a) + \frac{2}{2}f''(a)(x-a) + \frac{3}{3!}f'''(a)(x-a)^2 + \cdots + \frac{k}{k!}f^{(k)}(a)(x-a)^{k-1} \right)$$

$$= f'(x) - \left(f'(a) + f''(a)(x-a) + \frac{1}{2}f'''(a)(x-a)^2 + \cdots + \frac{1}{(k-1)!}f^{(k)}(a)(x-a)^{k-1} \right)$$

  Now notice that if we set $f'(x) = g(x)$ then this becomes

$$F'(x) = g(x) - \left(g(a) + g'(a)(x-a) + \frac{1}{2}g''(a)(x-a)^2 + \cdots + \frac{1}{(k-1)!}g^{(k-1)}(a)(x-a)^{k-1} \right)$$

  So $F'(x)$ is then exactly the remainder formula but for an order $k - 1$ approximation to the function $g(x) = f'(x)$.

- Hence the function $F'(q)$ is the remainder when we approximate $f'(q)$ with an order $k - 1$ Taylor polynomial. The remainder formula, equation (3.4.33), then tells us that there is a number c between a and q so that

$$F'(q) = g(q) - \left(g(a) + g'(a)(q - a) + \frac{1}{2}g''(a)(q - a)^2 + \cdots + \frac{1}{(k-1)!}g^{(k-1)}(a)(q - a)^{k-1} \right)$$
$$= \frac{1}{k!}g^{(k)}(c)(q - a)^k = \frac{1}{k!}f^{(k+1)}(c)(q - a)^k$$

Notice that here we have assumed that $f^{(k+1)}(x)$ exists.

- Now substitute this back into our equation above

$$\begin{aligned} F(x) &= \frac{(x - a)^{k+1}}{(k + 1)(q - a)^k} \cdot F'(q) \\ &= \frac{(x - a)^{k+1}}{(k + 1)(q - a)^k} \cdot \frac{1}{k!}f^{(k+1)}(c)(q - a)^k \\ &= \frac{1}{(k + 1)k!} \cdot f^{(k+1)}(c) \cdot \frac{(x - a)^{k+1}(q - a)^k}{(q - a)^k} \\ &= \frac{1}{(k + 1)!} \cdot f^{(k+1)}(c) \cdot (x - a)^{k+1} \end{aligned}$$

as required.

So we now know that

- if, for some k, the remainder formula (with $n = k$) is true for all k times differentiable functions,

- then the remainder formula is true (with $n = k + 1$) for all $k + 1$ times differentiable functions.

Repeatedly applying this for $k = 1, 2, 3, 4, \cdots$ (and recalling that we have shown the remainder formula is true when $n = 0, 1$) gives equation (3.4.33) for all $n = 0, 1, 2, \ldots$.   □

3.5 ▲ Optimisation

One important application of differential calculus is to find the maximum (or minimum) value of a function. This often finds real world applications in problems such as the following.

Example 3.5.1

A farmer has 400m of fencing materials. What is the largest rectangular paddock that can be enclosed?

Solution. We will describe a general approach to these sorts of problems in Sections 3.5.2 and 3.5.3 below, but here we can take a stab at starting the problem.

- Begin by defining variables and their units (more generally we might draw a picture too); let the dimensions of the paddock be x by y metres.

- The area enclosed is then Am^2 where

$$A = x \cdot y$$

At this stage we cannot apply the calculus we have developed since the area is a function of two variables and we only know how to work with functions of a single variable. We need to eliminate one variable.

- We know that the perimeter of the rectangle (and hence the dimensions x and y) are constrained by the amount of fencing materials the farmer has to hand:

$$2x + 2y \leqslant 400$$

and so we have

$$y \leqslant 200 - x$$

Clearly the area of the paddock is maximised when we use all the fencing possible, so

$$y = 200 - x$$

- Now substitute this back into our expression for the area

$$A = x \cdot (200 - x)$$

Since the area cannot be negative (and our lengths x, y cannot be negative either), we must also have

$$0 \leqslant x \leqslant 200$$

- Thus the question of the largest paddock enclosed becomes the problem of finding the maximum value of

$$A = x \cdot (200 - x) \qquad \text{subject to the constraint } 0 \leqslant x \leqslant 200.$$

Example 3.5.1

The above example is sufficiently simple that we can likely determine the answer by several different methods. In general, we will need more systematic methods for solving problems of the form

Find the maximum value of $y = f(x)$ subject to $a \leqslant x \leqslant b$

To do this we need to examine what a function looks like near its maximum and minimum values.

3.5.1 ▸▸ Local and Global Maxima and Minima

We start by asking:

> Suppose that the maximum (or minimum) value of $f(x)$ is $f(c)$ then what does that tell us about c?

Notice that we have not yet made the ideas of maximum and minimum very precise. For the moment think of maximum as "the biggest value" and minimum as "the smallest value".

Warning 3.5.2.

It is important to distinguish between "the smallest value" and "the smallest magnitude". For example, because

$$-5 < -1$$

the number -5 is smaller than -1. But the magnitude of -1, which is $|-1| = 1$, is smaller than the magnitude of -5, which is $|-5| = 5$. Thus the smallest number in the set $\{-1, -5\}$ is -5, while the number in the set $\{-1, -5\}$ that has the smallest magnitude is -1.

Now back to thinking about what happens around a maximum. Suppose that the maximum value of $f(x)$ is $f(c)$, then for all "nearby" points, the function should be smaller.
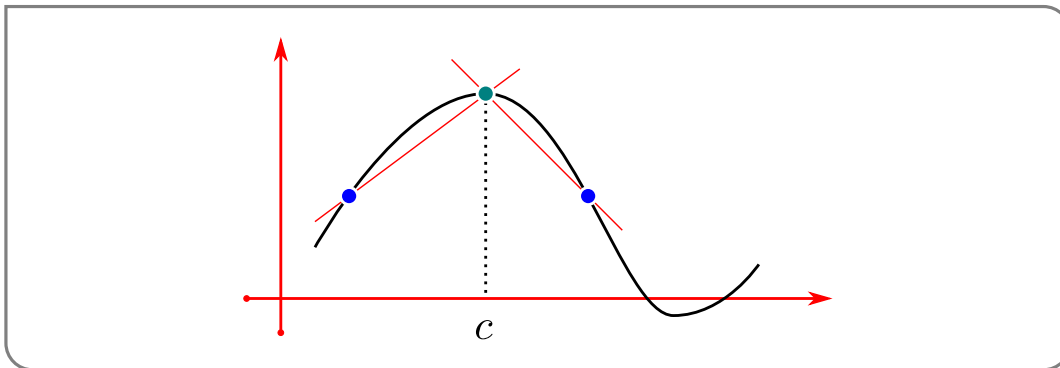
Consider the derivative of $f'(c)$:

$$f'(c) = \lim_{h \to 0} \frac{f(c+h) - f(c)}{h}.$$

Split the above limit into the left and right limits:

- Consider points to the right of $x = c$, For all $h > 0$,

$$
\begin{aligned}
f(c+h) &\leqslant f(c) && \text{which implies that} \\
f(c+h) - f(c) &\leqslant 0 && \text{which also implies} \\
\frac{f(c+h) - f(c)}{h} &\leqslant 0 && \text{since } \frac{\text{negative}}{\text{positive}} = \text{negative.}
\end{aligned}
$$

But now if we squeeze $h \to 0$ we get

$$\lim_{h \to 0^+} \frac{f(c+h) - f(c)}{h} \leqslant 0$$

(provided the limit exists).

- Consider points to the left of $x = c$. For all $h < 0$,

$$f(c + h) \leqslant f(c) \qquad \text{which implies that}$$
$$f(c + h) - f(c) \leqslant 0 \qquad \text{which also implies}$$
$$\frac{f(c + h) - f(c)}{h} \geqslant 0 \qquad \text{since } \frac{\text{negative}}{\text{negative}} = \text{positive.}$$

But now if we squeeze $h \to 0$ we get

$$\lim_{h \to 0^-} \frac{f(c+h) - f(c)}{h} \geqslant 0$$

(provided the limit exists).

- So if the derivative $f'(c)$ exists, then the above right- and left-hand limits must agree, which forces $f'(c) = 0$.

Thus we can conclude that

If the maximum value of $f(x)$ is $f(c)$ and $f'(c)$ exists, then $f'(c) = 0$.

Using similar reasoning one can also see that

If the minimum value of $f(x)$ is $f(c)$ and $f'(c)$ exists, then $f'(c) = 0$.

Notice two things about the above reasoning:

- Firstly, in order for the argument to work we only need that $f(x) < f(c)$ for x close to c — it does not matter what happens for x values far from c.

- Secondly, in the above argument we needed to consider $f(x)$ for x both to the left of and to the right of c. If the function $f(x)$ is defined on a closed interval $[a, b]$, then the above argument only applies when $a < c < b$ — not when c is either of the endpoints a and b.

Consider the function below
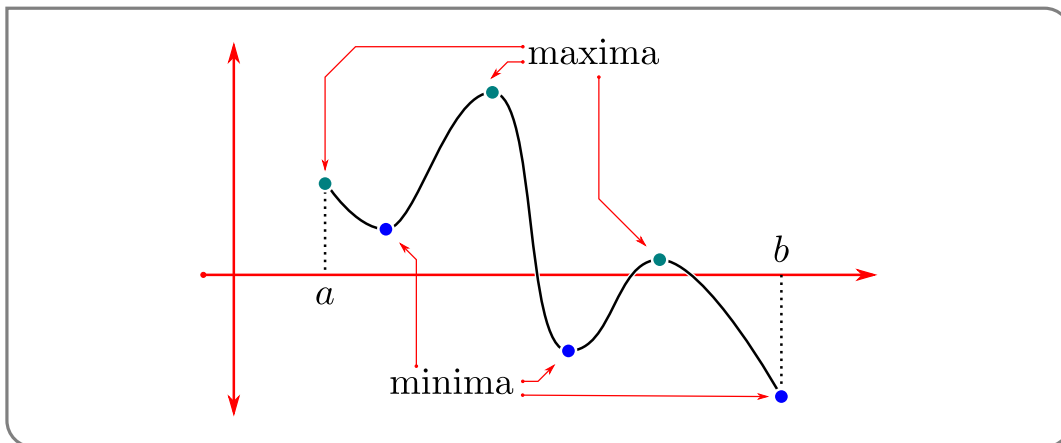
This function has only 1 maximum value (the middle green point in the graph) and 1 minimum value (the rightmost blue point), however it has 4 points at which the derivative is zero. In the small intervals around those points where the derivative is zero, we can see that function is *locally* a maximum or minimum, even if it is not the *global* maximum or minimum. We clearly need to be more careful distinguishing between these cases.

> **Definition 3.5.3.**
>
> Let I be an interval, like (a, b) or $[a, b]$ for example, and let the function $f(x)$ be defined for all $x \in I$. Now let $c \in I$. Then
>
> - we say that $f(x)$ has a *global (or absolute) minimum on the interval I* at the point $x = c$ if $f(x) \geqslant f(c)$ for all $x \in I$.
>
> - Similarly, we say that $f(x)$ has a *global (or absolute) maximum on I* at $x = c$ if $f(x) \leqslant f(c)$ for all $x \in I$.
>
> - We say that $f(x)$ has a *local*[51] *minimum on I* at $x = c$ if $f(x) \geqslant f(c)$ for all $x \in I$ that are near c. Precisely, if there is a $\delta > 0$ such that $f(x) \geqslant f(c)$ for all $x \in I$ that are within a distance δ of c.
>
> - Similarly, we say that $f(x)$ has a *local maximum on I* at $x = c$ if $f(x) \leqslant f(c)$ for all $x \in I$ that are near c. Precisely, if there is a $\delta > 0$ such that $f(x) \leqslant f(c)$ for all $x \in I$ that are within a distance δ of c.
>
> The global maxima and minima of a function are called the global extrema of the function, while the local maxima and minima are called the local extrema.

Consider again the function we showed in the figure above
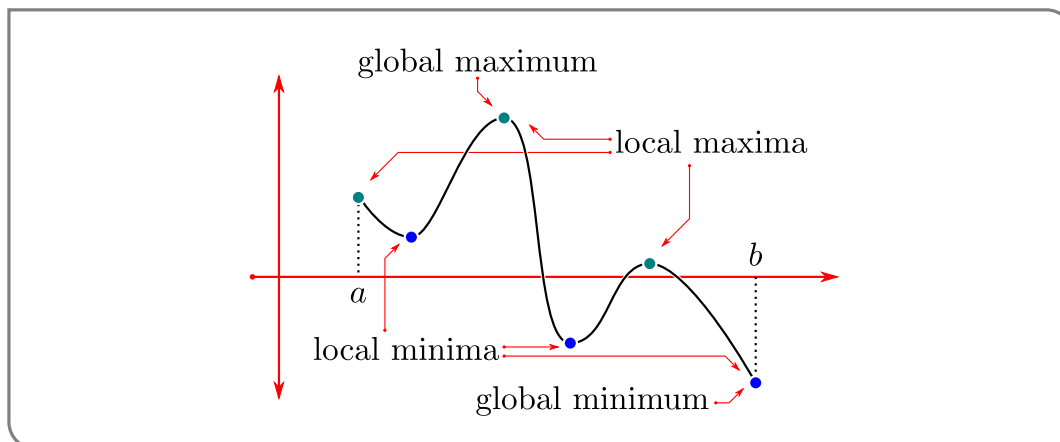
It has 3 local maxima and 3 local minima on the interval $[a, b]$. The global maximum occurs at the middle green point (which is also a local maximum), and the global minimum occurs at the rightmost blue point (which is also a local minimum).

---

51   Beware that, while many textbooks use these definitions of local minimum and maximum, some textbooks exclude the endpoints a, b of the interval $[a, b]$ from their definitions. Our definitions allow the endpoints a and b to be local minima and maxima. Note that, under our definitions, every global minimum (maximum) is also a local minimum (maximum).

Using the above definition we can summarise what we have learned above as the following theorem[52]:

Theorem 3.5.4.

Let the function $f(x)$ be defined on the interval I and let a, b, c be points in I with $a < c < b$. If $f(x)$ has a local maximum or local minimum at $x = c$ and if $f'(c)$ exists, then $f'(c) = 0$.

- It is often (but not always) the case that, when $f(x)$ has a local maximum at $x = c$, the function $f(x)$ increases strictly as x approaches c from the left and decreases strictly as x leaves c to the right. That is, $f'(x) > 0$ for x just to the left of c and $f'(x) < 0$ for x just to the right of c. Then, it is often the case, because $f'(x)$ is decreasing as x increases through c, that $f''(c) < 0$.

- Conversely, if $f'(c) = 0$ and $f''(c) < 0$, then, just to the right of c, $f'(x)$ must be negative, so that $f(x)$ is decreasing, and just to the left of c, $f'(x)$ must be positive, so that $f(x)$ is increasing. So $f(x)$ has a local maximum at c.

- Similarly, it is often the case that, when $f(x)$ has a local minimum at $x = c$, $f'(x) < 0$ for x just to the left of c and $f'(x) > 0$ for x just to the right of c and $f''(x) > 0$.

- Conversely, if $f'(c) = 0$ and $f''(c) > 0$, then, just to the right of c, $f'(x)$ must be positive, so that $f(x)$ is increasing, and, just to the left of c, $f'(x)$ must be negative, so that $f(x)$ is decreasing. So $f(x)$ has a local minimum at c.

Theorem 3.5.5.

Let $f(x)$ be defined on the interval I and let $a, b, c \in I$ with $a < c < b$.
If $f'(c) = 0$ and $f''(c) < 0$, then $f(x)$ has a local maximum at c.
If $f'(c) = 0$ and $f''(c) > 0$, then $f(x)$ has a local minimum at c.
Note the strict inequalities.

Theorem 3.5.4 says that, when $f(x)$ has a local maximum or minimum on an interval I at the point $x = c$, there are three possibilities.

- The derivative $f'(c) = 0$. This case is illustrated in the following figure.

---

[52]   This is one of several important mathematical contributions made by Pierre de Fermat, a French government lawyer and amateur mathematician, who lived in the first half of the seventeenth century.
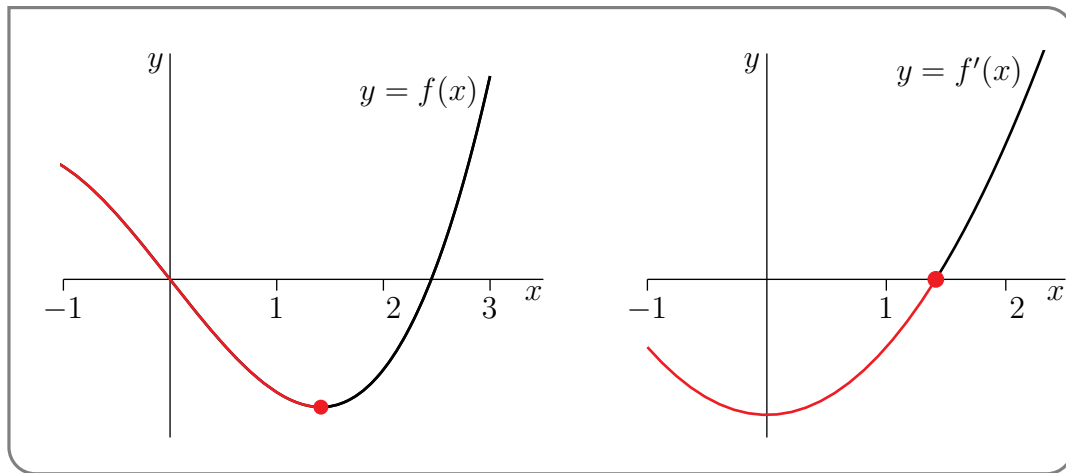
Observe that, in this example, $f'(x)$ changes continuously from negative to positive at the local minimum, taking the value zero at the local minimum (the red dot).

• The derivative $f'(c)$ does not exist. This case is illustrated in the following figure.
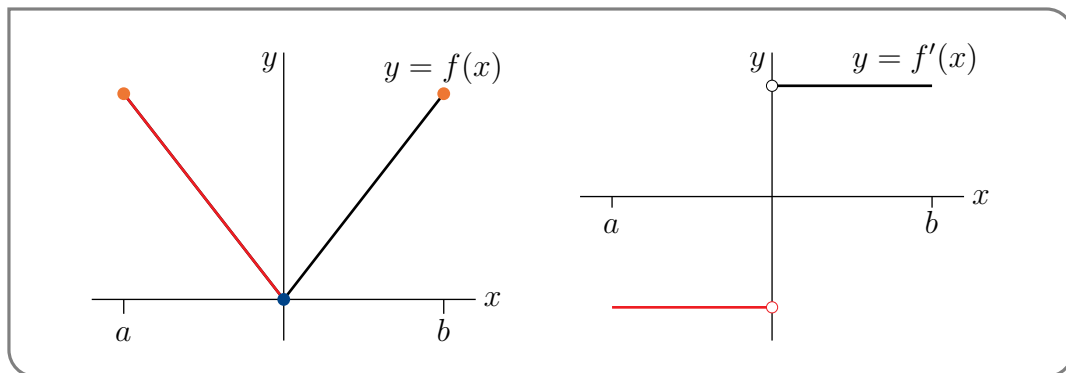
Observe that, in this example, $f'(x)$ changes discontinuously from negative to positive at the local minimum ($x = 0$) and $f'(0)$ does not exist.

• The point c is an endpoint of the interval $I = [a, b]$. This case is also illustrated in the above figure. The endpoints a and b are both local maxima. But $f'(a)$ and $f'(b)$ are not zero.

This theorem demonstrates that the points at which the derivative is zero or does not exist are very important. It simplifies the discussion that follows if we give these points names.

Definition 3.5.6.

Let $f(x)$ be a function that is defined on the interval $a < x < b$ and let $a < c < b$. Then

• if $f'(c)$ exists and is zero we call $x = c$ a critical point of the function, and

• if $f'(c)$ does not exist then we call $x = c$ a singular point[53] of the function.

> **Warning 3.5.7.**
>
> Note that some people (and texts) will combine both of these cases and call $x = c$ a critical point when either the derivative is zero or does not exist. The reader should be aware of the lack of convention on this point[54] and should be careful to understand whether the more inclusive definition of critical point is being used, or if the text is using the more precise definition that distinguishes critical and singular points.

We'll now look at a few simple examples involving local maxima and minima, critical points and singular points. Then we will move on to global maxima and minima.

Example 3.5.8

In this example, we'll look for local maxima and minima of the function $f(x) = x^3 - 6x$ on the interval $-2 \leqslant x \leqslant 3$.

- First compute the derivative

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

  Since this is a polynomial it is defined everywhere on the domain and so there will not be any singular points. So we now look for critical points.

- To do so we look for zeroes of the derivative

$$f'(x) = 3x^2 - 6 = 3(x^2 - 2) = 3(x - \sqrt{2})(x + \sqrt{2}).$$

  This derivative takes the value 0 at two different values of x. Namely $x = c_- = -\sqrt{2}$ and $x = c_+ = \sqrt{2}$. Here is a sketch of the graph of $f(x)$.

---
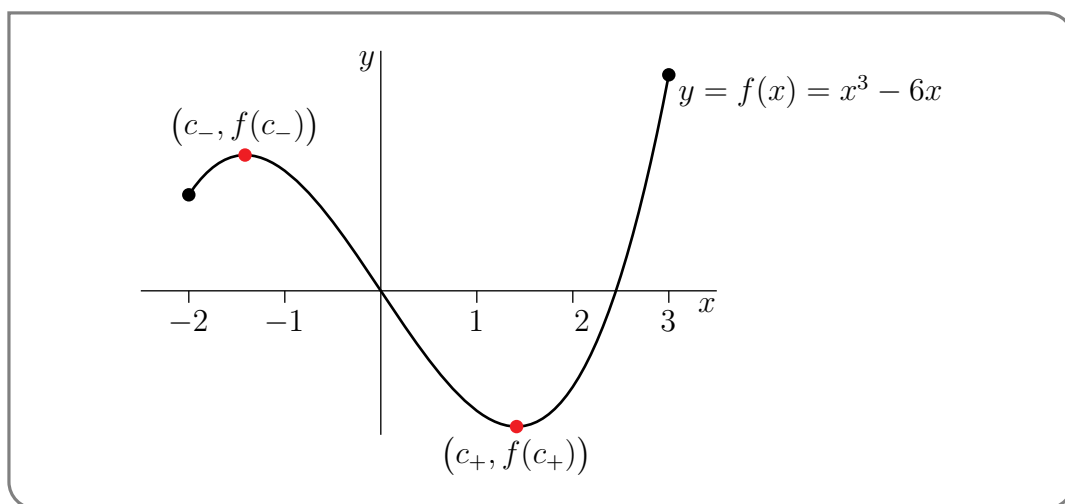
53 For c to be a local maximum or minimum of f, the function f must obviously be defined at c. So here we are considering only points c in the domain of f. We will later, in Section 3.6.2, extend the definition of singular points of f to points that are not in the domain of f.
54 No pun intended.

From the figure we see that

- $f(x)$ has a local minimum at the endpoint $x = -2$ (i.e. we have $f(x) \geqslant f(-2)$ whenever $x \geqslant -2$ is close to -2) and

- $f(x)$ has a local minimum at $x = c_+$ (i.e. we have $f(x) \geqslant f(c_+)$ whenever x is close to c_+) and

- $f(x)$ has a local maximum at $x = c_-$ (i.e. we have $f(x) \leqslant f(c_-)$ whenever x is close to c_-) and

- $f(x)$ has a local maximum at the endpoint $x = 3$ (i.e. we have $f(x) \leqslant f(3)$ whenever $x \leqslant 3$ is close to 3) and

- the global minimum of $f(x)$, for x in the interval $-2 \leqslant x \leqslant 3$, is at $x = c_+$ (i.e. we have $f(x) \geqslant f(c_+)$ whenever $-2 \leqslant x \leqslant 3$) and

- the global maximum of $f(x)$, for x in the interval $-2 \leqslant x \leqslant 3$, is at $x = 3$ (i.e. we have $f(x) \leqslant f(3)$ whenever $-2 \leqslant x \leqslant 3$).

- Note that we have carefully constructed this example to illustrate that the global maximum (or minimum) of a function on an interval may or may not also be a critical point of the function.

Example 3.5.8

Example 3.5.9

In this example, we'll look for local maxima and minima of the function $f(x) = x^3$ on the interval $-1 < x < 1$.

- First compute the derivative:

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

Again, this is a polynomial and so defined on all of the domain. The function will not have singular points, but may have critical points.

- The derivative is zero only when $x = 0$, so $x = c = 0$ is the only critical point of the function.

- The graph of $f(x)$ is sketched below. From that sketch we see that $f(x)$ has *neither* a local maximum *nor* a local minimum at $x = c$ despite the fact that $f'(c) = 0$ — we have $f(x) < f(c) = 0$ for all $x < c = 0$ and $f(x) > f(c) = 0$ for all $x > c = 0$.
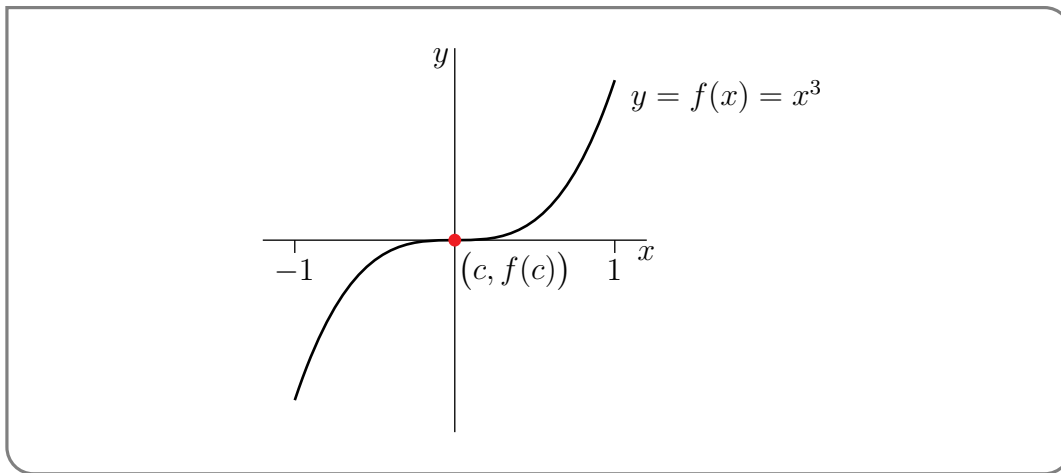
- Note that this example has been constructed to illustrate that a critical point (or singular point) of a function *need not be a local maximum or minimum* for the function.

- Reread Theorem 3.5.4. It says[55] "Let $\cdots$. If $f(x)$ has a local maximum/minimum at $x = c$ and if $f'(c)$ exists, then $f'(c) = 0$". It *does not say* that "if $f'(c) = 0$ then f has a local maximum/minimum at $x = c$".

Example 3.5.9

Example 3.5.10

In this example, we'll look for local maxima and minima of the function

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

on the interval $-1 < x < 1$ and we'll also look for local maxima and minima of the function

$$g(x) = x^{2/3}$$

on the interval $-1 < x < 1$.

- Again, start by computing the derivatives (reread Example 2.2.10):

$$f'(x) = \begin{cases} 1 & \text{if } x > 0 \\ \text{undefined} & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$$

$$g'(x) = \begin{cases} \frac{2}{3}x^{-1/3} & \text{if } x \neq 0 \\ \text{undefined} & \text{if } x = 0 \end{cases}$$

---

55   A very common error of logic that people make is "Affirming the consequent". When the statement "if P then Q" is true, observing Q does *not* imply P. ("Affirming the consequent" eliminates "not" from the previous sentence.) For example, "If he is Shakespeare, then he is dead," and "That man is dead." does not imply "He must be Shakespeare.". Or you may have also seen someone use this reasoning: "If a person is a genius before their time then they are misunderstood." "I am misunderstood." "So I must be a genius before my time.".

- These derivatives *never* take the value 0, so the functions $f(x)$ and $g(x)$ do not have any critical points. However both derivatives do not exist at the point $x = 0$, so that point is a singular point for both $f(x)$ and $g(x)$.

- Here is a sketch of the graph of $f(x)$
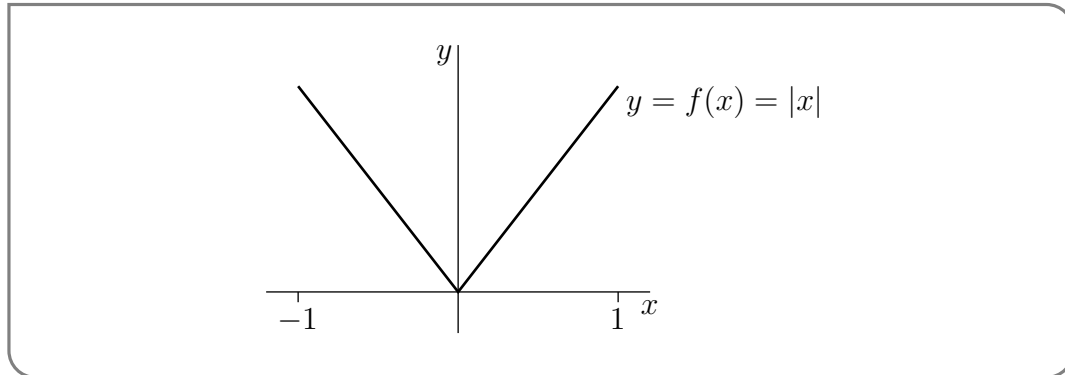
and a sketch of the graph of $g(x)$.
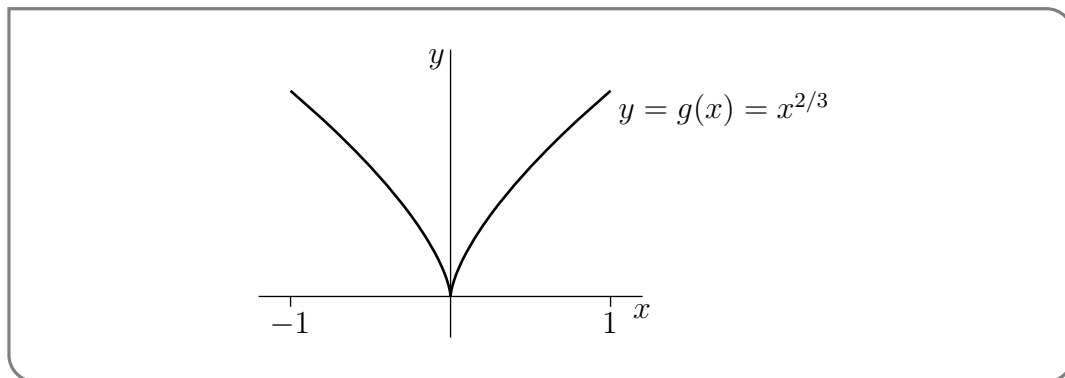
From the figures we see that both $f(x)$ and $g(x)$ have a local (and in fact global) minimum at $x = 0$ despite the fact that $x = 0$ is not a critical point.

- Reread Theorem 3.5.4 yet again. It says "Let $\cdots$. If $f(x)$ has a local maximum or local minimum at $x = c$ *and if f is differentiable at $x = c$, then $f'(c) = 0$*". It says nothing about what happens at points where the derivative does not exist. Indeed that is why we have to consider both critical points and singular points when we look for maxima and minima.

Example 3.5.10

3.5.2 ▸▸ Finding Global Maxima and Minima

We now have a technique for finding local maxima and minima — just look at endpoints of the interval of interest and for values of x for which either $f'(x) = 0$ or $f'(x)$ does not exist. What about finding global maxima and minima? We'll start by stating explicitly that, under appropriate hypotheses, global maxima and minima are guaranteed to exist.

> **Theorem 3.5.11.**
>
> Let the function $f(x)$ be defined and continuous on the closed, finite interval[56] $-\infty < a \leqslant x \leqslant b < \infty$. Then $f(x)$ attains a maximum and a minimum at least once. That is, there exist numbers $a \leqslant x_m, x_M \leqslant b$ such that
>
> $$f(x_m) \leqslant f(x) \leqslant f(x_M) \qquad \text{for all } a \leqslant x \leqslant b$$

So let's again consider the question

> Suppose that the maximum (or minimum) value of $f(x)$, for $a \leqslant x \leqslant b$, is $f(c)$. What does that tell us about c?

If c obeys $a < c < b$ (note the strict inequalities), then f has a local maximum (or minimum) at $x = c$ and Theorem 3.5.4 tells us that either $f'(c) = 0$ or $f'(c)$ does not exist. The only other place that a maximum or minimum can occur are at the ends of the interval. We can summarise this as:

> **Theorem 3.5.12.**
>
> If $f(x)$ has a global maximum or global minimum, for $a \leqslant x \leqslant b$, at $x = c$ then there are 3 possibilities. Either
>
> - $f'(c) = 0$, or
>
> - $f'(c)$ does not exist, or
>
> - $c = a$ or $c = b$.
>
> That is, a global maximum or minimum must occur either at a critical point, a singular point or at the endpoints of the interval.

This theorem provides the basis for a method to find the maximum and minimum values of $f(x)$ for $a \leqslant x \leqslant b$:

---

56   The hypotheses that $f(x)$ be continuous and that the interval be finite and closed are all essential. We suggest that you find three functions $f_1(x)$, $f_2(x)$ and $f_3(x)$ with f_1 defined but not continuous on $0 \leqslant x \leqslant 1$, f_2 defined and continuous on $-\infty < x < \infty$, and f_3 defined and continuous on $0 < x < 1$, and with none of f_1, f_2 and f_3 attaining either a global maximum or a global minimum.

> **Corollary 3.5.13.**
>
> Let $f(x)$ be a function on the interval $a \leqslant x \leqslant b$. Then to find the global maximum and minimum of the function:
>
> - Make a list of all values of c, with $a \leqslant c \leqslant b$, for which
>
>   - $f'(c) = 0$, or
>   - $f'(c)$ does not exist, or
>   - $c = a$ or $c = b$.
>
>   That is — compute the function at all the critical points, singular points, and endpoints.
>
> - Evaluate $f(c)$ for each c in that list. The largest (or smallest) of those values is the largest (or smallest) value of $f(x)$ for $a \leqslant x \leqslant b$.

Let's now demonstrate how to use this strategy. The function in this first example is not too simple — but it is a good example of a function that contains both a singular point and a critical point.

Example 3.5.14

Find the largest and smallest values of the function $f(x) = 2x^{5/3} + 3x^{2/3}$ for $-1 \leqslant x \leqslant 1$.

Solution. We will apply the method in Corollary 3.5.13. It is perhaps easiest to find the values at the endpoints of the intervals and then move on to the values at any critical or singular points.

- Before we get into things, notice that we can rewrite the function by factoring it:

$$f(x) = 2x^{5/3} + 3x^{2/3} = x^{2/3} \cdot (2x + 3)$$

- Let's compute the function at the endpoints of the interval:

$$f(1) = 2 + 3 = 5$$
$$f(-1) = 2 \cdot (-1)^{5/3} + 3 \cdot (-1)^{2/3} = -2 + 3 = 1$$

- To compute the function at the critical and singular points we first need to find the derivative:

$$f'(x) = 2 \cdot \frac{5}{3}x^{2/3} + 3 \cdot \frac{2}{3}x^{-1/3}$$
$$= \frac{10}{3}x^{2/3} + 2x^{-1/3}$$
$$= \frac{10x + 6}{3x^{1/3}}$$

- Notice that the numerator and denominator are defined for all x. The only place the derivative is undefined is when the denominator is zero. Hence the only singular point is at $x = 0$. The corresponding function value is

$$f(0) = 0$$

- To find the critical points we need to solve $f'(x) = 0$:

$$0 = \frac{10x + 6}{3x^{1/3}}$$

Hence we must have $10x = -6$ or $x = -3/5$. The corresponding function value is

$$f(x) = x^{2/3} \cdot (2x + 3) \qquad\qquad \text{recall this from above, then}$$

$$f(-3/5) = (-3/5)^{2/3} \cdot \left(2 \cdot \frac{-3}{5} + 3\right)$$

$$= \left(\frac{9}{25}\right)^{1/3} \cdot \frac{-6 + 15}{5}$$

$$= \left(\frac{9}{25}\right)^{1/3} \cdot \frac{9}{5} \approx 1.28$$

Note that if we do not want to approximate the root (if, for example, we do not have a calculator handy), then we can also write

$$f(-3/5) = \left(\frac{9}{25}\right)^{1/3} \cdot \frac{9}{5}$$

$$= \left(\frac{9}{25}\right)^{1/3} \cdot \frac{9}{25} \cdot 5$$

$$= 5 \cdot \left(\frac{9}{25}\right)^{4/3}$$

Since $0 < 9/25 < 1$, we know that $0 < \left(\frac{9}{25}\right)^{4/3} < 1$, and hence

$$0 < f(-3/5) = 5 \cdot \left(\frac{9}{25}\right)^{4/3} < 5.$$

- We summarise our work in this table

c	$-\frac{3}{5}$	0	-1	1
type	critical point	singular point	endpoint	endpoint
$f(c)$	$\frac{9}{5}\sqrt[3]{\frac{9}{25}} \approx 1.28$	0	1	5

- The largest value of f in the table is 5 and the smallest value of f in the table is 0.

- Thus on the interval $-1 \leqslant x \leqslant 1$ the global maximum of f is 5, and is taken at $x = 1$, while the global minimum value of $f(x)$ is 0, and is taken at $x = 0$.

- For completeness we also sketch the graph of this function on the same interval.
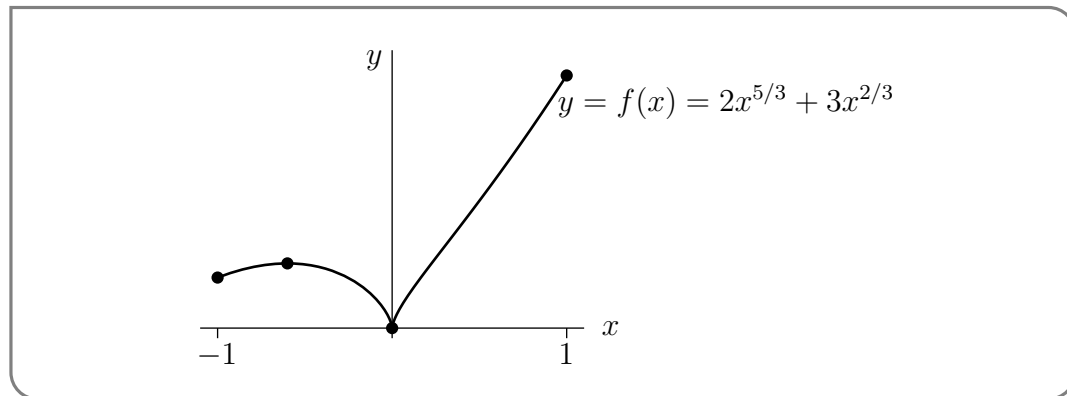
Later (in Section 3.6) we will see how to construct such a sketch without using a calculator or computer.

Example 3.5.14

3.5.3 ⇥ Max/Min Examples

As noted at the beginning of this section, the problem of finding maxima and minima is a very important application of differential calculus in the real world. We now turn to a number of examples of this process. But to guide the reader we will describe a general procedure to follow for these problems.

(1) Read — read the problem carefully. Work out what information is given in the statement of the problem and what we are being asked to compute.

(2) Diagram — draw a diagram. This will typically help you to identify what you know about the problem and what quantities you need to work out.

(3) Variables — assign variables to the quantities in the problem along with their units. It is typically a good idea to make sensible choices of variable names: A for area, h for height, t for time etc.

(4) Relations — find relations between the variables. By now you should know the quantity we are interested in (the one we want to maximise or minimise) and we need to establish a relation between it and the other variables.

(5) Reduce — the relation down to a function of one variable. In order to apply the calculus we know, we must have a function of a single variable. To do this we need to use all the information we have to eliminate variables. We should also work out the domain of the resulting function.

(6) Maximise or minimise — we can now apply the methods of Corollary 3.5.13 to find the maximum or minimum of the quantity we need (as the problem dictates).

(7) Be careful — make sure your answer makes sense. Make sure quantities are physical. For example, lengths and areas cannot be negative.

(8) Answer the question — be sure your answer really answers the question asked in the problem.

Let us start with a relatively simple problem:

Example 3.5.15

A closed rectangular container with a square base is to be made from two different materials. The material for the base costs \$5 per square meter, while the material for the other five sides costs \$1 per square meter. Find the dimensions of the container which has the largest possible volume if the total cost of materials is \$72.

Solution. We can follow the steps we outlined above to find the solution.

- We need to determine the area of the two types of materials used and the corresponding total cost.

- Draw a picture of the box.
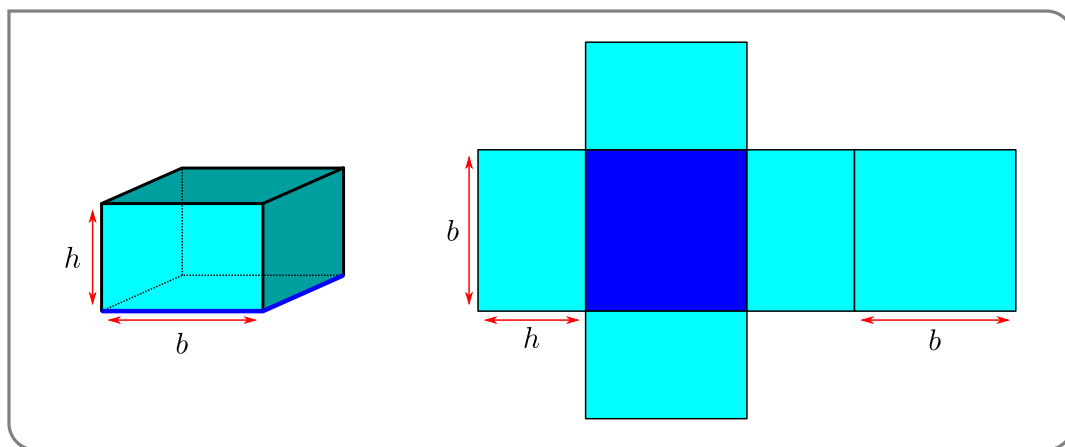
The more useful picture is the unfolded box on the right.

- In the picture we have already introduced two variables. The square base has side-length b metres and it has height h metres. Let the area of the base be A_b and the area of the other fives sides be A_s (both in m^2), and the total cost be C (in dollars). Finally let the volume enclosed be Vm^3.

- Some simple geometry tells us that

$$A_b = b^2$$
$$A_s = 4bh + b^2$$
$$V = b^2h$$
$$C = 5 \cdot A_b + 1 \cdot A_s = 5b^2 + 4bh + b^2 = 6b^2 + 4bh.$$

- To eliminate one of the variables we use the fact that the total cost is \$72.

$$C = 6b^2 + 4bh = 72 \qquad\qquad \text{rearrange}$$
$$4bh = 72 - 6b^2 \qquad\qquad \text{isolate } h$$
$$h = \frac{72 - 6b^2}{4b} = \frac{3}{2} \cdot \frac{12 - b^2}{b}$$

Substituting this into the volume gives

$$V = b^2 h = \frac{3b}{2}(12 - b^2) = 18b - \frac{3}{2}b^3$$

Now note that since b is a length it cannot be negative, so $b \geqslant 0$. Further since volume cannot be negative, we must also have

$$12 - b^2 \geqslant 0$$

and so $b \leqslant \sqrt{12}$.

- Now we can apply Corollary 3.5.13 on the above expression for the volume with $0 \leqslant b \leqslant \sqrt{12}$. The endpoints give:

$$V(0) = 0$$
$$V(\sqrt{12}) = 0$$

The derivative is

$$V'(b) = 18 - \frac{9b^2}{2}$$

Since this is a polynomial there are no singular points. However we can solve $V'(b) = 0$ to find critical points:

$$18 - \frac{9b^2}{2} = 0 \qquad\qquad \text{divide by 9 and multiply by 2}$$
$$4 - b^2 = 0$$

Hence $b = \pm 2$. Thus the only critical point in the domain is $b = 2$. The corresponding volume is

$$V(2) = 18 \times 2 - \frac{3}{2} \times 2^3$$
$$= 36 - 12 = 24.$$

So by Corollary 3.5.13, the maximum volume is when 24 when $b = 2$ and

$$h = \frac{3}{2} \cdot \frac{12 - b^2}{b} = \frac{3}{2}\frac{12 - 4}{2} = 6.$$

- All our quantities make sense; lengths, areas and volumes are all non-negative.

- Checking the question again, we see that we are asked for the dimensions of the container (rather than its volume) so we can answer with

  The container with dimensions $2 \times 2 \times 6m$ will be the largest possible.

Example 3.5.16

A rectangular sheet of cardboard is 6 inches by 9 inches. Four identical squares are cut from the corners of the cardboard, as shown in the figure below, and the remaining piece is folded into an open rectangular box. What should the size of the cut out squares be in order to maximize the volume of the box?

Solution. This one is quite similar to the previous one, so we perhaps don't need to go into so much detail.

- After reading carefully we produce the following picture:
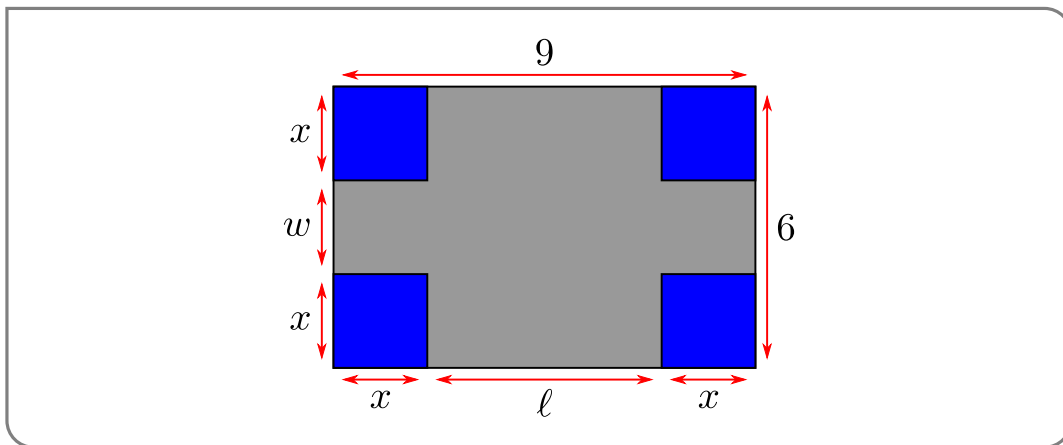
- Let the height of the box be x inches, and the base be $\ell \times w$ inches. The volume of the box is then V cubic inches.

- Some simple geometry tells us that $\ell = 9 - 2x, w = 6 - 2x$ and so

$$V = x(9 - 2x)(6 - 2x)\text{cubic inches}$$
$$= 54x - 30x^2 + 4x^3.$$

Notice that since all lengths must be non-negative, we must have

$$x, \ell, w \geqslant 0$$

and so $0 \leqslant x \leqslant 3$ (if $x > 3$ then $w < 0$).

- We can now apply Corollary 3.5.13. First the endpoints of the interval give

$$V(0) = 0 \qquad\qquad V(3) = 0$$

The derivative is

$$V'(x) = 54 - 60x + 12x^2$$
$$= 6(9 - 10x + 2x^2)$$

Since this is a polynomial there are no singular points. To find critical points we solve $V'(x) = 0$ to get

$$x_{\pm} = \frac{10 \pm \sqrt{100 - 4 \times 2 \times 9}}{4}$$
$$= \frac{10 \pm \sqrt{28}}{4} = \frac{10 \pm 2\sqrt{7}}{4} = \frac{5 \pm \sqrt{7}}{2}$$

We can then use a calculator to approximate

$$x_+ \approx 3.82 \qquad\qquad\qquad x_- \approx 1.18.$$

So x_- is inside the domain, while x_+ lies outside.

Alternatively[57], we can bound $x_{\pm}$ by first noting that $2 \leqslant \sqrt{7} \leqslant 3$. From this we know that

$$1 = \frac{5-3}{2} \leqslant x_- = \frac{5 - \sqrt{7}}{2} \leqslant \frac{5-2}{2} = 1.5$$
$$3.5 = \frac{5+2}{2} \leqslant x_+ = \frac{5 + \sqrt{7}}{2} \leqslant \frac{5+3}{2} = 4$$

- Since the volume is zero when $x = 0, 3$, it must be the case that the volume is maximised when $x = x_- = \frac{5-\sqrt{7}}{2}$.

- Notice that since $0 < x_- < 3$ we know that the other lengths are positive, so our answer makes sense. Further, the question only asks for the length x and not the resulting volume so we have answered the question.

Example 3.5.16

There is a new wrinkle in the next two examples. Each involves finding the minimum value of a function $f(x)$ with x running over all real numbers, rather than just over a finite interval as in Corollary 3.5.13. Both in Example 3.5.18 and in Example 3.5.19 the function $f(x)$ tends to $+\infty$ as x tends to either $+\infty$ or $-\infty$. So the minimum value of $f(x)$ will be achieved for some finite value of x, which will be a local minimum as well as a global minimum.

---

[57] Say if we do not have a calculator to hand, or your instructor insists that the problem be done without one.

Theorem 3.5.17.

Let $f(x)$ be defined and continuous for all $-\infty < x < \infty$. Let c be a finite real number.

(a) If $\lim_{x \to +\infty} f(x) = +\infty$ and $\lim_{x \to -\infty} f(x) = +\infty$ and if $f(x)$ has a global minimum at $x = c$, then there are 2 possibilities. Either

- $f'(c) = 0$, or
- $f'(c)$ does not exist

That is, a global minimum must occur either at a critical point or at a singular point.

(b) If $\lim_{x \to +\infty} f(x) = -\infty$ and $\lim_{x \to -\infty} f(x) = -\infty$ and if $f(x)$ has a global maximum at $x = c$, then there are 2 possibilities. Either

- $f'(c) = 0$, or
- $f'(c)$ does not exist

That is, a global maximum must occur either at a critical point or at a singular point.

Example 3.5.18

Find the point on the line $y = 6 - 3x$ that is closest to the point $(7, 5)$.

Solution. In this problem

- A simple picture
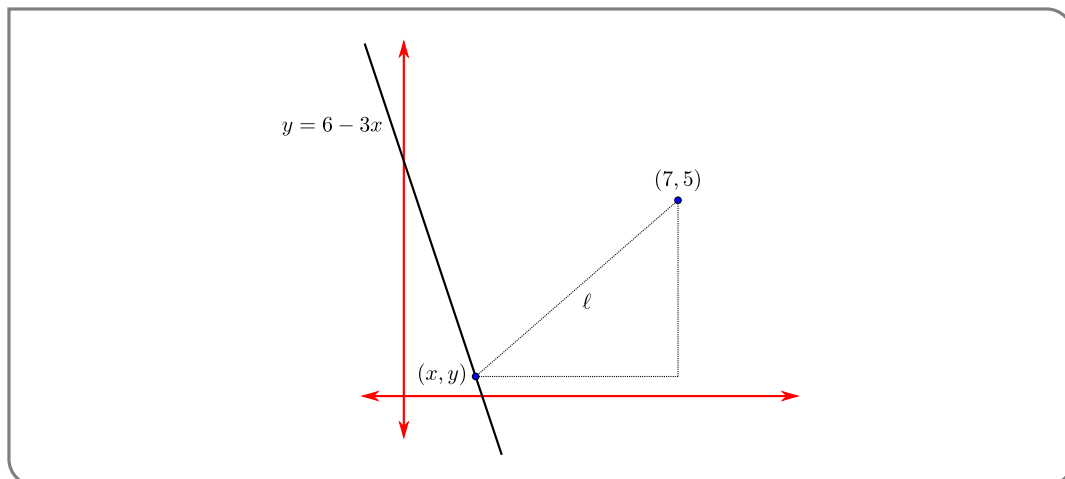
- Some notation is already given to us. Let a point on the line have coordinates (x, y), and we do not need units. And let ℓ be the distance from the point (x, y) to the point $(7, 5)$.

- Since the points are on the line the coordinates (x, y) must obey

$$y = 6 - 3x$$

  Notice that x and y have no further constraints. The distance ℓ is given by

$$\ell^2 = (x - 7)^2 + (y - 5)^2$$

- We can now eliminate the variable y:

$$\begin{aligned}
\ell^2 &= (x - 7)^2 + (y - 5)^2 \\
&= (x - 7)^2 + (6 - 3x - 5)^2 = (x - 7)^2 + (1 - 3x)^2 \\
&= x^2 - 14x + 49 + 1 - 6x + 9x^2 = 10x^2 - 20x + 50 \\
&= 10(x^2 - 2x + 5) \\
\ell &= \sqrt{10} \cdot \sqrt{x^2 - 2x + 5}
\end{aligned}$$

  Notice that as $x \to \pm\infty$ the distance $\ell \to +\infty$.

- We can now apply Theorem 3.5.17

  - Since the distance is defined for all real x, we do not have to check the endpoints of the domain — there are none.

  - Form the derivative:

$$\frac{\mathrm{d}\ell}{\mathrm{d}x} = \sqrt{10}\frac{2x - 2}{2\sqrt{x^2 - 2x + 5}}$$

    It is zero when $x = 1$, and undefined if $x^2 - 2x + 5 < 0$. However, since

$$x^2 - 2x + 5 = (x^2 - 2x + 1) + 4 = \underbrace{(x - 1)^2}_{\geqslant 0} + 4$$

    we know that $x^2 - 2x + 5 \geqslant 4$. Thus the function has no singular points and the only critical point occurs at $x = 1$. The corresponding function value is then

$$\ell(1) = \sqrt{10}\sqrt{1 - 2 + 5} = 2\sqrt{10}.$$

  - Thus the minimum value of the distance is $\ell = 2\sqrt{10}$ and occurs at $x = 1$.

- This answer makes sense — the distance is not negative.

- The question asks for the point that minimises the distance, not that minimum distance. Hence the answer is $x = 1, y = 6 - 3 = 3$. I.e.

  The point that minimises the distance is $(1, 3)$.

Notice that we can make the analysis easier by observing that the point that minimises the distance also minimises the squared-distance. So that instead of minimising the function ℓ, we can just minimise ℓ^2:

$$\ell^2 = 10(x^2 - 2x + 5)$$

The resulting algebra is a bit easier and we don't have to hunt for singular points.

Example 3.5.18

Example 3.5.19

Find the minimum distance from $(2,0)$ to the curve $y^2 = x^2 + 1$.

Solution. This is very much like the previous question.

- After reading the problem carefully we can draw a picture
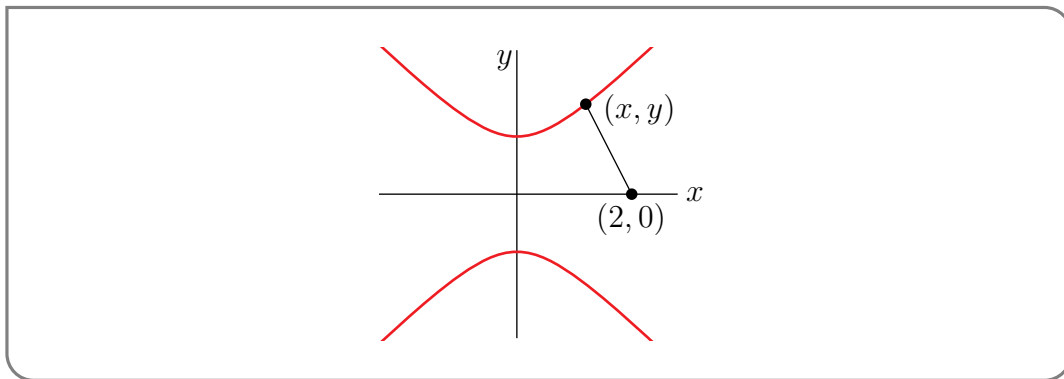
- In this problem we do not need units and the variables x, y are supplied. We define the distance to be ℓ and it is given by

$$\ell^2 = (x - 2)^2 + y^2.$$

As noted in the previous problem, we will minimise the squared-distance since that also minimises the distance.

- Since x, y satisfy $y^2 = x^2 + 1$, we can write the distance as a function of x:

$$\ell^2 = (x - 2)^2 + y^2 = (x - 2)^2 + (x^2 + 1)$$

Notice that as $x \to \pm\infty$ the squared-distance $\ell^2 \to +\infty$.

- Since the squared-distance is a polynomial it will not have any singular points, only critical points. The derivative is

$$\frac{d}{dx}\ell^2 = 2(x - 2) + 2x = 4x - 4$$

so the only critical point occurs at $x = 1$.

- When $x = 1, y = \pm\sqrt{2}$ and the distance is

$$\ell^2 = (1-2)^2 + (1+1) = 3 \qquad\qquad \ell = \sqrt{3}$$

and thus the minimum distance from the curve to $(2,0)$ is $\sqrt{3}$.

Example 3.5.19

Example 3.5.20

A water trough is to be constructed from a metal sheet of width 45 cm by bending up one third of the sheet on each side through an angle θ. Which θ will allow the trough to carry the maximum amount of water?

Solution. Clearly $0 \leqslant \theta \leqslant \pi$, so we are back in the domain[58] of Corollary 3.5.13.

- After reading the problem carefully we should realise that it is really asking us to maximise the cross-sectional area. A figure really helps.
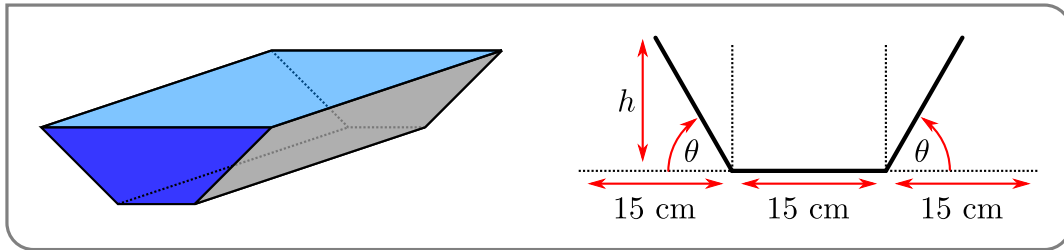
- From this we are led to define the height h *cm* and cross-sectional area A *cm*^{2}. Both are functions of θ.

$$h = 15\sin\theta$$

while the area can be computed as the sum of the central $15 \times h$ rectangle, plus two triangles. Each triangle has height h and base $15\cos\theta$. Hence

$$A = 15h + 2 \cdot \frac{1}{2} \cdot h \cdot 15\cos\theta$$
$$= 15h\,(1 + \cos\theta)$$

- Since $h = 15\sin\theta$ we can rewrite the area as a function of just θ:

$$A(\theta) = 225\sin\theta\,(1 + \cos\theta)$$

where $0 \leqslant \theta \leqslant \pi$.

---

58 Again, no pun intended.

- Now we use Corollary 3.5.13. The ends of the interval give

$$A(0) = 225 \sin 0(1 + \cos 0) = 0$$
$$A(\pi) = 225 \sin \pi (1 + \cos \pi) = 0$$

The derivative is

$$A'(\theta) = 225 \cos \theta \cdot (1 + \cos \theta) + 225 \sin \theta \cdot (-\sin \theta)$$
$$= 225 \left[\cos \theta + \cos^2 \theta - \sin^2 \theta \right] \qquad \text{recall } \sin^2 \theta = 1 - \cos^2 \theta$$
$$= 225 \left[\cos \theta + 2 \cos^2 \theta - 1 \right]$$

This is a continuous function, so there are no singular points. However we can still hunt for critical points by solving $A'(\theta) = 0$. That is

$$2 \cos^2 \theta + \cos \theta - 1 = 0 \qquad \text{factor carefully}$$
$$(2 \cos \theta - 1)(\cos \theta + 1) = 0$$

Hence we must have $\cos \theta = -1$ or $\cos \theta = \frac{1}{2}$. On the domain $0 \leqslant \theta \leqslant \pi$, this means $\theta = \pi/3$ or $\theta = \pi$.

$$A(\pi) = 0$$
$$A(\pi/3) = 225 \sin(\pi/3)(1 + \cos(\pi/3))$$
$$= 225 \cdot \frac{\sqrt{3}}{2} \cdot \left(1 + \frac{1}{2} \right)$$
$$= 225 \cdot \frac{3\sqrt{3}}{4} \approx 292.28$$

- Thus the cross-sectional area is maximised when $\theta = \dfrac{\pi}{3}$.

Example 3.5.20

Example 3.5.21

Find the points on the ellipse $\frac{x^2}{4} + y^2 = 1$ that are nearest to and farthest from the point $(1, 0)$.

Solution. While this is another distance problem, the possible values of x, y are bounded, so we need Corollary 3.5.13 rather than Theorem 3.5.17.

- We start by drawing a picture:
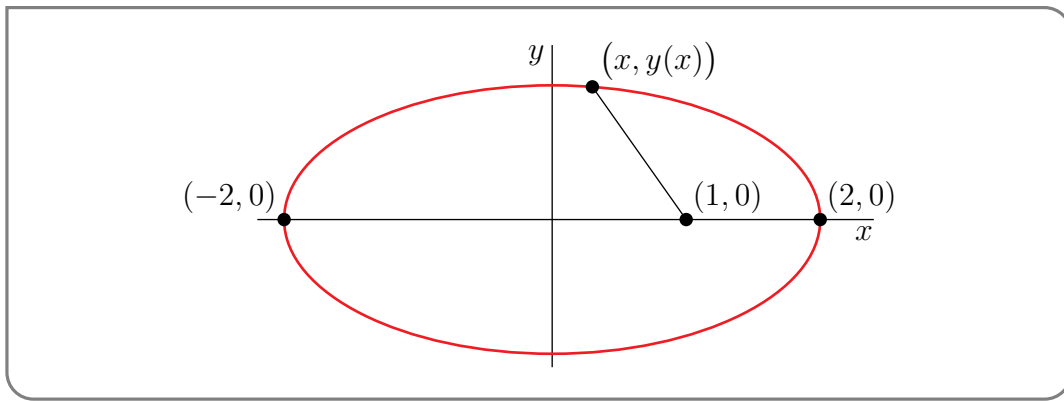
- Let ℓ be the distance from the point (x, y) on the ellipse to the point $(1, 0)$. As was the case above, we will maximise the squared-distance.

$$\ell^2 = (x - 1)^2 + y^2.$$

- Since (x, y) lies on the ellipse we have

$$\frac{x^2}{4} + y^2 = 1$$

Note that this also shows that $-2 \leqslant x \leqslant 2$ and $-1 \leqslant y \leqslant 1$.

Isolating y^2 and substituting this into our expression for ℓ^2 gives

$$\ell^2 = (x - 1)^2 + \underbrace{1 - x^2/4}_{=y^2}.$$

- Now we can apply Corollary 3.5.13. The endpoints of the domain give

$$\ell^2(-2) = (-2 - 1)^2 + 1 - (-2)^2/4 = 3^2 + 1 - 1 = 9$$
$$\ell^2(2) = (2 - 1)^2 + 1 - 2^2/4 = 1 + 1 - 1 = 1$$

The derivative is

$$\frac{\mathrm{d}}{\mathrm{d}x}\ell^2 = 2(x - 1) - x/2 = \frac{3x}{2} - 2$$

Thus there are no singular points, but there is a critical point at $x = 4/3$. The corresponding squared-distance is

$$\ell^2(4/3) = \left(\frac{4}{3} - 1\right)^2 + 1 - \frac{(4/3)^2}{4}$$
$$= (1/3)^2 + 1 - (4/9) = 6/9 = 2/3.$$

- To summarise (and giving distances and coordinates of points):

---

x	(x, y)	ℓ
-2	$(-2, 0)$	3
$4/3$	$\left(4/3, \pm\sqrt{5}/3\right)$	$\sqrt{2/3}$
2	$(2, 0)$	1

The point of maximum distance is $(-2, 0)$, and the point of minimum distance is $\left(4/3, \pm\sqrt{5}/3\right)$.

Example 3.5.21

Example 3.5.22

Find the dimensions of the rectangle of largest area that can be inscribed in an equilateral triangle of side a if one side of the rectangle lies on the base of the triangle.

Solution. Since the rectangle must sit inside the triangle, its dimensions are bounded and we will end up using Corollary 3.5.13.

- Carefully draw a picture:
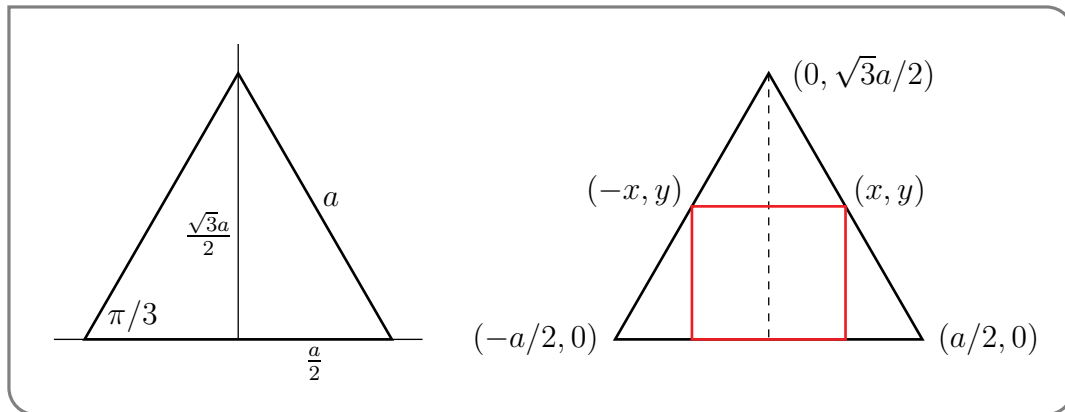
We have drawn (on the left) the triangle in the xy-plane with its base on the x-axis. The base has been drawn running from $(-a/2, 0)$ to $(a/2, 0)$ so its centre lies at the origin. A little Pythagoras (or a little trigonometry) tells us that the height of the triangle is

$$\sqrt{a^2 - (a/2)^2} = \frac{\sqrt{3}}{2} \cdot a = a \cdot \sin\frac{\pi}{3}$$

Thus the vertex at the top of the triangle lies at $\left(0, \frac{\sqrt{3}}{2} \cdot a\right)$.

- If we construct a rectangle that does not touch the sides of the triangle, then we can increase the dimensions of the rectangle until it touches the triangle and so make its area larger. Thus we can assume that the two top corners of the rectangle touch the triangle as drawn in the right-hand figure above.

- Now let the rectangle be $2x$ wide and y high. And let A denote its area. Clearly

$$A = 2xy.$$

where $0 \leqslant x \leqslant a/2$ and $0 \leqslant y \leqslant \frac{\sqrt{3}}{2}a$.

- Our construction means that the top-right corner of the rectangle will have coordinates (x, y) and lie on the line joining the top vertex of the triangle at $(0, \sqrt{3}a/2)$ to the bottom-right vertex at $(a/2, 0)$. In order to write the area as a function of x alone, we need the equation for this line since it will tell us how to write y as a function of x. The line has slope

$$\text{slope} = \frac{\sqrt{3}a/2 - 0}{0 - a/2} = -\sqrt{3}.$$

and passes through the point $(0, \sqrt{3}a/2)$, so any point (x, y) on that line satisfies:

$$y = -\sqrt{3}x + \frac{\sqrt{3}}{2}a.$$

- We can now write the area as a function of x alone

$$A(x) = 2x\left(-\sqrt{3}x + \frac{\sqrt{3}}{2}a\right)$$
$$= \sqrt{3}x(a - 2x).$$

with $0 \leqslant x \leqslant a/2$.

- The ends of the domain give:

$$A(0) = 0 \qquad\qquad\qquad A(a/2) = 0.$$

The derivative is

$$A'(x) = \sqrt{3}\left(x \cdot (-2) + 1 \cdot (a - 2x)\right) = \sqrt{3}(a - 4x).$$

Since this is a polynomial there are no singular points, but there is a critical point at $x = a/4$. There

$$A(a/4) = \sqrt{3} \cdot \frac{a}{4} \cdot (a - a/2) = \sqrt{3} \cdot \frac{a^2}{8}.$$
$$y = -\sqrt{3} \cdot (a/4) + \frac{\sqrt{3}}{2}a = \sqrt{3} \cdot \frac{a}{4}.$$

- Checking the question again, we see that we are asked for the dimensions rather than the area, so the answer is $2x \times y$:

The largest such rectangle has dimensions $\frac{a}{2} \times \frac{\sqrt{3}a}{4}$.

Example 3.5.22

This next one is a good physics example. In it we will derive Snell's Law[59] from Fermat's principle[60].

Example 3.5.23

Consider the figure below which shows the trajectory of a ray of light as it passes through two different mediums (say air and water).
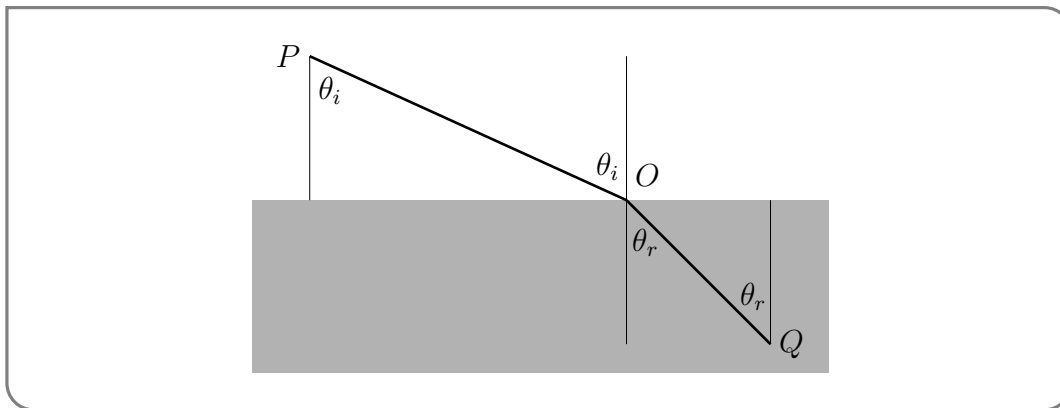
Let c_a be the speed of light in air and c_w be the speed of light in water. Fermat's principle states that a ray of light will always travel along a path that minimises the time taken. So if a ray of light travels from P (in air) to Q (in water) then it will "choose" the point O (on the interface) so as to minimise the total time taken. Use this idea to show Snell's law,

$$\frac{\sin \theta_i}{\sin \theta_r} = \frac{c_a}{c_w}$$

where θ_i is the angle of incidence and θ_r is the angle of refraction (as illustrated in the figure above).

Solution. This problem is a little more abstract than the others we have examined, but we can still apply Theorem 3.5.17.

- We are given a figure in the statement of the problem and it contains all the relevant points and angles. However it will simplify things if we decide on a coordinate system. Let's assume that the point O lies on the x-axis, at coordinates $(x, 0)$. The point P then lies above the axis at $(X_P, +Y_P)$, while Q lies below the axis at $(X_Q, -Y_Q)$. This is drawn below.

---

59   Snell's law is named after the Dutch astronomer Willebrord Snellius who derived it in around 1621, though it was first stated accurately in 984 by Ibn Sahl.
60   Named after Pierre de Fermat who described it in a letter in 1662. The beginnings of the idea, however, go back as far as Hero of Alexandria in around 60CE. Hero is credited with many inventions including the first vending machine, and a precursor of the steam engine called an aeolipile.
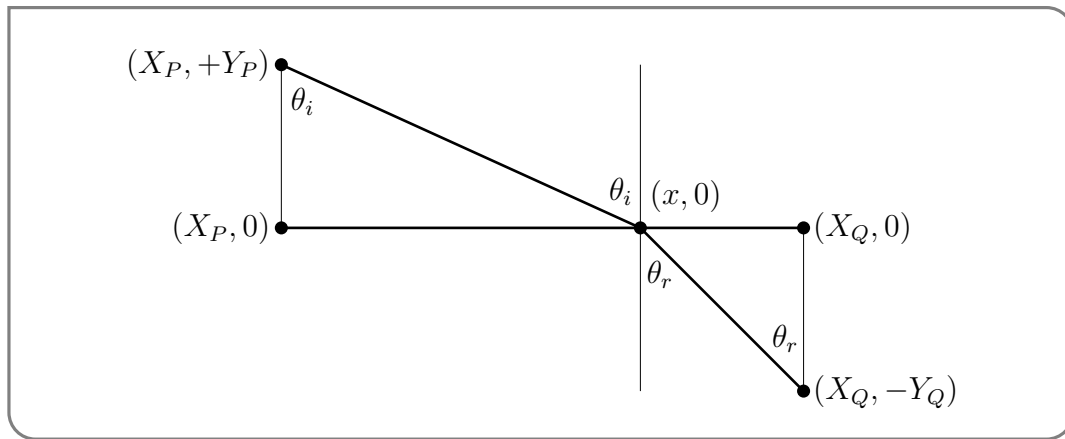
- The statement of Snell's law contains terms $\sin\theta_i$ and $\sin\theta_r$, so it is a good idea for us to see how to express these in terms of the coordinates we have just introduced:

$$\sin\theta_i = \frac{\text{opposite}}{\text{hypotenuse}} = \frac{(x - X_P)}{\sqrt{(X_P - x)^2 + Y_P^2}}$$

$$\sin\theta_r = \frac{\text{opposite}}{\text{hypotenuse}} = \frac{(X_Q - x)}{\sqrt{(X_Q - x)^2 + Y_Q^2}}$$

- Let ℓ_P denote the distance PO, and ℓ_Q denote the distance OQ. Then we have

$$\ell_P = \sqrt{(X_P - x)^2 + Y_P^2}$$

$$\ell_Q = \sqrt{(X_Q - x)^2 + Y_Q^2}$$

If we then denote the total time taken by T, then

$$T = \frac{\ell_P}{c_a} + \frac{\ell_Q}{c_w} = \frac{1}{c_a}\sqrt{(X_P - x)^2 + Y_P^2} + \frac{1}{c_w}\sqrt{(X_Q - x)^2 + Y_Q^2}$$

which is written as a function of x since all the other terms are constants.

- Notice that as $x \to +\infty$ or $x \to -\infty$ the total time $T \to \infty$ and so we can apply Theorem 3.5.17. The derivative is

$$\frac{\mathrm{d}T}{\mathrm{d}x} = \frac{1}{c_a}\frac{-2(X_P - x)}{2\sqrt{(X_P - x)^2 + Y_P^2}} + \frac{1}{c_w}\frac{-2(X_Q - x)}{2\sqrt{(X_Q - x)^2 + Y_Q^2}}$$

Notice that the terms inside the square-roots cannot be zero or negative since they are both sums of squares and $Y_P, Y_Q > 0$. So there are no singular points, but there is a critical point when $T'(x) = 0$, namely when

$$0 = \frac{1}{c_a}\frac{X_P - x}{\sqrt{(X_P - x)^2 + Y_P^2}} + \frac{1}{c_w}\frac{X_Q - x}{\sqrt{(X_Q - x)^2 + Y_Q^2}}$$

$$= \frac{-\sin\theta_i}{c_a} + \frac{\sin\theta_r}{c_w}$$

Rearrange this to get

$$\frac{\sin \theta_i}{c_a} = \frac{\sin \theta_r}{c_w} \qquad\qquad \text{move sines to one side}$$

$$\frac{\sin \theta_i}{\sin \theta_r} = \frac{c_a}{c_w}$$

which is exactly Snell's law.

Example 3.5.23

Example 3.5.24

The Statue of Liberty has height 46m and stands on a 47m tall pedestal. How far from the statue should an observer stand to maximize the angle subtended by the statue at the observer's eye, which is 1.5m above the base of the pedestal?

Solution. Obviously if we stand too close then all the observer sees is the pedestal, while if they stand too far then everything is tiny. The best spot for taking a photograph is somewhere in between.

- Draw a careful picture[61]
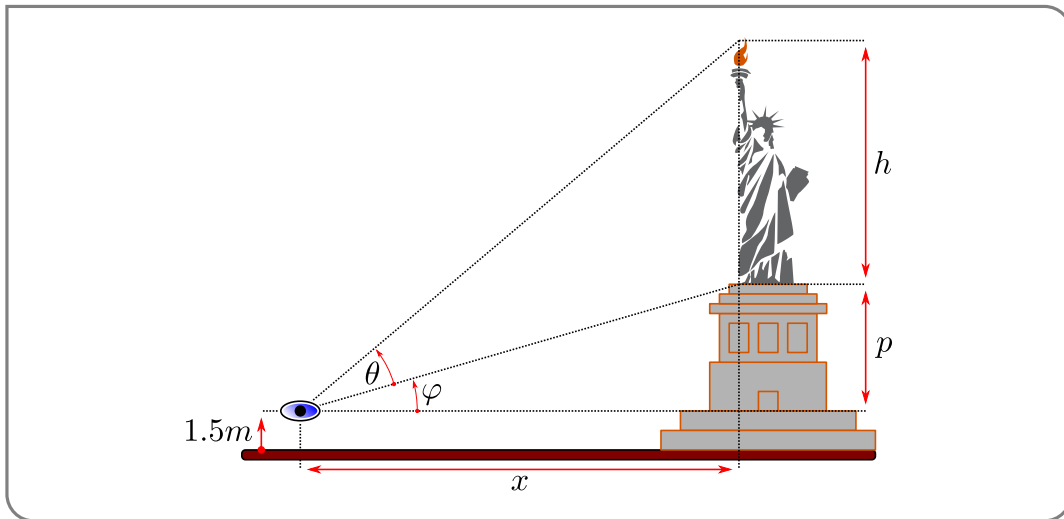
and we can put in the relevant lengths and angles.

- The height of the statue is $h = 46$m, and the height of the pedestal (above the eye) is $p = 47 - 1.5 = 45.5$m. The horizontal distance from the statue to the eye is x. There are two relevant angles. First θ is the angle subtended by the statue, while φ is the angle subtended by the portion of the pedestal above the eye.

---

61   And make some healthy use of public domain clip art.

- Some trigonometry gives us

$$\tan \varphi = \frac{p}{x}$$

$$\tan(\varphi + \theta) = \frac{p + h}{x}$$

Thus

$$\varphi = \arctan \frac{p}{x}$$

$$\varphi + \theta = \arctan \frac{p + h}{x}$$

and so

$$\theta = \arctan \frac{p + h}{x} - \arctan \frac{p}{x}.$$

- If we allow the viewer to stand at any point in front of the statue, then $0 \leqslant x < \infty$. Further observe that as $x \to \infty$ or $x \to 0$ the angle $\theta \to 0$, since

$$\lim_{x \to \infty} \arctan \frac{p + h}{x} = \lim_{x \to \infty} \arctan \frac{p}{x} = 0$$

and

$$\lim_{x \to 0^+} \arctan \frac{p + h}{x} = \lim_{x \to 0^+} \arctan \frac{p}{x} = \frac{\pi}{2}$$

Clearly the largest value of θ will be strictly positive and so has to be taken for some $0 < x < \infty$. (Note the strict inequalities.) This x will be a local maximum as well as a global maximum. As θ is not singular at any $0 < x < \infty$, we need only search for critical points. A careful application of the chain rule shows that the derivative is

$$\frac{d\theta}{dx} = \frac{1}{1 + (\frac{p+h}{x})^2} \cdot \left(\frac{-(p+h)}{x^2} \right) - \frac{1}{1 + (\frac{p}{x})^2} \cdot \left(\frac{-p}{x^2} \right)$$

$$= \frac{-(p+h)}{x^2 + (p+h)^2} + \frac{p}{x^2 + p^2}$$

So a critical point occurs when

$$\frac{(p+h)}{x^2 + (p+h)^2} = \frac{p}{x^2 + p^2} \qquad\qquad \text{cross multiply}$$

$$(p+h)(x^2 + p^2) = p(x^2 + (p+h)^2) \qquad\qquad \text{collect } x \text{ terms}$$

$$x^2(p+h-p) = p(p+h)^2 - p^2(p+h) \qquad\qquad \text{clean up}$$

$$hx^2 = p(p+h)(p+h-p) = ph(p+h) \qquad \text{cancel common factors}$$

$$x^2 = p(p+h)$$

$$x = \pm\sqrt{p(p+h)} \approx \pm 64.9m$$

- Thus the best place to stand approximately 64.9m in front or behind the statue. At that point $\theta \approx 0.348$ radians or $19.9°$.

Example 3.5.24

Example 3.5.25

Find the length of the longest rod that can be carried horizontally (no tilting allowed) from a corridor 3m wide into a corridor 2m wide. The two corridors are perpendicular to each other.

Solution.

- Suppose that we are carrying the rod around the corner, then if the rod is as long as possible it must touch the corner and the outside walls of both corridors. A picture of this is show below.
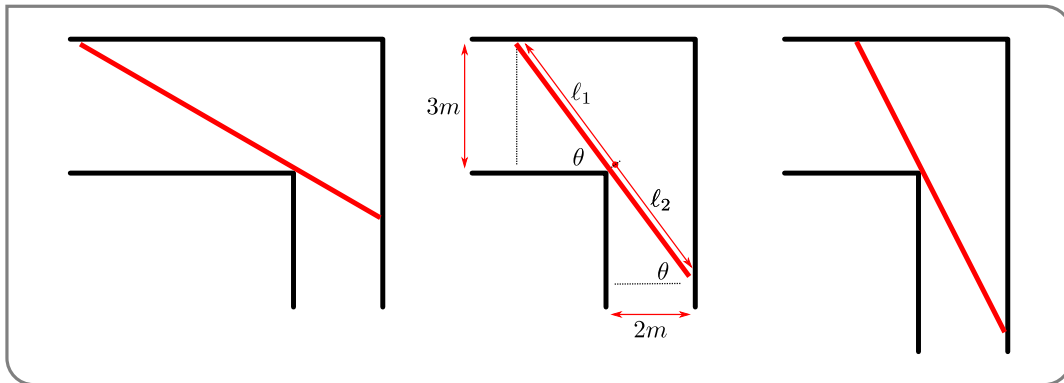
You can see that this gives rise to two similar triangles, one inside each corridor. Also the maximum length of the rod changes with the angle it makes with the walls of the corridor.

- Suppose that the angle between the rod and the inner wall of the 3m corridor is θ, as illustrated in the figure above. At the same time it will make an angle of $\frac{\pi}{2} - \theta$ with the outer wall of the 3m corridor. Denote by $\ell_1(\theta)$ the length of the part of the rod forming the hypotenuse of the upper triangle in the figure above. Similarly, denote by $\ell_2(\theta)$ the length of the part of the rod forming the hypotenuse of the lower triangle in the figure above. Then

$$\ell_1(\theta) = \frac{3}{\sin \theta} \qquad \ell_2(\theta) = \frac{2}{\cos \theta}$$

and the total length is

$$\ell(\theta) = \ell_1(\theta) + \ell_2(\theta) = \frac{3}{\sin \theta} + \frac{2}{\cos \theta}$$

where $0 \leqslant \theta \leqslant \frac{\pi}{2}$.

- The length of the longest rod we can move through the corridor in this way is the minimum of $\ell(\theta)$. Notice that $\ell(\theta)$ is not defined at $\theta = 0, \frac{\pi}{2}$. Indeed we find that as $\theta \to 0^+$ or $\theta \to \frac{\pi}{2}^-$, the length $\ell \to +\infty$. (You should be able to picture what happens to our rod in those two limits). Clearly the minimum allowed $\ell(\theta)$ is going to be finite and will be achieved for some $0 < \theta < \frac{\pi}{2}$ (note the strict inequalities) and so will be a local minimum as well as a global minimum. So we only need to find zeroes of $\ell'(\theta)$. Differentiating ℓ gives

$$\frac{\mathrm{d}\ell}{\mathrm{d}\theta} = -\frac{3\cos\theta}{\sin^2\theta} + \frac{2\sin\theta}{\cos^2\theta} = \frac{-3\cos^3\theta + 2\sin^3\theta}{\sin^2\theta\cos^2\theta}.$$

This does not exist at $\theta = 0, \frac{\pi}{2}$ (which we have already analysed) but does exist at every $0 < \theta < \frac{\pi}{2}$ and is equal to zero when the numerator is zero. Namely when

$$2\sin^3\theta = 3\cos^3\theta \qquad\qquad \text{divide by } \cos^3\theta$$
$$2\tan^3\theta = 3$$
$$\tan\theta = \sqrt[3]{\frac{3}{2}}$$

- From this we can recover $sin\theta$ and $cos\theta$, without having to compute θ itself. We can, for example, construct a right-angle triangle with adjacent length $\sqrt[3]{2}$ and opposite length $\sqrt[3]{3}$ (so that $\tan\theta = \sqrt[3]{3/2}$):
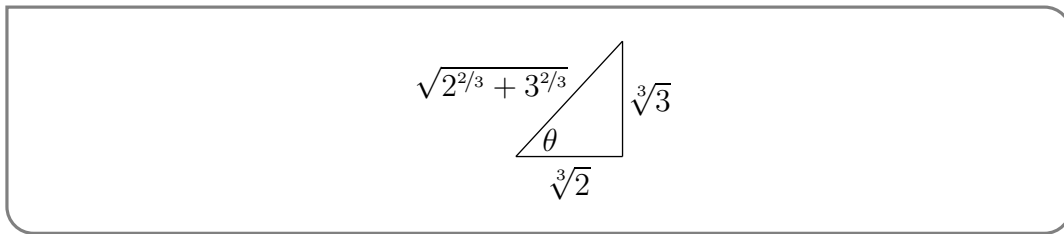
It has hypotenuse $\sqrt{3^{2/3} + 2^{2/3}}$, and so

$$\sin\theta = \frac{3^{1/3}}{\sqrt{3^{2/3} + 2^{2/3}}}$$
$$\cos\theta = \frac{2^{1/3}}{\sqrt{3^{2/3} + 2^{2/3}}}$$

Alternatively could use the identities:

$$1 + \tan^2\theta = \sec^2\theta \qquad\qquad 1 + \cot^2\theta = \csc^2\theta$$

to obtain expressions for $1/\cos\theta$ and $1/\sin\theta$.

- Using the above expressions for $\sin\theta, \cos\theta$ we find the minimum of ℓ (which is the longest rod that we can move):

$$\ell = \frac{3}{\sin\theta} + \frac{2}{\cos\theta} = \frac{3}{\frac{\sqrt[3]{3}}{\sqrt{2^{2/3}+3^{2/3}}}} + \frac{2}{\frac{\sqrt[3]{2}}{\sqrt{2^{2/3}+3^{2/3}}}}$$

$$= \sqrt{2^{2/3}+3^{2/3}}\left[3^{2/3}+2^{2/3}\right]$$

$$= \left[2^{2/3}+3^{2/3}\right]^{3/2} \approx 7.02\text{m}$$

Example 3.5.25

3.6 ▴ Sketching Graphs

One of the most obvious applications of derivatives is to help us understand the shape of the graph of a function. In this section we will use our accumulated knowledge of derivatives to identify the most important qualitative features of graphs $y = f(x)$. The goal of this section is to highlight features of the graph $y = f(x)$ that are easily

- determined from $f(x)$ itself, and

- deduced from $f'(x)$, and

- read from $f''(x)$.

We will then use the ideas to sketch several examples.

3.6.1 ▸▸ Domain, Intercepts and Asymptotes

Given a function $f(x)$, there are several important features that we can determine from that expression before examining its derivatives.

- The domain of the function — take note of values where f does not exist. If the function is rational, look for where the denominator is zero. Similarly be careful to look for roots of negative numbers or other possible sources of discontinuities.

- Intercepts — examine where the function crosses the x-axis and the y-axis by solving $f(x) = 0$ and computing $f(0)$.

- Vertical asymptotes — look for values of x at which $f(x)$ blows up. If $f(x)$ approaches either $+\infty$ or $-\infty$ as x approaches a (or possibly as x approaches a from one side) then $x = a$ is a vertical asymptote to $y = f(x)$. When $f(x)$ is a rational function (written so that common factors are cancelled), then $y = f(x)$ has vertical asymptotes at the zeroes of the denominator.

- Horizontal asymptotes — examine the limits of $f(x)$ as $x \to +\infty$ and $x \to -\infty$. Often $f(x)$ will tend to $+\infty$ or to $-\infty$ or to a finite limit L. If, for example, $\lim\limits_{x \to +\infty} f(x) = L$, then $y = L$ is a horizontal asymptote to $y = f(x)$ as $x \to \infty$.

Example 3.6.1

Consider the function

$$f(x) = \frac{x+1}{(x+3)(x-2)}$$

- We see that it is defined on all real numbers except $x = -3, +2$.

- Since $f(0) = -1/6$ and $f(x) = 0$ only when $x = -1$, the graph has y-intercept $(0, -1/6)$ and x-intercept $(-1, 0)$.

- Since the function is rational and its denominator is zero at $x = -3, +2$ it will have vertical asymptotes at $x = -3, +2$. To determine the shape around those asymptotes we need to examine the limits

$$\lim_{x \to -3} f(x) \qquad\qquad \lim_{x \to 2} f(x)$$

  Notice that when x is close to -3, the factors $(x+1)$ and $(x-2)$ are both negative, so the sign of $f(x) = \frac{x+1}{x-2} \cdot \frac{1}{x+3}$ is the same as the sign of $x + 3$. Hence

$$\lim_{x \to -3^+} f(x) = +\infty \qquad\qquad \lim_{x \to -3^-} f(x) = -\infty$$

  A similar analysis when x is near 2 gives

$$\lim_{x \to 2^+} f(x) = +\infty \qquad\qquad \lim_{x \to 2^-} f(x) = -\infty$$

- Finally since the numerator has degree 1 and the denominator has degree 2, we see that as $x \to \pm\infty$, $f(x) \to 0$. So $y = 0$ is a horizontal asymptote.

- Since we know the behaviour around the asymptotes and we know the locations of the intercepts (as shown in the left graph below), we can then join up the pieces and smooth them out to get the a good sketch of this function (below right).
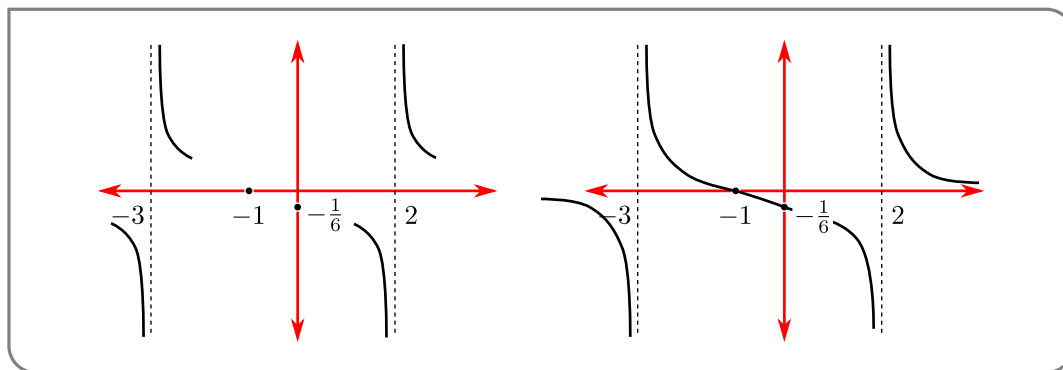
Example 3.6.1

3.6.2 ⇥ First Derivative — Increasing or Decreasing

Now we move on to the first derivative, $f'(x)$. This is a good time to revisit the mean-value theorem (Theorem 2.13.4) and some of its consequences (Corollary 2.13.11). There we considered any function $f(x)$ that is continuous on an interval $A \leqslant x \leqslant B$ and is differentiable on $A < x < B$. Then

- if $f'(x) > 0$ for all $A < x < B$, then $f(x)$ is increasing on $[A, B]$ — that is, for all $A \leqslant a < b \leqslant B$, $f(a) < f(b)$.

- if $f'(x) < 0$ for all $A < x < B$, then $f(x)$ is decreasing on $[A, B]$ — that is, for all $A \leqslant a < b \leqslant B$, $f(a) > f(b)$.

Thus the sign of the derivative indicates to us whether the function is increasing or decreasing. Further, as we discussed in Section 3.5.1, we should also examine points at which the derivative is zero — critical points — and points where the derivative does not exist. These points may indicate a local maximum or minimum.

We will now consider a function $f(x)$ that is defined on an interval I, *except possibly* at finitely many points of I. If f or its derivative f' is not defined at a point a of I, then we call a a *singular point*[62] of f.

After studying the function $f(x)$ as described above, we should compute its derivative $f'(x)$.

- Critical points — determine where $f'(x) = 0$. At a critical point, f has a horizontal tangent.

- Singular points — determine where $f'(x)$ is not defined. If $f'(x)$ approaches $\pm\infty$ as x approaches a singular point a, then f has a vertical tangent there when f approaches a finite value as x approaches a (or possibly approaches a from one side) and a vertical asymptote when $f(x)$ approaches $\pm\infty$ as x approaches a (or possibly approaches a from one side).

- Increasing and decreasing — where is the derivative positive and where is it negative. Notice that in order for the derivative to change sign, it must either pass through zero (a critical point) or have a singular point. Thus neighbouring regions of increase and decrease will be separated by critical and singular points.

Example 3.6.2

Consider the function

$$f(x) = x^4 - 6x^3$$

- Before we move on to derivatives, let us first examine the function itself as we did above.

  – As $f(x)$ is a polynomial its domain is all real numbers.

---

62 This is the extension of the definition of "singular point" that was mentioned in the footnote in Definition 3.5.6.

– Its y-intercept is at $(0,0)$. We find its x-intercepts by factoring

$$f(x) = x^4 - 6x^3 = x^3(x-6)$$

So it crosses the x-axis at $x = 0, 6$.

– Again, since the function is a polynomial it does not have any vertical asymptotes. And since

$$\lim_{x \to \pm\infty} f(x) = \lim_{x \to \pm\infty} x^4(1 - 6/x) = +\infty$$

it does not have horizontal asymptotes — it blows up to $+\infty$ as x goes to $\pm\infty$.

– We can also determine where the function is positive or negative since we know it is continuous everywhere and zero at $x = 0, 6$. Thus we must examine the intervals

$$(-\infty, 0) \qquad\qquad (0, 6) \qquad\qquad (6, \infty)$$

When $x < 0$, $x^3 < 0$ and $x - 6 < 0$ so $f(x) = x^3(x-6) = $ (negative)(negative) > 0. Similarly when $x > 6$, $x^3 > 0$, $x - 6 > 0$ we must have $f(x) > 0$. Finally when $0 < x < 6$, $x^3 > 0$ but $x - 6 < 0$ so $f(x) < 0$. Thus

interval	$(-\infty, 0)$	0	$(0, 6)$	6	$(6, \infty)$
$f(x)$	positive	0	negative	0	positive

– Based on this information we can already construct a rough sketch.
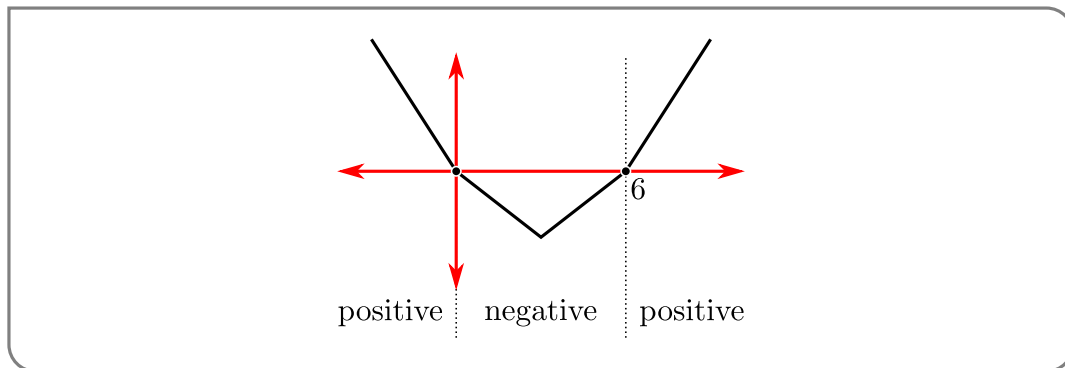
• Now we compute its derivative

$$f'(x) = 4x^3 - 18x^2 = 2x^2(2x - 9)$$

• Since the function is a polynomial, it does not have any singular points, but it does have two critical points at $x = 0, 9/2$. These two critical points split the real line into 3 open intervals

$$(-\infty, 0) \qquad\qquad (0, 9/2) \qquad\qquad (9/2, \infty)$$

We need to determine the sign of the derivative in each intervals.

– When $x < 0$, $x^2 > 0$ but $(2x - 9) < 0$, so $f'(x) < 0$ and the function is decreasing.

– When $0 < x < 9/2$, $x^2 > 0$ but $(2x - 9) < 0$, so $f'(x) < 0$ and the function is still decreasing.

– When $x > 9/2$, $x^2 > 0$ and $(2x - 9) > 0$, so $f'(x) > 0$ and the function is increasing.

We can then summarise this in the following table

interval	$(-\infty, 0)$	0	$(0, 9/2)$	$9/2$	$(9/2, \infty)$
$f'(x)$	negative	0	negative	0	positive
	decreasing	horizontal tangent	decreasing	minimum	increasing

Since the derivative changes sign from negative to positive at the critical point $x = 9/2$, this point is a minimum. Its y-value is

$$y = f(9/2) = \frac{9^3}{2^3}\left(\frac{9}{2} - 6\right)$$

$$= \frac{3^6}{2^3} \cdot \left(\frac{-3}{2}\right) = -\frac{3^7}{2^4}$$

On the other hand, at $x = 0$ the derivative does not change sign; while this point has a horizontal tangent line it is not a minimum or maximum.

• Putting this information together we arrive at a quite reasonable sketch.
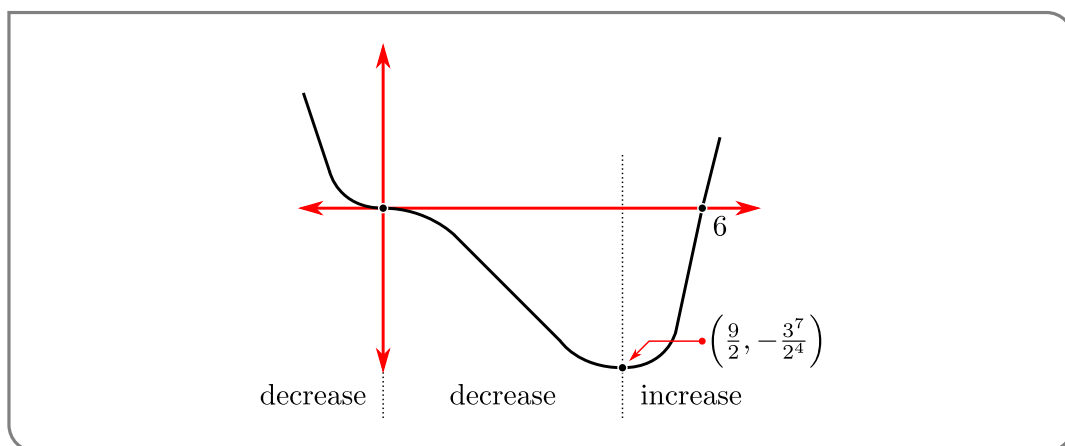
To improve upon this further we will examine the second derivative.

Example 3.6.2

3.6.3 ‣ Second Derivative — Concavity

The second derivative $f''(x)$ tells us the rate at which the derivative changes. Perhaps the easiest way to understand how to interpret the sign of the second derivative is to think about what it implies about the slope of the tangent line to the graph of the function. Consider the following sketches of $y = 1 + x^2$ and $y = -1 - x^2$.

$$f''(x) > 0 \qquad\qquad f''(x) < 0$$
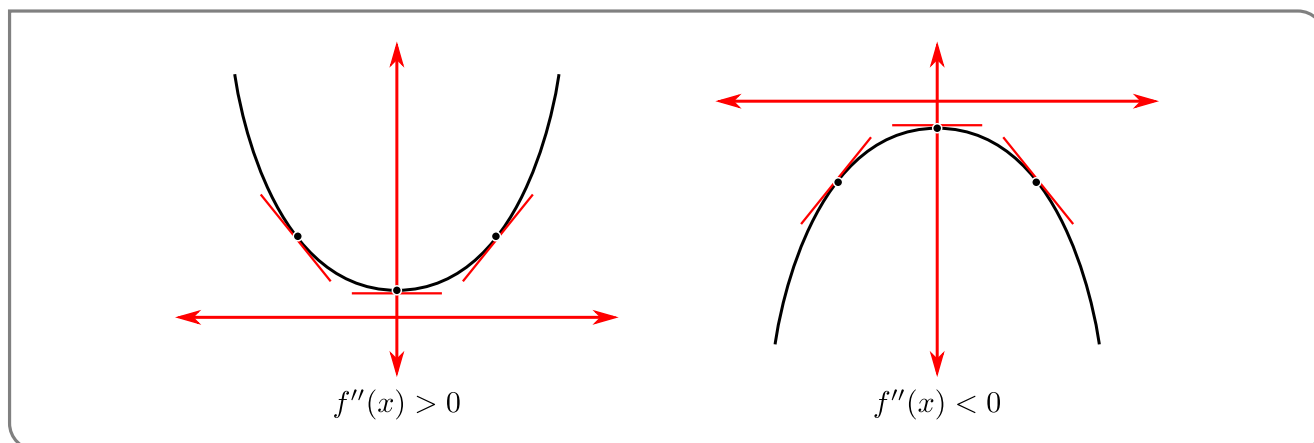
- In the case of $y = f(x) = 1 + x^2$, $f''(x) = 2 > 0$. Notice that this means the slope, $f'(x)$, of the line tangent to the graph at x increases as x increases. Looking at the figure on the left above, we see that the graph always lies above the tangent lines.

- For $y = f(x) = -1 - x^2$, $f''(x) = -2 < 0$. The slope, $f'(x)$, of the line tangent to the graph at x decreases as x increases. Looking at the figure on the right above, we see that the graph always lies below the tangent lines.

Similarly consider the following sketches of $y = x^{-1/2}$ and $y = \sqrt{4 - x}$:

$$y = \frac{1}{\sqrt{x}} \qquad\qquad y = \sqrt{4 - x}$$

$$f''(x) > 0 \qquad\qquad f''(x) < 0$$
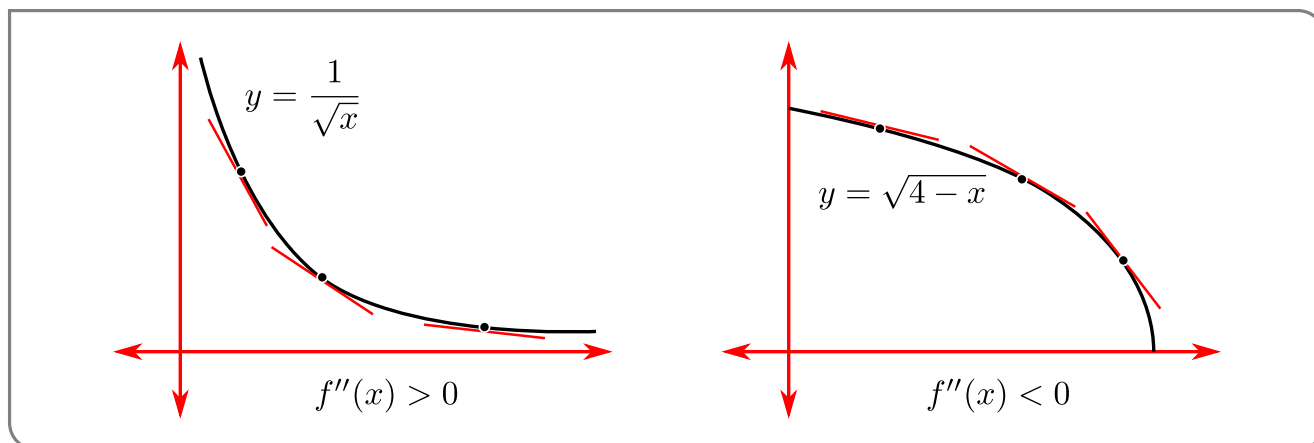
Both of their derivatives, $-\frac{1}{2}x^{-3/2}$ and $-\frac{1}{2}(4 - x)^{-1/2}$, are negative, so they are decreasing functions. Examining second derivatives shows some differences.

- For the first function, $y''(x) = \frac{3}{4}x^{-5/2} > 0$, so the slopes of tangent lines are increasing with x and the graph lies above its tangent lines.

- However, the second function has $y''(x) = -\frac{1}{4}(4-x)^{-3/2} < 0$ so the slopes of the tangent lines are decreasing with x and the graph lies below its tangent lines.

More generally

Definition 3.6.3.

Let $f(x)$ be a continuous function on the interval $[a, b]$ and suppose its first and second derivatives exist on that interval.

- If $f''(x) > 0$ for all $a < x < b$, then the graph of f lies above its tangent lines for $a < x < b$ and it is said to be concave up.
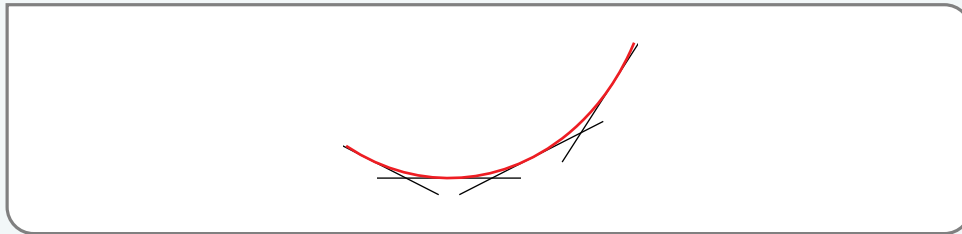
- If $f''(x) < 0$ for all $a < x < b$, then the graph of f lies below its tangent lines for $a < x < b$ and it is said to be concave down.
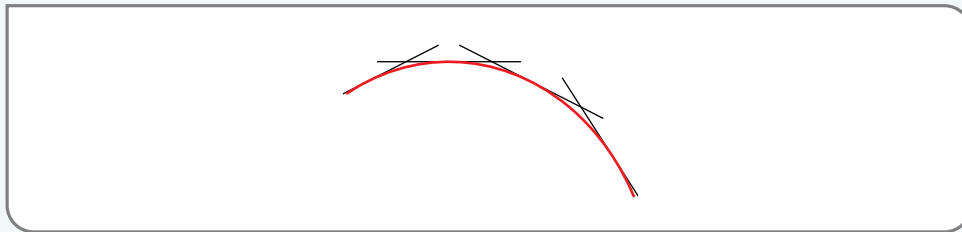
- If $f''(c) = 0$ for some $a < c < b$, and the concavity of f changes across $x = c$, then we call $(c, f(c))$ an inflection point.
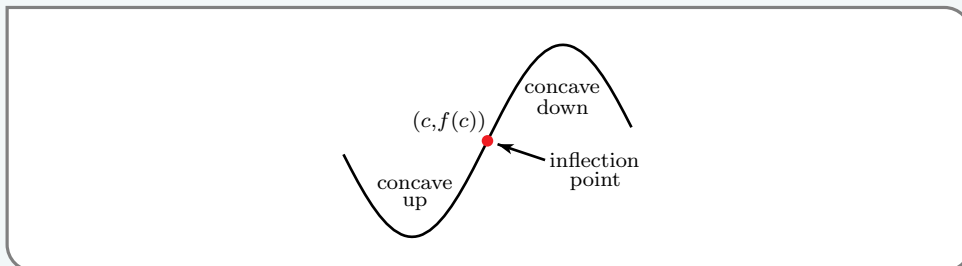
Note that one might also see the terms

- "convex" or "convex up" used in place of "concave up", and

- "concave" or "convex down" used to mean "concave down".

To avoid confusion we recommend the reader stick with the terms "concave up" and "concave down".

Let's now continue Example 3.6.2 by discussing the concavity of the curve.

Example 3.6.4 (Continuation of Example 3.6.2)

Consider again the function

$$f(x) = x^4 - 6x^3$$

- Its first derivative is $f'(x) = 4x^3 - 18x^2$, so

$$f''(x) = 12x^2 - 36x = 12x(x - 3)$$

- Thus the second derivative is zero (and potentially changes sign) at $x = 0, 3$. Thus we should consider the sign of the second derivative on the following intervals

$$(-\infty, 0) \qquad\qquad (0, 3) \qquad\qquad (3, \infty)$$

A little algebra gives us

interval	$(-\infty, 0)$	0	$(0, 3)$	3	$(3, \infty)$
$f''(x)$	positive	0	negative	0	positive
concavity	up	inflection	down	inflection	up

Since the concavity changes at both $x = 0$ and $x = 3$, the following are inflection points

$$(0, 0) \qquad\qquad (3, 3^4 - 6 \times 3^3) = (3, -3^4)$$

- Putting this together with the information we obtained earlier gives us the following sketch
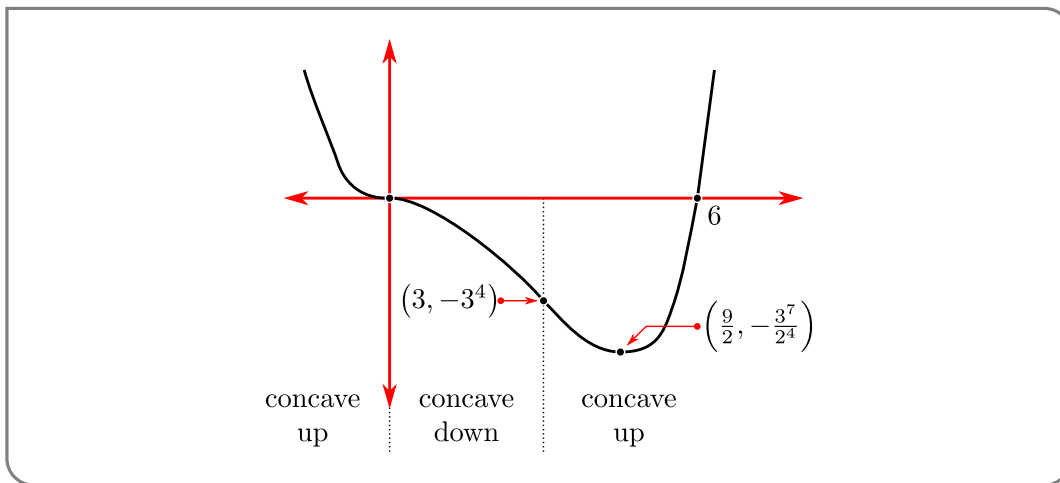
Example 3.6.4

Example 3.6.5 (Optional — $y = x^{1/3}$ and $y = x^{2/3}$)

In our Definition 3.6.3, concerning concavity and inflection points, we considered only functions having first and second derivatives on the entire interval of interest. In this example, we will consider the functions

$$f(x) = x^{1/3} \qquad g(x) = x^{2/3}$$

We shall see that $x = 0$ is a singular point for both of those functions. There is no universal agreement as to precisely when a singular point should also be called an inflection point. We choose to extend our definition of inflection point in Definition 3.6.3 as follows. If

- the function $f(x)$ is defined and continuous on an interval $a < x < b$ and if

- the first and second derivatives $f'(x)$ and $f''(x)$ exist on $a < x < b$ except possibly at the single point $a < c < b$ and if

- f is concave up on one side of c and is concave down on the other side of c

then we say that $(c, f(c))$ is an inflection point of $y = f(x)$. Now let's check out $y = f(x)$ and $y = g(x)$ from this point of view.

(1) Features of $y = f(x)$ and $y = g(x)$ that are read off of $f(x)$ and $g(x)$:

- Since $f(0) = 0^{1/3} = 0$ and $g(0) = 0^{2/3} = 0$, the origin $(0,0)$ lies on both $y = f(x)$ and $y = g(x)$.

- For example, $1^3 = 1$ and $(-1)^3 = -1$ so that the cube root of 1 is $1^{1/3} = 1$ and the cube root of -1 is $(-1)^{1/3} = -1$. In general,

$$x^{1/3} \begin{cases} < 0 & \text{if } x < 0 \\ = 0 & \text{if } x = 0 \\ > 0 & \text{if } x > 0 \end{cases}$$

  Consequently the graph $y = f(x) = x^{1/3}$ lies below the x-axis when $x < 0$ and lies above the x-axis when $x > 0$. On the other hand, the graph $y = g(x) = x^{2/3} = \left[x^{1/3}\right]^2$ lies on or above the x-axis for all x.

- As $x \to +\infty$, both $y = f(x) = x^{1/3}$ and $y = g(x) = x^{2/3}$ tend to $+\infty$.
- As $x \to -\infty$, $y = f(x) = x^{1/3}$ tends to $-\infty$ and $y = g(x) = x^{2/3}$ tends to $+\infty$.

(2) Features of $y = f(x)$ and $y = g(x)$ that are read off of $f'(x)$ and $g'(x)$:

$$f'(x) = \begin{cases} \frac{1}{3}x^{-2/3} & \text{if } x \neq 0 \\ \text{undefined} & \text{if } x = 0 \end{cases} \implies f'(x) > 0 \text{ for all } x \neq 0$$

$$g'(x) = \begin{cases} \frac{2}{3}x^{-1/3} & \text{if } x \neq 0 \\ \text{undefined} & \text{if } x = 0 \end{cases} \implies g'(x) \begin{cases} < 0 & \text{if } x < 0 \\ > 0 & \text{if } x > 0 \end{cases}$$

So the graph $y = f(x)$ is increasing on both sides of the singular point $x = 0$, while the graph $y = g(x)$ is decreasing to the left of $x = 0$ and is increasing to the right of $x = 0$. As $x \to 0$, $f'(x)$ and $g'(x)$ become infinite. That is, the slopes of the tangent lines at $(x, f(x))$ and $(x, g(x))$ become infinite and the tangent lines become vertical.

(3) Features of $y = f(x)$ and $y = g(x)$ that are read off of $f''(x)$ and $g''(x)$:

$$f''(x) = \begin{cases} -\frac{2}{9}x^{-5/3} = -\frac{2}{9}\left[x^{-1/3}\right]^5 & \text{if } x \neq 0 \\ \text{undefined} & \text{if } x = 0 \end{cases} \implies f''(x) \begin{cases} > 0 & \text{if } x < 0 \\ < 0 & \text{if } x > 0 \end{cases}$$

$$g''(x) = \begin{cases} -\frac{2}{9}x^{-4/3} = -\frac{2}{9}\left[x^{-1/3}\right]^4 & \text{if } x \neq 0 \\ \text{undefined} & \text{if } x = 0 \end{cases} \implies g''(x) < 0 \text{ for all } x \neq 0$$

So the graph $y = g(x)$ is concave down on both sides of the singular point $x = 0$, while the graph $y = f(x)$ is concave up to the left of $x = 0$ and is concave down to the right of $x = 0$.

By way of summary, we have, for $f(x)$,

interval	$(-\infty, 0)$	0	$(0, \infty)$
$f(x)$	negative	0	positive
$f'(x)$	positive	undefined	positive
	increasing		increasing
$f''(x)$	positive	undefined	negative
	concave up	inflection	concave down

and for $g(x)$,

interval	$(-\infty, 0)$	0	$(0, \infty)$
$g(x)$	positive	0	positive
$g'(x)$	negative	undefined	positive
	decreasing		increasing
$g''(x)$	negative	undefined	negative
	concave down		concave down

Since the concavity changes at $x = 0$ for $y = f(x)$, but not for $y = g(x)$, $(0,0)$ is an inflection point for $y = f(x)$, but not for $y = g(x)$. We have the following sketch for $y = f(x) = x^{1/3}$,
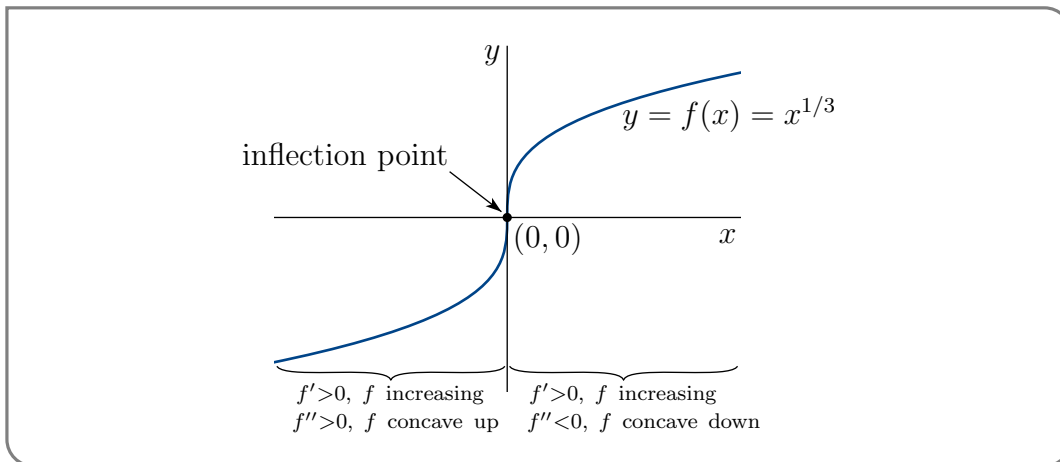
and the following sketch for $y = g(x) = x^{2/3}$.
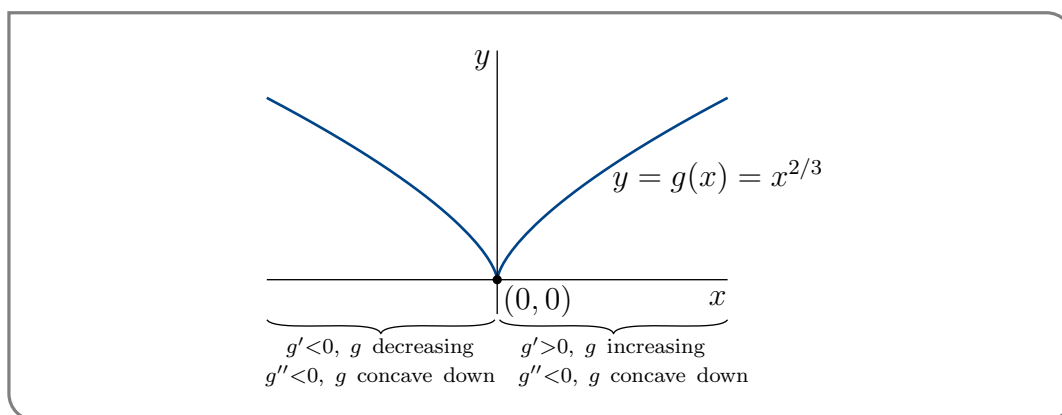
Note that the curve $y = f(x) = x^{1/3}$ looks perfectly smooth, even though $f'(x) \to \infty$ as $x \to 0$. There is no kink or discontinuity at $(0,0)$. The singularity at $x = 0$ has caused the y-axis to be a vertical tangent to the curve, but has not prevented the curve from looking smooth.

Example 3.6.5

3.6.4 ▸▸ Symmetries

Before we proceed to some examples, we should examine some simple symmetries possessed by some functions. We'll look at three symmetries — evenness, oddness and periodicity. If a function possesses one of these symmetries then it can be exploited to reduce the amount of work required to sketch the graph of the function.

Let us start with even and odd functions.

Definition 3.6.6.

A function $f(x)$ is said to be even if $f(-x) = f(x)$ for all x.

Definition 3.6.7.

A function $f(x)$ is said to be odd if $f(-x) = -f(x)$ for all x.

Example 3.6.8

Let $f(x) = x^2$ and $g(x) = x^3$. Then

$$f(-x) = (-x)^2 = x^2 = f(x)$$
$$g(-x) = (-x)^3 = -x^3 = -g(x)$$

Hence $f(x)$ is even and $g(x)$ is odd.

Notice any polynomial involving only even powers of x will be even

$$f(x) = 7x^6 + 2x^4 - 3x^2 + 5 \qquad\qquad \text{remember that } 5 = 5x^0$$
$$f(-x) = 7(-x)^6 + 2(-x)^4 - 3(-x)^2 + 5$$
$$= 7x^6 + 2x^4 - 3x^2 + 5 = f(x)$$

Similarly any polynomial involving only odd powers of x will be odd

$$g(x) = 2x^5 - 8x^3 - 3x$$
$$g(-x) = 2(-x)^5 - 8(-x)^3 - 3(-x)$$
$$= -2x^5 + 8x^3 + 3x = -g(x)$$

Example 3.6.8

Not all even and odd functions are polynomials. For example

$$|x| \qquad\qquad \cos x \qquad\qquad \text{and } (e^x + e^{-x})$$

are all even, while

$$\sin x \qquad\qquad \tan x \qquad\qquad \text{and } (e^x - e^{-x})$$

are all odd. Indeed, given any function $f(x)$, the function

$$g(x) = f(x) + f(-x) \qquad\qquad \text{will be even, and}$$
$$h(x) = f(x) - f(-x) \qquad\qquad\qquad \text{will be odd.}$$

Now let us see how we can make use of these symmetries to make graph sketching easier. Let $f(x)$ be an even function. Then

the point (x_0, y_0) lies on the graph of $y = f(x)$

if and only if $y_0 = f(x_0) = f(-x_0)$ which is the case if and only if

the point $(-x_0, y_0)$ lies on the graph of $y = f(x)$.
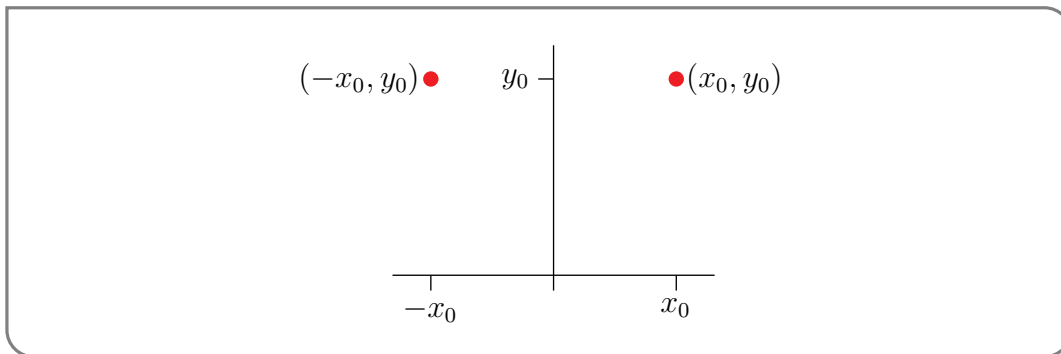
Notice that the points (x_0, y_0) and $(-x_0, y_0)$ are just reflections of each other across the y-axis. Consequently, to draw the graph $y = f(x)$, it suffices to draw the part of the graph with $x \geqslant 0$ and then reflect it in the y–axis. Here is an example. The part with $x \geqslant 0$ is on the left and the full graph is on the right.
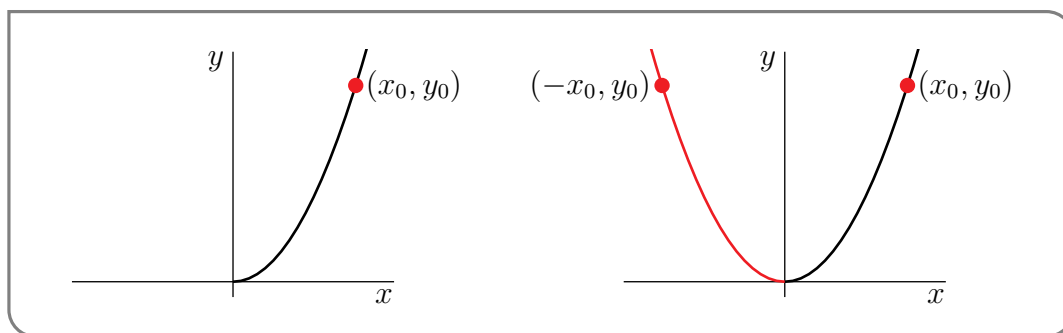
Very similarly, when $f(x)$ is an odd function then

$$(x_0, y_0) \text{ lies on the graph of } y = f(x)$$

if and only if

$$(-x_0, -y_0) \text{ lies on the graph of } y = f(x)$$
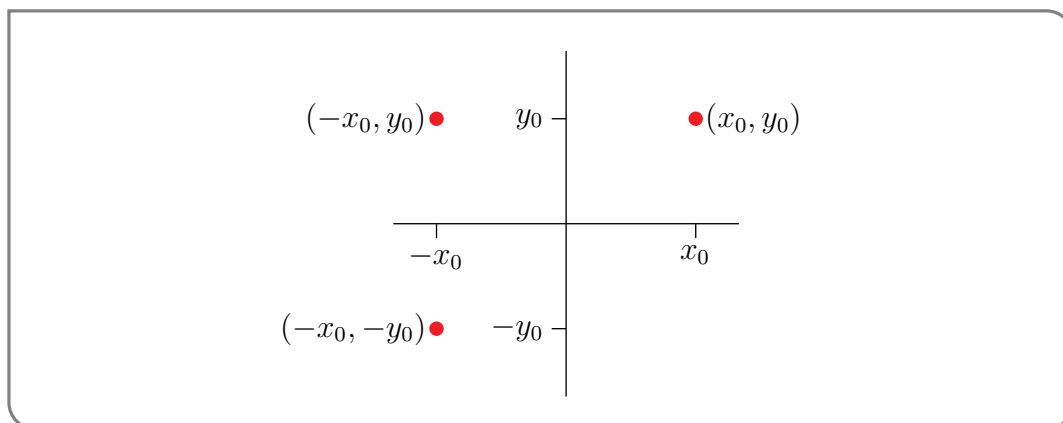
Now the symmetry is a little harder to interpret pictorially. To get from (x_0, y_0) to $(-x_0, -y_0)$ one can first reflect (x_0, y_0) in the y–axis to get to $(-x_0, y_0)$ and then reflect the result in the x–axis to get to $(-x_0, -y_0)$. Consequently, to draw the graph $y = f(x)$, it suffices to draw the part of the graph with $x \geqslant 0$ and then reflect it first in the y–axis and then in the x–axis. Here is an example. First, here is the part of the graph with $x \geqslant 0$.
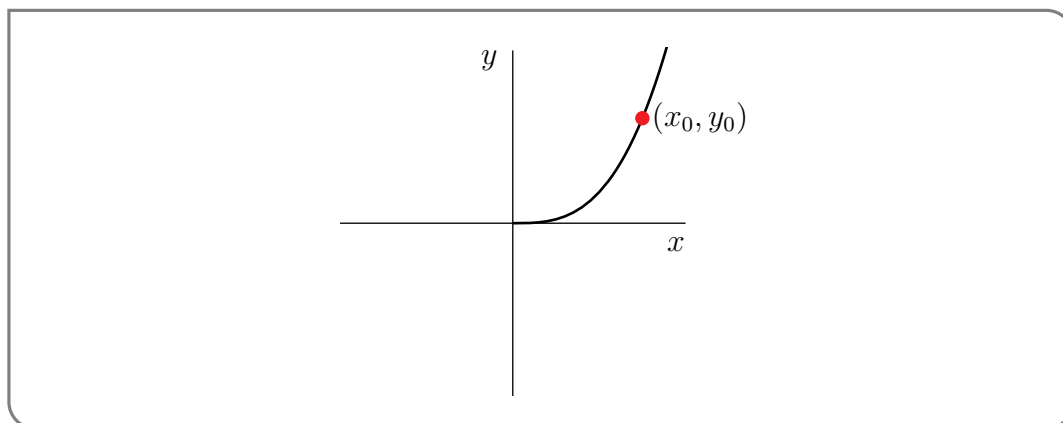
Next, as an intermediate step (usually done in our heads rather than on paper), we add in the reflection in the y–axis.
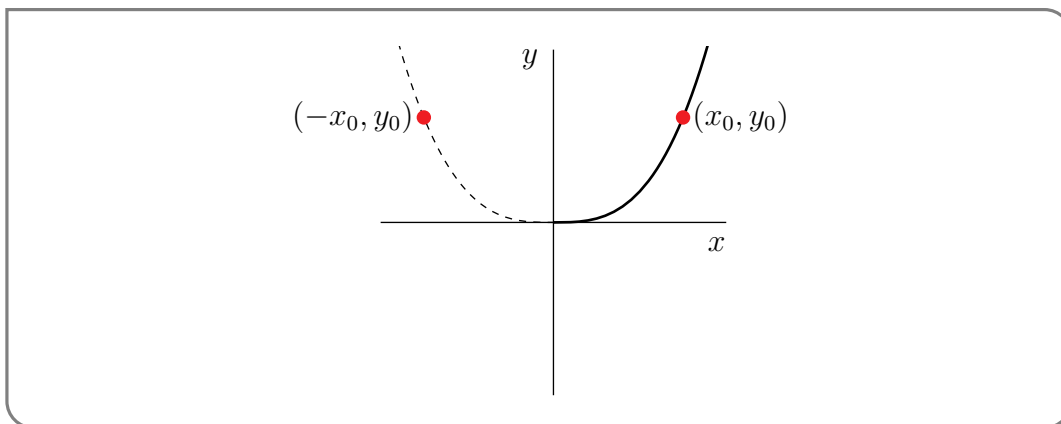
Finally to get the full graph, we reflect the dashed line in the x–axis
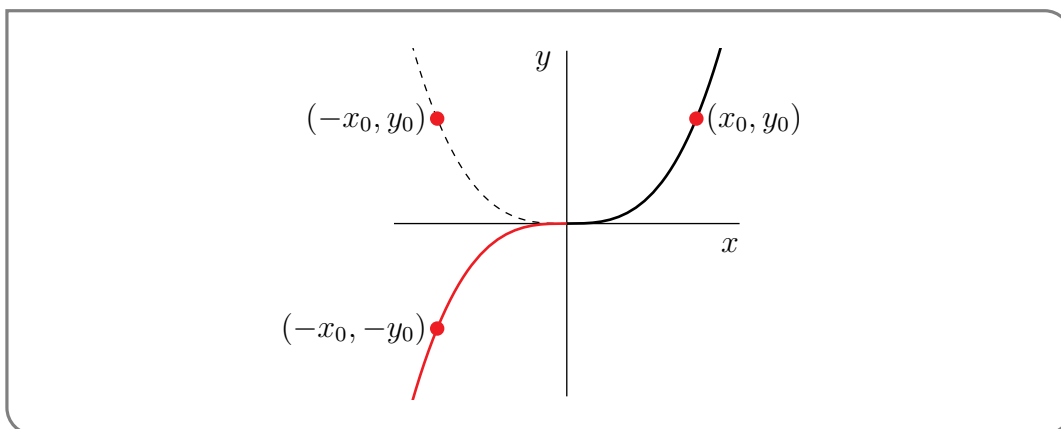
and then remove the dashed line.
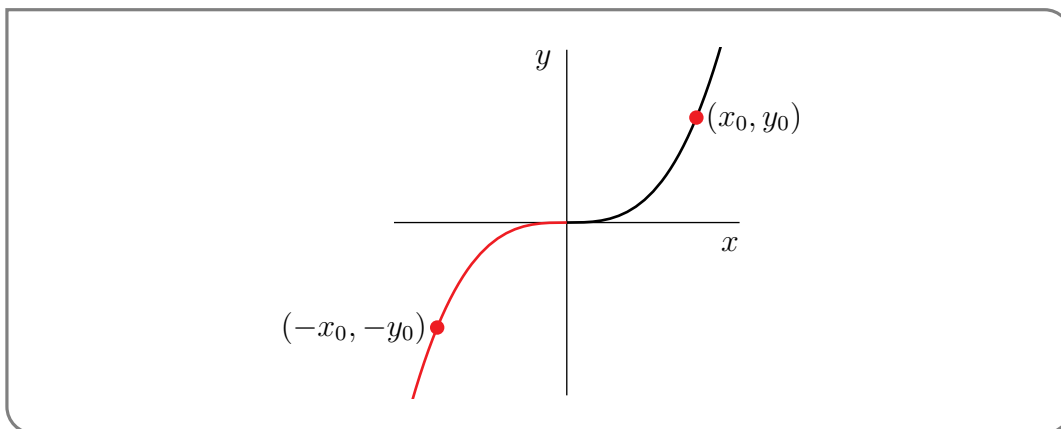
Let's do a more substantial example of an even function

Example 3.6.9

Consider the function

$$g(x) = \frac{x^2 - 9}{x^2 + 3}$$

- The function is even since

$$g(-x) = \frac{(-x)^2 - 9}{(-x)^2 + 3} = \frac{x^2 - 9}{x^2 + 3} = g(x)$$

  Thus it suffices to study the function for $x \geqslant 0$ because we can then use the even symmetry to understand what happens for $x < 0$.

- The function is defined on all real numbers since its denominator $x^2 + 3$ is never zero. Hence it has no vertical asymptotes.

- The y-intercept is $g(0) = \frac{-9}{3} = -3$. And x-intercepts are given by the solution of $x^2 - 9 = 0$, namely $x = \pm 3$. Note that we only need to establish $x = 3$ as an intercept. Then since g is even, we know that $x = -3$ is also an intercept.

- To find the horizontal asymptotes we compute the limit as $x \to +\infty$

$$\begin{aligned}
\lim_{x \to \infty} g(x) &= \lim_{x \to \infty} \frac{x^2 - 9}{x^2 + 3} \\
&= \lim_{x \to \infty} \frac{x^2(1 - 9/x^2)}{x^2(1 + 3/x^2)} \\
&= \lim_{x \to \infty} \frac{1 - 9/x^2}{1 + 3/x^2} = 1
\end{aligned}$$

  Thus $y = 1$ is a horizontal asymptote. Indeed, this is also the asymptote as $x \to -\infty$ since by the even symmetry

$$\lim_{x \to -\infty} g(x) = \lim_{x \to \infty} g(-x) = \lim_{x \to \infty} g(x).$$

- We can already produce a quite reasonable sketch just by putting in the horizontal asymptote and the intercepts and drawing a smooth curve between them.
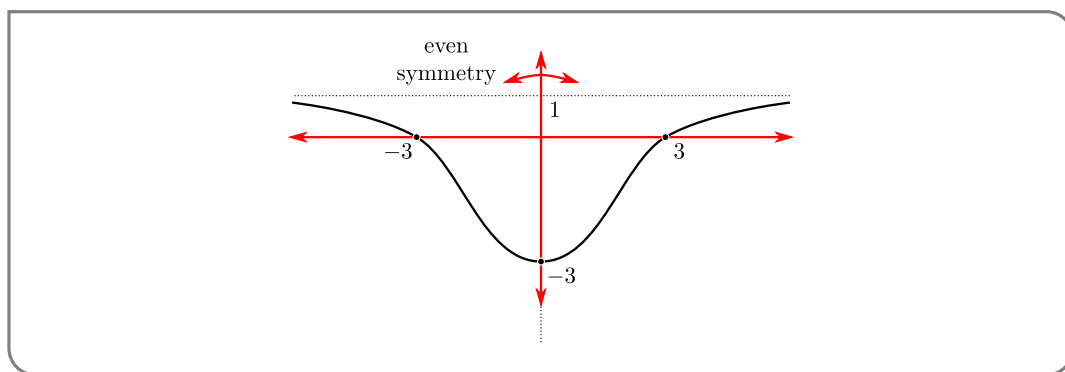
Note that we have drawn the function as never crossing the asymptote $y = 1$, however we have not yet proved that. We could by trying to solve $g(x) = 1$.

$$\begin{aligned}
\frac{x^2 - 9}{x^2 + 3} &= 1 \\
x^2 - 9 &= x^2 + 3 \\
-9 &= 3 \text{ so no solutions.}
\end{aligned}$$

Alternatively we could analyse the first derivative to see how the function approaches the asymptote.

- Now we turn to the first derivative:

$$g'(x) = \frac{(x^2 + 3)(2x) - (x^2 - 9)(2x)}{(x^2 + 3)^2}$$

$$= \frac{24x}{(x^2 + 3)^2}$$

There are no singular points since the denominator is nowhere zero. The only critical point is at $x = 0$. Thus we must find the sign of $g'(x)$ on the intervals

$$(-\infty, 0) \qquad\qquad\qquad (0, \infty)$$

- When $x > 0$, $24x > 0$ and $(x^2 + 3) > 0$, so $g'(x) > 0$ and the function is increasing. By even symmetry we know that when $x < 0$ the function must be decreasing. Hence the critical point $x = 0$ is a local minimum of the function.

- Notice that since the function is increasing for $x > 0$ and the function must approach the horizontal asymptote $y = 1$ from below. Thus the sketch above is quite accurate.

- Now consider the second derivative:

$$g''(x) = \frac{d}{dx} \frac{24x}{(x^2 + 3)^2}$$

$$= \frac{(x^2 + 3)^2 \cdot 24 - 24x \cdot 2(x^2 + 3) \cdot 2x}{(x^2 + 3)^4} \qquad \text{cancel a factor of } (x^2 + 3)$$

$$= \frac{(x^2 + 3) \cdot 24 - 96x^2}{(x^2 + 3)^3}$$

$$= \frac{72(1 - x^2)}{(x^2 + 3)^3}$$

- It is clear that $g''(x) = 0$ when $x = \pm 1$. Note that, again, we can infer the zero at $x = -1$ from the zero at $x = 1$ by the even symmetry. Thus we need to examine the sign of $g''(x)$ the intervals

$$(-\infty, -1) \qquad\qquad (-1, 1) \qquad\qquad (1, \infty)$$

- When $|x| < 1$ we have $(1 - x^2) > 0$ so that $g''(x) > 0$ and the function is concave up. When $|x| > 1$ we have $(1 - x^2) < 0$ so that $g''(x) < 0$ and the function is concave down. Thus the points $x = \pm 1$ are inflection points. Their coordinates are $(\pm 1, g(\pm 1)) = (\pm 1, -2)$.

- Putting this together gives the following sketch:

concave
down

concave
up

concave
down
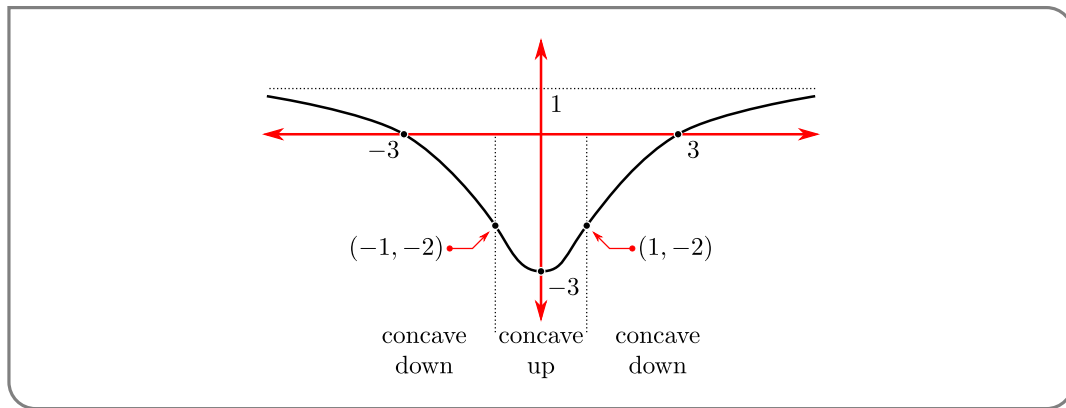
Another symmetry we should consider is periodicity.

Definition 3.6.10.

A function $f(x)$ is said to be periodic, with period $P > 0$, if $f(x + P) = f(x)$ for all x.

Note that if $f(x + P) = f(x)$ for all x, then replacing x by $x + P$, we have

$$f(x + 2P) = f(x + P + P) = f(x + P) = f(x).$$

More generally $f(x + kP) = f(x)$ for all integers k. Thus if f has period P, then it also has period nP for all natural numbers n. The smallest period is called the fundamental period.

The classic example of a periodic function is $f(x) = \sin x$, which has period 2π since $f(x + 2\pi) = \sin(x + 2\pi) = \sin x = f(x)$.

If $f(x)$ has period P then

$$(x_0, y_0) \text{ lies on the graph of } y = f(x)$$

if and only if $y_0 = f(x_0) = f(x_0 + P)$ which is the case if and only if

$$(x_0 + P, y_0) \text{ lies on the graph of } y = f(x)$$

and, more generally,

$$(x_0, y_0) \text{ lies on the graph of } y = f(x)$$

if and only if

$$(x_0 + nP, y_0) \text{ lies on the graph of } y = f(x)$$

for all integers n.

Note that the point $(x_0 + P, y_0)$ can be obtained by translating (x_0, y_0) horizontally by P. Similarly the point $(x_0 + nP, y_0)$ can be found by repeatedly translating (x_0, y_0) horizontally by P.
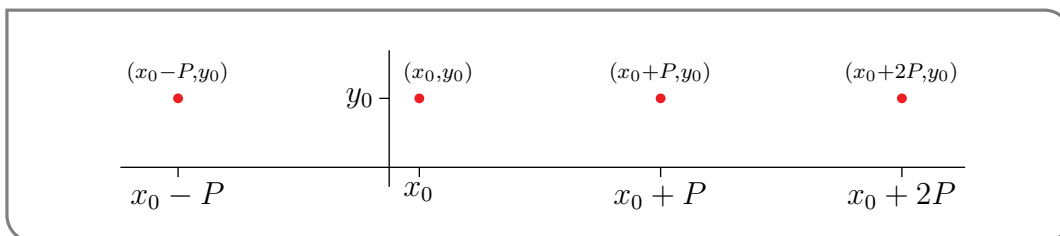
Consequently, to draw the graph $y = f(x)$, it suffices to draw one period of the graph, say the part with $0 \leqslant x \leqslant P$, and then translate it repeatedly. Here is an example. Here is a sketch of one period
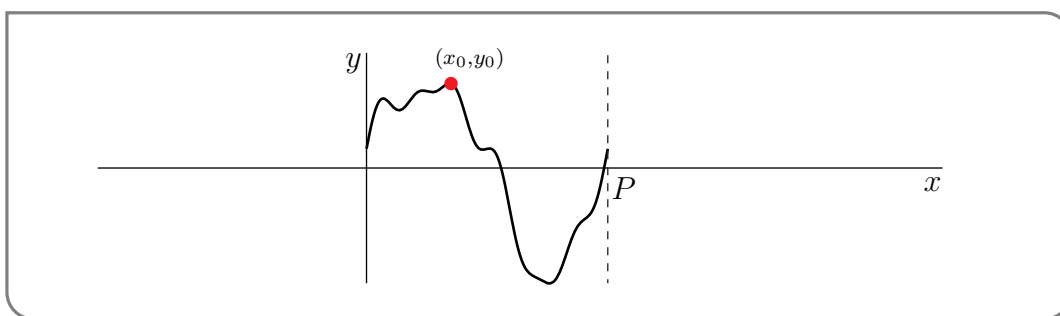
and here is the full sketch.
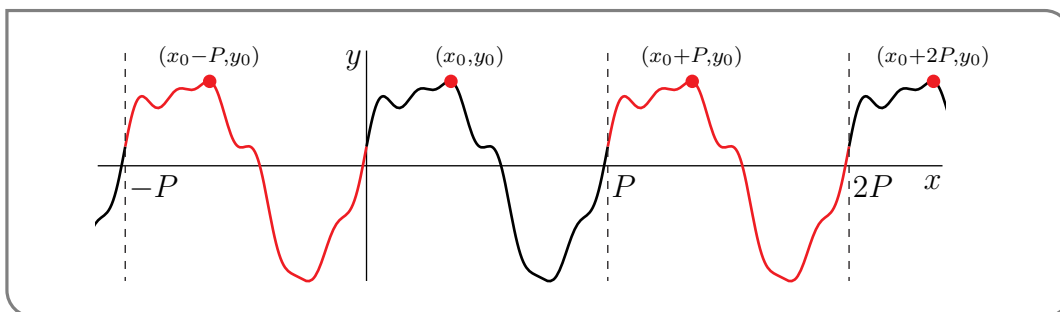
3.6.5 ▸ A Checklist for Sketching

Above we have described how we can use our accumulated knowledge of derivatives to quickly identify the most important qualitative features of graphs $y = f(x)$. Here we give the reader a quick checklist of things to examine in order to produce an accurate sketch based on properties that are easily read off from $f(x)$, $f'(x)$ and $f''(x)$.

▸▸▸ **A Sketching Checklist.**

(1) Features of $y = f(x)$ that are read off of $f(x)$:

- First check where $f(x)$ is defined. Then

- $y = f(x)$ is plotted only for x's in the domain of $f(x)$, i.e. where $f(x)$ is defined.

- $y = f(x)$ has vertical asymptotes at the points where $f(x)$ blows up to $\pm\infty$.

- Next determine whether the function is even, odd, or periodic.

- $y = f(x)$ is first plotted for $x \geqslant 0$ if the function is even or odd. The rest of the sketch is then created by reflections.

- $y = f(x)$ is first plotted for a single period if the function is periodic. The rest of the sketch is then created by translations.

- Next compute $f(0)$, $\lim_{x \to \infty} f(x)$ and $\lim_{x \to -\infty} f(x)$ and look for solutions to $f(x) = 0$ that you can easily find. Then

- $y = f(x)$ has y–intercept $(0, f(0))$.

- $y = f(x)$ has x–intercept $(a, 0)$ whenever $f(a) = 0$

- $y = f(x)$ has horizontal asymptote $y = Y$ if $\lim_{x \to \infty} f(x) = L$ or $\lim_{x \to -\infty} f(x) = L$.

(2) Features of $y = f(x)$ that are read off of $f'(x)$:

- Compute $f'(x)$ and determine its critical points and singular points, then

- $y = f(x)$ has a horizontal tangent at the points where $f'(x) = 0$.

- $y = f(x)$ is increasing at points where $f'(x) > 0$.

- $y = f(x)$ is decreasing at points where $f'(x) < 0$.

- $y = f(x)$ has vertical tangents or vertical asymptotes at the points where $f'(x) = \pm\infty$.

(3) Features of $y = f(x)$ that are read off of $f''(x)$:

- Compute $f''(x)$ and determine where $f''(x) = 0$ or does not exist, then

- $y = f(x)$ is concave up at points where $f''(x) > 0$.

- $y = f(x)$ is concave down at points where $f''(x) < 0$.

- $y = f(x)$ may or may not have inflection points where $f''(x) = 0$.

3.6.6 ▸▸ Sketching Examples

Example 3.6.12 (Sketch $f(x) = x^3 - 3x + 1$)

(1) Reading from $f(x)$:

- The function is a polynomial so it is defined everywhere.

- Since $f(-x) = -x^3 + 3x + 1 \neq \pm f(x)$, it is not even or odd. Nor is it periodic.

- The y-intercept is $y = 1$. The x-intercepts are not easily computed since it is a cubic polynomial that does not factor nicely[63]. So for this example we don't worry about finding them.

- Since it is a polynomial it has no vertical asymptotes.

- For very large x, both positive and negative, the x^3 term in $f(x)$ dominates the other two terms so that

$$f(x) \to \begin{cases} +\infty & \text{as } x \to +\infty \\ -\infty & \text{as } x \to -\infty \end{cases}$$

and there are no horizontal asymptotes.

(2) We now compute the derivative:

$$f'(x) = 3x^2 - 3 = 3(x^2 - 1) = 3(x+1)(x-1)$$

- The critical points (where $f'(x) = 0$) are at $x = \pm 1$. Further since the derivative is a polynomial it is defined everywhere and there are no singular points. The critical points split the real line into the intervals $(-\infty, -1), (-1, 1)$ and $(1, \infty)$.

- When $x < -1$, both factors $(x+1), (x-1) < 0$ so $f'(x) > 0$.

- Similarly when $x > 1$, both factors $(x+1), (x-1) > 0$ so $f'(x) > 0$.

- When $-1 < x < 1$, $(x-1) < 0$ but $(x+1) > 0$ so $f'(x) < 0$.

- Summarising all this

	$(-\infty, -1)$	-1	(-1,1)	1	$(1, \infty)$
$f'(x)$	positive	0	negative	0	positive
	increasing	maximum	decreasing	minimum	increasing

So $(-1, f(-1)) = (-1, 3)$ is a local maximum and $(1, f(1)) = (1, -1)$ is a local minimum.

(3) Compute the second derivative:

$$f''(x) = 6x$$

- The second derivative is zero when $x = 0$, and the problem is quite easy to analyse. Clearly, $f''(x) < 0$ when $x < 0$ and $f''(x) > 0$ when $x > 0$.

- Thus f is concave down for $x < 0$, concave up for $x > 0$ and has an inflection point at $x = 0$.

Putting this all together gives:

---

63 With the aid of a computer we can find the x-intercepts numerically: $x \approx -1.879385242, 0.3472963553$, and 1.532088886. If you are interested in more details check out Appendix C.

$$y = x^3 - 3x + 1$$

$(-1, 3)$

$(0, 1)$

x

$(1, -1)$

$f' > 0, \ f$ increasing $f' < 0, \ f$ decreasing $f' > 0, \ f$ increasing

$f'' < 0, \ f$ convex down $f'' > 0, \ f$ convex up
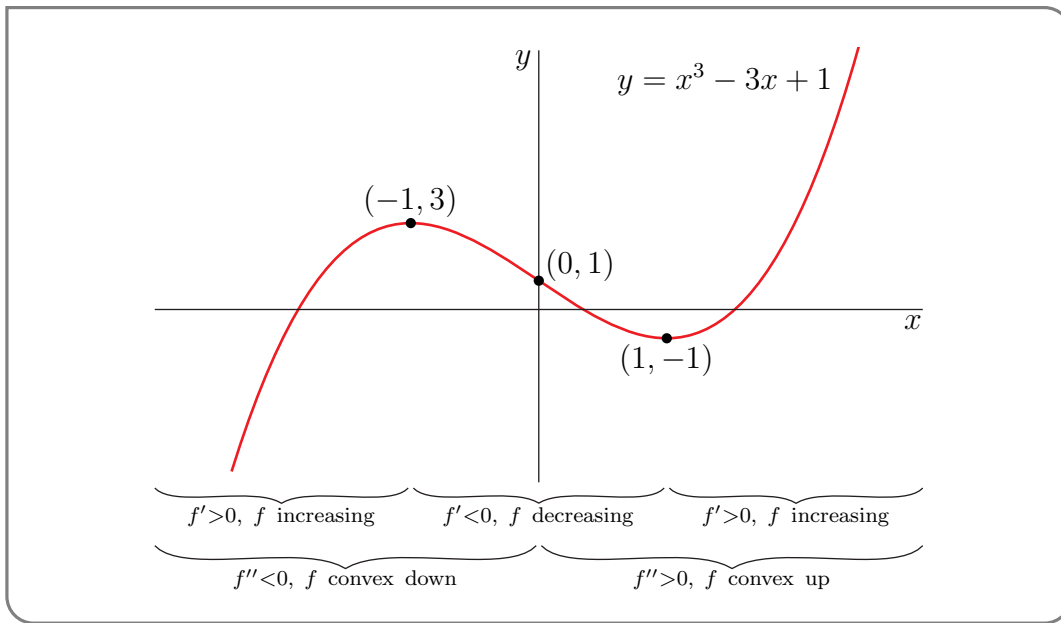
Example 3.6.12

Example 3.6.13 (Sketch $f(x) = x^4 - 4x^3$)

(1) Reading from $f(x)$:

- The function is a polynomial so it is defined everywhere.
- Since $f(-x) = x^4 + 4x^3 \neq \pm f(x)$, it is not even or odd. Nor is it periodic.
- The y-intercept is $y = f(0) = 0$, while the x-intercepts are given by the solution of

$$f(x) = x^4 - 4x^3 = 0$$
$$x^3(x - 4) = 0$$

Hence the x-intercepts are $0, 4$.

- Since f is a polynomial it does not have any vertical asymptotes.
- For very large x, both positive and negative, the x^4 term in $f(x)$ dominates the other term so that

$$f(x) \to \begin{cases} +\infty & \text{as } x \to +\infty \\ +\infty & \text{as } x \to -\infty \end{cases}$$

and the function has no horizontal asymptotes.

(2) Now compute the derivative $f'(x)$:

$$f'(x) = 4x^3 - 12x^2 = 4(x - 3)x^2$$

- The critical points are at $x = 0, 3$. Since the function is a polynomial there are no singular points. The critical points split the real line into the intervals $(-\infty, 0)$, $(0, 3)$ and $(3, \infty)$.
- When $x < 0$, $x^2 > 0$ and $x - 3 < 0$, so $f'(x) < 0$.
- When $0 < x < 3$, $x^2 > 0$ and $x - 3 < 0$, so $f'(x) < 0$.
- When $3 < x$, $x^2 > 0$ and $x - 3 > 0$, so $f'(x) > 0$.
- Summarising all this

	$(-\infty, 0)$	0	$(0,3)$	3	$(3, \infty)$
$f'(x)$	negative	0	negative	0	positive
	decreasing	horizontal tangent	decreasing	minimum	increasing

So the point $(3, f(3)) = (3, -27)$ is a local minimum. The point $(0, f(0)) = (0, 0)$ is neither a minimum nor a maximum, even though $f'(0) = 0$.

(3) Now examine $f''(x)$:

$$f''(x) = 12x^2 - 24x = 12x(x - 2)$$

- So $f''(x) = 0$ when $x = 0, 2$. This splits the real line into the intervals $(-\infty, 0)$, $(0, 2)$ and $(2, \infty)$.
- When $x < 0$, $x - 2 < 0$ and so $f''(x) > 0$.
- When $0 < x < 2$, $x > 0$ and $x - 2 < 0$ and so $f''(x) < 0$.
- When $2 < x$, $x > 0$ and $x - 2 > 0$ and so $f''(x) > 0$.
- Thus the function is convex up for $x < 0$, then convex down for $0 < x < 2$, and finally convex up again for $x > 2$. Hence $(0, f(0)) = (0, 0)$ and $(2, f(2)) = (2, -16)$ are inflection points.

Putting all this information together gives us the following sketch.
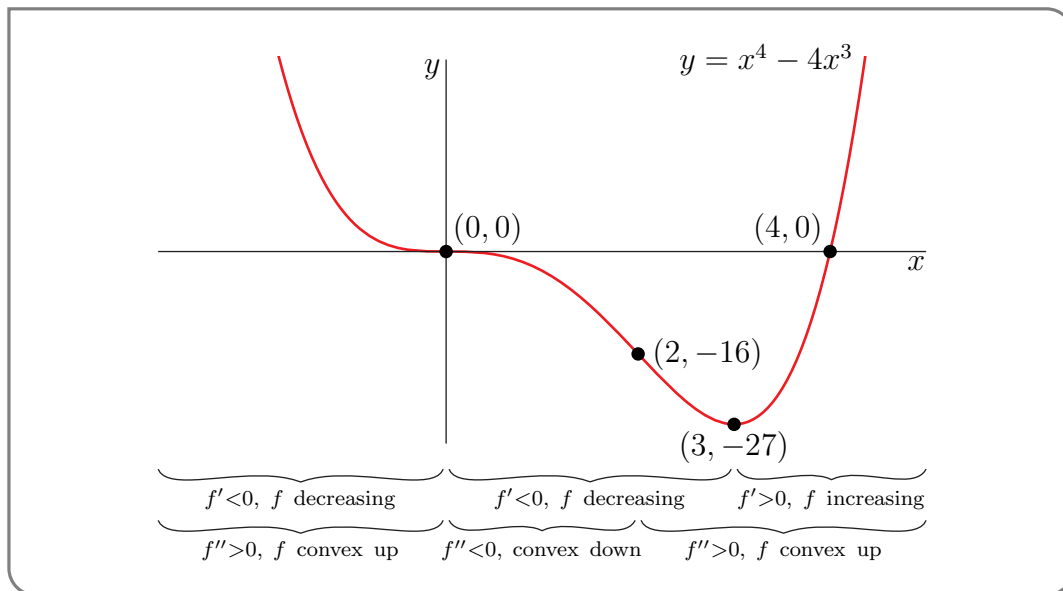
Example 3.6.13

Example 3.6.14 $(f(x) = x^3 - 6x^2 + 9x - 54)$

(1) Reading from $f(x)$:

- The function is a polynomial so it is defined everywhere.
- Since $f(-x) = -x^3 - 6x^2 - 9x - 54 \neq \pm f(x)$, it is not even or odd. Nor is it periodic.
- The y-intercept is $y = f(0) = -54$, while the x-intercepts are given by the solution of

$$f(x) = x^3 - 6x^2 + 9x - 54 = 0$$
$$x^2(x - 6) + 9(x - 6) = 0$$
$$(x^2 + 9)(x - 6) = 0$$

  Hence the only x-intercept is 6.
- Since f is a polynomial it does not have any vertical asymptotes.
- For very large x, both positive and negative, the x^3 term in $f(x)$ dominates the other term so that

$$f(x) \to \begin{cases} +\infty & \text{as } x \to +\infty \\ -\infty & \text{as } x \to -\infty \end{cases}$$

  and the function has no horizontal asymptotes.

(2) Now compute the derivative $f'(x)$:

$$f'(x) = 3x^2 - 12x + 9$$
$$= 3(x^2 - 4x + 3) = 3(x - 3)(x - 1)$$

- The critical points are at $x = 1, 3$. Since the function is a polynomial there are no singular points. The critical points split the real line into the intervals $(-\infty, 1)$, $(1, 3)$ and $(3, \infty)$.
- When $x < 1$, $(x - 1) < 0$ and $(x - 3) < 0$, so $f'(x) > 0$.
- When $1 < x < 3$, $(x - 1) > 0$ and $(x - 3) < 0$, so $f'(x) < 0$.
- When $3 < x$, $(x - 1) > 0$ and $(x - 3) > 0$, so $f'(x) > 0$.
- Summarising all this

	$(-\infty, 1)$	1	$(1,3)$	3	$(3, \infty)$
$f'(x)$	positive	0	negative	0	positive
	increasing	maximum	decreasing	minimum	increasing

So the point $(1, f(1)) = (1, -50)$ is a local maximum. The point $(3, f(3)) = (3, -54)$ is a local minimum.

(3) Now examine $f''(x)$:

$$f''(x) = 6x - 12$$

- So $f''(x) = 0$ when $x = 2$. This splits the real line into the intervals $(-\infty, 2)$ and $(2, \infty)$.
- When $x < 2$, $f''(x) < 0$.
- When $x > 2$, $f''(x) > 0$.
- Thus the function is convex down for $x < 2$, then convex up for $x > 2$. Hence $(2, f(2)) = (2, -52)$ is an inflection point.

Putting all this information together gives us the following sketch.
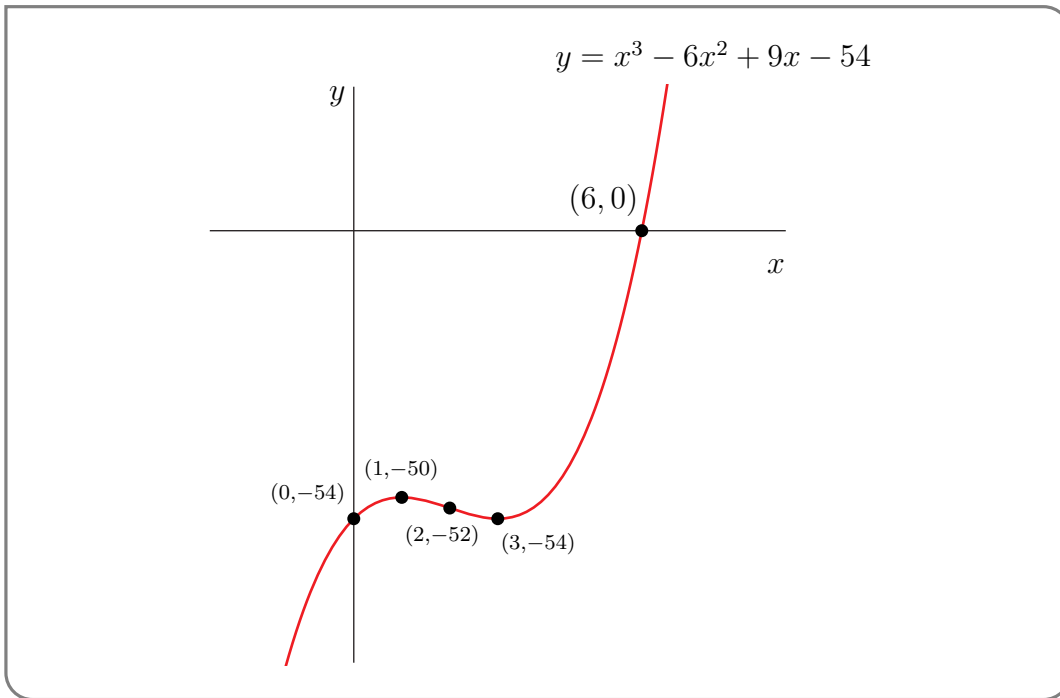
and if we zoom in around the interesting points (minimum, maximum and inflection point), we have
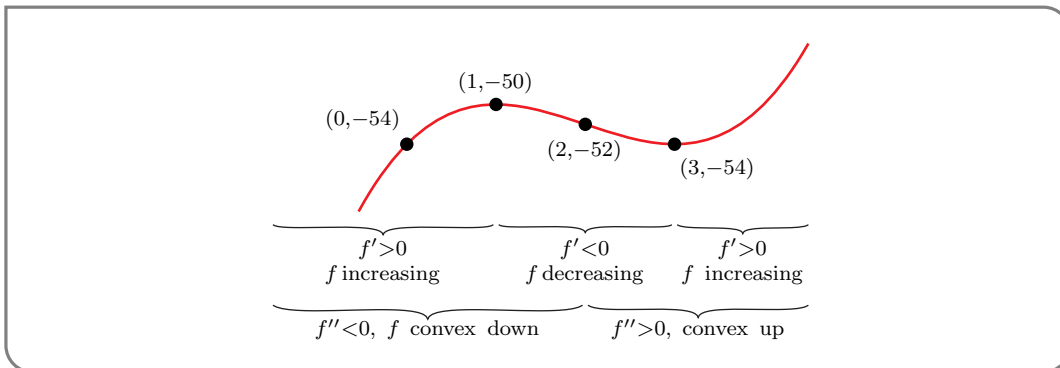
Example 3.6.14

An example of sketching a simple rational function.

Example 3.6.15 $\left(f(x) = \dfrac{x}{x^2 - 4} \right)$

(1) Reading from $f(x)$:

- The function is rational so it is defined except where its denominator is zero — namely at $x = \pm 2$.

- Since $f(-x) = \dfrac{-x}{x^2 - 4} = -f(x)$, it is odd. Indeed this means that we only need to examine what happens to the function for $x \geqslant 0$ and we can then infer what happens for $x \leqslant 0$ using $f(-x) = -f(x)$. In practice we will sketch the graph for $x \geqslant 0$ and then infer the rest from this symmetry.

- The y-intercept is $y = f(0) = 0$, while the x-intercepts are given by the solution of $f(x) = 0$. So the only x-intercept is 0.

- Since f is rational, it may have vertical asymptotes where its denominator is zero — at $x = \pm 2$. Since the function is odd, we only have to analyse the asymptote at $x = 2$ and we can then infer what happens at $x = -2$ by symmetry.

$$\lim_{x \to 2^+} f(x) = \lim_{x \to 2^+} \frac{x}{(x - 2)(x + 2)} = +\infty$$

$$\lim_{x \to 2^-} f(x) = \lim_{x \to 2^-} \frac{x}{(x - 2)(x + 2)} = -\infty$$

- We now check for horizontal asymptotes:

$$\lim_{x \to +\infty} f(x) = \lim_{x \to +\infty} \frac{x}{x^2 - 4}$$

$$= \lim_{x \to +\infty} \frac{1}{x - 4/x} = 0$$

(2) Now compute the derivative $f'(x)$:

$$f'(x) = \frac{(x^2 - 4) \cdot 1 - x \cdot 2x}{(x^2 - 4)^2}$$

$$= \frac{-(x^2 + 4)}{(x^2 - 4)^2}$$

- Hence there are no critical points. There are singular points where the denominator is zero, namely $x = \pm 2$. Before we proceed, notice that the numerator is always negative and the denominator is always positive. Hence $f'(x) < 0$ except at $x = \pm 2$ where it is undefined.

- The function is decreasing except at $x = \pm 2$.
- We already know that at $x = 2$ we have a vertical asymptote and that $f'(x) < 0$ for all x. So

$$\lim_{x \to 2} f'(x) = -\infty$$

- Summarising all this

	[0,2)	2	$(2, \infty)$
$f'(x)$	negative	DNE	negative
	decreasing	vertical asymptote	decreasing

Remember — we will draw the graph for $x \geqslant 0$ and then use the odd symmetry to infer the graph for $x < 0$.

(3) Now examine $f''(x)$:

$$
\begin{aligned}
f''(x) &= -\frac{(x^2 - 4)^2 \cdot (2x) - (x^2 + 4) \cdot 2 \cdot 2x \cdot (x^2 - 4)}{(x^2 - 4)^4} \\
&= -\frac{(x^2 - 4) \cdot (2x) - (x^2 + 4) \cdot 4x}{(x^2 - 4)^3} \\
&= -\frac{2x^3 - 8x - 4x^3 - 16x}{(x^2 - 4)^3} \\
&= \frac{2x(x^2 + 12)}{(x^2 - 4)^3}
\end{aligned}
$$

- So $f''(x) = 0$ when $x = 0$ and does not exist when $x = \pm 2$. This splits the real line into the intervals $(-\infty, -2), (-2, 0), (0, 2)$ and $(2, \infty)$. However we only need to consider $x \geqslant 0$ (because of the odd symmetry).
- When $0 < x < 2$, $x > 0$, $(x^2 + 12) > 0$ and $(x^2 - 4) < 0$ so $f''(x) < 0$.
- When $x > 2$, $x > 0$, $(x^2 + 12) > 0$ and $(x^2 - 4) > 0$ so $f''(x) > 0$.

Putting all this information together gives the following sketch for $x \geqslant 0$:
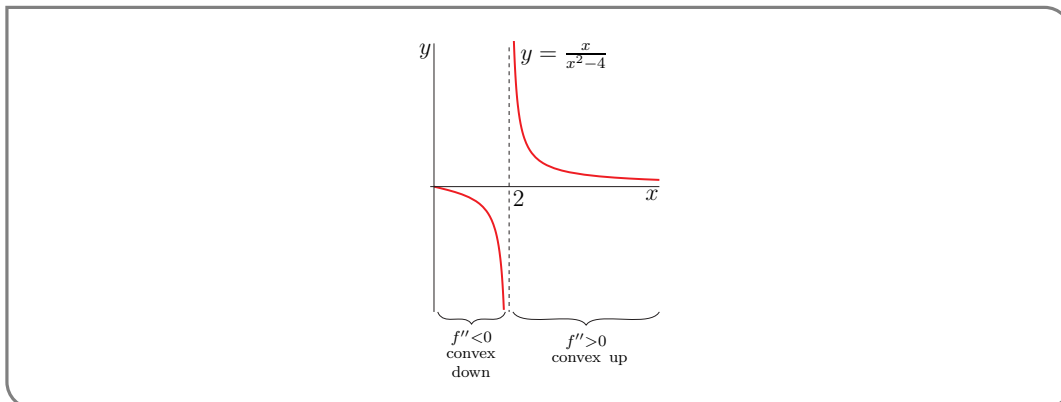
We can then draw in the graph for $x < 0$ using $f(-x) = -f(x)$:
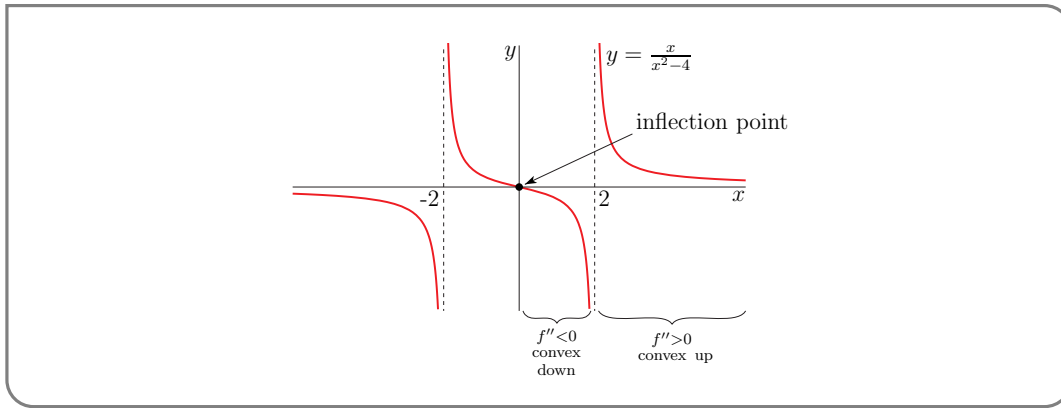
Notice that this means that the concavity changes at $x = 0$, so the point $(0, f(0)) = (0, 0)$ is a point of inflection (as indicated).

Example 3.6.15

This final example is more substantial since the function has singular points (points where the derivative is undefined). The analysis is more involved.

Example 3.6.16 $\left(f(x) = \sqrt[3]{\frac{x^2}{(x-6)^2}} \right)$

(1) Reading from $f(x)$:

- First notice that we can rewrite
$$f(x) = \sqrt[3]{\frac{x^2}{(x-6)^2}} = \sqrt[3]{\frac{x^2}{x^2 \cdot (1 - 6/x)^2}} = \sqrt[3]{\frac{1}{(1 - 6/x)^2}}$$

- The function is the cube root of a rational function. The rational function is defined except at $x = 6$, so the domain of f is all reals except $x = 6$.

- Clearly the function is not periodic, and examining
$$f(-x) = \sqrt[3]{\frac{1}{(1 - 6/(-x))^2}}$$
$$= \sqrt[3]{\frac{1}{(1 + 6/x)^2}} \neq \pm f(x)$$

shows the function is neither even nor odd.

- To compute horizontal asymptotes we examine the limit of the portion of the function inside the cube-root
$$\lim_{x \to \pm\infty} \frac{1}{(1 - \frac{6}{x})^2} = 1$$

This means we have
$$\lim_{x \to \pm\infty} f(x) = 1$$

That is, the line $y = 1$ will be a horizontal asymptote to the graph $y = f(x)$ both for $x \to +\infty$ and for $x \to -\infty$.

- Our function $f(x) \to +\infty$ as $x \to 6$, because of the $(1 - 6/x)^2$ in its denominator. So $y = f(x)$ has $x = 6$ as a vertical asymptote.

(2) Now compute $f'(x)$. Since we rewrote

$$f(x) = \sqrt[3]{\frac{1}{(1 - 6/x)^2}} = \left(1 - \frac{6}{x}\right)^{-2/3}$$

we can use the chain rule

$$f'(x) = -\frac{2}{3}\left(1 - \frac{6}{x}\right)^{-5/3} \frac{6}{x^2}$$

$$= -4\left(\frac{x - 6}{x}\right)^{-5/3} \frac{1}{x^2}$$

$$= -4\left(\frac{1}{x - 6}\right)^{5/3} \frac{1}{x^{1/3}}$$

- Notice that the derivative is nowhere equal to zero, so the function has no critical points. However there are two places the derivative is undefined. The terms

$$\left(\frac{1}{x - 6}\right)^{5/3} \qquad\qquad \frac{1}{x^{1/3}}$$

are undefined at $x = 6, 0$ respectively. Hence $x = 0, 6$ are singular points. These split the real line into the intervals $(-\infty, 0), (0, 6)$ and $(6, \infty)$.

- When $x < 0$, $(x - 6) < 0$, we have that $(x - 6)^{-5/3} < 0$ and $x^{-1/3} < 0$ and so $f'(x) = -4 \cdot (\text{negative}) \cdot (\text{negative}) < 0$.

- When $0 < x < 6$, $(x - 6) < 0$, we have that $(x - 6)^{-5/3} < 0$ and $x^{-1/3} > 0$ and so $f'(x) > 0$.

- When $x > 6$, $(x - 6) > 0$, we have that $(x - 6)^{-5/3} > 0$ and $x^{-1/3} > 0$ and so $f'(x) < 0$.

- We should also examine the behaviour of the derivative as $x \to 0$ and $x \to 6$.

$$\lim_{x \to 0^-} f'(x) = -4\left(\lim_{x \to 0^-} (x - 6)^{-5/3}\right)\left(\lim_{x \to 0^-} x^{-1/3}\right) = -\infty$$

$$\lim_{x \to 0^+} f'(x) = -4\left(\lim_{x \to 0^+} (x - 6)^{-5/3}\right)\left(\lim_{x \to 0^+} x^{-1/3}\right) = +\infty$$

$$\lim_{x \to 6^-} f'(x) = -4\left(\lim_{x \to 6^-} (x - 6)^{-5/3}\right)\left(\lim_{x \to 6^-} x^{-1/3}\right) = +\infty$$

$$\lim_{x \to 6^+} f'(x) = -4\left(\lim_{x \to 6^+} (x - 6)^{-5/3}\right)\left(\lim_{x \to 6^+} x^{-1/3}\right) = -\infty$$

We already know that $x = 6$ is a vertical asymptote of the function, so it is not surprising that the lines tangent to the graph become vertical as we approach 6.

The behavior around $x = 0$ is less standard, since the lines tangent to the graph become vertical, but $x = 0$ is not a vertical asymptote of the function. Indeed the function takes a finite value $y = f(0) = 0$.

- Summarising all this

	$(-\infty, 0)$	0	$(0,6)$	6	$(6, \infty)$
$f'(x)$	negative	DNE	positive	DNE	negative
	decreasing	vertical tangents	increasing	vertical asymptote	decreasing

(3) Now look at $f''(x)$:

$$f''(x) = -4\frac{d}{dx}\left[\left(\frac{1}{x-6}\right)^{5/3}\frac{1}{x^{1/3}}\right] = -4\left[-\frac{5}{3}\left(\frac{1}{x-6}\right)^{8/3}\frac{1}{x^{1/3}} - \frac{1}{3}\left(\frac{1}{x-6}\right)^{5/3}\frac{1}{x^{4/3}}\right]$$

$$= \frac{4}{3}\left(\frac{1}{x-6}\right)^{8/3}\frac{1}{x^{4/3}}\left[5x + (x-6)\right]$$

$$= 8\left(\frac{1}{x-6}\right)^{8/3}\frac{1}{x^{4/3}}\left[x - 1\right]$$

Oof!

- Both of the factors $\left(\frac{1}{x-6}\right)^{8/3} = \left(\frac{1}{\sqrt[3]{x-6}}\right)^{8}$ and $\frac{1}{x^{4/3}} = \left(\frac{1}{\sqrt[3]{x}}\right)^{4}$ are even powers and so are positive (though possibly infinite). So the sign of $f''(x)$ is the same as the sign of the factor $x - 1$. Thus

	$(-\infty, 1)$	1	$(1, \infty)$
$f''(x)$	negative	0	positive
	concave down	inflection point	concave up

Here is a sketch of the graph $y = f(x)$.

$$y = \sqrt[3]{\frac{x^2}{(x-6)^2}}$$

$f' < 0$, f decreasing $f' > 0$ $f' < 0$, f decreasing
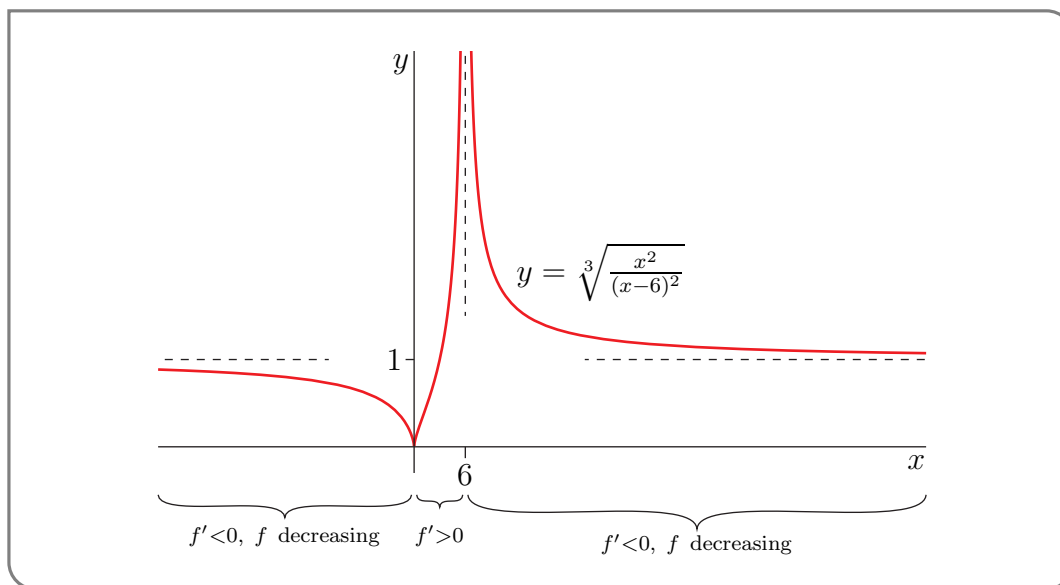
It is hard to see the inflection point at $x = 1$, $y = f(1) = \frac{1}{\sqrt[3]{25}}$ in the above sketch. So here is a blow up of the part of the sketch around $x = 1$.
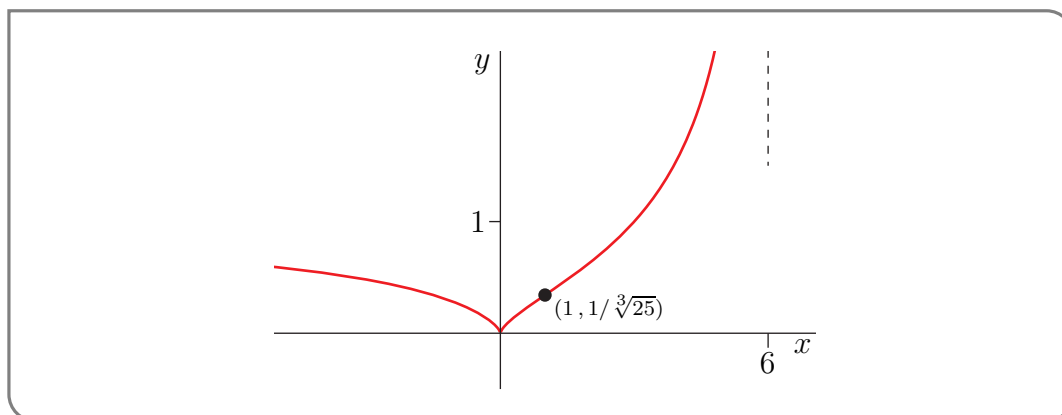
And if we zoom in even more we have
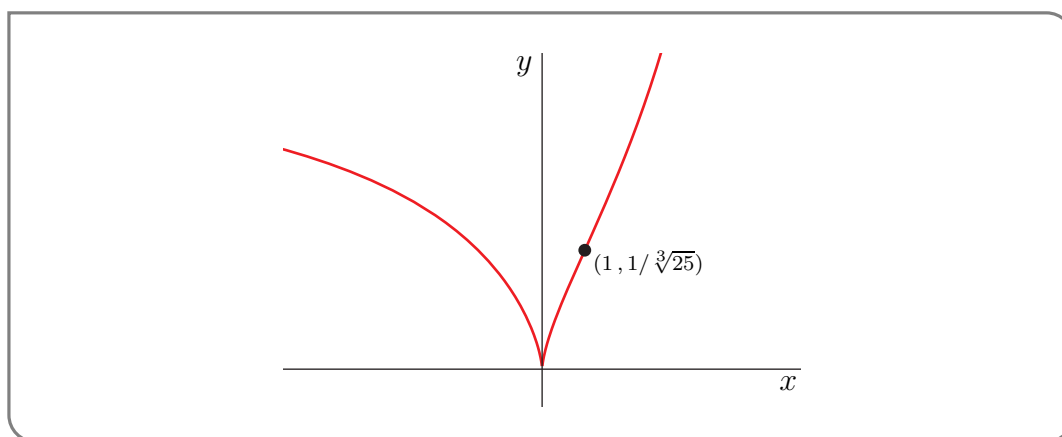
Example 3.6.16

3.7 ▴ L'Hôpital's Rule and Indeterminate Forms

Let us return to limits (Chapter 1) and see how we can use derivatives to simplify certain families of limits called indeterminate forms. We know, from Theorem 1.4.2 on the arithmetic of limits, that if

$$\lim_{x \to a} f(x) = F \qquad\qquad \lim_{x \to a} g(x) = G$$

and $G \neq 0$, then

$$\lim_{x \to a} \frac{f(x)}{g(x)} = \frac{F}{G}$$

The requirement that $G \neq 0$ is critical — we explored this in Example 1.4.6. Please reread that example.

Of course[64] it is not surprising that if $F \neq 0$ and $G = 0$, then

$$\lim_{x \to a} \frac{f(x)}{g(x)} = DNE$$

and if $F = 0$ but $G \neq 0$ then

$$\lim_{x \to a} \frac{f(x)}{g(x)} = 0$$

However when both $F, G = 0$ then, as we saw in Example 1.4.6, almost anything can happen

$$f(x) = x \qquad\qquad g(x) = x^2 \qquad\qquad \lim_{x \to 0} \frac{x}{x^2} = \lim_{x \to 0} \frac{1}{x} = DNE$$

$$f(x) = x^2 \qquad\qquad g(x) = x \qquad\qquad \lim_{x \to 0} \frac{x^2}{x} = \lim_{x \to 0} x = 0$$

$$f(x) = x \qquad\qquad g(x) = x \qquad\qquad \lim_{x \to 0} \frac{x}{x} = \lim_{x \to 0} 1 = 1$$

$$f(x) = 7x^2 \qquad\qquad g(x) = 3x^2 \qquad\qquad \lim_{x \to 0} \frac{7x^2}{3x^2} = \lim_{x \to 0} \frac{7}{3} = \frac{7}{3}$$

Indeed after exploring Example 1.4.11 and 1.4.13 we gave ourselves the rule of thumb that if we found $0/0$, then there must be something that cancels.

Because the limit that results from these $0/0$ situations is not immediately obvious, but also leads to some interesting mathematics, we should give it a name.

> **Definition 3.7.1** (First indeterminate forms).
>
> Let $a \in \mathbb{R}$ and let $f(x)$ and $g(x)$ be functions. If
>
> $$\lim_{x \to a} f(x) = 0 \qquad \text{and} \qquad \lim_{x \to a} g(x) = 0$$
>
> then the limit
>
> $$\lim_{x \to a} \frac{f(x)}{g(x)}$$
>
> is called a $0/0$ indeterminate form.

There are quite a number of mathematical tools for evaluating such indeterminate forms — Taylor series for example. A simpler method, which works in quite a few cases, is L'Hôpital's rule[65].

---

64   Now it is not so surprising, but perhaps back when we started limits, this was not so obvious.
65   Named for the 17th century mathematician, Guillaume de l'Hôpital, who published the first textbook on differential calculus. The eponymous rule appears in that text, but is believed to have been developed by Johann Bernoulli. The book was the source of some controversy since it contained many results by Bernoulli, which l'Hôpital acknowledged in the preface, but Bernoulli felt that l'Hôpital got undue credit.

> **Theorem 3.7.2** (L'Hôpital's Rule).
>
> Let $a \in \mathbb{R}$ and assume that
>
> $$\lim_{x \to a} f(x) = \lim_{x \to a} g(x) = 0$$
>
> Then
>
> (a) if $f'(a)$ and $g'(a)$ exist and $g'(a) \neq 0$, then
>
> $$\lim_{x \to a} \frac{f(x)}{g(x)} = \frac{f'(a)}{g'(a)},$$
>
> (b) while, if $f'(x)$ and $g'(x)$ exist, with $g'(x)$ nonzero, on an open interval that contains a, except possibly at a itself, and if the limit
>
> $$\lim_{x \to a} \frac{f'(x)}{g'(x)} \text{ exists or is } +\infty \text{ or is } -\infty$$
>
> then
>
> $$\lim_{x \to a} \frac{f(x)}{g(x)} = \lim_{x \to a} \frac{f'(x)}{g'(x)}$$

Proof. We only give the proof for part (a). The proof of part (b) is not very difficult, but uses the Generalised Mean–Value Theorem (Theorem 3.4.38), which is optional and most readers have not seen it.

- First note that we must have $f(a) = g(a) = 0$. To see this note that since derivative $f'(a)$ exists, we know that the limit

$$\lim_{x \to a} \frac{f(x) - f(a)}{x - a} \text{ exists}$$

Since we know that the denominator goes to zero, we must also have that the numerator goes to zero (otherwise the limit would be undefined). Hence we must have

$$\lim_{x \to a} (f(x) - f(a)) = \left(\lim_{x \to a} f(x) \right) - f(a) = 0$$

We are told that $\lim_{x \to a} f(x) = 0$ so we must have $f(a) = 0$. Similarly we know that $g(a) = 0$.

Note that around that time l'Hôpital's name was commonly spelled l'Hospital, but the spelling of silent s in French was changed subsequently; many texts spell his name l'Hospital. If you find yourself in Paris, you can hunt along Boulevard de l'Hôpital for older street signs carved into the sides of buildings which spell it "l'Hospital" — though arguably there are better things to do there.

- Now consider the indeterminate form

$$\lim_{x \to a} \frac{f(x)}{g(x)} = \lim_{x \to a} \frac{f(x) - 0}{g(x) - 0} \qquad \text{use } 0 = f(a) = g(a)$$

$$= \lim_{x \to a} \frac{f(x) - f(a)}{g(x) - g(a)} \qquad \text{multiply by } 1 = \frac{(x - a)^{-1}}{(x - a)^{-1}}$$

$$= \lim_{x \to a} \frac{f(x) - f(a)}{g(x) - g(a)} \cdot \frac{(x - a)^{-1}}{(x - a)^{-1}} \qquad \text{rearrange}$$

$$= \lim_{x \to a} \left[\frac{\dfrac{f(x) - f(a)}{x - a}}{\dfrac{g(x) - g(a)}{x - a}} \right] \qquad \text{use arithmetic of limits}$$

$$= \frac{\displaystyle\lim_{x \to a} \frac{f(x) - f(a)}{x - a}}{\displaystyle\lim_{x \to a} \frac{g(x) - g(a)}{x - a}} = \frac{f'(a)}{g'(a)}$$

We can justify this step and apply Theorem 1.4.2, since the limits in the numerator and denominator exist, because they are just $f'(a)$ and $g'(a)$.

$\square$

▸▸▸ Optional — Proof of Part (b) of l'Hôpital's Rule

To prove part (b) we must work around the possibility that $f'(a)$ and $g'(a)$ do not exist or that $f'(x)$ and $g'(x)$ are not continuous at $x = a$. To do this, we make use of the Generalised Mean-Value Theorem (Theorem 3.4.38) that was used to prove Equation (3.4.33). We recommend you review the GMVT before proceeding.

For simplicity we consider the limit

$$\lim_{x \to a^+} \frac{f(x)}{g(x)}$$

By assumption, we know that

$$\lim_{x \to a^+} f(x) = \lim_{x \to a^+} g(x) = 0$$

For simplicity, we also assume that $f(a) = g(a) = 0$. This allows us to write

$$\frac{f(x)}{g(x)} = \frac{f(x) - f(a)}{g(x) - g(a)}$$

which is the right form for an application of the GMVT.

By assumption $f'(x)$ and $g'(x)$ exist, with $g'(x)$ nonzero, in some open interval around a, except possibly at a itself. So we know that they exist, with $g'(x) \neq 0$, in some interval $(a, b]$ with $b > a$. Then the GMVT (Theorem 3.4.38) tells us that for $x \in (a, b]$

$$\frac{f(x)}{g(x)} = \frac{f(x) - f(a)}{g(x) - g(a)} = \frac{f'(c)}{g'(c)}$$

where $c \in (a, x)$. As we take the limit as $x \to a$, we also have that $c \to a$, and so

$$\lim_{x \to a^+} \frac{f(x)}{g(x)} = \lim_{x \to a^+} \frac{f'(c)}{g'(c)} = \lim_{c \to a^+} \frac{f'(c)}{g'(c)}$$

as required.

3.7.1 ▸▸ Standard Examples

Here are some simple examples using L'Hôpital's rule.

Example 3.7.3

Consider the limit

$$\lim_{x \to 0} \frac{\sin x}{x}$$

- Notice that

$$\lim_{x \to 0} \sin x = 0$$
$$\lim_{x \to 0} x = 0$$

so this is a $0/0$ indeterminate form, and suggests we try l'Hôpital's rule.

- To apply the rule we must first check the limits of the derivatives.

$$
\begin{aligned}
f(x) &= \sin x & f'(x) &= \cos x & \text{and} & & f'(0) &= 1 \\
g(x) &= x & g'(x) &= 1 & \text{and} & & g'(0) &= 1
\end{aligned}
$$

- So by l'Hôpital's rule

$$\lim_{x \to 0} \frac{\sin x}{x} = \frac{f'(0)}{g'(0)} = \frac{1}{1} = 1.$$

Example 3.7.3

Example 3.7.4

Consider the limit

$$\lim_{x \to 0} \frac{\sin(x)}{\sin(2x)}$$

- First check

$$\lim_{x \to 0} \sin 2x = 0$$

$$\lim_{x \to 0} \sin x = 0$$

so we again have a $0/0$ indeterminate form.

- Set $f(x) = \sin x$ and $g(x) = \sin 2x$, then

$$f'(x) = \cos x \qquad\qquad f'(0) = 1$$
$$g'(x) = 2\cos 2x \qquad\qquad g'(0) = 2$$

- And by l'Hôpital's rule

$$\lim_{x\to 0} \frac{\sin x}{\sin 2x} = \frac{f'(0)}{g'(0)} = \frac{1}{2}.$$

Example 3.7.4

Example 3.7.5

Let $q > 1$ and compute the limit

$$\lim_{x\to 0} \frac{q^x - 1}{x}$$

This limit arose in our discussion of exponential functions in Section 2.7.

- First check

$$\lim_{x\to 0} (q^x - 1) = 1 - 1 = 0$$
$$\lim_{x\to 0} x = 0$$

so we have a $0/0$ indeterminate form.

- Set $f(x) = q^x - 1$ and $g(x) = x$, then (maybe after a quick review of Section 2.7)

$$f'(x) = \frac{\mathrm{d}}{\mathrm{d}x}(q^x - 1) = q^x \cdot \log q \qquad\qquad f'(0) = \log q$$
$$g'(x) = 1 \qquad\qquad g'(0) = 1$$

- And by l'Hôpital's rule[66]

$$\lim_{h\to 0} \frac{q^h - 1}{h} = \log q.$$

---

[66] While it might not be immediately obvious, this example relies on circular reasoning. In order to apply l'Hôpital's rule, we need to compute the derivative of q^x. However in order to compute that limit (see Section 2.7) we needed to evaluate this limit.

A more obvious example of this sort of circular reasoning can be seen if we use l'Hôpital's rule to compute the derivative of $f(x) = x^n$ at $x = a$ using the limit

$$f'(a) = \lim_{x\to a} \frac{x^n - a^n}{x - a} = \lim_{x\to a} \frac{nx^{n-1} - 0}{1 - 0} = na^{n-1}.$$

We have used the result $\frac{\mathrm{d}}{\mathrm{d}x} x^n = nx^{n-1}$ to prove itself!

Example 3.7.5

In this example, we shall apply L'Hôpital's rule twice before getting the answer.

Example 3.7.6

Compute the limit

$$\lim_{x \to 0} \frac{\sin(x^2)}{1 - \cos x}$$

- Again we should check

$$\lim_{x \to 0} \sin(x^2) = \sin 0 = 0$$
$$\lim_{x \to 0} (1 - \cos x) = 1 - \cos 0 = 0$$

and we have a $0/0$ indeterminate form.

- Let $f(x) = \sin(x^2)$ and $g(x) = 1 - \cos x$ then

$$f'(x) = 2x \cos(x^2) \qquad\qquad f'(0) = 0$$
$$g'(x) = \sin x \qquad\qquad g'(0) = 0$$

So if we try to apply l'Hôpital's rule naively we will get

$$\lim_{x \to 0} \frac{\sin(x^2)}{1 - \cos x} = \frac{f'(0)}{g'(0)} = \frac{0}{0}.$$

which is another $0/0$ indeterminate form.

- It appears that we are stuck until we remember that l'Hôpital's rule (as stated in Theorem 3.7.2) has a part (b) — now is a good time to reread it.

- It says that

$$\lim_{x \to 0} \frac{f(x)}{g(x)} = \lim_{x \to 0} \frac{f'(x)}{g'(x)}$$

provided this second limit exists. In our case this requires us to compute

$$\lim_{x \to 0} \frac{2x \cos(x^2)}{\sin(x)}$$

which we can do using l'Hôpital's rule again. Now

$$h(x) = 2x \cos(x^2) \qquad h'(x) = 2\cos(x^2) - 4x^2 \sin(x^2) \qquad h'(0) = 2$$
$$\ell(x) = \sin(x) \qquad \ell'(x) = \cos(x) \qquad \ell'(0) = 1$$

By l'Hôpital's rule

$$\lim_{x \to 0} \frac{2x \cos(x^2)}{\sin(x)} = \frac{h'(0)}{\ell'(0)} = 2$$

- Thus our original limit is

$$\lim_{x \to 0} \frac{\sin(x^2)}{1 - \cos x} = \lim_{x \to 0} \frac{2x \cos(x^2)}{\sin(x)} = 2.$$

- We can succinctly summarise the two applications of L'Hôpital's rule in this example by

$$\lim_{x \to 0} \underbrace{\frac{\sin(x^2)}{1 - \cos x}}_{\substack{\text{num} \to 0 \\ \text{den} \to 0}} = \lim_{x \to 0} \underbrace{\frac{2x \cos(x^2)}{\sin x}}_{\substack{\text{num} \to 0 \\ \text{den} \to 0}} = \lim_{x \to 0} \underbrace{\frac{2 \cos(x^2) - 4x^2 \sin(x^2)}{\cos x}}_{\substack{\text{num} \to 2 \\ \text{den} \to 1}} = 2$$

Here "num" and "den" are used as abbreviations of "numerator" and "denominator" respectively."

Example 3.7.6

One must be careful to ensure that the hypotheses of l'Hôpital's rule are satisfied before applying it. The following "warnings" show the sorts of things that can go wrong.

Warning 3.7.7 (Denominator limit nonzero).

If

$$\lim_{x \to a} f(x) = 0 \qquad \text{but} \qquad \lim_{x \to a} g(x) \neq 0$$

then

$$\lim_{x \to a} \frac{f(x)}{g(x)} \qquad \text{need not be the same as} \qquad \frac{f'(a)}{g'(a)} \text{ or } \lim_{x \to a} \frac{f'(x)}{g'(x)}.$$

Here is an example. Take

$$a = 0 \qquad f(x) = 3x \qquad g(x) = 4 + 5x$$

Then

$$\lim_{x \to 0} \frac{f(x)}{g(x)} = \lim_{x \to 0} \frac{3x}{4 + 5x} = \frac{3 \times 0}{4 + 5 \times 0} = 0$$

$$\lim_{x \to 0} \frac{f'(x)}{g'(x)} = \frac{f'(0)}{g'(0)} = \frac{3}{5}$$

Warning 3.7.8 (Numerator limit nonzero).

If

$$\lim_{x \to a} g(x) = 0 \qquad \text{but} \qquad \lim_{x \to a} f(x) \neq 0$$

then

$$\lim_{x \to a} \frac{f(x)}{g(x)} \qquad \text{need not be the same as} \qquad \lim_{x \to a} \frac{f'(x)}{g'(x)}.$$

Here is an example. Take

$$a = 0 \qquad f(x) = 4 + 5x \qquad g(x) = 3x$$

Then

$$\lim_{x \to 0} \frac{f(x)}{g(x)} = \lim_{x \to 0} \frac{4 + 5x}{3x} = \text{DNE}$$

$$\lim_{x \to 0} \frac{f'(x)}{g'(x)} = \lim_{x \to 0} \frac{5}{3} = \frac{5}{3}$$

This next one is more subtle; the limits of the original numerator and denominator functions both go to zero, but the limit of the ratio their derivatives does not exist.

Warning 3.7.9 (Limit of ratio of derivatives DNE).

If

$$\lim_{x \to a} f(x) = 0 \qquad \text{and} \qquad \lim_{x \to a} g(x) = 0$$

but

$$\lim_{x \to a} \frac{f'(x)}{g'(x)} \text{ does not exist}$$

then it is still possible that

$$\lim_{x \to a} \frac{f(x)}{g(x)} \text{ exists.}$$

Here is an example. Take

$$a = 0 \qquad f(x) = x^2 \sin \frac{1}{x} \qquad g(x) = x$$

Then (with an application of the squeeze theorem)

$$\lim_{x \to 0} f(x) = 0 \qquad \text{and} \qquad \lim_{x \to 0} g(x) = 0.$$

If we attempt to apply l'Hôpital's rule then we have $g'(x) = 1$ and

$$f'(x) = 2x \sin \frac{1}{x} - \cos \frac{1}{x}$$

and we then try to compute the limit

$$\lim_{x \to 0} \frac{f'(x)}{g'(x)} = \lim_{x \to 0} \left(2x \sin \frac{1}{x} - \cos \frac{1}{x} \right)$$

However, this limit does not exist. The first term converges to 0 (by the squeeze theorem), but the second term $\cos(1/x)$ just oscillates wildly between ± 1. All we can conclude from this is

Since the limit of the ratio of derivatives does not exist, we cannot apply l'Hôpital's rule.

Instead we should go back to the original limit and apply the squeeze theorem:

$$\lim_{x \to 0} \frac{f(x)}{g(x)} = \lim_{x \to 0} \frac{x^2 \sin \frac{1}{x}}{x} = \lim_{x \to 0} x \sin \frac{1}{x} = 0,$$

since $|x \sin(1/x)| < |x|$ and $|x| \to 0$ as $x \to 0$.

It is also easy to construct an example in which the limits of numerator and denominator are both zero, but the limit of the ratio and the limit of the ratio of the derivatives do not exist. A slight change of the previous example shows that it is possible that

$$\lim_{x \to a} f(x) = 0 \qquad\qquad \text{and} \qquad\qquad \lim_{x \to a} g(x) = 0$$

but neither of the limits

$$\lim_{x \to a} \frac{f(x)}{g(x)} \qquad\qquad \text{or} \qquad\qquad \lim_{x \to a} \frac{f'(x)}{g'(x)}$$

exist. Take

$$a = 0 \qquad\qquad f(x) = x \sin \frac{1}{x} \qquad g(x) = x$$

Then (with a quick application of the squeeze theorem)

$$\lim_{x \to 0} f(x) = 0 \qquad\qquad \text{and} \qquad\qquad \lim_{x \to 0} g(x) = 0.$$

However,

$$\lim_{x \to 0} \frac{f(x)}{g(x)} = \lim_{x \to 0} \frac{x \sin \frac{1}{x}}{x} = \lim_{x \to 0} \sin \frac{1}{x}$$

does not exist. And similarly

$$\lim_{x \to 0} \frac{f'(x)}{g'(x)} = \lim_{x \to 0} \frac{\sin \frac{1}{x} - \frac{1}{x} \cos \frac{1}{x}}{x^2}$$

does not exist.

3.7.2 ▸ Variations

Theorem 3.7.2 is the basic form of L'Hôpital's rule, but there are also many variations. Here are a bunch of them.

(a) L'Hôpital's rule also applies when the limit of $x \to a$ is replaced by $\lim_{x \to a+}$ or by $\lim_{x \to a-}$ or by $\lim_{x \to +\infty}$ or by $\lim_{x \to -\infty}$.

We can justify adapting the rule to the limits to $\pm\infty$ via the following reasoning

$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = \lim_{y \to 0+} \frac{f(1/y)}{g(1/y)} \qquad\qquad \text{substitute } x = 1/y$$

$$= \lim_{y \to 0+} \frac{-\frac{1}{y^2} f'(1/y)}{-\frac{1}{y^2} g'(1/y)},$$

where we have used l'Hôpital's rule (assuming this limit exists) and the fact that $\frac{d}{dy}f(1/y) = -\frac{1}{y^2}f'(1/y)$ (and similarly for g). Cleaning this up and substituting $y = 1/x$ gives the required result:

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = \lim_{y\to 0^+} \frac{f'(1/y)}{g'(1/y)} = \lim_{x\to\infty} \frac{f'(x)}{g'(x)}.$$

Example 3.7.10

Consider the limit

$$\lim_{x\to\infty} \frac{\arctan x - \frac{\pi}{2}}{1/x}$$

Both numerator and denominator go to 0 as $x \to \infty$, so this is an $0/0$ indeterminate form. We find

$$\lim_{x\to+\infty} \underbrace{\frac{\arctan x - \frac{\pi}{2}}{\frac{1}{x}}}_{\substack{\text{num}\to 0 \\ \text{den}\to 0}} = \lim_{x\to+\infty} \frac{\frac{1}{1+x^2}}{-\frac{1}{x^2}} = -\lim_{x\to+\infty} \underbrace{\frac{1}{1 + \frac{1}{x^2}}}_{\substack{\text{num}\to 1 \\ \text{den}\to 1}} = -1$$

We have applied L'Hôpital's rule with

$$f(x) = \arctan x - \frac{\pi}{2} \qquad g(x) = \frac{1}{x}$$

$$f'(x) = \frac{1}{1+x^2} \qquad g'(x) = -\frac{1}{x^2}$$

Example 3.7.10

(b) $\frac{\infty}{\infty}$ *indeterminate form*: L'Hôpital's rule also applies when $\lim_{x\to a} f(x) = 0$, $\lim_{x\to a} g(x) = 0$ is replaced by $\lim_{x\to a} f(x) = \pm\infty$, $\lim_{x\to a} g(x) = \pm\infty$.

Example 3.7.11

Consider the limit

$$\lim_{x\to\infty} \frac{\log x}{x}$$

The numerator and denominator both blow up towards infinity so this is an ∞/∞ indeterminate form. An application of l'Hôpital's rule gives

$$\lim_{x\to\infty} \underbrace{\frac{\log x}{x}}_{\substack{\text{num}\to\infty \\ \text{den}\to\infty}} = \lim_{x\to\infty} \frac{1/x}{1}$$

$$= \lim_{x\to\infty} \frac{1}{x} = 0$$

Consider the limit

$$\lim_{x\to\infty} \frac{5x^2 + 3x - 3}{x^2 + 1}$$

Then by two applications of l'Hôpital's rule we get

$$\lim_{x\to\infty} \underbrace{\frac{5x^2 + 3x - 3}{x^2 + 1}}_{\substack{\text{num}\to\infty \\ \text{den}\to\infty}} = \lim_{x\to\infty} \underbrace{\frac{10x + 3}{2x}}_{\substack{\text{num}\to\infty \\ \text{den}\to\infty}} = \lim_{x\to\infty} \frac{10}{2} = 5.$$

Compute the limit

$$\lim_{x\to 0+} \frac{\log x}{\tan\left(\frac{\pi}{2} - x\right)}$$

We can compute this using l'Hôpital's rule twice:

$$\lim_{x\to 0+} \underbrace{\frac{\log x}{\tan\left(\frac{\pi}{2} - x\right)}}_{\substack{\text{num}\to -\infty \\ \text{den}\to +\infty}} = \lim_{x\to 0+} \frac{\frac{1}{x}}{-\sec^2\left(\frac{\pi}{2} - x\right)} = -\lim_{x\to 0+} \underbrace{\frac{\cos^2\left(\frac{\pi}{2} - x\right)}{x}}_{\substack{\text{num}\to 0 \\ \text{den}\to 0}}$$

$$= -\lim_{x\to 0+} \underbrace{\frac{2\cos\left(\frac{\pi}{2} - x\right)\sin\left(\frac{\pi}{2} - x\right)}{1}}_{\substack{\text{num}\to 0 \\ \text{den}\to 1}} = 0$$

The first application of L'Hôpital's was with

$$f(x) = \log x \quad g(x) = \tan\left(\frac{\pi}{2} - x\right)$$

$$f'(x) = \frac{1}{x} \qquad g'(x) = -\sec^2\left(\frac{\pi}{2} - x\right)$$

and the second time with

$$f(x) = \cos^2\left(\frac{\pi}{2} - x\right) \qquad\qquad\qquad g(x) = x$$

$$f'(x) = 2\cos\left(\frac{\pi}{2} - x\right)\left[-\sin\left(\frac{\pi}{2} - x\right)\right](-1) \qquad g'(x) = 1$$

Example 3.7.13

Sometimes things don't quite work out as we would like and l'Hôpital's rule can get stuck in a loop. Remember to think about the problem before you apply any rule.

Example 3.7.14

Consider the limit

$$\lim_{x\to\infty} \frac{e^x + e^{-x}}{e^x - e^{-x}}$$

Clearly both numerator and denominator go to ∞, so we have a ∞/∞ indeterminate form. Naively applying l'Hôpital's rule gives

$$\lim_{x\to\infty} \frac{e^x + e^{-x}}{e^x - e^{-x}} = \lim_{x\to\infty} \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

which is again a ∞/∞ indeterminate form. So apply l'Hôpital's rule again:

$$\lim_{x\to\infty} \frac{e^x - e^{-x}}{e^x + e^{-x}} = \lim_{x\to\infty} \frac{e^x + e^{-x}}{e^x - e^{-x}}$$

which is right back where we started!

The correct approach to such a limit is to apply the methods we learned in Chapter 1 and rewrite

$$\frac{e^x + e^{-x}}{e^x - e^{-x}} = \frac{e^x(1 + e^{-2x})}{e^x(1 - e^{-2x})} = \frac{1 + e^{-2x}}{1 - e^{-2x}}$$

and then take the limit.

A similar sort of l'Hôpital-rule-loop will occur if you naively apply l'Hôpital's rule to the limit

$$\lim_{x\to\infty} \frac{\sqrt{4x^2 + 1}}{5x - 1}$$

which appeared in Example 1.5.6.

Example 3.7.14

▸▸▸ **Optional — Proof of l'Hôpital's Rule for ∞/∞ forms**

We can justify this generalisation of l'Hôpital's rule with some careful manipulations. Since the derivatives f', g' exist in some interval around a, we know that f, g are con-

tinuous in some interval around a; let x, t be points inside that interval. Now rewrite[67]

$$\frac{f(x)}{g(x)} = \frac{f(x)}{g(x)} + \underbrace{\left(\frac{f(t)}{g(x)} - \frac{f(t)}{g(x)}\right)}_{=0} + \underbrace{\left(\frac{f(x)-f(t)}{g(x)-g(t)} - \frac{f(x)-f(t)}{g(x)-g(t)}\right)}_{=0}$$

$$= \underbrace{\frac{f(x)-f(t)}{g(x)-g(t)}}_{\text{ready for GMVT}} + \frac{f(t)}{g(x)} + \underbrace{\left(\frac{f(x)}{g(x)} - \frac{f(t)}{g(x)} - \frac{f(x)-f(t)}{g(x)-g(t)}\right)}_{\text{we can clean it up}}$$

$$= \frac{f(x)-f(t)}{g(x)-g(t)} + \frac{f(t)}{g(x)} + \left(\frac{f(x)-f(t)}{g(x)} - \frac{f(x)-f(t)}{g(x)-g(t)}\right)$$

$$= \frac{f(x)-f(t)}{g(x)-g(t)} + \frac{f(t)}{g(x)} + \left(\frac{1}{g(x)} - \frac{1}{g(x)-g(t)}\right) \cdot (f(x)-f(t))$$

$$= \frac{f(x)-f(t)}{g(x)-g(t)} + \frac{f(t)}{g(x)} + \left(\frac{g(x)-g(t)-g(x)}{g(x)(g(x)-g(t))}\right) \cdot (f(x)-f(t))$$

$$= \underbrace{\frac{f(x)-f(t)}{g(x)-g(t)}}_{\text{ready for GMVT}} + \frac{f(t)}{g(x)} - \frac{g(t)}{g(x)} \cdot \underbrace{\frac{f(x)-f(t)}{g(x)-g(t)}}_{\text{ready for GMVT}}$$

Oof! Now the generalised mean-value theorem (Theorem 3.4.38) tells us there is a c between x and t so that

$$\frac{f(x)-f(t)}{g(x)-g(t)} = \frac{f'(c)}{g'(c)}$$

Now substitute this into the large expression we derived above:

$$\frac{f(x)}{g(x)} = \frac{f'(c)}{g'(c)} + \frac{1}{g(x)}\left(f(t) - \frac{f'(c)}{g'(c)} \cdot g(t)\right)$$

At first glance this does not appear so useful, however if we fix t and take the limit as $x \to a$, then it becomes

$$\lim_{x \to a} \frac{f(x)}{g(x)} = \lim_{x \to a} \frac{f'(c)}{g'(c)} + \lim_{x \to a} \frac{1}{g(x)}\left(f(t) - \frac{f'(c)}{g'(c)} \cdot g(t)\right)$$

Since $g(x) \to \infty$ as $x \to a$, this last term goes to zero

$$= \lim_{x \to a} \frac{f'(c)}{g'(c)} + 0$$

Now take the limit as $t \to a$. The left-hand side is unchanged since it is independent of t. The right-hand side, however, does change; the number c is trapped between x and t. Since we have already taken the limit $x \to a$, so when we take the limit $t \to a$, we are effectively taking the limit $c \to a$. Hence

$$\lim_{x \to a} \frac{f(x)}{g(x)} = \lim_{c \to a} \frac{f'(c)}{g'(c)}$$

which is the desired result.

---

67 This is quite a clever argument, but it is not immediately obvious why one rewrites things this way. After the fact it becomes clear that it is done to massage the expression into the form where we can apply the generalised mean-value theorem (Theorem 3.4.38).

(c) $0 \cdot \infty$ *indeterminate form*:   When $\lim_{x \to a} f(x) = 0$ and $\lim_{x \to a} g(x) = \infty$. We can use a little algebra to manipulate this into either a $\frac{0}{0}$ or $\frac{\infty}{\infty}$ form:

$$\lim_{x \to a} \frac{f(x)}{1/g(x)} \qquad\qquad \lim_{x \to a} \frac{g(x)}{1/f(x)}$$

Example 3.7.15

Consider the limit

$$\lim_{x \to 0^+} x \cdot \log x$$

Here the function $f(x) = x$ goes to zero, while $g(x) = \log x$ goes to $-\infty$. If we rewrite this as the fraction

$$x \cdot \log x = \frac{\log x}{1/x}$$

then the $0 \cdot \infty$ form has become an ∞/∞ form.

The result is then

$$\lim_{x \to 0+} \underbrace{x}_{\to 0} \underbrace{\log x}_{\to -\infty} = \lim_{x \to 0+} \underbrace{\frac{\log x}{\frac{1}{x}}}_{\substack{\text{num} \to -\infty \\ \text{den} \to \infty}} = \lim_{x \to 0+} \frac{\frac{1}{x}}{-\frac{1}{x^2}} = -\lim_{x \to 0+} x = 0$$

Example 3.7.15

Example 3.7.16

In this example we'll evaluate $\lim_{x \to +\infty} x^n e^{-x}$, for all natural numbers n. We'll start with $n = 1$ and $n = 2$ and then, using what we have learned from those cases, move on to general n.

$$\lim_{x \to +\infty} \underbrace{x}_{\to \infty} \underbrace{e^{-x}}_{\to 0} = \lim_{x \to +\infty} \underbrace{\frac{x}{e^x}}_{\substack{\text{num} \to +\infty \\ \text{den} \to +\infty}} = \lim_{x \to +\infty} \underbrace{\frac{1}{e^x}}_{\substack{\text{num} \to 1 \\ \text{den} \to +\infty}} = \lim_{x \to +\infty} e^{-x} = 0$$

Applying l'Hôpital twice,

$$\lim_{x \to +\infty} \underbrace{x^2}_{\to \infty} \underbrace{e^{-x}}_{\to 0} = \lim_{x \to +\infty} \underbrace{\frac{x^2}{e^x}}_{\substack{\text{num} \to +\infty \\ \text{den} \to +\infty}} = \lim_{x \to +\infty} \underbrace{\frac{2x}{e^x}}_{\substack{\text{num} \to \infty \\ \text{den} \to +\infty}} = \lim_{x \to +\infty} \underbrace{\frac{2}{e^x}}_{\substack{\text{num} \to 2 \\ \text{den} \to +\infty}} = \lim_{x \to +\infty} 2e^{-x} = 0$$

Indeed, for any natural number n, applying l'Hôpital n times gives

$$\lim_{x\to+\infty} \underbrace{x^n}_{\to\infty} \underbrace{e^{-x}}_{\to 0} = \lim_{x\to+\infty} \underbrace{\frac{x^n}{e^x}}_{\substack{\text{num}\to+\infty \\ \text{den}\to+\infty}}$$

$$= \lim_{x\to+\infty} \underbrace{\frac{nx^{n-1}}{e^x}}_{\substack{\text{num}\to\infty \\ \text{den}\to+\infty}}$$

$$= \lim_{x\to+\infty} \underbrace{\frac{n(n-1)x^{n-2}}{e^x}}_{\substack{\text{num}\to\infty \\ \text{den}\to+\infty}}$$

$$= \cdots = \lim_{x\to+\infty} \underbrace{\frac{n!}{e^x}}_{\substack{\text{num}\to n! \\ \text{den}\to+\infty}} = 0$$

Example 3.7.16

(d) $\infty - \infty$ *indeterminate form*: When $\lim_{x\to a} f(x) = \infty$ and $\lim_{x\to a} g(x) = \infty$. We rewrite the difference as a fraction using a common denominator

$$f(x) - g(x) = \frac{h(x)}{\ell(x)}$$

which is then a $0/0$ or ∞/∞ form.

Example 3.7.17

Consider the limit

$$\lim_{x\to\frac{\pi}{2}^-} (\sec x - \tan x)$$

Since the limit of both $\sec x$ and $\tan x$ is $+\infty$ as $x \to \frac{\pi}{2}^-$, this is an $\infty - \infty$ indeterminate form. However we can rewrite this as

$$\sec x - \tan x = \frac{1}{\cos x} - \frac{\sin x}{\cos x} = \frac{1 - \sin x}{\cos x}$$

which is then a $0/0$ indeterminate form. This then gives

$$\lim_{x\to\frac{\pi}{2}^-} \left(\underbrace{\sec x}_{\to+\infty} - \underbrace{\tan x}_{\to+\infty} \right) = \lim_{x\to\frac{\pi}{2}^-} \underbrace{\frac{1 - \sin x}{\cos x}}_{\substack{\text{num}\to 0 \\ \text{den}\to 0}} = \lim_{x\to\frac{\pi}{2}^-} \underbrace{\frac{-\cos x}{-\sin x}}_{\substack{\text{num}\to 0 \\ \text{den}\to -1}} = 0$$

Example 3.7.17

In the last example, Example 3.7.17, we converted an $\infty - \infty$ indeterminate form into a $\frac{0}{0}$ indeterminate form by exploiting the fact that the two terms, $\sec x$ and $\tan x$, in the $\infty - \infty$ indeterminate form shared a common denominator, namely $\cos x$. In the "real world" that will, of course, almost never happen. However as the next couple of examples show, you can often massage these expressions into suitable forms.

Here is another, much more complicated, example, where it doesn't happen.

Example 3.7.18

In this example, we evaluate the $\infty - \infty$ indeterminate form

$$\lim_{x \to 0} \Big(\underbrace{\frac{1}{x}}_{\to \pm\infty} - \underbrace{\frac{1}{\log(1+x)}}_{\to \pm\infty} \Big)$$

We convert it into a $\frac{0}{0}$ indeterminate form simply by putting the two fractions, $\frac{1}{x}$ and $\frac{1}{\log(1+x)}$ over a common denominator.

$$\lim_{x \to 0} \Big(\underbrace{\frac{1}{x}}_{\to \pm\infty} - \underbrace{\frac{1}{\log(1+x)}}_{\to \pm\infty} \Big) = \lim_{x \to 0} \underbrace{\frac{\log(1+x) - x}{x \log(1+x)}}_{\substack{\text{num} \to 0 \\ \text{den} \to 0}} \tag{E1}$$

Now we apply L'Hôpital's rule, and simplify

$$\lim_{x \to 0} \underbrace{\frac{\log(1+x) - x}{x \log(1+x)}}_{\substack{\text{num} \to 0 \\ \text{den} \to 0}} = \lim_{x \to 0} \frac{\frac{1}{1+x} - 1}{\log(1+x) + \frac{x}{1+x}} = \lim_{x \to 0} \frac{1 - (1+x)}{(1+x)\log(1+x) + x}$$

$$= -\lim_{x \to 0} \underbrace{\frac{x}{(1+x)\log(1+x) + x}}_{\substack{\text{num} \to 0 \\ \text{den} \to 1 \times 0 + 0 = 0}} \tag{E2}$$

Then we apply L'Hôpital's rule a second time

$$-\lim_{x \to 0} \underbrace{\frac{x}{(1+x)\log(1+x) + x}}_{\substack{\text{num} \to 0 \\ \text{den} \to 1 \times 0 + 0 = 0}} = -\lim_{x \to 0} \underbrace{\frac{1}{\log(1+x) + \frac{1+x}{1+x} + 1}}_{\substack{\text{num} \to 1 \\ \text{den} \to 0 + 1 + 1 = 2}} = -\frac{1}{2} \tag{E3}$$

Combining (E1), (E2) and (E3) gives our final answer

$$\lim_{x \to 0} \Big(\frac{1}{x} - \frac{1}{\log(1+x)} \Big) = -\frac{1}{2}$$

Example 3.7.18

The following example can be done by l'Hôpital's rule, but it is actually far simpler to multiply by the conjugate and take the limit using the tools of Chapter 1.

Example 3.7.19

Consider the limit

$$\lim_{x \to \infty} \sqrt{x^2 + 4x} - \sqrt{x^2 - 3x}$$

Neither term is a fraction, but we can write

$$\sqrt{x^2 + 4x} - \sqrt{x^2 - 3x} = x\sqrt{1 + 4/x} - x\sqrt{1 - 3/x} \qquad \text{assuming } x > 0$$
$$= x\left(\sqrt{1 + 4/x} - \sqrt{1 - 3/x}\right)$$
$$= \frac{\sqrt{1 + 4/x} - \sqrt{1 - 3/x}}{1/x}$$

which is now a $0/0$ form with $f(x) = \sqrt{1 + 4/x} - \sqrt{1 - 3/x}$ and $g(x) = 1/x$. Then

$$f'(x) = \frac{-4/x^2}{2\sqrt{1 + 4/x}} - \frac{3/x^2}{2\sqrt{1 - 3/x}} \qquad\qquad g'(x) = -\frac{1}{x^2}$$

Hence

$$\frac{f'(x)}{g'(x)} = \frac{4}{2\sqrt{1 + 4/x}} + \frac{3}{\sqrt{1 - 3/x}}$$

And so in the limit as $x \to \infty$

$$\lim_{x \to \infty} \frac{f'(x)}{g'(x)} = \frac{4}{2} + \frac{3}{2} = \frac{7}{2}$$

and so our original limit is also $7/2$.

By comparison, if we multiply by the conjugate we have

$$\sqrt{x^2 + 4x} - \sqrt{x^2 - 3x} = \left(\sqrt{x^2 + 4x} - \sqrt{x^2 - 3x}\right) \cdot \frac{\sqrt{x^2 + 4x} + \sqrt{x^2 - 3x}}{\sqrt{x^2 + 4x} + \sqrt{x^2 - 3x}}$$
$$= \frac{x^2 + 4x - (x^2 - 3x)}{\sqrt{x^2 + 4x} + \sqrt{x^2 - 3x}}$$
$$= \frac{7x}{\sqrt{x^2 + 4x} + \sqrt{x^2 - 3x}}$$
$$= \frac{7}{\sqrt{1 + 4/x} + \sqrt{1 - 3/x}} \qquad \text{assuming } x > 0$$

Now taking the limit as $x \to \infty$ gives $7/2$ as required. Just because we know l'Hôpital's rule, it does not mean we should use it everywhere it might be applied.

Example 3.7.19

(e) 1^∞ *indeterminate form*:   We can use l'Hôpital's rule on limits of the form

$$\lim_{x \to a} f(x)^{g(x)} \text{ with}$$

$$\lim_{x \to a} f(x) = 1 \qquad \text{and} \qquad \lim_{x \to a} g(x) = \infty$$

by considering the logarithm of the limit[68]:

$$\log\left(\lim_{x \to a} f(x)^{g(x)}\right) = \lim_{x \to a} \log\left(f(x)^{g(x)}\right) = \lim_{x \to a} \log\left(f(x)\right) \cdot g(x)$$

which is now an $0 \cdot \infty$ form. This can be further transformed into a $0/0$ or ∞/∞ form:

$$\log\left(\lim_{x \to a} f(x)^{g(x)}\right) = \lim_{x \to a} \log\left(f(x)\right) \cdot g(x)$$
$$= \lim_{x \to a} \frac{\log\left(f(x)\right)}{1/g(x)}.$$

Example 3.7.20

The following limit appears quite naturally when considering systems which display exponential growth or decay.

$$\lim_{x \to 0} (1 + x)^{a/x} \qquad \text{with the constant } a \neq 0$$

Since $(1 + x) \to 1$ and $a/x \to \infty$ this is an 1^∞ indeterminate form.

By considering its logarithm we have

$$\log\left(\lim_{x \to 0} (1 + x)^{a/x}\right) = \lim_{x \to 0} \log\left((1 + x)^{a/x}\right)$$
$$= \lim_{x \to 0} \frac{a}{x} \log(1 + x)$$
$$= \lim_{x \to 0} \frac{a \log(1 + x)}{x}$$

which is now a $0/0$ form. Applying l'Hôpital's rule gives

$$\lim_{x \to 0} \underbrace{\frac{a \log(1 + x)}{x}}_{\substack{\text{num} \to 0 \\ \text{den} \to 0}} = \lim_{x \to 0} \underbrace{\frac{\frac{a}{1+x}}{1}}_{\substack{\text{num} \to a \\ \text{den} \to 1}} = a$$

Since $(1 + x)^{a/x} = \exp\left[\log\left((1 + x)^{a/x}\right)\right]$ and the exponential function is continuous, our original limit is e^a.

Example 3.7.20

---

68   We are using the fact that the logarithm is a continuous function and Theorem 1.6.10.

Here is a more complicated example of a 1^{∞} indeterminate form.

Example 3.7.21

In the limit

$$\lim_{x \to 0} \left(\frac{\sin x}{x} \right)^{1/x^2}$$

the base, $\frac{\sin x}{x}$, converges to 1 (see Example 3.7.3) and the exponent, $\frac{1}{x^2}$, goes to ∞. But if we take logarithms then

$$\log \left(\frac{\sin x}{x} \right)^{1/x^2} = \frac{\log \frac{\sin x}{x}}{x^2}$$

then, in the limit $x \to 0$, we have a $0/0$ indeterminate form. One application of l'Hôpital's rule gives

$$\lim_{x \to 0} \underbrace{\frac{\log \frac{\sin x}{x}}{x^2}}_{\substack{\text{num} \to 0 \\ \text{den} \to 0}} = \lim_{x \to 0} \frac{\frac{x}{\sin x} \frac{x \cos x - \sin x}{x^2}}{2x} = \lim_{x \to 0} \frac{\frac{x \cos x - \sin x}{x \sin x}}{2x} = \lim_{x \to 0} \frac{x \cos x - \sin x}{2x^2 \sin x}$$

which is another $0/0$ form. Applying l'Hôpital's rule again gives:

$$\lim_{x \to 0} \underbrace{\frac{x \cos x - \sin x}{2x^2 \sin x}}_{\substack{\text{num} \to 0 \\ \text{den} \to 0}} = \lim_{x \to 0} \frac{\cos x - x \sin x - \cos x}{4x \sin x + 2x^2 \cos x}$$

$$= -\lim_{x \to 0} \frac{x \sin x}{4x \sin x + 2x^2 \cos x} = -\lim_{x \to 0} \frac{\sin x}{4 \sin x + 2x \cos x}$$

which is yet another $0/0$ form. Once more with l'Hôpital's rule:

$$-\lim_{x \to 0} \underbrace{\frac{\sin x}{4 \sin x + 2x \cos x}}_{\substack{\text{num} \to 0 \\ \text{den} \to 0}} = -\lim_{x \to 0} \underbrace{\frac{\cos x}{4 \cos x + 2 \cos x - 2x \sin x}}_{\substack{\text{num} \to 1 \\ \text{den} \to 6}}$$

$$= -\frac{1}{6}$$

Oof! We have just shown that the logarithm of our original limit is $-1/6$. Hence the original limit itself is $e^{-1/6}$.

This was quite a complicated example. However it does illustrate the importance of cleaning up your algebraic expressions. This will both reduce the amount of work you have to do and will also reduce the number of errors you make.

Example 3.7.21

(f) 0^0 *indeterminate form*:   Like the 1^∞ form, this can be treated by considering its logarithm.

Example 3.7.22

For example, in the limit

$$\lim_{x\to 0+} x^x$$

both the base, x, and the exponent, also x, go to zero. But if we consider the logarithm then we have

$$\log x^x = x \log x$$

which is a $0 \cdot \infty$ indeterminate form, which we already know how to treat. In fact, we already found, in Example 3.7.15, that

$$\lim_{x\to 0+} x \log x = 0$$

Since the exponential is a continuous function

$$\lim_{x\to 0+} x^x = \lim_{x\to 0+} \exp\left(x \log x\right) = \exp\left(\lim_{x\to 0+} x \log x\right) = e^0 = 1$$

Example 3.7.22

(g) ∞^0 *indeterminate form*:   Again, we can treat this form by considering its logarithm.

Example 3.7.23

For example, in the limit

$$\lim_{x\to +\infty} x^{1/x}$$

the base, x, goes to infinity and the exponent, $\frac{1}{x}$, goes to zero. But if we take logarithms

$$\log x^{1/x} = \frac{\log x}{x}$$

which is an ∞/∞ form, which we know how to treat.

$$\lim_{x\to +\infty} \underbrace{\frac{\log x}{x}}_{\substack{\text{num}\to\infty \\ \text{den}\to\infty}} = \lim_{x\to +\infty} \underbrace{\frac{\frac{1}{x}}{1}}_{\substack{\text{num}\to 0 \\ \text{den}\to 1}} = 0$$

Since the exponential is a continuous function

$$\lim_{x\to +\infty} x^{1/x} = \lim_{x\to +\infty} \exp\left(\frac{\log x}{x}\right) = \exp\left(\lim_{x\to\infty} \frac{\log x}{x}\right) = e^0 = 1$$

Example 3.7.23

TOWARDS INTEGRAL CALCULUS

4.1 ▲ Introduction to Antiderivatives

We have now come to the final topic of the course — antiderivatives. This is only a short section since it is really just to give a taste of the next calculus subject: integral calculus.

So far in the course we have learned how to determine the rate of change (i.e. the derivative) of a given function. That is

$$\text{given a function } f(x) \text{ find } \frac{\mathrm{d}f}{\mathrm{d}x}.$$

Along the way we developed an understanding of limits, which allowed us to define instantaneous rates of change — the derivative. We then went on to develop a number of applications of derivatives to modelling and approximation. In this last section we want to just introduce the idea of antiderivatives. That is

$$\text{given a derivative } \frac{\mathrm{d}f}{\mathrm{d}x} \text{ find the original function } f(x).$$

For example — say we know that

$$\frac{\mathrm{d}f}{\mathrm{d}x} = x^2$$

and we want to find $f(x)$. From our previous experience differentiating we know that derivatives of polynomials are again polynomials. So we guess that our unknown function $f(x)$ is a polynomial. Further we know that when we differentiate x^n we get nx^{n-1} — multiply by the exponent and reduce the exponent by 1. So to end up with a derivative of x^2 we need to have differentiated an x^3. But $\frac{\mathrm{d}}{\mathrm{d}x}x^3 = 3x^2$, so we need to divide both sides by 3 to get the answer we want. That is

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{1}{3}x^3\right) = x^2$$

However we know that the derivative of a constant is zero, so we also have

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{1}{3}x^3 + 1\right) = x^2$$

and

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{1}{3}x^3 - \pi\right) = x^2$$

At this point it will really help the discussion to give a name to what we are doing.

Definition 4.1.1.

A function $F(x)$ that satisfies

$$\frac{\mathrm{d}}{\mathrm{d}x}F(x) = f(x)$$

is called an antiderivative of $f(x)$.

Notice the use of the indefinite article there — *an* antiderivative. This is precisely because we can always add or subtract a constant to an antiderivative and when we differentiate we'll get the same answer. We can write this as a lemma, but it is actually just Corollary 2.13.12 (from back in the section on the mean-value theorem) in disguise.

Lemma 4.1.2.

Let $F(x)$ be an antiderivative of $f(x)$, then for any constant c, the function $F(x) + c$ is also an antiderivative of $f(x)$.

Because of this lemma we typically write antiderivatives with "$+c$" tacked on the end. That is, if we know that $F'(x) = f(x)$, then we would state that *the* antiderivative of $f(x)$ is

$$F(x) + c$$

where this "$+c$" is there to remind us that we can always add or subtract some constant and it will still be an antiderivative of $f(x)$. Hence the antiderivative of x^2 is

$$\frac{1}{3}x^3 + c$$

Similarly, the antiderivative of x^4 is

$$\frac{1}{5}x^5 + c$$

and for $\sqrt{x} = x^{1/2}$ it is

$$\frac{2}{3}x^{3/2} + c$$

This last one is tricky (at first glance) — but we can always check our answer by differentiating.

$$\frac{d}{dx}\left(\frac{2}{3}x^{3/2} + c\right) = \frac{2}{3} \cdot \frac{3}{2}x^{1/2} + 0 \qquad\qquad \checkmark$$

Now in order to determine the value of c we need more information. For example, we might be asked

Given that $g'(t) = t^2$ and $g(3) = 7$ find $g(t)$.

We are given the derivative and one piece of additional information and from these two facts we need to find the original function. From our work above we know that

$$g(t) = \frac{1}{3}t^3 + c$$

and we can find c from the other piece of information

$$7 = g(3) = \frac{1}{3} \cdot 27 + c = 9 + c$$

Hence $c = -2$ and so

$$g(t) = \frac{1}{3}t^3 - 2$$

We can then very easily check our answer by recomputing $g(3)$ and $g'(t)$. This is a good habit to get into.

Finding antiderivatives of polynomials is generally not too hard. We just need to use the rule

$$\text{if } f(x) = x^n \text{ then } F(x) = \frac{1}{n+1}x^{n+1} + c.$$

Of course this breaks down when $n = -1$. In order to find an antiderivative for $f(x) = \frac{1}{x}$ we need to remember that $\frac{d}{dx}\log x = \frac{1}{x}$, and more generally that $\frac{d}{dx}\log|x| = \frac{1}{x}$. See Example 2.10.4. So

$$\text{if } f(x) = \frac{1}{x} \text{ then } F(x) = \log|x| + c$$

Example 4.1.3

Let $f(x) = 3x^5 - 7x^2 + 2x + 3 + x^{-1} - x^{-2}$. Then the antiderivative of $f(x)$ is

$$F(x) = \frac{3}{6}x^6 - \frac{7}{3}x^3 + \frac{2}{2}x^2 + 3x + \log|x| - \frac{1}{-1}x^{-1} + c \qquad\qquad \text{clean it up}$$

$$= \frac{1}{2}x^6 - \frac{7}{3}x^3 + x^2 + 3x + \log|x| + x^{-1} + c$$

Now to check we should differentiate and hopefully we get back to where we started

$$F'(x) = \frac{6}{2}x^5 - \frac{7}{3} \cdot 3x^2 + 2x + 3 + \frac{1}{x} - x^{-2}$$

$$= 3x^5 - 7x^2 + 2x + 3 + \frac{1}{x} - x^{-2} \qquad \checkmark$$

Example 4.1.3

In your next calculus course you will develop a lot of machinery to help you find antiderivatives. At this stage about all that we can do is continue the sort of thing we have done. Think about the derivatives we know and work backwards. So, for example, we can take a list of derivatives

$F(x)$	1	x^n	$\sin x$	$\cos x$	$\tan x$	e^x	$\ln|x|$	$\arcsin x$	$\arctan x$
$f(x) = \frac{d}{dx}F(x)$	0	nx^{n-1}	$\cos x$	$-\sin x$	$\sec^2 x$	e^x	$\frac{1}{x}$	$\frac{1}{\sqrt{1-x^2}}$	$\frac{1}{1+x^2}$

and flip it upside down to give the tables

$f(x) = \frac{d}{dx}F(x)$	0	nx^{n-1}	$\cos x$	$-\sin x$	$\sec^2 x$	e^x	$\frac{1}{x}$
$F(x)$	c	$x^n + c$	$\sin x + c$	$\cos x + c$	$\tan x + c$	$e^x + c$	$\ln|x| + c$

$f(x) = \frac{d}{dx}F(x)$	$\frac{1}{\sqrt{1-x^2}}$	$\frac{1}{1+x^2}$
$F(x)$	$\arcsin x + c$	$\arctan x + c$

of antiderivatives. Here c is just a constant — any constant. But we can do a little more; clean up x^n by dividing by n and then replacing n by $n + 1$. Similarly we can tweak $\sin x$ by multiplying by -1:

$f(x) = \frac{d}{dx}F(x)$	0	x^n	$\cos x$	$\sin x$	$\sec^2 x$	e^x	$\frac{1}{x}$
$F(x)$	c	$\frac{1}{n+1}x^{n+1} + c$	$\sin x + c$	$-\cos x + c$	$\tan x + c$	$e^x + c$	$\ln|x| + c$

$f(x) = \frac{d}{dx}F(x)$	$\frac{1}{\sqrt{1-x^2}}$	$\frac{1}{1+x^2}$
$F(x)$	$\arcsin x + c$	$\arctan x + c$

Here are a couple more examples.

Example 4.1.4

Consider the functions

$$f(x) = \sin x + \cos 2x \qquad\qquad g(x) = \frac{1}{1 + 4x^2}$$

Find their antiderivatives.

Solution. The first one we can almost just look up our table. Let F be the antiderivative of f, then

$$F(x) = -\cos x + \sin 2x + c \qquad\qquad \text{is not quite right.}$$

When we differentiate to check things, we get a factor of two coming from the chain rule. Hence to compensate for that we multiply $\sin 2x$ by $\frac{1}{2}$:

$$F(x) = -\cos x + \frac{1}{2}\sin 2x + c$$

Differentiating this shows that we have the right answer.

Similarly, if we use G to denote the antiderivative of g, then it appears that G is nearly $\arctan x$. To get this extra factor of 4 we need to substitute $x \mapsto 2x$. So we try

$$G(x) = \arctan(2x) + c \qquad\qquad \text{which is nearly correct.}$$

Differentiating this gives us

$$G'(x) = \frac{2}{1 + (2x)^2} = 2g(x)$$

Hence we should multiply by $\frac{1}{2}$. This gives us

$$G(x) = \frac{1}{2}\arctan(2x) + c.$$

We can then check that this is, in fact, correct just by differentiating.

Example 4.1.4

Now let's do a more substantial example.

Example 4.1.5

Suppose that we are driving to class. We start at $x = 0$ at time $t = 0$. Our velocity is $v(t) = 50\sin(t)$. The class is at $x = 25$. When do we get there?

Solution. Let's denote by $x(t)$ our position at time t. We are told that

- $x(0) = 0$

- $x'(t) = 50\sin t$

We have to determine $x(t)$ and then find the time T that obeys $x(T) = 25$. Now armed with our table above we know that the antiderivative of $\sin t$ is just $-\cos t$. We can check this:

$$\frac{d}{dt}(-\cos t) = \sin t$$

We can then get the factor of 50 by multiplying both sides of the above equation by 50:

$$\frac{d}{dt}(-50\cos t) = 50\sin t$$

And of course, this is just *an* antiderivative of $50 \sin t$; to write down the general antiderivative we just add a constant c:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(-50 \cos t + c\right) = 50 \sin t$$

Since $v(t) = \frac{\mathrm{d}}{\mathrm{d}t}x(t)$, this antiderivative is $x(t)$:

$$x(t) = -50 \cos t + c$$

To determine c we make use of the other piece of information we are given, namely

$$x(0) = 0.$$

Substituting this in gives us

$$x(0) = -50 \cos 0 + c = -50 + c$$

Hence we must have $c = 50$ and so

$$x(t) = -50 \cos t + 50 = 50(1 - \cos t).$$

Now that we have our position as a function of time, we can determine how long it takes us to arrive there. That is, we can find the time T so that $x(T) = 25$.

$$25 = x(T) = 50(1 - \cos T) \qquad\qquad \text{so}$$
$$\frac{1}{2} = 1 - \cos T$$
$$-\frac{1}{2} = - \cos T$$
$$\frac{1}{2} = \cos T.$$

Recalling our special triangles, we see that $T = \frac{\pi}{3}$.

Example 4.1.5

The example below shows how antiderivatives arise naturally when studying differential equations.

Example 4.1.6 (Theorem 3.3.2 revisited.)

Back in Section 3.3 we encountered a simple differential equation, namely equation 3.3.1. We were able to solve this equation by guessing the answer and then checking it carefully. We can derive the solution more systematically by using antiderivatives.

Recall equation 3.3.1:

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = -kQ$$

where $Q(t)$ is the amount of radioactive material at time t and we assume $Q(t) > 0$. Take this equation and divide both sides by $Q(t)$ to get

$$\frac{1}{Q(t)} \frac{dQ}{dt} = -k$$

At this point we should[1] think that the left-hand side is familiar. Now is a good moment to look back at logarithmic differentiation in Section 2.10.

The left-hand side is just the derivative of $\log Q(t)$:

$$\frac{d}{dt} \left(\log Q(t) \right) = \frac{1}{Q(t)} \frac{dQ}{dt}$$
$$= -k$$

So to solve this equation, we are really being asked to find all functions $\log Q(t)$ having derivative $-k$. That is, we need to find all antiderivatives of $-k$. Of course that is just $-kt + c$. Hence we must have

$$\log Q(t) = -kt + c$$

and then taking the exponential of both sides gives

$$Q(t) = e^{-kt+c} = e^c \cdot e^{-kt} = Ce^{-kt}$$

where $C = e^c$. This is precisely Theorem 3.3.2.

Example 4.1.6

The above is a small example of the interplay between antiderivatives and differential equations.

Here is another example of how we might use antidifferentiation to compute areas or volumes.

Example 4.1.7

We know (especially if one has revised the material in the appendix and Appendix B.5.2 in particular) that the volume of a right-circular cone is

$$V = \frac{\pi}{3} r^2 h$$

where h is the height of the cone and r is the radius of its base. Now, the derivation of this formula given in Appendix B.5.2 is not too simple. We present an alternate proof here that uses antiderivatives.

---

1   Well — perhaps it is better to say "notice that". Let's not make this a moral point.
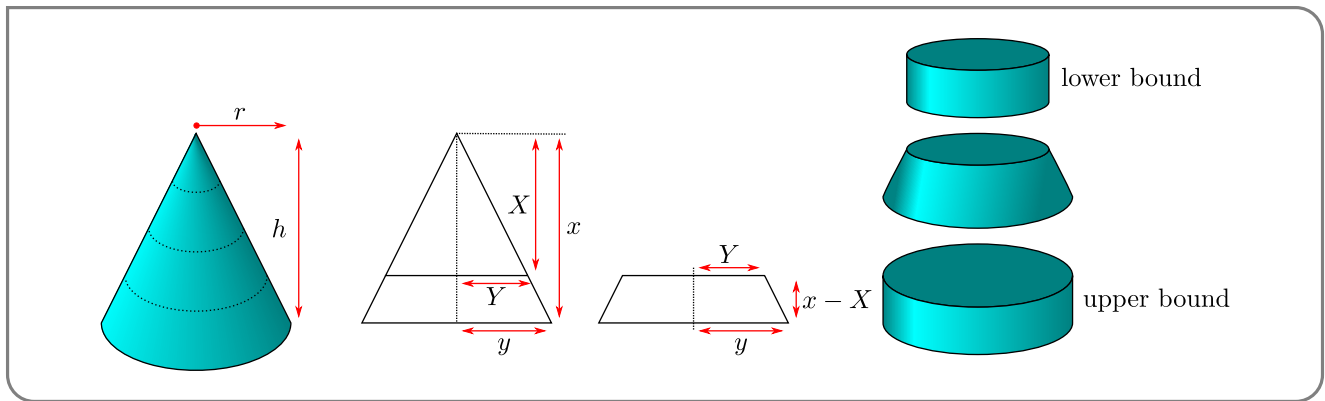
Consider cutting off a portion of the cone so that its new height is x (rather than h). Call the volume of the resulting smaller cone $V(x)$. We are going to determine $V(x)$ for all $x \geqslant 0$, including $x = h$, by first evaluating $V'(x)$ and $V(0)$ (which is obviously 0).

Call the radius of the base of the new smaller cone y (rather than r). By similar triangles we know that

$$\frac{r}{h} = \frac{y}{x}.$$

Now keep x and y fixed and consider cutting off a little more of the cone so its height is X. When we do so, the radius of the base changes from y to Y and again by similar triangles we know that

$$\frac{Y}{X} = \frac{y}{x} = \frac{r}{h}$$

The change in volume is then

$$V(x) - V(X)$$

Of course if we knew the formula for the volume of a cone, then we could compute the above exactly. However, even without knowing the volume of a cone, it is easy to derive upper and lower bounds on this quantity. The piece removed has bottom radius y and top radius Y. Hence its volume is bounded above and below by the cylinders of height $x - X$ and with radius y and Y respectively. Hence

$$\pi Y^2 (x - X) \leqslant V(x) - V(X) \leqslant \pi y^2 (x - X)$$

since the volume of a cylinder is just the area of its base times its height. Now massage this expression a little

$$\pi Y^2 \leqslant \frac{V(x) - V(X)}{x - X} \leqslant \pi y^2$$

The middle term now looks like a derivative; all we need to do is take the limit as $X \to x$:

$$\lim_{X \to x} \pi Y^2 \leqslant \lim_{X \to x} \frac{V(x) - V(X)}{x - X} \leqslant \lim_{X \to x} \pi y^2$$

The rightmost term is independent of X and so is just πy^2. In the leftmost term, as $X \to x$, we must have that $Y \to y$. Hence the leftmost term is just πy^2. Then by the squeeze theorem (Theorem 1.4.17) we know that

$$\frac{\mathrm{d}V}{\mathrm{d}x} = \lim_{X \to x} \frac{V(x) - V(X)}{x - X} = \pi y^2.$$

But we know that

$$y = \frac{r}{h} \cdot x$$

so

$$\frac{\mathrm{d}V}{\mathrm{d}x} = \pi \left(\frac{r}{h}\right)^2 x^2$$

Now we can antidifferentiate to get back to V:

$$V(x) = \frac{\pi}{3} \left(\frac{r}{h}\right)^2 x^3 + c$$

To determine c notice that when $x = 0$ the volume of the cone is just zero and so $c = 0$. Thus

$$V(x) = \frac{\pi}{3} \left(\frac{r}{h}\right)^2 x^3$$

and so when $x = h$ we are left with

$$V(h) = \frac{\pi}{3} \left(\frac{r}{h}\right)^2 h^3 = \frac{\pi}{3} r^2 h$$

as required.

Example 4.1.7

HIGH SCHOOL MATERIAL

This chapter is really split into two parts.

- Sections A.1 to A.13 we expect you to understand and know.

- The very last section, Section A.14, contains results that we don't expect you to memorise, but that we think you should be able to quickly derive from other results you know.

A.1 ▴ Similar Triangles
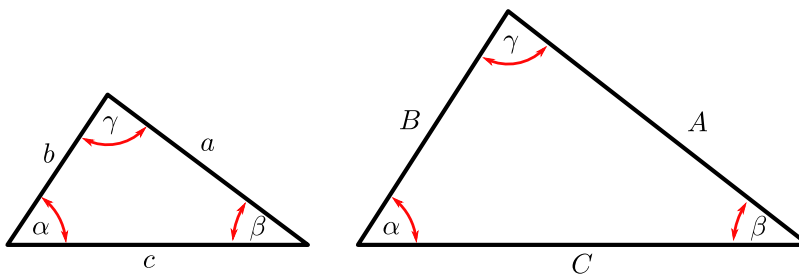
Two triangles T_1, T_2 are similar when

- (AAA — angle angle angle) The angles of T_1 are the same as the angles of T_2.

- (SSS — side side side) The ratios of the side lengths are the same. That is

$$\frac{A}{a} = \frac{B}{b} = \frac{C}{c}$$

- (SAS — side angle side) Two sides have lengths in the same ratio and the angle between them is the same. For example

$$\frac{A}{a} = \frac{C}{c} \text{ and angle } \beta \text{ is same}$$

A.2 ▴ Pythagoras

For a right-angled triangle the length of the hypotenuse is related to the lengths of the other two sides by
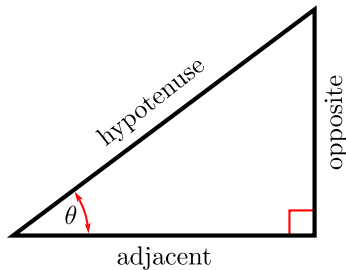
$$(\text{adjacent})^2 + (\text{opposite})^2 = (\text{hypotenuse})^2$$

A.3 ▴ Trigonometry — Definitions
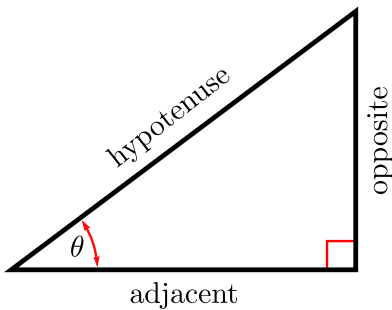
$$\sin\theta = \frac{\text{opposite}}{\text{hypotenuse}} \qquad \csc\theta = \frac{1}{\sin\theta}$$

$$\cos\theta = \frac{\text{adjacent}}{\text{hypotenuse}} \qquad \sec\theta = \frac{1}{\cos\theta}$$

$$\tan\theta = \frac{\text{opposite}}{\text{adjacent}} \qquad \cot\theta = \frac{1}{\tan\theta}$$

A.4 ▴ Radians, Arcs and Sectors
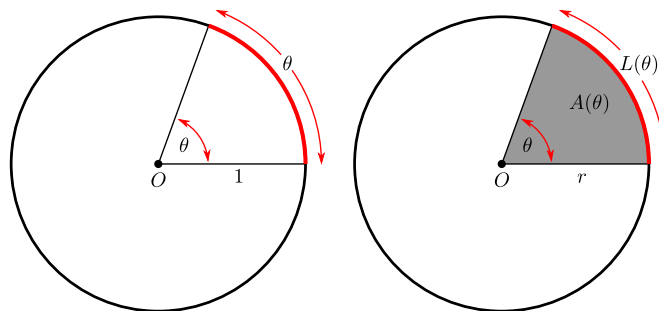
For a circle of radius r and angle of θ radians:

- Arc length $L(\theta) = r\theta$.

- Area of sector $A(\theta) = \frac{\theta}{2}r^2$.

A.5 ▴ Trigonometry — Graphs
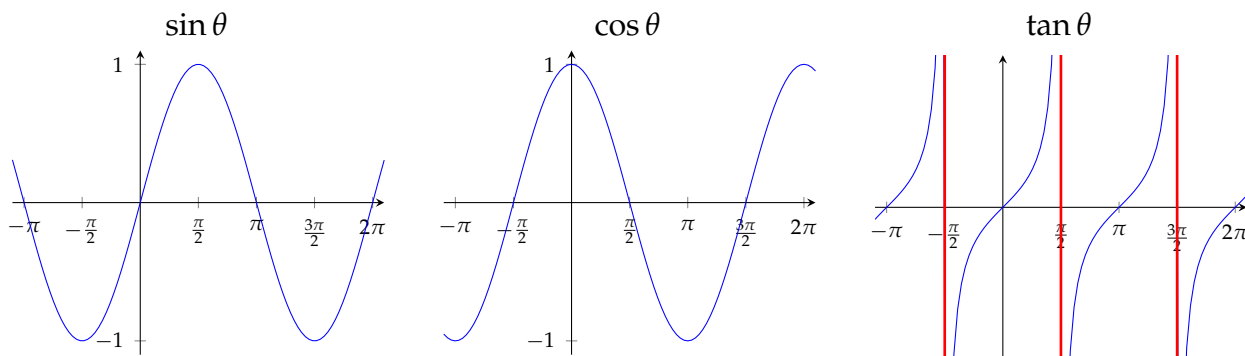
A.6 ▴ Trigonometry — Special Triangles
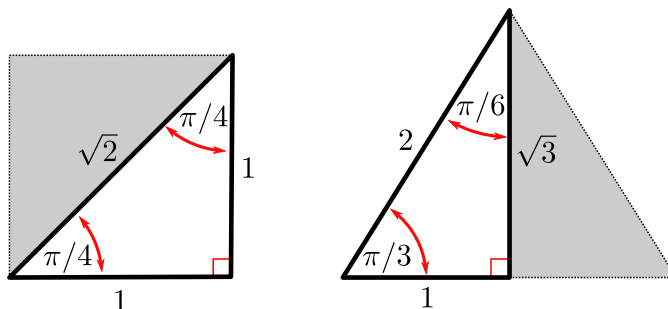
From the above pair of special triangles we have

$$\sin\frac{\pi}{4} = \frac{1}{\sqrt{2}} \qquad\qquad \sin\frac{\pi}{6} = \frac{1}{2} \qquad\qquad \sin\frac{\pi}{3} = \frac{\sqrt{3}}{2}$$

$$\cos\frac{\pi}{4} = \frac{1}{\sqrt{2}} \qquad\qquad \cos\frac{\pi}{6} = \frac{\sqrt{3}}{2} \qquad\qquad \cos\frac{\pi}{3} = \frac{1}{2}$$

$$\tan\frac{\pi}{4} = 1 \qquad\qquad \tan\frac{\pi}{6} = \frac{1}{\sqrt{3}} \qquad\qquad \tan\frac{\pi}{3} = \sqrt{3}$$

A.7 ▴ Trigonometry — Simple Identities

- Periodicity

$$\sin(\theta + 2\pi) = \sin(\theta) \qquad\qquad \cos(\theta + 2\pi) = \cos(\theta)$$

- Reflection

$$\sin(-\theta) = -\sin(\theta) \qquad\qquad \cos(-\theta) = \cos(\theta)$$

- Reflection around $\pi/4$

$$\sin\left(\tfrac{\pi}{2} - \theta\right) = \cos\theta \qquad\qquad \cos\left(\tfrac{\pi}{2} - \theta\right) = \sin\theta$$

- Reflection around $\pi/2$

$$\sin(\pi - \theta) = \sin\theta \qquad\qquad \cos(\pi - \theta) = -\cos\theta$$

- Rotation by π

$$\sin(\theta + \pi) = -\sin\theta \qquad\qquad \cos(\theta + \pi) = -\cos\theta$$

- Pythagoras

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

A.8 ▴ Trigonometry — Add and Subtract Angles

- Sine

$$\sin(\alpha \pm \beta) = \sin(\alpha)\cos(\beta) \pm \cos(\alpha)\sin(\beta)$$
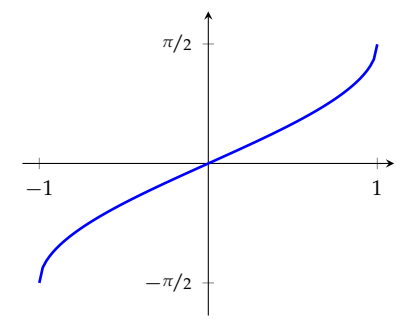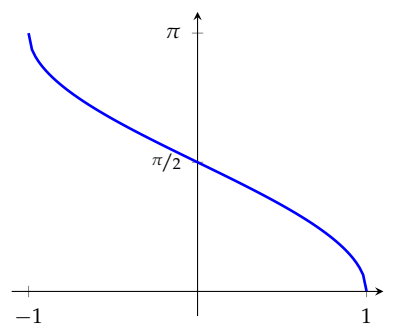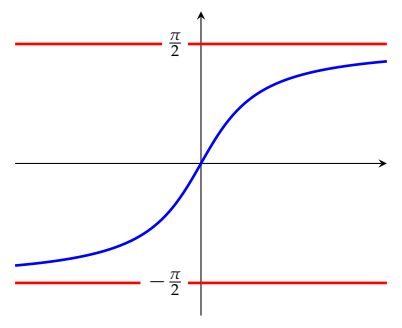
- Cosine

$$\cos(\alpha \pm \beta) = \cos(\alpha)\cos(\beta) \mp \sin(\alpha)\sin(\beta)$$

A.9 ▴ Inverse Trigonometric Functions

Some of you may not have studied inverse trigonometric functions in highschool, however we still expect you to know them by the end of the course.

| $\arcsin x$	$\arccos x$	$\arctan x$
Domain: $-1 \leqslant x \leqslant 1$	Domain: $-1 \leqslant x \leqslant 1$	Domain: all real numbers
Range: $-\frac{\pi}{2} \leqslant \arcsin x \leqslant \frac{\pi}{2}$	Range: $0 \leqslant \arccos x \leqslant \pi$	Range: $-\frac{\pi}{2} < \arctan x < \frac{\pi}{2}$

Since these functions are inverses of each other we have

$$\arcsin(\sin\theta) = \theta \qquad\qquad -\frac{\pi}{2} \leqslant \theta \leqslant \frac{\pi}{2}$$
$$\arccos(\cos\theta) = \theta \qquad\qquad 0 \leqslant \theta \leqslant \pi$$
$$\arctan(\tan\theta) = \theta \qquad\qquad -\frac{\pi}{2} \leqslant \theta \leqslant \frac{\pi}{2}$$

and also

$$\sin(\arcsin x) = x \qquad -1 \leqslant x \leqslant 1$$
$$\cos(\arccos x) = x \qquad -1 \leqslant x \leqslant 1$$
$$\tan(\arctan x) = x \qquad \text{any real } x$$

arccsc x	arcsec x	arccot x
Domain: $\lvert x \rvert \geqslant 1$	Domain: $\lvert x \rvert \geqslant 1$	Domain: all real numbers
Range: $-\frac{\pi}{2} \leqslant \operatorname{arccsc} x \leqslant \frac{\pi}{2}$ $\operatorname{arccsc} x \neq 0$	Range: $0 \leqslant \operatorname{arcsec} x \leqslant \pi$ $\operatorname{arcsec} x \neq \frac{\pi}{2}$	Range: $0 < \operatorname{arccot} x < \pi$

Again

$$\operatorname{arccsc}(\csc \theta) = \theta \qquad -\frac{\pi}{2} \leqslant \theta \leqslant \frac{\pi}{2},\ \theta \neq 0$$

$$\operatorname{arcsec}(\sec \theta) = \theta \qquad 0 \leqslant \theta \leqslant \pi,\ \theta \neq \frac{\pi}{2}$$

$$\operatorname{arccot}(\cot \theta) = \theta \qquad 0 < \theta < \pi$$

and

$$\csc(\operatorname{arccsc} x) = x \qquad \lvert x \rvert \geqslant 1$$
$$\sec(\operatorname{arcsec} x) = x \qquad \lvert x \rvert \geqslant 1$$
$$\cot(\operatorname{arccot} x) = x \qquad \text{any real } x$$

A.10 ▴ Areas
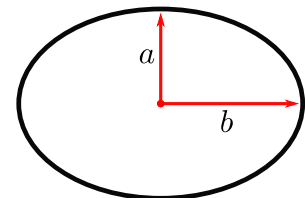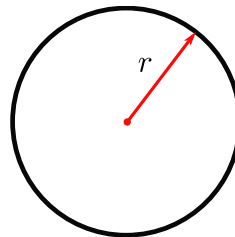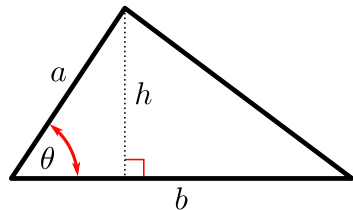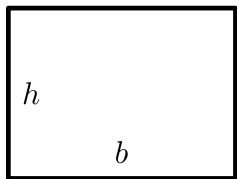
- Area of a rectangle

$$A = bh$$

- Area of a triangle

$$A = \frac{1}{2}bh = \frac{1}{2}ab\sin\theta$$

- Area of a circle

$$A = \pi r^2$$

- Area of an ellipse

$$A = \pi ab$$

A.11 ▴ Volumes
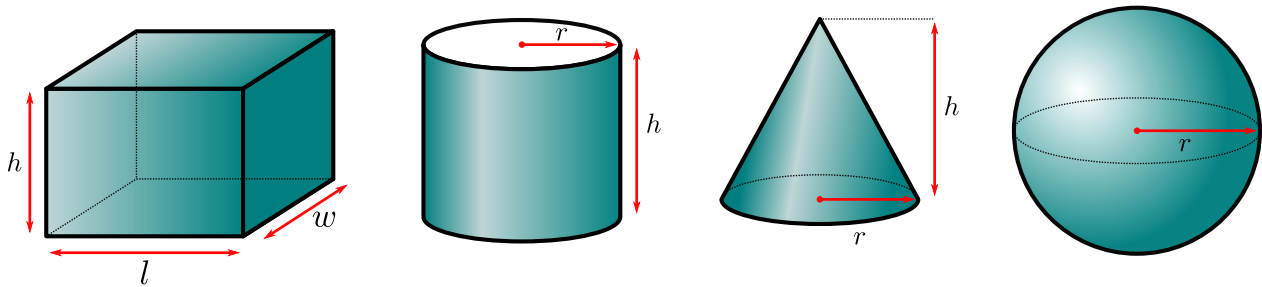
- Volume of a rectangular prism

$$V = lwh$$

- Volume of a cylinder

$$V = \pi r^2 h$$

- Volume of a cone
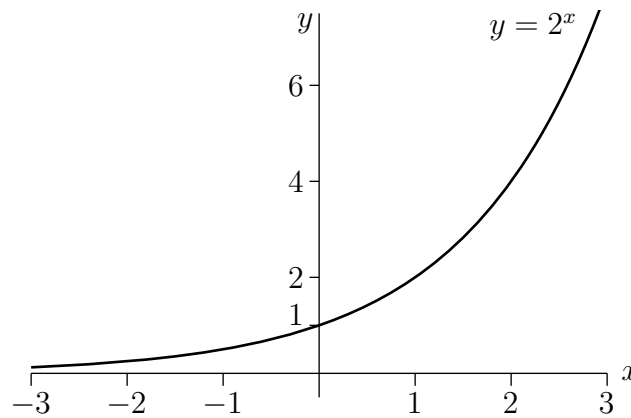
$$V = \frac{1}{3}\pi r^2 h$$

- Volume of a sphere

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

A.12 ▴ Powers

In the following, x and y are arbitrary real numbers, and q is an arbitrary constant that is strictly bigger than zero.

- $q^0 = 1$

- $q^{x+y} = q^x q^y, \, q^{x-y} = \frac{q^x}{q^y}$

- $q^{-x} = \frac{1}{q^x}$

- $(q^x)^y = q^{xy}$

- $\lim\limits_{x \to \infty} q^x = \infty, \, \lim\limits_{x \to -\infty} q^x = 0$ if $q > 1$

- $\lim\limits_{x \to \infty} q^x = 0, \, \lim\limits_{x \to -\infty} q^x = \infty$ if $0 < q < 1$

- The graph of 2^x is given below. The graph of q^x, for any $q > 1$, is similar.

A.13 ▴ Logarithms

In the following, x and y are arbitrary real numbers that are strictly bigger than 0, and p and q are arbitrary constants that are strictly bigger than one.

- $q^{\log_q x} = x, \, \log_q (q^x) = x$

- $\log_q x = \frac{\log_p x}{\log_p q}$

- $\log_q 1 = 0, \, \log_q q = 1$

- $\log_q (xy) = \log_q x + \log_q y$

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

- $\log_q \left(\frac{1}{y}\right) = -\log_q y,$

- $\log_q (x^y) = y \log_q x$

- $\lim\limits_{x \to \infty} \log_q x = \infty, \quad \lim\limits_{x \to 0+} \log_q x = -\infty$

- The graph of $\log_{10} x$ is given below. The graph of $\log_q x$, for any $q > 1$, is similar.
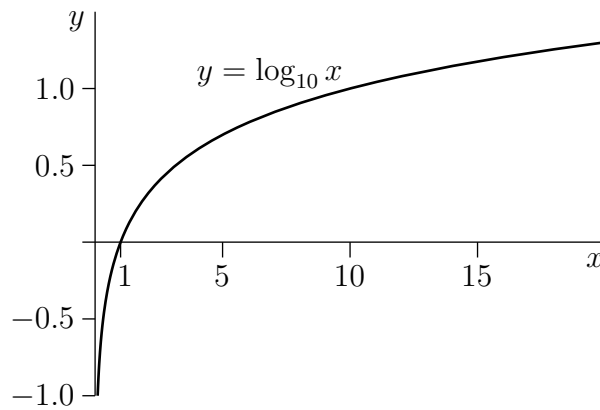
A.14 ▴ Highschool Material You Should be Able to Derive

- Graphs of $\csc\theta$, $\sec\theta$ and $\cot\theta$:
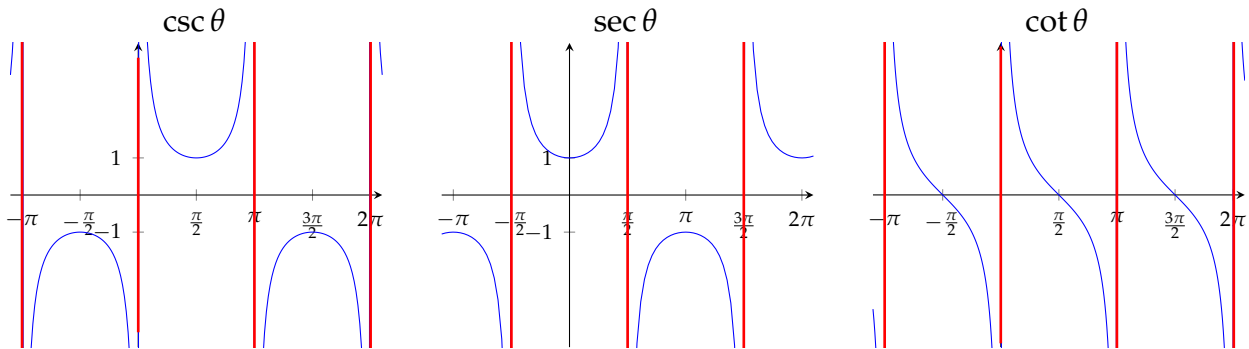
- More Pythagoras

$$\sin^2\theta + \cos^2\theta = 1 \qquad \xrightarrow{\text{divide by } \cos^2\theta} \qquad \tan^2\theta + 1 = \sec^2\theta$$

$$\sin^2\theta + \cos^2\theta = 1 \qquad \xrightarrow{\text{divide by } \sin^2\theta} \qquad 1 + \cot^2\theta = \csc^2\theta$$

- Sine — double angle (set $\beta = \alpha$ in sine angle addition formula)

$$\sin(2\alpha) = 2\sin(\alpha)\cos(\alpha)$$

- Cosine — double angle (set $\beta = \alpha$ in cosine angle addition formula)

$$
\begin{aligned}
\cos(2\alpha) &= \cos^2(\alpha) - \sin^2(\alpha) \\
&= 2\cos^2(\alpha) - 1 && (\text{use } \sin^2(\alpha) = 1 - \cos^2(\alpha)) \\
&= 1 - 2\sin^2(\alpha) && (\text{use } \cos^2(\alpha) = 1 - \sin^2(\alpha))
\end{aligned}
$$

- Composition of trigonometric and inverse trigonometric functions:

$$\cos(\arcsin x) = \sqrt{1 - x^2} \qquad\qquad \sec(\arctan x) = \sqrt{1 + x^2}$$

and similar expressions.

378

ORIGIN OF TRIG, AREA AND VOLUME FORMULAS

B.1 ▲ Theorems about Triangles

B.1.1 ➤ Thales' Theorem

We want to get at right-angled triangles. A classic construction for this is to draw a triangle inside a circle, so that all three corners lie on the circle and the longest side forms the diameter of the circle. See the figure below in which we have scaled the circle to have radius 1 and the triangle has longest side 2.
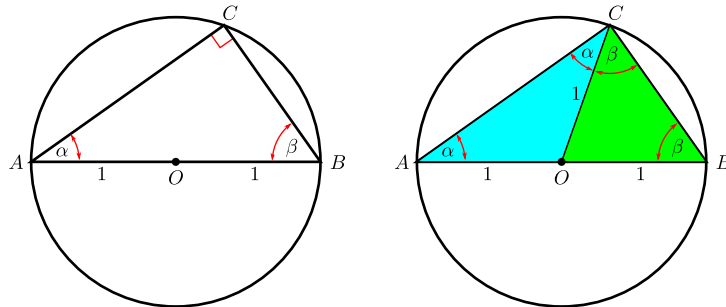
Thales theorem states that the angle at C is always a right-angle. The proof is quite straight-forward and relies on two facts:

- the angles of a triangle add to π, and

- the angles at the base of an isosceles triangle are equal.

So we split the triangle ABC by drawing a line from the centre of the circle to C. This creates two isosceles triangles OAC and OBC. Since they are isosceles, the angles at their bases α and β must be equal (as shown). Adding the angles of the original triangle now gives

$$\pi = \alpha + (\alpha + \beta) + \beta = 2(\alpha + \beta)$$

So the angle at $C = \pi - (\alpha + \beta) = \pi/2$.

B.1.2 ‣ Pythagoras

Since trigonometry, at its core, is the study of lengths and angles in right-angled triangles, we must include a result you all know well, but likely do not know how to prove.
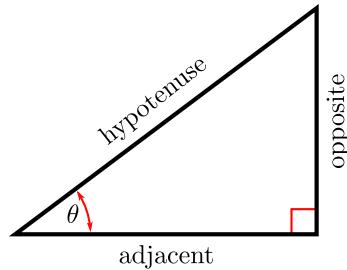
The lengths of the sides of any right-angled triangle are related by the famous result due to Pythagoras

$$c^2 = a^2 + b^2.$$

There are many ways to prove this, but we can do so quite simply by studying the following diagram:

white area $= a^2 + b^2$ white area $= c^2$
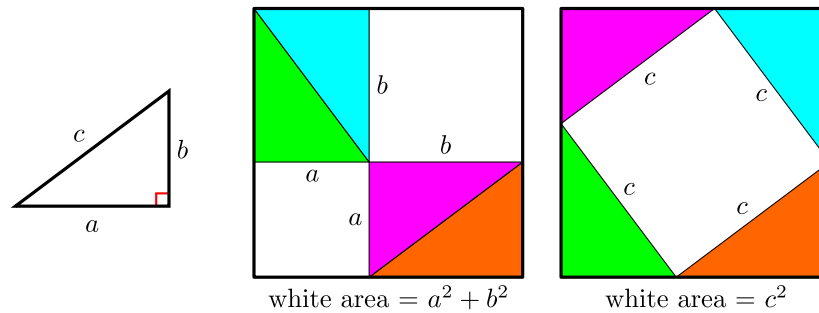
We start with a right-angled triangle with sides labeled a, b and c. Then we construct a square of side-length $a + b$ and draw inside it 4 copies of the triangle arranged as shown in the centre of the above figure. The area in white is then $a^2 + b^2$. Now move the triangles around to create the arrangement shown on the right of the above figure. The area in white is bounded by a square of side-length c and so its area is c^2. The area of the outer square didn't change when the triangles were moved, nor did the area of the triangles, so the white area cannot have changed either. This proves $a^2 + b^2 = c^2$.

B.2 ▴ Trigonometry

B.2.1 ‣ Angles — Radians vs Degrees

For mathematics, and especially in calculus, it is much better to measure angles in units called radians rather than degrees. By definition, an arc of length θ on a circle of radius one subtends an angle of θ radians at the centre of the circle.
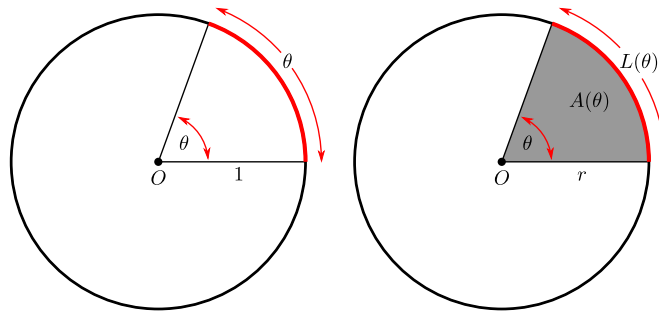
The circle on the left has radius 1, and the arc swept out by an angle of θ radians has length θ. Because a circle of radius one has circumference 2π we have

$$2\pi \text{ radians} = 360° \qquad \pi \text{ radians} = 180° \qquad \pi/2 \text{ radians} = 90°$$

$$\frac{\pi}{3} \text{ radians} = 60° \qquad \frac{\pi}{4} \text{ radians} = 45° \qquad \frac{\pi}{6} \text{ radians} = 30°$$

More generally, consider a circle of radius r. Let $L(\theta)$ denote the length of the arc swept out by an angle of θ radians and let $A(\theta)$ denote the area of the sector (or wedge) swept out by the same angle. Since the angle sweeps out the fraction $\theta/2\pi$ of a whole circle, we have

$$L(\theta) = 2\pi r \cdot \frac{\theta}{2\pi} = \theta r \qquad\qquad \text{and}$$

$$A(\theta) = \pi r^2 \cdot \frac{\theta}{2\pi} = \frac{\theta}{2} r^2$$

B.2.2 ▸▸ Trig Function Definitions

The trigonometric functions are defined as ratios of the lengths of the sides of a right-angle triangle as shown in the left of the diagram below . These ratios depend only on the angle θ.
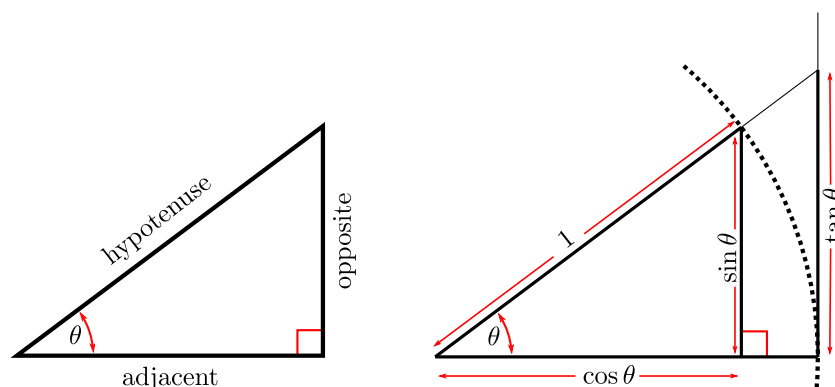
The trigonometric functions sine, cosine and tangent are defined as ratios of the lengths of the sides

$$\sin\theta = \frac{\text{opposite}}{\text{hypotenuse}} \qquad \cos\theta = \frac{\text{adjacent}}{\text{hypotenuse}} \qquad \tan\theta = \frac{\text{opposite}}{\text{adjacent}} = \frac{\sin\theta}{\cos\theta}.$$

These are frequently abbreviated as

$$\sin\theta = \frac{o}{h} \qquad\qquad \cos\theta = \frac{a}{h} \qquad\qquad \tan\theta = \frac{o}{a}$$
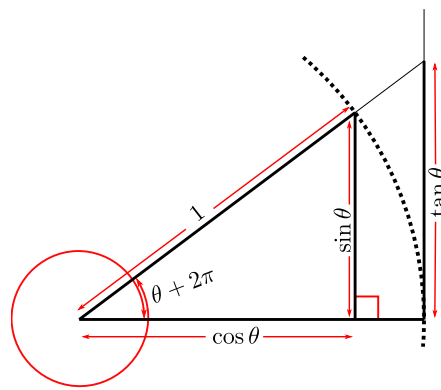
which gives rise to the mnemonic

SOH CAH TOA

If we scale the triangle so that they hypotenuse has length 1 then we obtain the diagram on the right. In that case, $\sin\theta$ is the height of the triangle, $\cos\theta$ the length of its base and $\tan\theta$ is the length of the line tangent to the circle of radius 1 as shown.
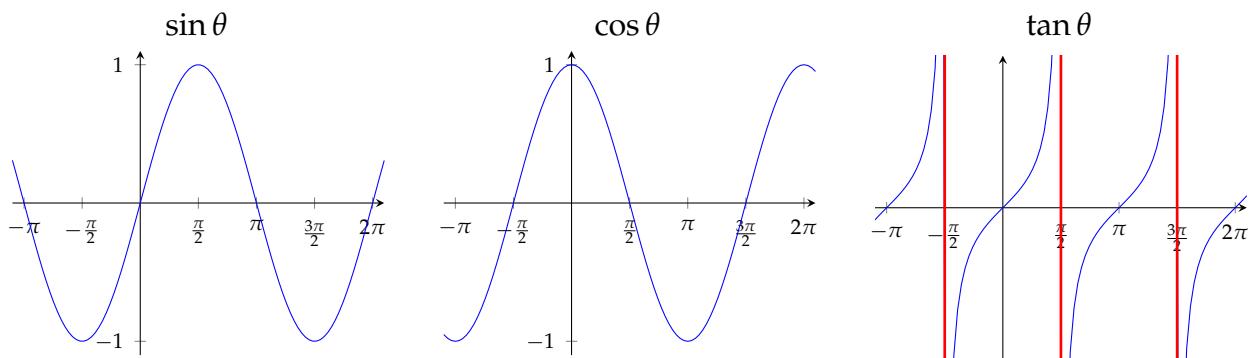
Since the angle 2π sweeps out a full circle, the angles θ and $\theta + 2\pi$ are really the same.

Hence all the trigonometric functions are periodic with period 2π. That is

$$\sin(\theta + 2\pi) = \sin(\theta) \qquad \cos(\theta + 2\pi) = \cos(\theta) \qquad \tan(\theta + 2\pi) = \tan(\theta)$$
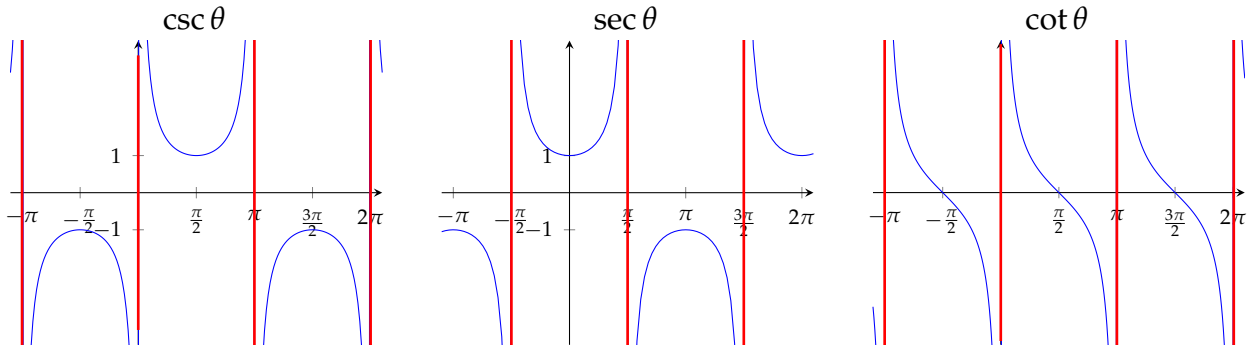
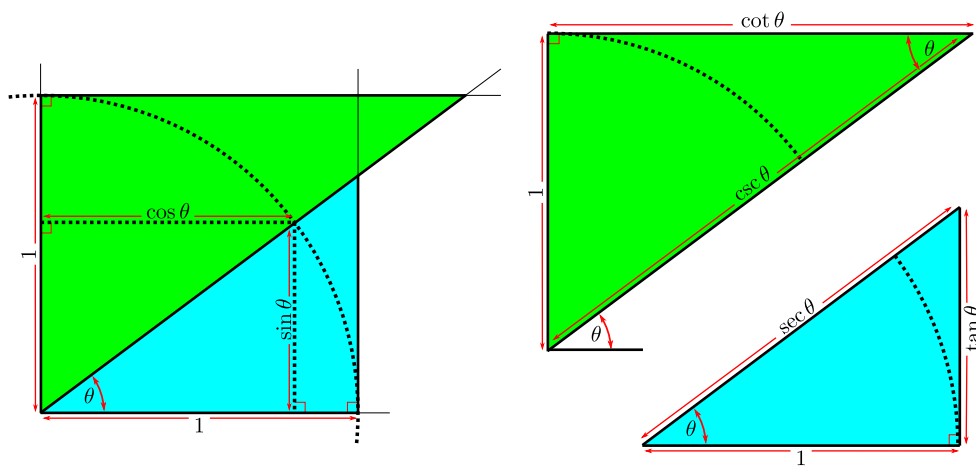The plots of these functions are shown below

The reciprocals (cosecant, secant and cotangent) of these functions also play important roles in trigonometry and calculus:

$$\csc\theta = \frac{1}{\sin\theta} = \frac{h}{o} \qquad \sec\theta = \frac{1}{\cos\theta} = \frac{h}{a} \qquad \cot\theta = \frac{1}{\tan\theta} = \frac{\cos\theta}{\sin\theta} = \frac{a}{o}$$

The plots of these functions are shown below

These reciprocal functions also have geometric interpretations:

Since these are all right-angled triangles we can use Pythagoras to obtain the following identities:

$$\sin^2 \theta + \cos^2 \theta = 1 \qquad \tan^2 \theta + 1 = \sec^2 \theta \qquad 1 + \cot^2 \theta = \csc^2 \theta$$

Of these it is only necessary to remember the first

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

The second can then be obtained by dividing this by $\cos^2 \theta$ and the third by dividing by $\sin^2 \theta$.

B.2.3 ▸ Important Triangles

Computing sine and cosine is non-trivial for general angles — we need Taylor series (or similar tools) to do this. However there are some special angles (usually small integer fractions of π) for which we can use a little geometry to help. Consider the following two triangles.
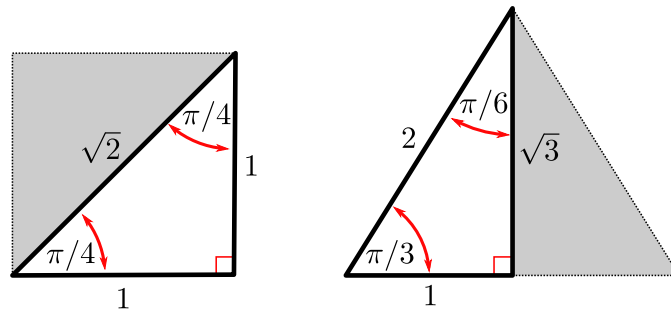
The first results from cutting a square along its diagonal, while the second is obtained by cutting an equilateral triangle from one corner to the middle of the opposite side. These, together with the angles 0, $\frac{\pi}{2}$ and π give the following table of values

θ	$\sin\theta$	$\cos\theta$	$\tan\theta$	$\csc\theta$	$\sec\theta$	$\cot\theta$
0 rad	0	1	0	DNE	1	DNE
$\frac{\pi}{2}$ rad	1	0	DNE	1	DNE	0
π rad	0	-1	0	DNE	-1	DNE
$\frac{\pi}{4}$ rad	$\frac{1}{\sqrt{2}}$	$\frac{1}{\sqrt{2}}$	1	$\sqrt{2}$	$\sqrt{2}$	1
$\frac{\pi}{6}$ rad	$\frac{1}{2}$	$\frac{\sqrt{3}}{2}$	$\frac{1}{\sqrt{3}}$	2	$\frac{2}{\sqrt{3}}$	$\sqrt{3}$
$\frac{\pi}{3}$ rad	$\frac{\sqrt{3}}{2}$	$\frac{1}{2}$	$\sqrt{3}$	$\frac{2}{\sqrt{3}}$	2	$\frac{1}{\sqrt{3}}$

B.2.4 ▸▸ Some More Simple Identities

Consider the figure below
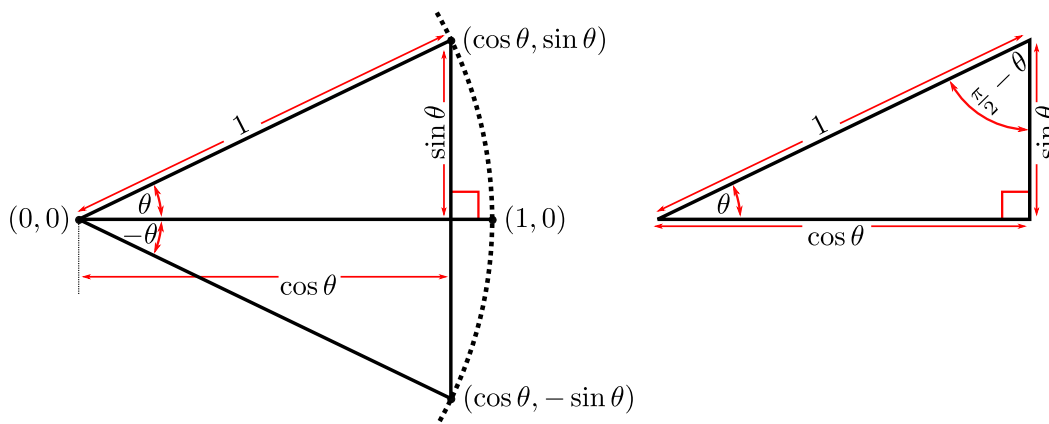
The pair triangles on the left shows that there is a simple relationship between trigonometric functions evaluated at θ and at $-\theta$:

$$\sin(-\theta) = -\sin(\theta) \qquad\qquad \cos(-\theta) = \cos(\theta)$$

That is — sine is an odd function, while cosine is even. Since the other trigonometric functions can be expressed in terms of sine and cosine we obtain

$$\tan(-\theta) = -\tan(\theta) \quad \csc(-\theta) = -\csc(\theta) \quad \sec(-\theta) = \sec(\theta) \quad \cot(-\theta) = -\cot(\theta)$$
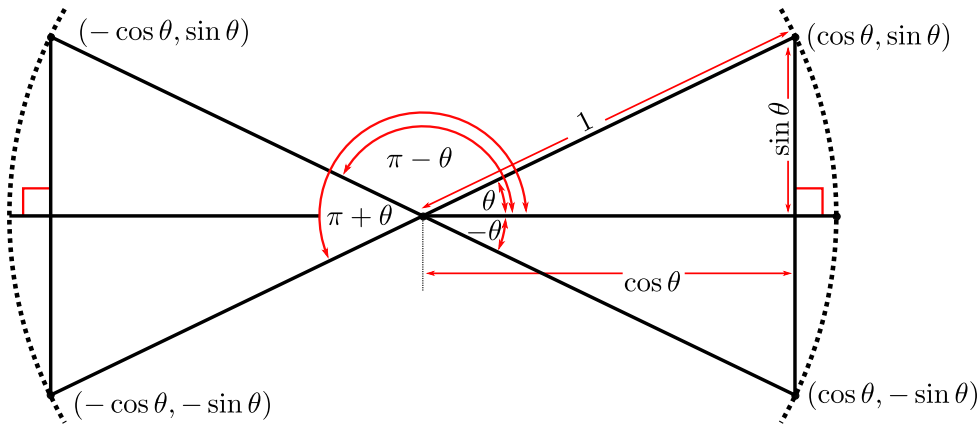
Now consider the triangle on the right — if we consider the angle $\frac{\pi}{2} - \theta$ the side-lengths of the triangle remain unchanged, but the roles of "opposite" and "adjacent" are swapped. Hence we have

$$\sin\left(\tfrac{\pi}{2} - \theta\right) = \cos\theta \qquad\qquad \cos\left(\tfrac{\pi}{2} - \theta\right) = \sin\theta$$

Again these imply that

$$\tan\left(\tfrac{\pi}{2} - \theta\right) = \cot\theta \quad \csc\left(\tfrac{\pi}{2} - \theta\right) = \sec\theta \quad \sec\left(\tfrac{\pi}{2} - \theta\right) = \csc\theta \quad \cot\left(\tfrac{\pi}{2} - \theta\right) = \tan\theta$$

We can go further. Consider the following diagram:

This implies that

$$\begin{aligned}
\sin(\pi - \theta) &= \sin(\theta) & \cos(\pi - \theta) &= -\cos(\theta) \\
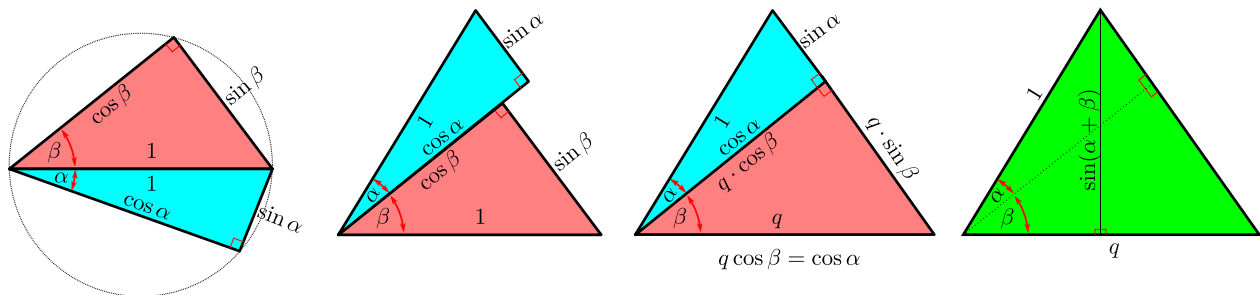\sin(\pi + \theta) &= -\sin(\theta) & \cos(\pi + \theta) &= -\cos(\theta)
\end{aligned}$$

From which we can get the rules for the other four trigonometric functions.

B.2.5 ▸▸ Identities — Adding Angles

We wish to explain the origins of the identity

$$\sin(\alpha + \beta) = \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta).$$

A very geometric demonstration uses the figure below and an observation about areas.

- The left-most figure shows two right-angled triangles with angles α and β and both with hypotenuse length 1.

- The next figure simply rearranges the triangles — translating and rotating the lower triangle so that it lies adjacent to the top of the upper triangle.

- Now scale the lower triangle by a factor of q so that edges opposite the angles α and β are flush. This means that $q \cos \beta = \cos \alpha$. ie

$$q = \frac{\cos \alpha}{\cos \beta}$$

Now compute the areas of these (blue and red) triangles

$$A_{\text{red}} = \frac{1}{2} q^2 \sin \beta \cos \beta$$

$$A_{\text{blue}} = \frac{1}{2} \sin \alpha \cos \alpha$$

So twice the total area is

$$2A_{\text{total}} = \sin \alpha \cos \alpha + q^2 \sin \beta \cos \beta$$

- But we can also compute the total area using the rightmost triangle:

$$2A_{\text{total}} = q \sin(\alpha + \beta)$$

Since the total area must be the same no matter how we compute it we have

$$q \sin(\alpha + \beta) = \sin \alpha \cos \alpha + q^2 \sin \beta \cos \beta$$
$$\sin(\alpha + \beta) = \frac{1}{q} \sin \alpha \cos \alpha + q \sin \beta \cos \beta$$
$$= \frac{\cos \beta}{\cos \alpha} \sin \alpha \cos \alpha + \frac{\cos \alpha}{\cos \beta} \sin \beta \cos \beta$$
$$= \sin \alpha \cos \beta + \cos \alpha \sin \beta$$

as required.

We can obtain the angle addition formula for cosine by substituting $\alpha \mapsto \pi/2 - \alpha$ and $\beta \mapsto -\beta$ into our sine formula:

$$\sin(\alpha + \beta) = \sin(\alpha) \cos(\beta) + \cos(\alpha) \sin(\beta) \qquad \text{becomes}$$
$$\underbrace{\sin(\pi/2 - \alpha - \beta)}_{\cos(\alpha+\beta)} = \underbrace{\sin(\pi/2 - \alpha)}_{\cos(\alpha)} \cos(-\beta) + \underbrace{\cos(\pi/2 - \alpha)}_{\sin(\alpha)} \sin(-\beta)$$
$$\cos(\alpha + \beta) = \cos(\alpha) \cos(\beta) - \sin(\alpha) \sin(\beta)$$

where we have used $\sin(\pi/2 - \theta) = \cos(\theta)$ and $\cos(\pi/2 - \theta) = \sin(\theta)$.

It is then a small step to the formulas for the difference of angles. From the relation

$$\sin(\alpha + \beta) = \sin(\alpha) \cos(\beta) + \cos(\alpha) \sin(\beta)$$

we can substitute $\beta \mapsto -\beta$ and so obtain

$$\sin(\alpha - \beta) = \sin(\alpha) \cos(-\beta) + \cos(\alpha) \sin(-\beta)$$
$$= \sin(\alpha) \cos(\beta) - \cos(\alpha) \sin(\beta)$$

The formula for cosine can be obtained in a similar manner. To summarise

$$\sin(\alpha \pm \beta) = \sin(\alpha)\cos(\beta) \pm \cos(\alpha)\sin(\beta)$$
$$\cos(\alpha \pm \beta) = \cos(\alpha)\cos(\beta) \mp \sin(\alpha)\sin(\beta)$$

The formulas for tangent are a bit more work, but

$$
\begin{aligned}
\tan(\alpha + \beta) &= \frac{\sin(\alpha + \beta)}{\cos(\alpha + \beta)} \\
&= \frac{\sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta)}{\cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta)} \\
&= \frac{\sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta)}{\cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta)} \cdot \frac{\sec(\alpha)\sec(\beta)}{\sec(\alpha)\sec(\beta)} \\
&= \frac{\sin(\alpha)\sec(\alpha) + \sin(\beta)\sec(\beta)}{1 - \sin(\alpha)\sec(\alpha)\sin(\beta)\sec(\beta)} \\
&= \frac{\tan(\alpha) + \tan(\beta)}{1 - \tan(\alpha)\tan(\beta)}
\end{aligned}
$$

and similarly we get

$$\tan(\alpha - \beta) = \frac{\tan(\alpha) - \tan(\beta)}{1 + \tan(\alpha)\tan(\beta)}$$

B.2.6 ▸▸ Identities — Double-angle Formulas

If we set $\beta = \alpha$ in the angle-addition formulas we get

$$
\begin{aligned}
\sin(2\alpha) &= 2\sin(\alpha)\cos(\alpha) \\
\cos(2\alpha) &= \cos^2(\alpha) - \sin^2(\alpha) \\
&= 2\cos^2(\alpha) - 1 && \text{since } \sin^2\theta = 1 - \cos^2\theta \\
&= 1 - 2\sin^2(\alpha) && \text{since } \cos^2\theta = 1 - \sin^2\theta \\
\tan(2\alpha) &= \frac{2\tan(\alpha)}{1 - \tan^2(\alpha)} \\
&= \frac{2}{\cot(\alpha) - \tan(\alpha)} && \text{divide top and bottom by } \tan(\alpha)
\end{aligned}
$$

B.2.7 ▸▸ Identities — Extras

▸▸▸ Sums to Products

Consider the identities

$$\sin(\alpha + \beta) = \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta) \quad \sin(\alpha - \beta) = \sin(\alpha)\cos(\beta) - \cos(\alpha)\sin(\beta)$$

If we add them together some terms on the right-hand side cancel:

$$\sin(\alpha + \beta) + \sin(\alpha - \beta) = 2\sin(\alpha)\cos(\beta).$$

If we now set $u = \alpha + \beta$ and $v = \alpha - \beta$ (i.e. $\alpha = \frac{u+v}{2}, \beta = \frac{u-v}{2}$) then

$$\sin(u) + \sin(v) = 2\sin\left(\frac{u+v}{2}\right)\cos\left(\frac{u-v}{2}\right)$$

This transforms a sum into a product. Similarly:

$$\sin(u) - \sin(v) = 2\sin\left(\frac{u-v}{2}\right)\cos\left(\frac{u+v}{2}\right)$$

$$\cos(u) + \cos(v) = 2\cos\left(\frac{u+v}{2}\right)\cos\left(\frac{u-v}{2}\right)$$

$$\cos(u) - \cos(v) = -2\sin\left(\frac{u+v}{2}\right)\sin\left(\frac{u-v}{2}\right)$$

▸▸▸ **Products to Sums**

Again consider the identities

$$\sin(\alpha + \beta) = \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta) \quad \sin(\alpha - \beta) = \sin(\alpha)\cos(\beta) - \cos(\alpha)\sin(\beta)$$

and add them together:

$$\sin(\alpha + \beta) + \sin(\alpha - \beta) = 2\sin(\alpha)\cos(\beta).$$

Then rearrange:

$$\sin(\alpha)\cos(\beta) = \frac{\sin(\alpha + \beta) + \sin(\alpha - \beta)}{2}$$

In a similar way, start with the identities

$$\cos(\alpha + \beta) = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta) \quad \cos(\alpha - \beta) = \cos(\alpha)\cos(\beta) + \sin(\alpha)\sin(\beta)$$

If we add these together we get

$$2\cos(\alpha)\cos(\beta) = \cos(\alpha + \beta) + \cos(\alpha - \beta)$$

while taking their difference gives

$$2\sin(\alpha)\sin(\beta) = \cos(\alpha - \beta) - \cos(\alpha + \beta)$$

Hence

$$\sin(\alpha)\sin(\beta) = \frac{\cos(\alpha - \beta) - \cos(\alpha + \beta)}{2}$$

$$\cos(\alpha)\cos(\beta) = \frac{\cos(\alpha - \beta) + \cos(\alpha + \beta)}{2}$$

B.3 ▴ Inverse Trigonometric Functions

In order to construct inverse trigonometric functions we first have to restrict their domains so as to make them one-to-one (or injective). We do this as shown below

$\sin \theta$	$\cos \theta$	$\tan \theta$
Domain: $-\frac{\pi}{2} \leqslant \theta \leqslant \frac{\pi}{2}$	Domain: $0 \leqslant \theta \leqslant \pi$	Domain: $-\frac{\pi}{2} < \theta < \frac{\pi}{2}$
Range: $-1 \leqslant \sin \theta \leqslant 1$	Range: $-1 \leqslant \cos \theta \leqslant 1$	Range: all real numbers

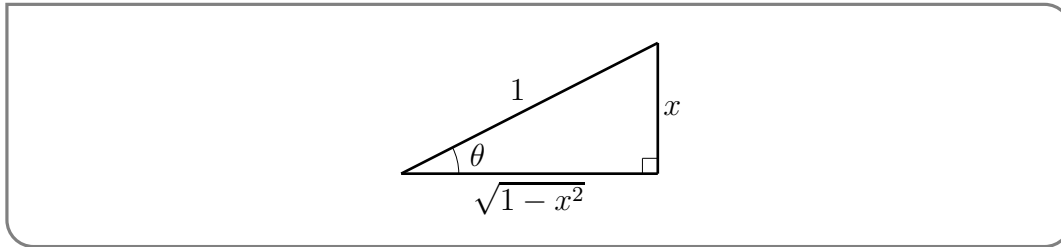| $\arcsin x$	$\arccos x$	$\arctan x$
Domain: $-1 \leqslant x \leqslant 1$	Domain: $-1 \leqslant x \leqslant 1$	Domain: all real numbers
Range: $-\frac{\pi}{2} \leqslant \arcsin x \leqslant \frac{\pi}{2}$	Range: $0 \leqslant \arccos x \leqslant \pi$	Range: $-\frac{\pi}{2} < \arctan x < \frac{\pi}{2}$

Since these functions are inverses of each other we have

$$\arcsin(\sin \theta) = \theta \qquad\qquad -\frac{\pi}{2} \leqslant \theta \leqslant \frac{\pi}{2}$$
$$\arccos(\cos \theta) = \theta \qquad\qquad 0 \leqslant \theta \leqslant \pi$$
$$\arctan(\tan \theta) = \theta \qquad\qquad -\frac{\pi}{2} \leqslant \theta \leqslant \frac{\pi}{2}$$

and also

$$\sin(\arcsin x) = x \qquad\qquad -1 \leqslant x \leqslant 1$$
$$\cos(\arccos x) = x \qquad\qquad -1 \leqslant x \leqslant 1$$
$$\tan(\arctan x) = x \qquad\qquad \text{any real } x$$

We can read other combinations of trig functions and their inverses, like, for example, $\cos(\arcsin x)$, off of triangles like

We have chosen the hypotenuse and opposite sides of the triangle to be of length 1 and x, respectively, so that $\sin(\theta) = x$. That is, $\theta = \arcsin x$. We can then read off of the triangle that

$$\cos(\arcsin x) = \cos(\theta) = \sqrt{1 - x^2}$$

We can reach the same conclusion using trig identities, as follows.

- Write $\arcsin x = \theta$. We know that $\sin(\theta) = x$ and we wish to compute $\cos(\theta)$. So we just need to express $\cos(\theta)$ in terms of $\sin(\theta)$.

- To do this we make use of one of the Pythagorean identities

$$\sin^2 \theta + \cos^2 \theta = 1$$
$$\cos \theta = \pm\sqrt{1 - \sin^2 \theta}$$

- Thus

$$\cos(\arcsin x) = \cos \theta = \pm\sqrt{1 - \sin^2 \theta}$$

- To determine which branch we should use we need to consider the domain and range of $\arcsin x$:

$$\text{Domain: } -1 \leqslant x \leqslant 1 \qquad\qquad \text{Range: } -\frac{\pi}{2} \leqslant \arcsin x \leqslant \frac{\pi}{2}$$

Thus we are applying cosine to an angle that always lies between $-\frac{\pi}{2}$ and $\frac{\pi}{2}$. Cosine is non-negative on this range. Hence we should take the positive branch and

$$\cos(\arcsin x) = \sqrt{1 - \sin^2 \theta} = \sqrt{1 - \sin^2(\arcsin x)}$$
$$= \sqrt{1 - x^2}$$

In a very similar way we can simplify $\tan(\arccos x)$.

- Write $\arccos x = \theta$, and then

$$\tan(\arccos x) = \tan \theta = \frac{\sin \theta}{\cos \theta}$$

- Now the denominator is easy since $\cos\theta = \cos\arccos x = x$.

- The numerator is almost the same as the previous computation.

$$\sin\theta = \pm\sqrt{1 - \cos^2\theta}$$
$$= \pm\sqrt{1 - x^2}$$

- To determine which branch we again consider domains and and ranges:

$$\text{Domain: } -1 \leqslant x \leqslant 1 \qquad\qquad \text{Range: } 0 \leqslant \arccos x \leqslant \pi$$

Thus we are applying sine to an angle that always lies between 0 and π. Sine is non-negative on this range and so we take the positive branch.

- Putting everything back together gives

$$\tan(\arccos x) = \frac{\sqrt{1 - x^2}}{x}$$

Completing the 9 possibilities gives:

$$\sin(\arcsin x) = x \qquad \sin(\arccos x) = \sqrt{1 - x^2} \qquad \sin(\arctan x) = \frac{x}{\sqrt{1 + x^2}}$$

$$\cos(\arcsin x) = \sqrt{1 - x^2} \qquad \cos(\arccos x) = x \qquad \cos(\arctan x) = \frac{1}{\sqrt{1 + x^2}}$$

$$\tan(\arcsin x) = \frac{x}{\sqrt{1 - x^2}} \qquad \tan(\arccos x) = \frac{\sqrt{1 - x^2}}{x} \qquad \tan(\arctan x) = x$$
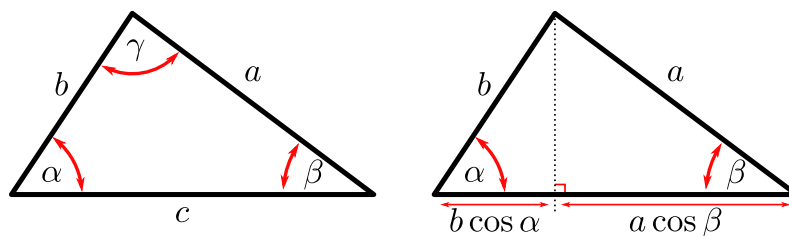
B.4 ▴ Cosine and Sine Laws

B.4.1 ▸▸ Cosine Law or Law of Cosines

The cosine law says that, if a triangle has sides of length a, b and c and the angle opposite the side of length c is γ, then

$$c^2 = a^2 + b^2 - 2ab\cos\gamma$$

Observe that, when $\gamma = \frac{\pi}{2}$, this reduces to, (surpise!) Pythagoras' theorem $c^2 = a^2 + b^2$. Let's derive the cosine law.

Consider the triangle on the left. Now draw a perpendicular line from the side of length c to the opposite corner as shown. This demonstrates that

$$c = a\cos\beta + b\cos\alpha$$

Multiply this by c to get an expression for c^2:

$$c^2 = ac\cos\beta + bc\cos\alpha$$

Doing similarly for the other corners gives

$$a^2 = ac\cos\beta + ab\cos\gamma$$
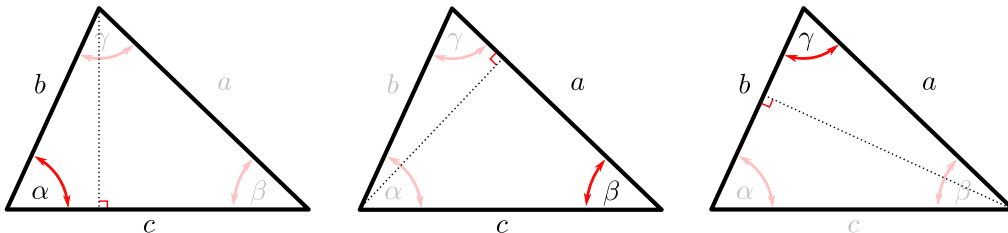$$b^2 = bc\cos\alpha + ab\cos\gamma$$

Now combining these:

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

as required.

B.4.2 ▸▸ Sine Law or Law of Sines

The sine law says that, if a triangle has sides of length a, b and c and the angles opposite those sides are α, β and γ, then

$$\frac{a}{\sin\alpha} = \frac{b}{\sin\beta} = \frac{c}{\sin\gamma}.$$

This rule is best understood by computing the area of the triangle using the formula $A = \frac{1}{2}ab\sin\theta$ of Appendix A.10. Doing this three ways gives

$$2A = bc\sin\alpha$$
$$2A = ac\sin\beta$$
$$2A = ab\sin\gamma$$

Dividing these expressions by abc gives

$$\frac{2A}{abc} = \frac{\sin\alpha}{a} = \frac{\sin\beta}{b} = \frac{\sin\gamma}{c}$$

as required.

B.5 ▴ Circles, cones and spheres

B.5.1 ➤ Where Does the Formula for the Area of a Circle Come From?

Typically when we come across π for the first time it is as the ratio of the circumference of a circle to its diameter

$$\pi = \frac{C}{d} = \frac{C}{2r}$$

Indeed this is typically the first definition we see of π. It is easy to build an intuition that the area of the circle should be propotional to the square of its radius. For example we can draw the largest possible square inside the circle (an *inscribed* square) and the smallest possible square outside the circle (a *circumscribed* square):
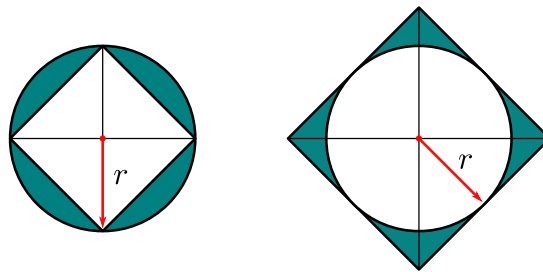
The smaller square has side-length $\sqrt{2}r$ and the longer has side-length $2r$. Hence

$$2r^2 \leqslant A \leqslant 4r^2 \qquad\qquad \text{or } 2 \leqslant \frac{A}{r^2} \leqslant 4$$

That is, the area of the circle is between 2 and 4 times the square of the radius. What is perhaps less obvious (if we had not been told this in school) is that the constant of propotionality for area is also π:

$$\pi = \frac{A}{r^2}.$$

We will show this using Archimedes' proof. He makes use of these inscribed and circumscribed polygons to make better and better approximations of the circle. The steps of the proof are somewhat involved and the starting point is to rewrite the area of a circle as

$$A = \frac{1}{2}Cr$$

where C is (still) the circumference of the circle. This suggests that this area is the same as that of a triangle of height r and base length C
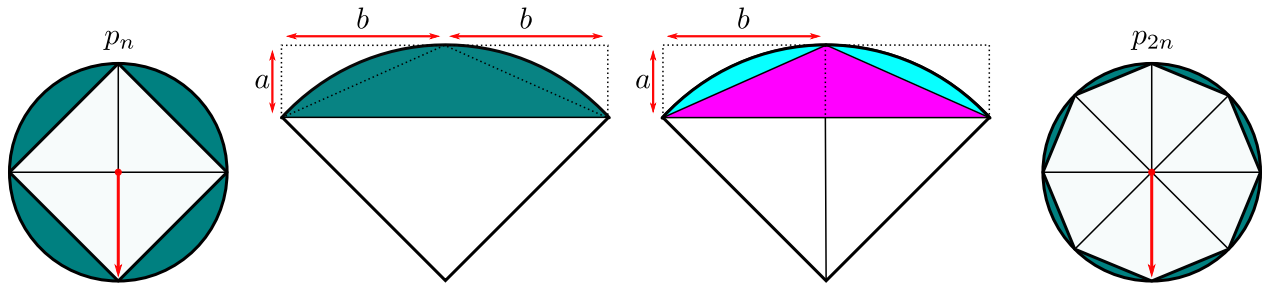
$$T = \frac{1}{2}Cr$$
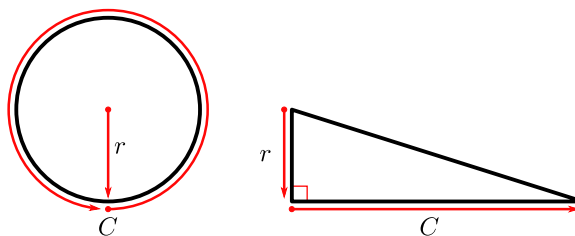
Archimedes' proof then demonstrates that indeed this triangle and the circle have the same area. It relies on a "proof by contradiction" — showing that $T < A$ and $T > A$ cannot be true and so the only possibility is that $A = T$.

We will first show that $T < A$ cannot happen. Construct an n-sided "inscribed" polygon as shown below:
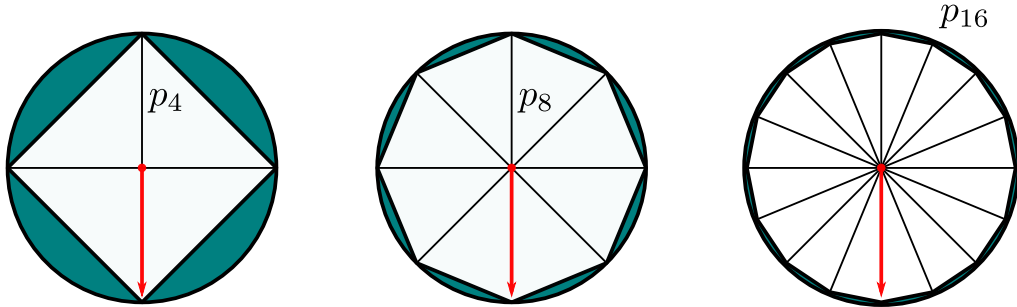
Let p_n be the inscribed polygon as shown.

We need 4 steps.

1. The area of p_n is smaller than that of the circle — this follows since we can construct p_n by cutting slices from the circle.

2. Let E_n be the difference between the area of the circle and p_n: $E_n = A - A(p_n)$ (see the left of the previous figure). By the previous point we know $E_n > 0$. Now as we increase the number of sides, this difference becomes smaller. To be more precise, we will now show that

$$E_{2n} \leqslant \frac{1}{2} E_n.$$

The error E_n is made up of n "lobes". In the centre-left of the previous figure we draw one such lobe and surround it by a rectangle of dimensions $a \times 2b$ — we could determine these more precisely using a little trigonometry, but it is not necessary.

This diagram shows the lobe is smaller than the rectangle of base $2b$ and height a Since there are n copies of the lobe, we have

$$E_n \leqslant n \times 2ab \qquad\qquad \text{rewrite as } \frac{E_n}{2} \leqslant nab$$

Now draw in the polygon p_{2n} and consider the associated "error" E_{2n}. If we focus on the two lobes shown then we see that the area of these two new lobes is equal to

that of the old lobe (shown in centre-left) minus the area of the triangle with base $2b$ and height a (drawn in purple). Since there are n copies of this picture we have

$$E_{2n} = E_n - nab \qquad\qquad \text{now use that } nab \geqslant E_n/2$$
$$\leqslant E_n - \frac{E_n}{2} = \frac{E_n}{2}$$

3. The area of p_n is smaller than T. To see this decompose p_n into n isosceles triangles. Each of these has base shorter than C/n; the straight line is shorter than the corresponding arc — though strictly speaking we should prove this. The height of each triangle is shorter than r. Thus

$$A(p_n) = n \times \frac{1}{2}(\text{base}) \times (\text{height})$$
$$\leqslant n \times \frac{Cr}{2n} = T$$

4. If we assume that $T < A$, then $A - T = d$ where d is some positive number. However we know from point 2 that we can make n large enough so that $E_n < d$ (each time we double n we halve the error). But now we have a contradiction to step 3, since we have just shown that

$$E_n = A - A(p_n) < d = A - T \qquad\qquad \text{which implies that}$$
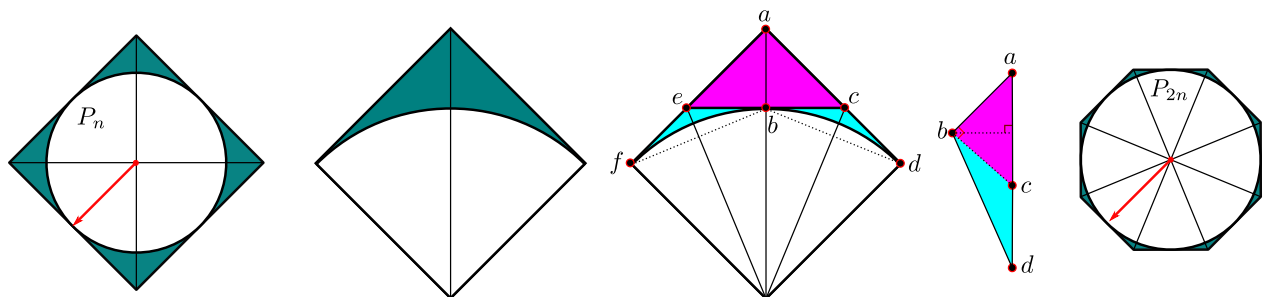$$A(p_n) > T.$$

Thus we cannot have $T < A$.

   If we now assume that $T > A$ we will get a similar contradiction by a similar construction. Now we use regular n-sided *circumscribed* polygons, P_n.
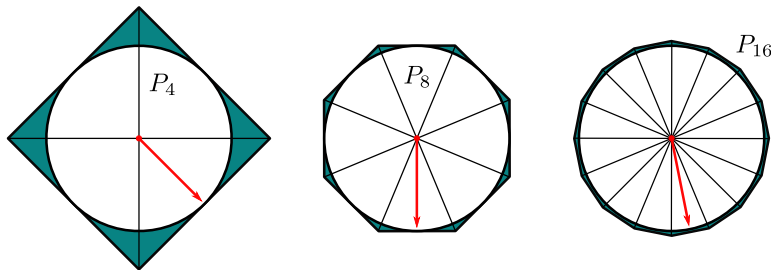
The proof can be broken into 4 similar steps.

1. The area of P_n is greater than that of the circle — this follows since we can construct the circle by trimming the polygon P_n.

2. Let E_n be the difference between the area of the polygon and the circle: $E_n = A(P_n) - A$ (see the left of the previous figure). By the previous point we know $E_n > 0$. Now as we increase the number of sides, this difference becomes smaller. To be more precise, we will now show

$$E_{2n} \leqslant \frac{1}{2}E_n.$$

The error E_n is made up of n "lobes". In the centre-left of the previous figure we draw one such lobe. Let L_n denote the area of one of these lobes, so $E_n = nL_n$. In the centre of the previous figure we have labelled this lobe carefully and also shown how it changes when we create the polygon P_{2n}. In particular, the original lobe is bounded by the straight lines $\vec{ad}, \vec{af}$ and the arc $\overset{\frown}{fbd}$. We create P_{2n} from P_n by cutting away the corner triangle $\triangle aec$. Accordingly the lines $\vec{ec}$ and $\vec{ba}$ are orthogonal and the segments $|bc| = |cd|$.

By the construction of P_{2n} from P_n, we have

$$2L_{2n} = L_n - A(\triangle aec) \qquad \text{or equivalently } L_{2n} = \frac{1}{2}L_n - A(\triangle abc)$$

And additionally

$$L_{2n} \leqslant A(\triangle bcd)$$

Now consider the triangle $\triangle abd$ (centre-right of the previous figure) and the two triangles within it $\triangle abc$ and $\triangle bcd$. We know that $\vec{ab}$ and $\vec{cb}$ form a right-angle. Consequently $\vec{ac}$ is the hypotenuse of a right-angled triangle, so $|ac| > |bc| = |cd|$. So now, the triangles $\triangle abc$ and $\triangle bcd$ have the same heights, but the base of $\vec{ac}$ is longer than $\vec{cd}$. Hence the area of $\triangle abc$ is strictly larger than that of $\triangle bcd$.

Thus we have

$$L_{2n} \leqslant A(\triangle bcd) < A(\triangle abc)$$

But now we can write

$$L_{2n} = \frac{1}{2}L_n - A(\triangle abc) < \frac{1}{2}L_n - L_{2n} \qquad\qquad \text{rearrange}$$

$$2L_{2n} < \frac{1}{2}L_n \qquad\qquad \text{there are } n \text{ such lobes, so}$$

$$2nL_{2n} < \frac{n}{2}L_n \qquad\qquad \text{since } E_n = nL_n, \text{ we have}$$

$$E_{2n} < \frac{1}{2}E_n \qquad\qquad \text{which is what we wanted to show.}$$

3. The area of P_n is greater than T. To see this decompose P_n into n isosceles triangles. The height of each triangle is r, while the base of each is longer than C/n (this is a subtle point and its proof is equivalent to showing that $\tan\theta > \theta$). Thus

$$A(P_n) = n \times \frac{1}{2}(\text{base}) \times (\text{height})$$

$$\geqslant n \times \frac{Cr}{2n} = T$$

4. If we assume that $T > A$, then $T - A = d$ where d is some positive number. However we know from point 2 that we can make n large enough so that $E_n < d$ (each time we double n we halve the error). But now we have a contradiction to step 3, since we have just shown that

$$E_n = A(P_n) - A < d = T - A \qquad \text{which implies that}$$
$$A(P_n) < T.$$

Thus we cannot have $T > A$. The only possibility that remains is that $T = A$.

B.5.2 ▸▸ Where Do These Volume Formulas Come From?

We can establish the volumes of cones and spheres from the formula for the volume of a cylinder and a little work with limits and some careful summations. We first need a few facts.

- Every square number can be written as a sum of consecutive odd numbers. More precisely

$$n^2 = 1 + 3 + \cdots + (2n - 1)$$

- The sum of the first n positive integers is $\frac{1}{2}n(n+1)$. That is

$$1 + 2 + 3 + \cdots + n = \frac{1}{2}n(n+1)$$

- The sum of the squares of the first n positive integers is $\frac{1}{6}n(n+1)(2n+1)$.

$$1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{1}{6}n(n+1)(2n+1)$$

We will not give completely rigorous proofs of the above identities (since we are not going to assume that the reader knows mathematical induction), rather we will explain them using pictorial arguments. The first two of these we can explain by some quite simple pictures:
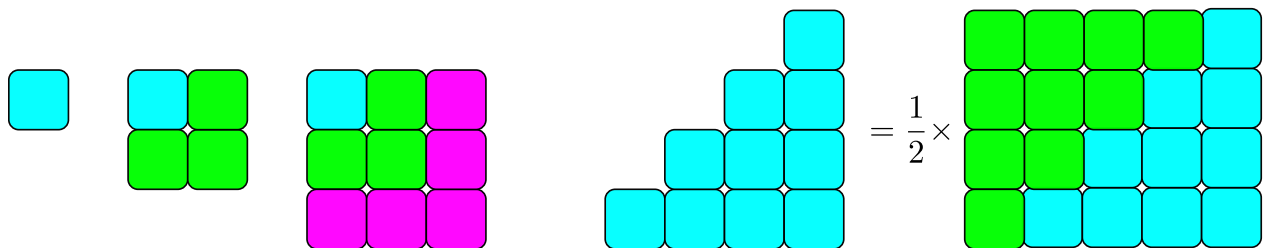

We see that we can decompose any square of unit-squares into a sequence of strips, each of which consists of an odd number of unit-squares. This is really just from the fact that

$$n^2 - (n - 1)^2 = 2n - 1$$

Similarly, we can represent the sum of the first n integers as a triangle of unit squares as shown. If we make a second copy of that triangle and arrange it as shown, it gives a rectangle of dimensions n by $n + 1$. Hence the rectangle, being exactly twice the size of the original triangle, contains $n(n + 1)$ unit squares.

The explanation of the last formula takes a little more work and a carefully constructed picture:
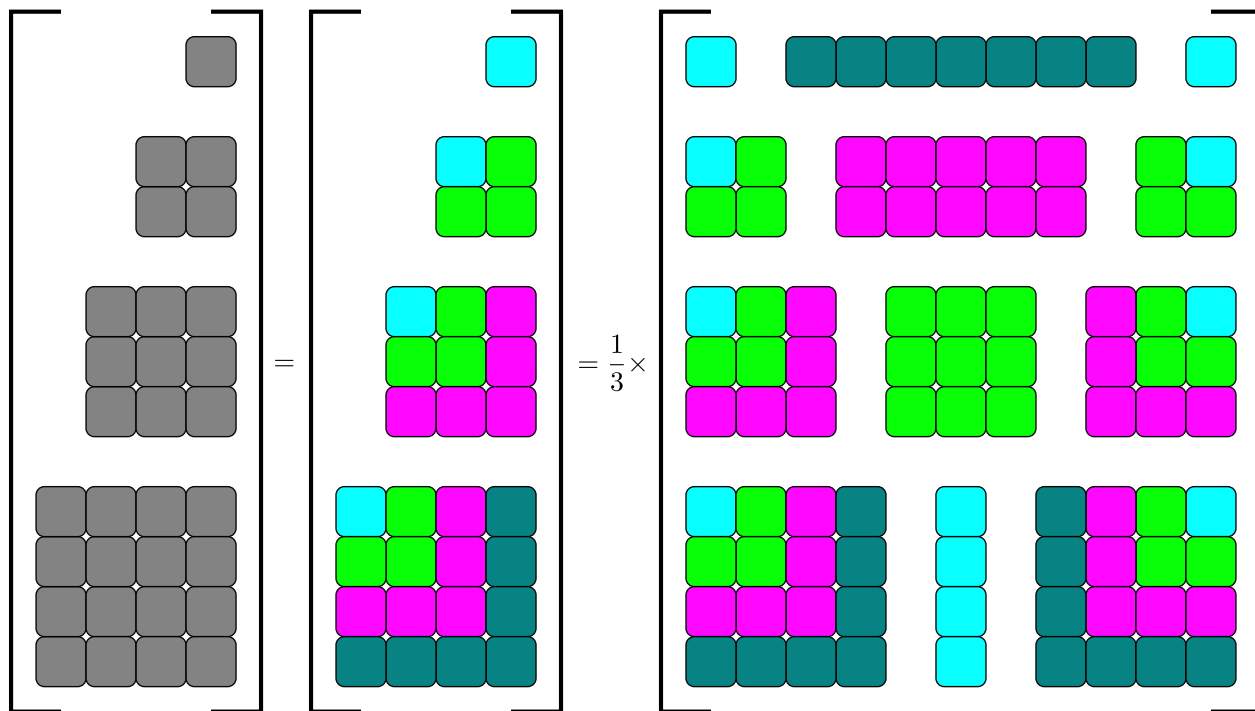
Let us break these pictures down step by step

- Leftmost represents the sum of the squares of the first n integers.

- Centre — We recall from above that each square number can be written as a sum of consecutive odd numbers, which have been represented as coloured bands of unit-squares.

- Make three copies of the sum and arrange them carefully as shown. The first and third copies are obvious, but the central copy is rearranged considerably; all bands of the same colour have the same length and have been arranged into rectangles as shown.

  Putting everything from the three copies together creates a rectangle of dimensions $(2n + 1) \times (1 + 2 + 3 + \cdots + n)$.

We know (from above) that $1 + 2 + 3 + \cdots + n = \frac{1}{2}n(n + 1)$ and so

$$(1^2 + 2^2 + \cdots + n^2) = \frac{1}{3} \times \frac{1}{2}n(n + 1)(2n + 1)$$

as required.

Now we can start to look at volumes. Let us start with the volume of a cone; consider the figure below. We bound the volume of the cone above and below by stacks of cylinders. The cross-sections of the cylinders and cone are also shown.
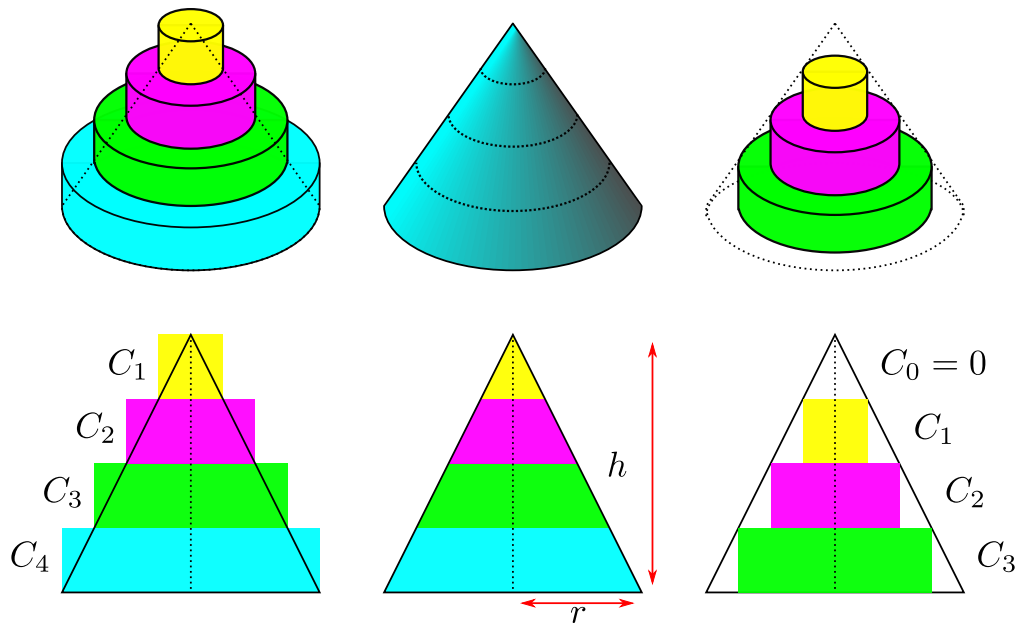
To obtain the bounds we will construct two stacks of n cylinders, $C_1, C_2, \ldots, C_n$. Each cylinder has height h/n and radius that varies with height. In particular, we define cylinder C_k to have height h/n and radius $k \times r/n$. This radius was determined using similar triangles so that cylinder C_n has radius r. Now cylinder C_k has volume

$$V_k = \pi \times \text{radius}^2 \times \text{height} = \pi \left(\frac{kr}{n}\right)^2 \cdot \frac{h}{n}$$

$$= \frac{\pi r^2 h}{n^3} k^2$$

We obtain an upper bound by stacking cylinders $C_1, C_2, \ldots, C_n$ as shown. This object has volume

$$V = V_1 + V_2 + \ldots V_n$$

$$= \frac{\pi r^2 h}{n^3} \left(1^2 + 2^2 + 3^2 + \cdots + n^2\right)$$

$$= \frac{\pi r^2 h}{n^3} \cdot \frac{n(n+1)(2n+1)}{6}$$

A similar lower bound is obtained by stacking cylinders $C_1, \ldots, C_{n-1}$ which gives a volume of

$$V = V_1 + V_2 + \ldots V_{n-1}$$

$$= \frac{\pi r^2 h}{n^3} \left(1^2 + 2^2 + 3^2 + \cdots + (n-1)^2\right)$$

$$= \frac{\pi r^2 h}{n^3} \cdot \frac{(n-1)(n)(2n-1)}{6}$$

Thus the true volume of the cylinder is bounded between

$$\frac{\pi r^2 h}{n^3} \cdot \frac{(n-1)(n)(2n-1)}{6} \leqslant \text{correct volume} \leqslant \frac{\pi r^2 h}{n^3} \cdot \frac{n(n+1)(2n+1)}{6}$$

We can now take the limit as the number of cylinders, n, goes to infinity. The upper bound becomes

$$\lim_{n\to\infty} \frac{\pi r^2 h}{n^3} \frac{n(n+1)(2n+1)}{6} = \frac{\pi r^2 h}{6} \lim_{n\to\infty} \frac{n(n+1)(2n+1)}{n^3}$$
$$= \frac{\pi r^2 h}{6} \lim_{n\to\infty} \frac{(1+1/n)(2+1/n)}{1}$$
$$= \frac{\pi r^2 h}{6} \times 2$$
$$= \frac{\pi r^2 h}{3}$$

The other limit is identical, so by the squeeze theorem we have

$$\text{Volume of cone } = \frac{1}{3}\pi r^2 h$$

Now the sphere — though we will do the analysis for a hemisphere of radius R. Again we bound the volume above and below by stacks of cylinders. The cross-sections of the cylinders and cone are also shown.
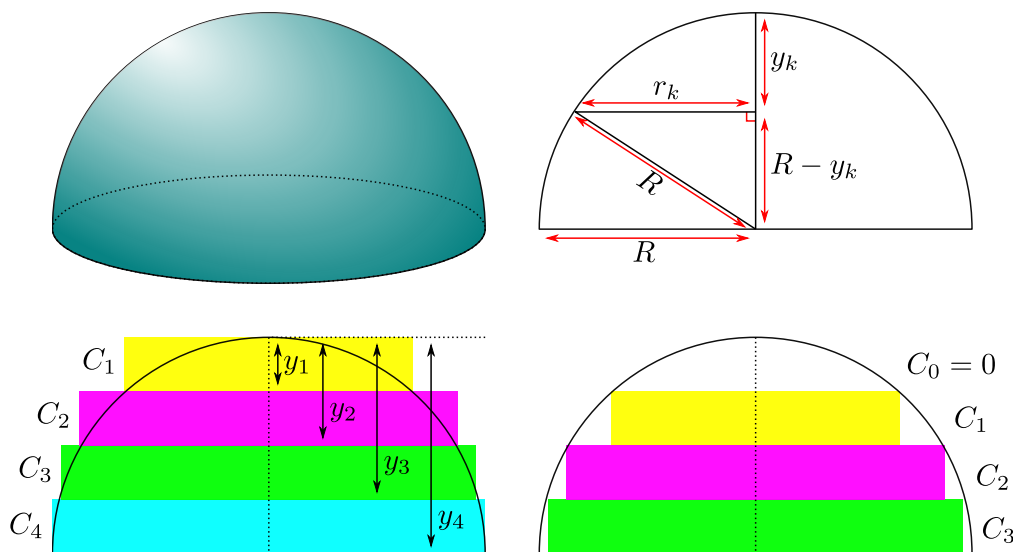
To obtain the bounds we will construct two stacks of n cylinders, $C_1, C_2, \ldots, C_n$. Each cylinder has height R/n and radius that varies with its position in the stack. To describe the position, define

$$y_k = k \times \frac{R}{n}$$

That is, y_k, is k steps of distance $\frac{R}{n}$ from the top of the hemisphere. Then we set the kth cylinder, C_k to have height R/n and radius r_k given by

$$r_k^2 = R^2 - (R - y_k)^2 = R^2 - R^2(1 - k/n)^2$$
$$= R^2(2k/n - k^2/n^2)$$

as shown in the top-right and bottom-left illustrations. The volume of C_k is then

$$V_k = \pi \times \text{radius}^2 \times \text{height} = \pi \times R^2 \left(2k/n - k^2/n^2\right) \times \frac{R}{n}$$

$$= \pi R^3 \cdot \left(\frac{2k}{n^2} - \frac{k^2}{n^3}\right)$$

We obtain an upper bound by stacking cylinders $C_1, C_2, \ldots, C_n$ as shown. This object has volume

$$V = V_1 + V_2 + \ldots V_n$$

$$= \pi R^3 \cdot \left(\frac{2}{n^2} (1 + 2 + 3 + \cdots + n) - \frac{1}{n^3} \left(1^2 + 2^2 + 3^2 + \cdots + n^2\right)\right)$$

Now recall from above that

$$1 + 2 + 3 + \cdots + n = \frac{1}{2} n(n+1) \qquad 1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{1}{6} n(n+1)(2n+1)$$

so

$$V = \pi R^3 \cdot \left(\frac{n(n+1)}{n^2} - \frac{n(n+1)(2n+1)}{6n^3}\right)$$

Again, a lower bound is obtained by stacking cylinders $C_1, \ldots, C_{n-1}$ and a similar analysis gives

$$V = \pi R^3 \cdot \left(\frac{n(n-1)}{(n-1)^2} - \frac{n(n-1)(2n-1)}{6(n-1)^3}\right)$$

Thus the true volume of the hemisphere is bounded between

$$\pi R^3 \cdot \left(\frac{n(n+1)}{n^2} - \frac{n(n+1)(2n+1)}{6n^3}\right) \leqslant \text{correct volume} \leqslant \pi R^3 \cdot \left(\frac{n(n+1)}{n^2} - \frac{n(n+1)(2n+1)}{6n^3}\right)$$

We can now take the limit as the number of cylinders, n, goes to infinity. The upper bound becomes

$$\lim_{n \to \infty} \pi R^3 \cdot \left(\frac{n(n+1)}{n^2} - \frac{n(n+1)(2n+1)}{6n^3}\right) = \pi R^3 \left(\lim_{n \to \infty} \frac{n(n+1)}{n^2} - \frac{n(n+1)(2n+1)}{6n^3}\right)$$

$$= \pi R^3 \left(1 - \frac{2}{6}\right) = \frac{2}{3} \pi R^3.$$

The other limit is identical, so by the squeeze theorem we have

$$\text{Volume of hemisphere} = \frac{2}{3} \pi R^3 \qquad\qquad \text{and so}$$

$$\text{Volume of sphere} = \frac{4}{3} \pi R^3$$

ROOT FINDING

To this point you have found solutions to equations almost exclusively by algebraic manipulation. This is possible only for the artificially simple equations of problem sets and tests. In the "real world" it is very common to encounter equations that cannot be solved by algebraic manipulation. For example, you found, by completing a square, that the solutions to the quadratic equation $ax^2 + bx + c = 0$ are $x = (-b \pm \sqrt{b^2 - 4ac})/2a$. But it is known that there simply does not exist a corresponding formula for the roots of a general polynomial of degree five or more. Fortunately, encountering such an equation is not the end of the world, because usually one does not need to know the solutions exactly. One only needs to know them to within some specified degree of accuracy. For example, one rarely needs to know π to more than a few decimal places. There is a whole subject, called numerical analysis, that concerns using algorithms to solve equations (and perform other tasks) approximately, to any desired degree of accuracy.

We have already had, in Examples 1.6.14 and 1.6.15, and the lead up to them, a really quick introduction to the bisection method, which is a crude, but effective, algorithm for finding approximate solutions to equations of the form $f(x) = 0$. We shall shortly use a little calculus to derive a very efficient algorithm for finding approximate solutions to such equations. But first here is a simple example which provides a review of some of the basic ideas of root finding and the bisection method.

Example C.0.1

Suppose that we are given some function $f(x)$ and we have to find solutions to the equation $f(x) = 0$. To be concrete, suppose that $f(x) = 8x^3 + 12x^2 + 6x - 15$. How do we go about solving $f(x) = 0$? To get a rough idea of the lay of the land, sketch the graph of $f(x)$. First observe that

- when x is very large and negative, $f(x)$ is very large and negative
- when x is very large and positive, $f(x)$ is very large and positive
- when $x = 0$, $f(x) = f(0) = -15 < 0$
- when $x = 1$, $f(x) = f(1) = 11 > 0$
- $f'(x) = 24x^2 + 24x + 6 = 24(x^2 + x + \frac{1}{4}) = 24(x + \frac{1}{2})^2 \geqslant 0$ for all x. So $f(x)$ increases monotonically with x. The graph has a tangent of slope 0 at $x = -\frac{1}{2}$ and tangents of strictly positive slope everywhere else.

This tells us that the graph of $f(x)$ looks like
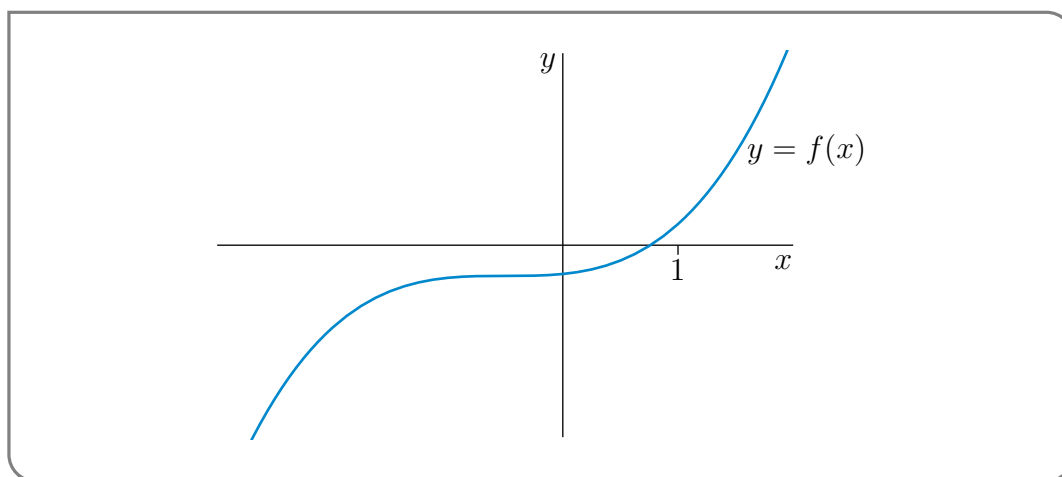
Since $f(x)$ strictly increases[1] as x increases, $f(x)$ can take the value zero for at most one value of x.

- Since $f(0) < 0$ and $f(1) > 0$ and f is continuous, $f(x)$ must pass through 0 as x travels from $x = 0$ to $x = 1$, by Theorem 1.6.12 (the intermediate value theorem). So $f(x)$ takes the value zero for some x between 0 and 1. We will often write this as "the root is $x = 0.5 \pm 0.5$" to indicate the uncertainty.
- To get closer to the root, we evaluate $f(x)$ halfway between 0 and 1.

$$f\left(\tfrac{1}{2}\right) = 8\left(\tfrac{1}{2}\right)^3 + 12\left(\tfrac{1}{2}\right)^2 + 6\left(\tfrac{1}{2}\right) - 15 = -8$$

Since $f\left(\tfrac{1}{2}\right) < 0$ and $f(1) > 0$ and f is continuous, $f(x)$ must take the value zero for some x between $\tfrac{1}{2}$ and 1. The root is 0.75 ± 0.25.

- To get still closer to the root, we evaluate $f(x)$ halfway between $\tfrac{1}{2}$ and 1.

$$f\left(\tfrac{3}{4}\right) = 8\left(\tfrac{3}{4}\right)^3 + 12\left(\tfrac{3}{4}\right)^2 + 6\left(\tfrac{3}{4}\right) - 15 = -\tfrac{3}{8}$$

Since $f\left(\tfrac{3}{4}\right) < 0$ and $f(1) > 0$ and f is continuous, $f(x)$ must take the value zero for some x between $\tfrac{3}{4}$ and 1. The root is 0.875 ± 0.125.

- And so on.

Example C.0.1

The root finding strategy used in Example C.0.1 is called the bisection method. The bisection method will home in on a root of the function $f(x)$ whenever

- $f(x)$ is continuous ($f(x)$ need not have a derivative) and

---

1   By "$f(x)$ is strictly increasing" we mean that $f(a) < f(b)$ whenever $a < b$. As $f'(x) > 0$ for all $x \neq -\tfrac{1}{2}$, $f(x)$ is strictly increasing even as x passes through $-\tfrac{1}{2}$. For example, for any $x > -\tfrac{1}{2}$, the mean value theorem (Theorem 2.13.4), tells us that there is a c strictly between $-\tfrac{1}{2}$ and x such that $f(x) - f\left(-\tfrac{1}{2}\right) = f'(c)\left(x + \tfrac{1}{2}\right) > 0$.

- you can find two numbers $a_1 < b_1$ with $f(a_1)$ and $f(b_1)$ being of opposite sign.

Denote by I_1 the interval $[a_1, b_1] = \{x \mid a_1 \leqslant x \leqslant b_1\}$. Once you have found the interval I_1, the bisection method generates a sequence I_1, I_2, I_3, $\cdots$ of intervals by the following rule.

Equation C.0.2 (bisection method).

Denote by $c_n = \frac{a_n + b_n}{2}$ the midpoint of the interval $I_n = [a_n, b_n]$. If $f(c_n)$ has the same sign as $f(a_n)$, then

$$I_{n+1} = [a_{n+1}, b_{n+1}] \quad \text{with} \quad a_{n+1} = c_n, \ b_{n+1} = b_n$$

and if $f(c_n)$ and $f(a_n)$ have opposite signs, then

$$I_{n+1} = [a_{n+1}, b_{n+1}] \quad \text{with} \quad a_{n+1} = a_n, \ b_{n+1} = c_n$$

This rule was chosen so that $f(a_n)$ and $f(b_n)$ have opposite sign for every n. Since $f(x)$ is continuous, $f(x)$ has a zero in each interval I_n. Thus each step reduces the error bars by a factor of 2. That isn't too bad, but we can come up with something that is much more efficient. We just need a little calculus.

C.1 ▲ Newton's Method

Newton's method[2], also known as the Newton-Raphson method, is another technique for generating numerical approximate solutions to equations of the form $f(x) = 0$. For example, one can easily get a good approximation to $\sqrt{2}$ by applying Newton's method to the equation $x^2 - 2 = 0$. This will be done in Example C.1.2, below.

Here is the derivation of Newton's method. We start by simply making a guess for the solution. For example, we could base the guess on a sketch of the graph of $f(x)$. Call the initial guess x_1. Next recall, from Theorem 2.3.2, that the tangent line to $y = f(x)$ at $x = x_1$ is $y = F(x)$, where

$$F(x) = f(x_1) + f'(x_1)(x - x_1)$$

Usually $F(x)$ is a pretty good approximation to $f(x)$ for x near x_1. So, instead of trying to solve $f(x) = 0$, we solve the linear equation $F(x) = 0$ and call the solution x_2.

$$0 = F(x) = f(x_1) + f'(x_1)(x - x_1) \iff x - x_1 = -\frac{f(x_1)}{f'(x_1)}$$

$$\iff x = x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$$

Note that if $f(x)$ were a linear function, then $F(x)$ would be exactly $f(x)$ and x_2 would solve $f(x) = 0$ exactly.

---

2   The algorithm that we are about to describe grew out of a method that Newton wrote about in 1669. But the modern method incorporates substantial changes introduced by Raphson in 1690 and Simpson in 1740.
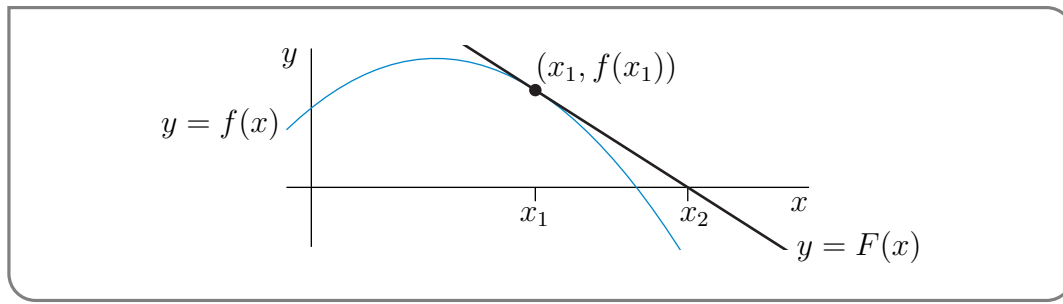
Now we repeat, but starting with the (second) guess x_2 rather than x_1. This gives the (third) guess $x_3 = x_2 - \frac{f(x_2)}{f'(x_2)}$. And so on. By way of summary, Newton's method is

1. Make a preliminary guess x_1.

2. Define $x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$.

3. Iterate. That is, for each natural number n, once you have computed x_n, define

Equation C.1.1 (Newton's method).

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

Example C.1.2 (Approximating $\sqrt{2}$)

In this example we compute, approximately, the square root of two. We will of course pretend that we do not already know that $\sqrt{2} = 1.41421\cdots$. So we cannot find it by solving, approximately, the equation $f(x) = x - \sqrt{2} = 0$. Instead we apply Newton's method to the equation

$$f(x) = x^2 - 2 = 0$$

Since $f'(x) = 2x$, Newton's method says that we should generate approximate solutions by iteratively applying

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^2 - 2}{2x_n} = \frac{x_n}{2} + \frac{1}{x_n}$$

We need a starting point. Since $1^2 = 1 < 2$ and $2^2 = 4 > 2$, the square root of two must be between 1 and 2, so let's start Newton's method with the initial guess $x_1 = 1.5$. Here goes[3]:

$$x_1 = 1.5$$
$$x_2 = \frac{1}{2}x_1 + \frac{1}{x_1} = \frac{1}{2}(1.5) + \frac{1}{1.5}$$
$$= 1.416666667$$

---

[3]   The following computations have been carried out in double precision, which is computer speak for about 15 significant digits. We are displaying each x_n rounded to 10 significant digits (9 decimal places). So each displayed x_n has not been impacted by roundoff error, and still contains more decimal places than are usually needed.

$$x_3 = \frac{1}{2}x_2 + \frac{1}{x_2} = \frac{1}{2}(1.416666667) + \frac{1}{1.416666667}$$
$$= 1.414215686$$

$$x_4 = \frac{1}{2}x_3 + \frac{1}{x_3} = \frac{1}{2}(1.414215686) + \frac{1}{1.414215686}$$
$$= 1.414213562$$

$$x_5 = \frac{1}{2}x_4 + \frac{1}{x_4} = \frac{1}{2}(1.414213562) + \frac{1}{1.414213562}$$
$$= 1.414213562$$

It looks like the x_n's, rounded to nine decimal places, have stabilized to 1.414213562. So it is reasonable to guess that $\sqrt{2}$, rounded to nine decimal places, is exactly 1.414213562. Recalling that all numbers $1.4142135615 \leqslant y < 1.4142135625$ round to 1.414213562, we can check our guess by evaluating $f(1.4142135615)$ and $f(1.4142135625)$. Since $f(1.4142135615) = -2.5 \times 10^{-9} < 0$ and $f(1.4142135625) = 3.6 \times 10^{-10} > 0$ the square root of two must indeed be between 1.4142135615 and 1.4142135625.

Example C.1.2

Example C.1.3 (Approximating π)

In this example we compute, approximately, π by applying Newton's method to the equation

$$f(x) = \sin x = 0$$

starting with $x_1 = 3$. Since $f'(x) = \cos x$, Newton's method says that we should generate approximate solutions by iteratively applying

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{\sin x_n}{\cos x_n} = x_n - \tan x_n$$

Here goes

$$x_1 = 3$$
$$x_2 = x_1 - \tan x_1 = 3 - \tan 3$$
$$= 3.142546543$$
$$x_3 = 3.142546543 - \tan 3.142546543$$
$$= 3.141592653$$
$$x_4 = 3.141592653 - \tan 3.141592653$$
$$= 3.141592654$$
$$x_5 = 3.141592654 - \tan 3.141592654$$
$$= 3.141592654$$

Since $f(3.1415926535) = 9.0 \times 10^{-11} > 0$ and $f(3.1415926545) = -9.1 \times 10^{-11} < 0$, π must be between 3.1415926535 and 3.1415926545. Of course to compute π in this way, we (or at least our computers) have to be able to evaluate $\tan x$ for various values of x. Taylor expansions can help us do that. See Example 3.4.22.

Example C.1.3

Example C.1.4 (wild instability)

This example illustrates how Newton's method can go badly wrong if your initial guess is not good enough. We'll try to solve the equation

$$f(x) = \arctan x = 0$$

starting with $x_1 = 1.5$. (Of course the solution to $f(x) = 0$ is just $x = 0$; we chose $x_1 = 1.5$ for demonstration purposes.) Since the derivative $f'(x) = \frac{1}{1+x^2}$, Newton's method gives

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - (1 + x_n^2)\arctan x_n$$

So[4]

$$x_1 = 1.5$$
$$x_2 = 1.5 - (1 + 1.5^2)\arctan 1.5 = -1.69$$
$$x_3 = -1.69 - (1 + 1.69^2)\arctan(-1.69) = 2.32$$
$$x_4 = 2.32 - (1 + 2.32^2)\arctan(2.32) = -5.11$$
$$x_5 = -5.11 - (1 + 5.11^2)\arctan(-5.11) = 32.3$$
$$x_6 = 32.3 - (1 + 32.3^2)\arctan(32.3) = -1575$$
$$x_7 = 3,894,976$$

Looks pretty bad! Our x_n's are not settling down at all!

The figure below shows what went wrong. In this figure, $y = F_1(x)$ is the tangent line to $y = \arctan x$ at $x = x_1$. Under Newton's method, this tangent line crosses the x–axis at $x = x_2$. Then $y = F_2(x)$ is the tangent to $y = \arctan x$ at $x = x_2$. Under Newton's method, this tangent line crosses the x–axis at $x = x_3$. And so on.

The problem arose because the x_n's were far enough from the solution, $x = 0$, that the tangent line approximations, while good approximations to $f(x)$ for $x \approx x_n$, were very

4   Once again, the following computations have been carried out in double precision. This time, it is clear that the x_n's are growing madly as n increases. So there is not much point to displaying many decimal places and we have not done so.
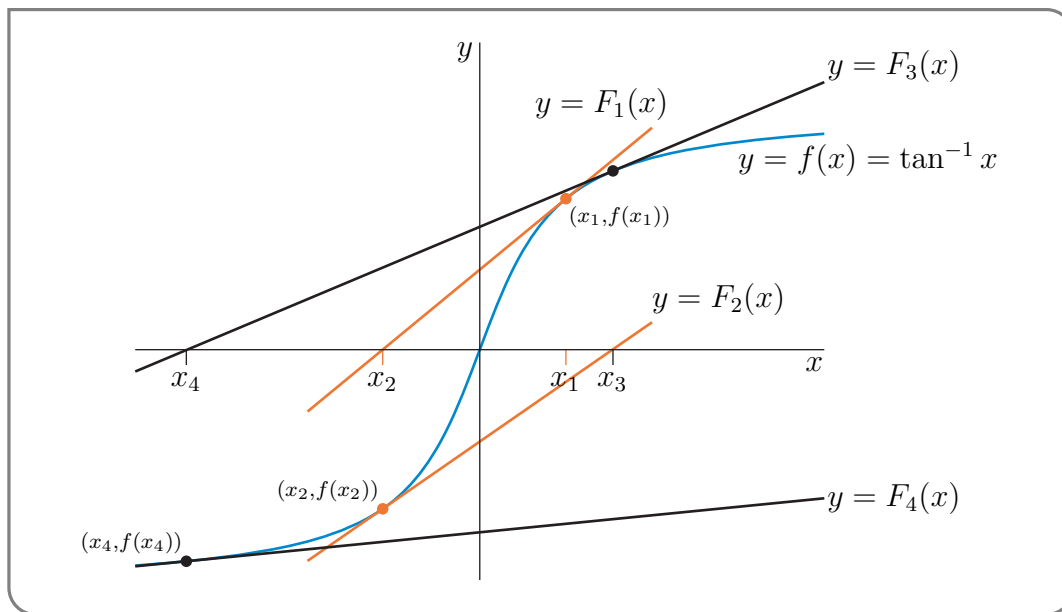
poor approximations to $f(x)$ for $x \approx 0$. In particular, $y = F_1(x)$ (i.e. the tangent line at $x = x_1$) was a bad enough approximation to $y = \arctan x$ for $x \approx 0$ that $x = x_2$ (i.e. the value of x where $y = F_1(x)$ crosses the x-axis) is farther from the solution $x = 0$ than our original guess $x = x_1$. If we had started with $x_1 = 0.5$ instead of $x_1 = 1.5$, Newton's method would have succeeded very nicely:

$$x_1 = 0.5 \qquad x_2 = -0.0796 \qquad x_3 = 0.000335 \qquad x_4 = -2.51 \times 10^{-11}$$

Example C.1.4

Example C.1.5 (interest rate)

A car dealer sells a new car for $23,520. He also offers to finance the same car for payments of $420 per month for five years. What interest rate is this dealer charging?

Solution. By way of preparation, we'll start with a simpler problem. Suppose that you will have to make a single $420 payment n months in the future. The simpler problem is to determine how much money you have to deposit now in an account that pays an interest rate of $100r\%$ per month, compounded monthly[5], in order to be able to make the $420 payment in n months.

Let's denote by P the initial deposit. Because the interest rate is $100r\%$ per month, compounded monthly,

- the first month's interest is $P \times r$. So at the end of month #1, the account balance is $P + Pr = P(1 + r)$.
- The second month's interest is $[P(1 + r)] \times r$. So at the end of month #2, the account balance is $P(1 + r) + P(1 + r)r = P(1 + r)^2$.
- And so on.

5   "Compounded monthly", means that, each month, interest is paid on the accumulated interest that was paid in all previous months.

- So at the end of n months, the account balance is $P(1+r)^n$.

In order for the balance at the end of n months, $P(1+r)^n$, to be \$420, the initial deposit has to be $P = 420(1+r)^{-n}$. That is what is meant by the statement "The present value[6] of a \$420 payment made n months in the future, when the interest rate is $100r\%$ per month, compounded monthly, is $420(1+r)^{-n}$."

Now back to the original problem. We will be making 60 monthly payments of \$420. The present value of all 60 payments is[7]

$$420(1+r)^{-1} + 420(1+r)^{-2} + \cdots + 420(1+r)^{-60} = 420\frac{(1+r)^{-1} - (1+r)^{-61}}{1 - (1+r)^{-1}}$$

$$= 420\frac{1 - (1+r)^{-60}}{(1+r) - 1} = 420\frac{1 - (1+r)^{-60}}{r}$$

The interest rate $100r\%$ being charged by the car dealer is such that the present value of 60 monthly payments of \$420 is \$23520. That is, the monthly interest rate being charged by the car dealer is the solution of

$$23520 = 420\frac{1 - (1+r)^{-60}}{r} \qquad \text{or} \qquad 56 = \frac{1 - (1+r)^{-60}}{r}$$

$$\text{or} \qquad 56r = 1 - (1+r)^{-60}$$

$$\text{or} \qquad 56r(1+r)^{60} = (1+r)^{60} - 1$$

$$\text{or} \qquad (1 - 56r)(1+r)^{60} = 1$$

Set $f(r) = (1 - 56r)(1+r)^{60} - 1$. Then

$$f'(r) = -56(1+r)^{60} + 60(1 - 56r)(1+r)^{59}$$

or

$$f'(r) = \left[-56(1+r) + 60(1 - 56r)\right](1+r)^{59} = (4 - 3416r)(1+r)^{59}$$

Apply Newton's method with an initial guess of $r_1 = .002$. (That's 0.2% per month or 2.4% per year.) Then

$$r_2 = r_1 - \frac{(1 - 56r_1)(1+r_1)^{60} - 1}{(4 - 3416r_1)(1+r_1)^{59}} = 0.002344$$

$$r_3 = r_2 - \frac{(1 - 56r_2)(1+r_2)^{60} - 1}{(4 - 3416r_2)(1+r_2)^{59}} = 0.002292$$

$$r_4 = r_3 - \frac{(1 - 56r_3)(1+r_3)^{60} - 1}{(4 - 3416r_3)(1+r_3)^{59}} = 0.002290$$

$$r_5 = r_4 - \frac{(1 - 56r_4)(1+r_4)^{60} - 1}{(4 - 3416r_4)(1+r_4)^{59}} = 0.002290$$

6   Inflation means that prices of goods (typically) increase with time, and hence \$100 now is worth more than \$100 in 10 years time. The term "present value" is widely used in economics and finance to mean "the current amount of money that will have a specified value at a specified time in the future". It takes inflation into account. If the money is invested, it takes into account the rate of return of the investment. We recommend that the interested reader do some search-engining to find out more.

7   Don't worry if you don't know how to evaluate such sums. They are called geometric sums, and will be covered in the CLP-2 text. (See (1.1.3) in the CLP-2 text.) In any event, you can check that this is correct, by multiplying the whole equation by $1 - (1+r)^{-1}$. When you simplify the left hand side, you should get the right hand side.

So the interest rate is 0.229% per month or 2.75% per year.

Example C.1.5

C.2 ▴ The Error Behaviour of Newton's Method

Newton's method usually works spectacularly well, provided your initial guess is reasonably close to a solution of $f(x) = 0$. A good way to select this initial guess is to sketch the graph of $y = f(x)$. We now explain why "Newton's method usually works spectacularly well, provided your initial guess is reasonably close to a solution of $f(x) = 0$".

Let r be any solution of $f(x) = 0$. Then $f(r) = 0$. Suppose that we have already computed x_n. The error in x_n is $|x_n - r|$. We now derive a formula that relates the error after the next step, $|x_{n+1} - r|$, to $|x_n - r|$. We have seen in (3.4.32) that

$$f(x) = f(x_n) + f'(x_n)(x - x_n) + \frac{1}{2}f''(c)(x - x_n)^2$$

for some c between x_n and x. In particular, choosing $x = r$,

$$0 = f(r) = f(x_n) + f'(x_n)(r - x_n) + \frac{1}{2}f''(c)(r - x_n)^2 \tag{E1}$$

Recall that x_{n+1} is the solution of $0 = f(x_n) + f'(x_n)(x - x_n)$. So

$$0 = f(x_n) + f'(x_n)(x_{n+1} - x_n) \tag{E2}$$

We need to get an expression for $x_{n+1} - r$. Subtracting (E2) from (E1) gives

$$0 = f'(x_n)(r - x_{n+1}) + \frac{1}{2}f''(c)(r - x_n)^2 \implies x_{n+1} - r = \frac{f''(c)}{2f'(x_n)}(x_n - r)^2$$

$$\implies |x_{n+1} - r| = \frac{|f''(c)|}{2|f'(x_n)|}|x_n - r|^2$$

If the guess x_n is close to r, then c, which must be between x_n and r, is also close to r and we will have $f''(c) \approx f''(r)$ and $f'(x_n) \approx f'(r)$ and

$$|x_{n+1} - r| \approx \frac{|f''(r)|}{2|f'(r)|}|x_n - r|^2 \tag{E3}$$

Even when x_n is not close to r, if we know that there are two numbers $L, M > 0$ such that f obeys:

(H1) $|f'(x_n)| \geqslant L$

(H2) $|f''(c)| \leqslant M$

(we'll see examples of this below) then we will have

$$|x_{n+1} - r| \leqslant \frac{M}{2L}|x_n - r|^2 \tag{E4}$$

Let's denote by ε_1 the error, $|x_1 - r|$, of our initial guess. In fact, let's denote by ε_n the error, $|x_n - r|$, in x_n. Then (E4) says

$$\varepsilon_{n+1} \leqslant \frac{M}{2L}\varepsilon_n^2$$

In particular

$$\varepsilon_2 \leqslant \frac{M}{2L}\varepsilon_1^2$$

$$\varepsilon_3 \leqslant \frac{M}{2L}\varepsilon_2^2 \leqslant \frac{M}{2L}\left(\frac{M}{2L}\varepsilon_1^2\right)^2 \quad = \left(\frac{M}{2L}\right)^3 \varepsilon_1^4$$

$$\varepsilon_4 \leqslant \frac{M}{2L}\varepsilon_3^2 \leqslant \frac{M}{2L}\left[\left(\frac{M}{2L}\right)^3 \varepsilon_1^4\right]^2 = \left(\frac{M}{2L}\right)^7 \varepsilon_1^8$$

$$\varepsilon_5 \leqslant \frac{M}{2L}\varepsilon_4^2 \leqslant \frac{M}{2L}\left[\left(\frac{M}{2L}\right)^7 \varepsilon_1^8\right]^2 = \left(\frac{M}{2L}\right)^{15} \varepsilon_1^{16}$$

By now we can see a pattern forming, that is easily verified by induction[8].

$$\varepsilon_n \leqslant \left(\frac{M}{2L}\right)^{2^{n-1}-1} \varepsilon_1^{2^{n-1}} = \frac{2L}{M}\left(\frac{M}{2L}\varepsilon_1\right)^{2^{n-1}} \tag{E5}$$

As long as $\frac{M}{2L}\varepsilon_1 < 1$ (which gives us a quantitative idea as to how good our first guess has to be in order for Newton's method to work), this goes to zero extremely quickly as n increases. For example, suppose that $\frac{M}{2L}\varepsilon_1 \leqslant \frac{1}{2}$. Then

$$\varepsilon_n \leqslant \frac{2L}{M}\left(\frac{1}{2}\right)^{2^{n-1}} \leqslant \frac{2L}{M} \cdot \begin{cases} 0.25 & \text{if } n = 2 \\ 0.0625 & \text{if } n = 3 \\ 0.0039 = 3.9 \times 10^{-3} & \text{if } n = 4 \\ 0.000015 = 1.5 \times 10^{-5} & \text{if } n = 5 \\ 0.00000000023 = 2.3 \times 10^{-10} & \text{if } n = 6 \\ 0.000000000000000000054 = 5.4 \times 10^{-20} & \text{if } n = 7 \end{cases}$$

Each time you increase n by one, the number of zeroes after the decimal place roughly doubles. You can see why from (E5). Since

$$\left(\frac{M}{2L}\varepsilon_1\right)^{2^{(n+1)-1}} = \left(\frac{M}{2L}\varepsilon_1\right)^{2^{n-1}\times 2} = \left[\left(\frac{M}{2L}\varepsilon_1\right)^{2^{n-1}}\right]^2$$

we have, *very* roughly speaking, $\varepsilon_{n+1} \approx \varepsilon_n^2$. This *quadratic* behaviour is the reason that Newton's method is so useful.

---

8   Mathematical induction is a technique for proving a sequence $S_1, S_2, S_3, \cdots$ of statements. That technique consists of first proving that S_1 is true, and then proving that, for any natural number n, if S_n is true then S_{n+1} is true.

Example C.2.1 (*Example C.1.2, continued*)

Let's consider, as we did in Example C.1.2, $f(x) = x^2 - 2$, starting with $x_1 = \frac{3}{2}$. Then

$$f'(x) = 2x \qquad f''(x) = 2$$

Recalling, from (H1) and (H2), that L is a lower bound on $|f'|$ and M is an upper bound on $|f''|$, we may certainly take $M = 2$ and if, for example, $x_n \geqslant 1$ for all n (as happened in Example C.1.2), we may take $L = 2$ too. While we do not know what r is, we do know that $1 \leqslant r \leqslant 2$ (since $f(1) = 1^1 - 2 < 0$ and $f(2) = 2^2 - 2 > 0$). As we took $x_1 = \frac{3}{2}$, we have $\varepsilon_1 = |x_1 - r| \leqslant \frac{1}{2}$, so that $\frac{M}{2L}\varepsilon_1 \leqslant \frac{1}{4}$ and

$$\varepsilon_{n+1} \leqslant \frac{2L}{M}\left(\frac{M}{2L}\varepsilon_1\right)^{2^{n-1}} \leqslant 2\left(\frac{1}{4}\right)^{2^{n-1}} \tag{E6}$$

This tends to zero very quickly as n increases. Furthermore this is an upper bound on the error and not the actual error. In fact (E6) is a very crude upper bound. For example, setting $n = 3$ gives the bound

$$\varepsilon_4 \leqslant 2\left(\frac{1}{4}\right)^{2^2} = 7 \times 10^{-3}$$

and we saw in Example C.1.2 that the actual error in x_4 was smaller than 5×10^{-10}.

Example C.2.1

Example C.2.2 (*Example C.1.3, continued*)

Let's consider, as we did in Example C.1.3, $f(x) = \sin x$, starting with $x_1 = 3$. Then

$$f'(x) = \cos x \qquad f''(x) = -\sin x$$

As $|-\sin x| \leqslant 1$, we may certainly take $M = 1$. In Example C.1.3, all x_n's were between 3 and 3.2. Since (to three decimal places)

$$\sin(3) = 0.141 > 0 \qquad \sin(3.2) = -0.058 < 0$$

the IVT (intermediate value theorem) tells us that $3 < r < 3.2$ and $\varepsilon_1 = |x_1 - r| < 0.2$.
So r and all x_n's and hence all c's lie in the interval $(3, 3.2)$. Since

$$-0.9990 = \cos(3) < \cos c < \cos(3.2) = -0.9983$$

we necessarily have $|f'(c)| = |\cos c| \geqslant 0.9$ and we may take $L = 0.9$. So

$$\varepsilon_{n+1} \leqslant \frac{2L}{M}\left(\frac{M}{2L}\varepsilon_1\right)^{2^{n-1}} \leqslant \frac{2 \times 0.9}{1}\left(\frac{1}{2 \times 0.9}0.2\right)^{2^{n-1}} \leqslant 2\left(\frac{1}{9}\right)^{2^{n-1}}$$

This tends to zero very quickly as n increases.

Example C.2.2

We have now seen two procedures for finding roots of a function $f(x)$ — the bisection method (which does not use the derivative of $f(x)$, but which is not very efficient) and Newton's method (which does use the derivative of $f(x)$, and which is very efficient). In fact, there is a whole constellation of other methods[9] and the interested reader should search engine their way to, for example, Wikipedia's article on root finding algorithms. Here, we will just mention two other methods, one being a variant of the bisection method and the other being a variant of Newton's method.

C.3 ▴ The false position (regula falsi) method

Let $f(x)$ be a continuous function and let $a_1 < b_1$ with $f(a_1)$ and $f(b_1)$ being of opposite sign.

As we have seen, the bisection method generates a sequence of intervals $I_n = [a_n, b_n]$, $n = 1, 2, 3, \cdots$ with, for each n, $f(a_n)$ and $f(b_n)$ having opposite sign (so that, by continuity, f has a root in I_n). Once we have I_n, we choose I_{n+1} based on the sign of f at the midpoint, $\frac{a_n + b_n}{2}$, of I_n. Since we always test the midpoint, the possible error decreases by a factor of 2 each step.

The false position method tries to make the whole procedure more efficient by testing the sign of f at a point that is closer to the end of I_n where the magnitude of f is smaller. To be precise, we approximate $y = f(x)$ by the equation of the straight line through $(a_n, f(a_n))$ and $(b_n, f(b_n))$.
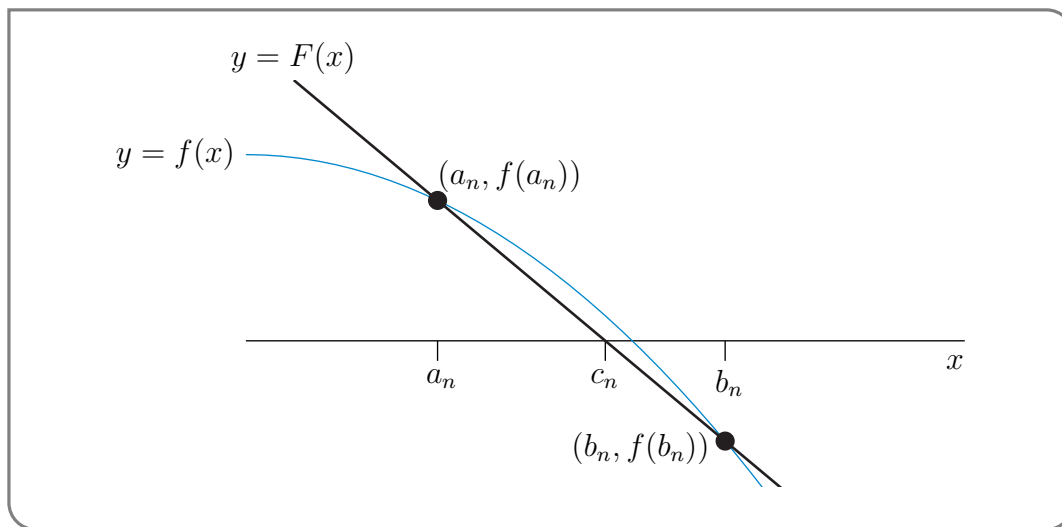
The equation of that straight line is

$$y = F(x) = f(a_n) + \frac{f(b_n) - f(a_n)}{b_n - a_n}(x - a_n)$$

---

9   What does it say about mathematicians that they have developed so many ways of finding zero?

Then the false position method tests the sign of $f(x)$ at the value of x where $F(x) = 0$.

$$F(x) = f(a_n) + \frac{f(b_n) - f(a_n)}{b_n - a_n}(x - a_n) = 0$$

$$\iff x = a_n - \frac{b_n - a_n}{f(b_n) - f(a_n)}f(a_n) = \frac{a_n f(b_n) - b_n f(a_n)}{f(b_n) - f(a_n)}$$

So once we have the interval I_n, the false position method generates the interval I_{n+1} by the following rule.[10]

Equation C.3.1 (false position method).

Set $c_n = \frac{a_n f(b_n) - b_n f(a_n)}{f(b_n) - f(a_n)}$. If $f(c_n)$ has the same sign as $f(a_n)$, then

$$I_{n+1} = [a_{n+1}, b_{n+1}] \quad \text{with} \quad a_{n+1} = c_n, \; b_{n+1} = b_n$$

and if $f(c_n)$ and $f(a_n)$ have opposite signs, then

$$I_{n+1} = [a_{n+1}, b_{n+1}] \quad \text{with} \quad a_{n+1} = a_n, \; b_{n+1} = c_n$$

C.4 ▲ The secant method

Let $f(x)$ be a continuous function. The secant method is a variant of Newton's method that avoids the use of the derivative of $f(x)$ — which can be very helpful when dealing with the derivative is not easy. It avoids the use of the derivative by approximating $f'(x)$ by $\frac{f(x+h) - f(x)}{h}$ for some h. That is, it approximates the tangent line to f at x by a secant line to f that passes through x. To limit the number of evaluations of $f(x)$ required, it uses $x = x_{n-1}$ and $x + h = x_n$. Here is how it works.

Suppose that we have already found x_n. Then we denote by $y = F(x)$ the equation of the (secant) line that passes through $(x_{n-1}, f(x_{n-1}))$ and $(x_n, f(x_n))$ and we choose x_{n+1} to be the value of x where $F(x) = 0$.

---
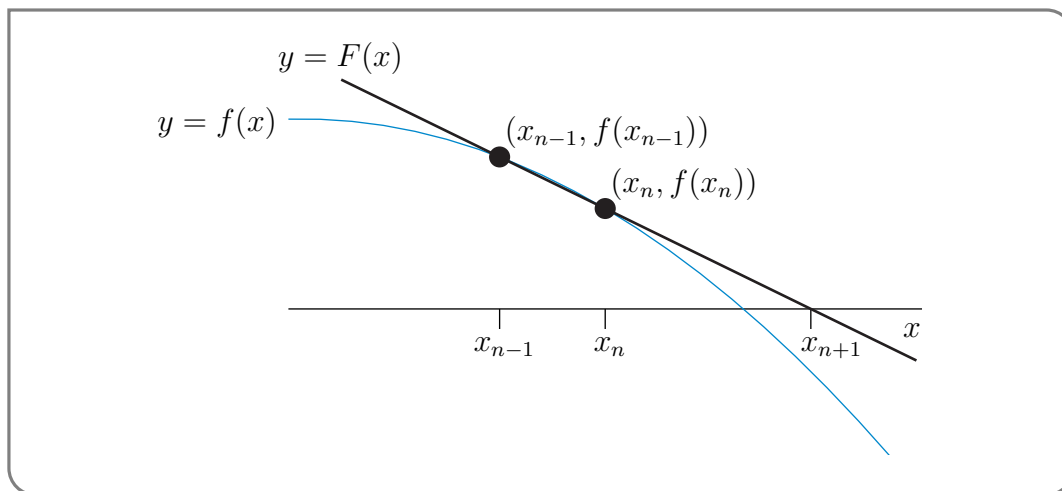
10   The convergence behaviour of the false position method is relatively complicated. So we do not discuss it here. As always, we invite the interested reader to visit their favourite search engine.

The equation of the secant line is

$$y = F(x) = f(x_{n-1}) + \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}(x - x_{n-1})$$

so that x_{n+1} is determined by

$$0 = F(x_{n+1}) = f(x_{n-1}) + \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}(x_{n+1} - x_{n-1})$$

$$\iff x_{n+1} = x_{n-1} - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}f(x_{n-1})$$

or, simplifying,

Equation C.4.1 (secant method).

$$x_{n+1} = \frac{x_{n-1}f(x_n) - x_nf(x_{n-1})}{f(x_n) - f(x_{n-1})}$$

Of course, to get started with $n = 1$, we need two initial guesses, x_0 and x_1, for the root.

Example C.4.2 (Approximating $\sqrt{2}$, again)

In this example we compute, approximately, the square root of two by applying the secant method to the equation

$$f(x) = x^2 - 2 = 0$$

and we'll compare the secant method results with the corresponding Newton's method results. (See Example C.1.2.)

Since $f'(x) = 2x$, (C.1.1) says that, under Newton's method, we should iteratively apply

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^2 - 2}{2x_n} = \frac{x_n}{2} + \frac{1}{x_n}$$

while (C.4.1) says that, under the secant method, we should iteratively apply (after a little simplifying algebra)

$$\begin{aligned}x_{n+1} &= \frac{x_{n-1}f(x_n) - x_nf(x_{n-1})}{f(x_n) - f(x_{n-1})} = \frac{x_{n-1}[x_n^2 - 2] - x_n[x_{n-1}^2 - 2]}{x_n^2 - x_{n-1}^2}\\ &= \frac{x_{n-1}x_n[x_n - x_{n-1}] + 2[x_n - x_{n-1}]}{x_n^2 - x_{n-1}^2}\\ &= \frac{x_{n-1}x_n + 2}{x_{n-1} + x_n}\end{aligned}$$

Here are the results, starting Newton's method with $x_1 = 4$ and starting the secant method

with $x_0 = 4$, $x_1 = 3$. (So we are giving the secant method a bit of a head start.)

	secant method	Newton's method
x_0	4	
x_1	3	4
x_2	2	2.25
x_3	1.6	1.57
x_4	1.444	1.422
x_5	1.4161	1.414234
x_6	1.414233	1.414213562525
x_7	1.414213575	1.414213562373095

For comparison purposes, the square root of 2, to 15 decimal places, is 1.414213562373095. So the secant method x_7 is accurate to 7 decimal places and the Newton's method x_7 is accurate to at least 15 decimal places.

Example C.4.2

The advantage that the secant method has over Newton's method is that it does not use the derivative of f. This can be a substantial advantage, for example when evaluation of the derivative is computationally difficult or expensive. On the other hand, the above example suggests that the secant method is not as fast as Newton's method. The following section shows that this is indeed the case.

C.5 ▴ The Error Behaviour of the Secant Method

Let $f(x)$ have two continuous derivatives, and let r be any solution of $f(x) = 0$. We will now get a pretty good handle on the error behaviour of the secant method near r.

Denote by $\tilde{\varepsilon}_n = x_n - r$ the (signed) error in x_n and by $\varepsilon_n = |x_n - r|$ the (absolute) error in x_n. Then, $x_n = r + \tilde{\varepsilon}_n$, and, by (C.4.1),

$$\tilde{\varepsilon}_{n+1} = \frac{x_{n-1}f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})} - r$$
$$= \frac{[r + \tilde{\varepsilon}_{n-1}]f(x_n) - [r + \tilde{\varepsilon}_n]f(x_{n-1})}{f(x_n) - f(x_{n-1})} - r$$
$$= \frac{\tilde{\varepsilon}_{n-1}f(x_n) - \tilde{\varepsilon}_n f(x_{n-1})}{f(x_n) - f(x_{n-1})}$$

By the Taylor expansion (3.4.32) and the mean value theorem (Theorem 2.13.4),

$$f(x_n) = f(r) + f'(r)\tilde{\varepsilon}_n + \frac{1}{2}f''(c_1)\tilde{\varepsilon}_n^2$$
$$= f'(r)\tilde{\varepsilon}_n + \frac{1}{2}f''(c_1)\tilde{\varepsilon}_n^2$$
$$f(x_n) - f(x_{n-1}) = f'(c_2)[x_n - x_{n-1}]$$
$$= f'(c_2)[\tilde{\varepsilon}_n - \tilde{\varepsilon}_{n-1}]$$

for some c_1 between r and x_n and some c_2 between x_{n-1} and x_n. So, for x_{n-1} and x_n near r, c_1 and c_2 also have to be near r and

$$f(x_n) \approx f'(r)\tilde{\varepsilon}_n + \frac{1}{2}f''(r)\tilde{\varepsilon}_n^2$$

$$f(x_{n-1}) \approx f'(r)\tilde{\varepsilon}_{n-1} + \frac{1}{2}f''(r)\tilde{\varepsilon}_{n-1}^2$$

$$f(x_n) - f(x_{n-1}) \approx f'(r)[\tilde{\varepsilon}_n - \tilde{\varepsilon}_{n-1}]$$

and

$$\begin{aligned}
\tilde{\varepsilon}_{n+1} &= \frac{\tilde{\varepsilon}_{n-1}f(x_n) - \tilde{\varepsilon}_n f(x_{n-1})}{f(x_n) - f(x_{n-1})} \\
&\approx \frac{\tilde{\varepsilon}_{n-1}[f'(r)\tilde{\varepsilon}_n + \frac{1}{2}f''(r)\tilde{\varepsilon}_n^2] - \tilde{\varepsilon}_n[f'(r)\tilde{\varepsilon}_{n-1} + \frac{1}{2}f''(r)\tilde{\varepsilon}_{n-1}^2]}{f'(r)[\tilde{\varepsilon}_n - \tilde{\varepsilon}_{n-1}]} \\
&= \frac{\frac{1}{2}\tilde{\varepsilon}_{n-1}\tilde{\varepsilon}_n f''(r)[\tilde{\varepsilon}_n - \tilde{\varepsilon}_{n-1}]}{f'(r)[\tilde{\varepsilon}_n - \tilde{\varepsilon}_{n-1}]} \\
&= \frac{f''(r)}{2f'(r)}\tilde{\varepsilon}_{n-1}\tilde{\varepsilon}_n
\end{aligned}$$

Taking absolute values, we have

$$\varepsilon_{n+1} \approx K\varepsilon_{n-1}\varepsilon_n \qquad \text{with } K = \left|\frac{f''(r)}{2f'(r)}\right| \tag{E7}$$

We have seen that Newton's method obeys a similar formula — (E3) says that, when x_n is near r, Newton's method obeys $\varepsilon_{n+1} \approx K\varepsilon_n^2$, also with $K = \left|\frac{f''(r)}{2f'(r)}\right|$. As we shall now see, the change from ε_n^2, in $\varepsilon_{n+1} \approx K\varepsilon_n^2$, to $\varepsilon_{n-1}\varepsilon_n$, in $\varepsilon_{n+1} \approx K\varepsilon_{n-1}\varepsilon_n$, does have a substantial impact on the behaviour of ε_n for large n.

To see the large n behaviour, we now iterate (E7). The formulae will look simpler if we multiply (E7) by K and write $\delta_n = K\varepsilon_n$. Then (E7) becomes $\delta_{n+1} \approx \delta_{n-1}\delta_n$ (and we have eliminated K). The first iterations are

$$\begin{aligned}
\delta_2 &\approx \delta_0\delta_1 \\
\delta_3 &\approx \delta_1\delta_2 \approx \delta_0\delta_1^2 \\
\delta_4 &\approx \delta_2\delta_3 \approx \delta_0^2\delta_1^3 \\
\delta_5 &\approx \delta_3\delta_4 \approx \delta_0^3\delta_1^5 \\
\delta_6 &\approx \delta_4\delta_5 \approx \delta_0^5\delta_1^8 \\
\delta_7 &\approx \delta_5\delta_6 \approx \delta_0^8\delta_1^{13}
\end{aligned}$$

Notice that every δ_n is of the form $\delta_0^{\alpha_n}\delta_1^{\beta_n}$. Substituting $\delta_n = \delta_0^{\alpha_n}\delta_1^{\beta_n}$ into $\delta_{n+1} \approx \delta_{n-1}\delta_n$ gives

$$\delta_0^{\alpha_{n+1}}\delta_1^{\beta_{n+1}} \approx \delta_0^{\alpha_{n-1}}\delta_1^{\beta_{n-1}}\delta_0^{\alpha_n}\delta_1^{\beta_n}$$

and we have

$$\alpha_{n+1} = \alpha_{n-1} + \alpha_n \qquad \beta_{n+1} = \beta_{n-1} + \beta_n \tag{E8}$$

The recursion rule in (E8) is famous[11]. The Fibonacci[12] sequence (which is 0, 1, 1, 2, 3, 5, 8, 13, $\cdots$), is defined by

$$F_0 = 0$$
$$F_1 = 1$$
$$F_n = F_{n-1} + F_{n-2} \qquad \text{for } n > 1$$

So, for $n \geqslant 2$, $\alpha_n = F_{n-1}$ and $\beta_n = F_n$ and

$$\delta_n \approx \delta_0^{\alpha_n} \delta_1^{\beta_n} = \delta_0^{F_{n-1}} \delta_1^{F_n}$$

One of the known properties of the Fibonacci sequence is that, for large n,

$$F_n \approx \frac{\varphi^n}{\sqrt{5}} \qquad \text{where } \varphi = \frac{1 + \sqrt{5}}{2} \approx 1.61803$$

This φ is the golden ratio[13]. So, for large n,

$$K\varepsilon_n = \delta_n \approx \delta_0^{F_{n-1}} \delta_1^{F_n} \approx \delta_0^{\frac{\varphi^{n-1}}{\sqrt{5}}} \delta_1^{\frac{\varphi^n}{\sqrt{5}}} = \delta_0^{\frac{1}{\sqrt{5}\varphi} \times \varphi^n} \delta_1^{\frac{1}{\sqrt{5}} \times \varphi^n}$$

$$= d^{\varphi^n} \qquad \text{where} \quad d = \delta_0^{\frac{1}{\sqrt{5}\varphi}} \delta_1^{\frac{1}{\sqrt{5}}}$$

$$\approx d^{1.6^n}$$

Assuming that $0 < \delta_0 = K\varepsilon_0 < 1$ and $0 < \delta_1 = K\varepsilon_1 < 1$, we will have $0 < d < 1$.

By way of contrast, for Newton's method, for large n,

$$K\varepsilon_n \approx d^{2^n} \qquad \text{where} \quad d = (K\varepsilon_1)^{1/2}$$

As 2^n grows quite a bit more quickly than 1.6^n (for example, when n=5, $2^n = 32$ and $1.6^n = 10.5$, and when $n = 10$, $2^n = 1024$ and $1.6^n = 110$) Newton's method homes in on the root quite a bit faster than the secant method, assuming that you start reasonably close to the root.

---

11   Plug "Fibonacci sequence in nature" into your search engine of choice.
12   Fibonacci (1170-1250) was an Italian mathematician who was also known as Leonardo of Pisa, Leonardo Bonacci and Leonardo Biglio Pisano.
13   Also worth a quick trip to your search engine.