CLP 1

DIFFERENTIAL CALCULUS

FELDMAN RECHNITZER YEAGER

CLP-1 Differential Calculus

Joel Feldman Andrew Rechnitzer Elyse Yeager

This document was typeset on Friday 16th August, 2024.

▸▸ Legal stuff

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

Contents

Chapter 0

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

$$\frac{2}{15} = 0.133333333\ldots$$
$$\frac{5}{17} = 0.\underline{2941176470588235}2941176470588235\underline{2941176470588235}294117647058823\ldots$$

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 $10q$:

$$\begin{aligned} q &= 0.99999999999\ldots \\ 10q &= 9.99999999999\ldots \end{aligned}$$

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

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

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 $1, 2, 3, -5, 18, \ldots$).
- The letters x, y, z, w usually denote real numbers (like 1.4323, π, $\sqrt{2}$, 6.0221415×10^{23} ... 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

- $C = \{1, 2, 3, \ldots, 40\}$ the set of all integers between 1 and 40 (inclusive).
- $A = \{1, 4, 9, 16, \ldots\}$ the set of all perfect squares[9]

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

- $B = \{3, 5, 7, \ldots\}$ — 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

  $$B = \{a \in A | a < 8\} = \{2, 3, 5, 7\}$$

  the set of elements of A that are strictly less than 8.

- Even and odd integers

  $$\begin{aligned} E &= \{n | n \text{ is an even integer}\} \\ &= \{n | n = 2k \text{ for some } k \in \mathbb{Z}\} \\ &= \{2n | n \in \mathbb{Z}\}, \end{aligned}$$

  and similarly

  $$\begin{aligned} O &= \{n | n \text{ is an odd integer}\} \\ &= \{2n + 1 | n \in \mathbb{Z}\}. \end{aligned}$$

- 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

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

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.

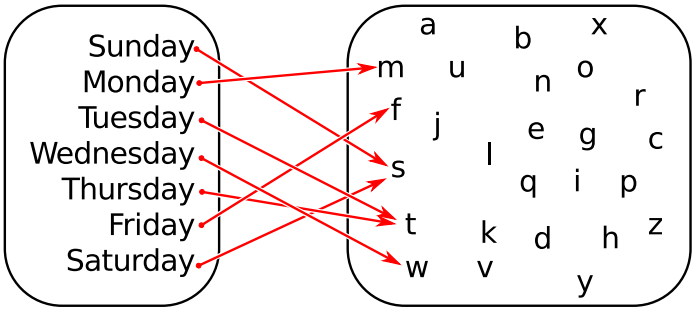
This view of functions — a function is a formula — was how mathematicians defined them up until the 19th century. As basic ideas of sets became better defined, people revised ideas surrounding functions. The more modern definition of a function between two sets is that it is a rule which assigns to each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

$A = \{\text{Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday}\}$
$B = \{\text{a,b,c,d,e,}\ldots\text{,x,y,z}\}$

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

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

  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

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

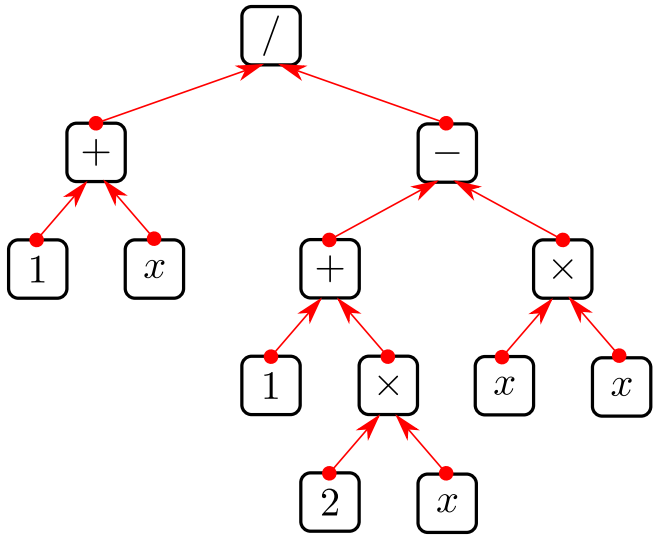
More important, however, is that we understand how we decompose this function into simpler pieces. Since much of your calculus course will involve creating and studying complicated functions by building them up from simple pieces, it is important that you really understand this point.

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.

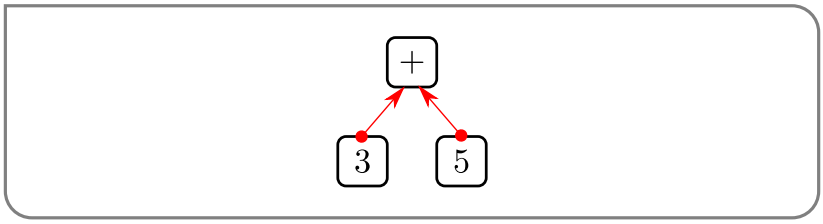
Figure 0.5.1.



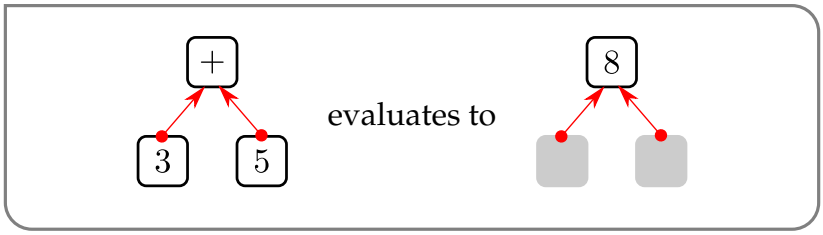
A parse tree of the function $\frac{1+x}{1+2x-x^2}$.

Let us explain the pieces here.

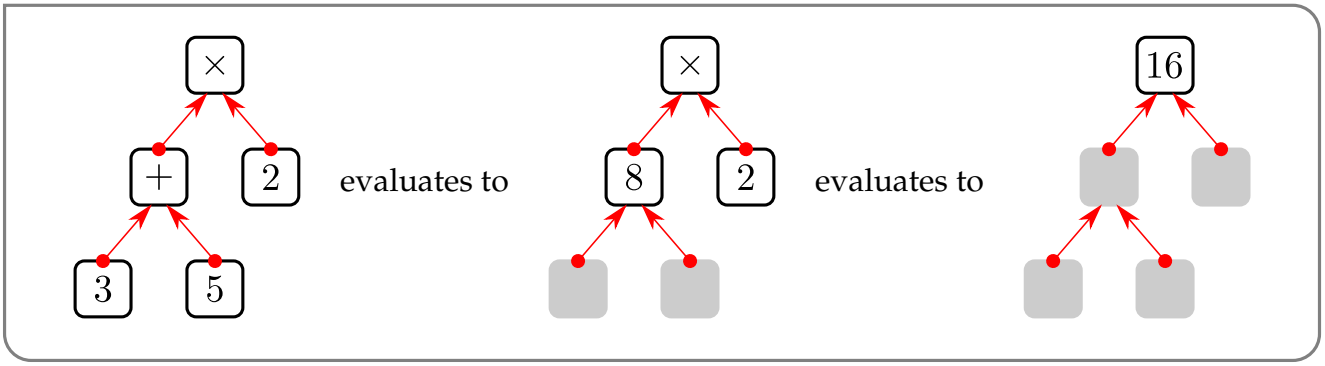
- The picture consists of boxes and arrows which are called "nodes" and "edges" respectively.
- There are two types of boxes, those containing numbers and the variable x, and those containing arithmetic operations "$+$","$-$", "$\times$" and "$/$".
- If we wish to represent the formula $3+5$, then we 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.

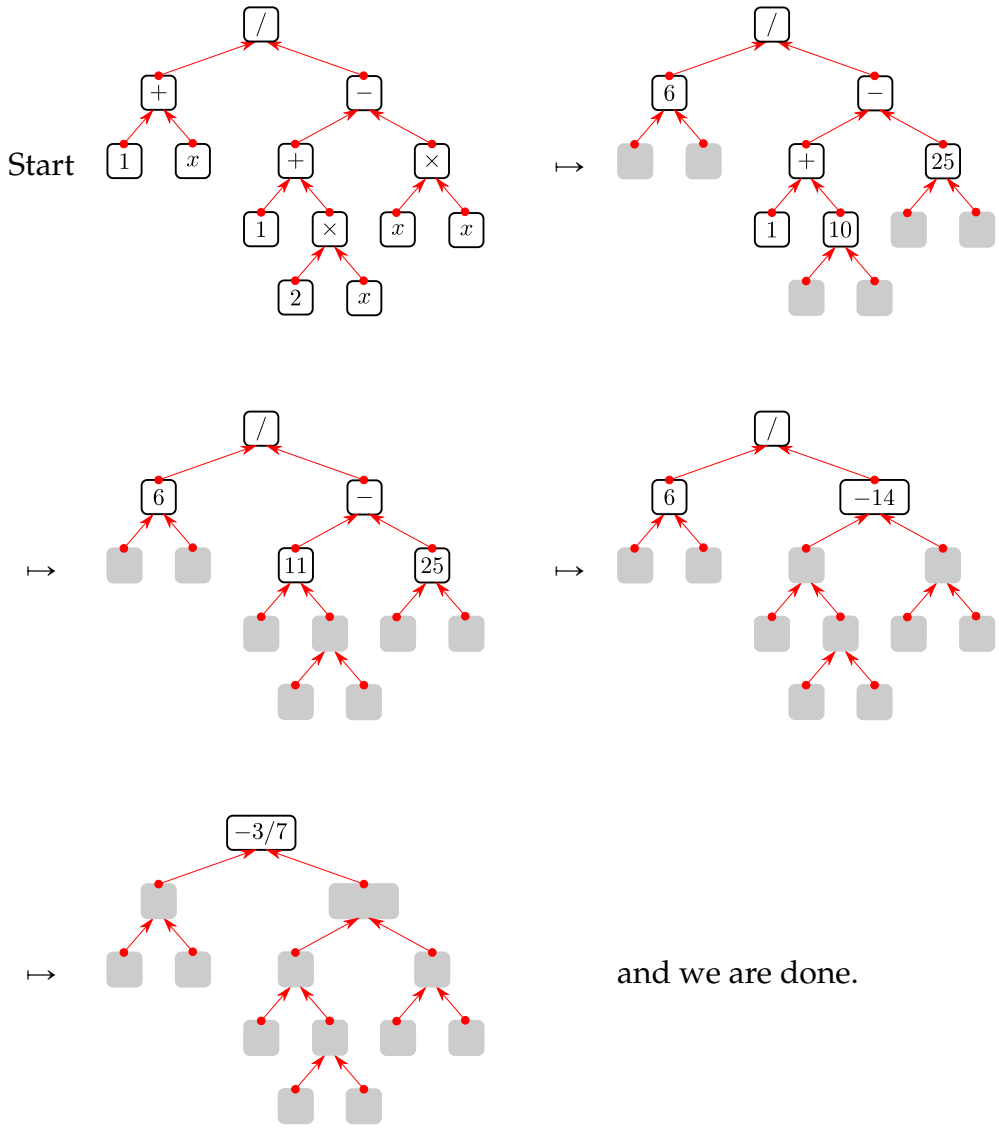


- By stringing such little "cherries" together we can describe more complicated formulas. For example, if we compute "$(3+5)\times 2$", we first compute "$(3+5)$" and then multiply the result by 2. The corresponding diagrams are



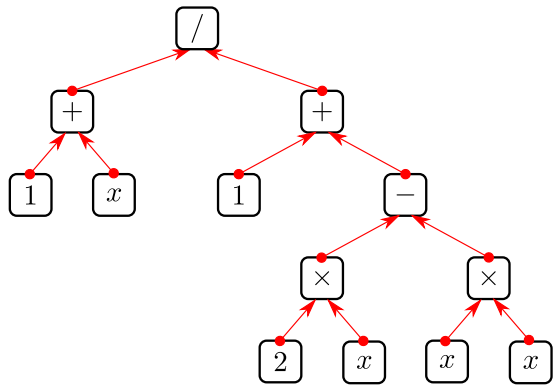
The tree we drew in Figure 0.5.1 above representing our formula has x in some of the boxes, and so when we want to compute the function at a particular value of x — say at $x = 5$ — then we replace those "x"s in the tree by that value and then compute back up the tree. See the example below

Figure 0.5.2.



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:

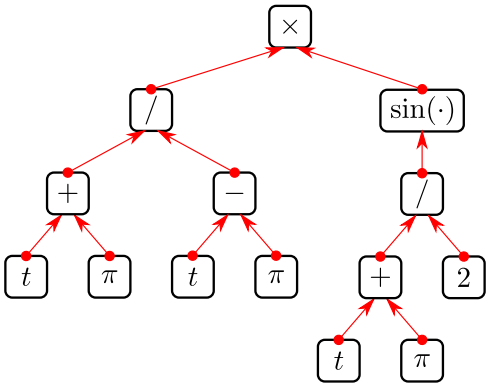
$$a \times b \times c = (a \times b) \times c = a \times (b \times c).$$

Example 0.5.1 (parsing a formula)

Consider the formula

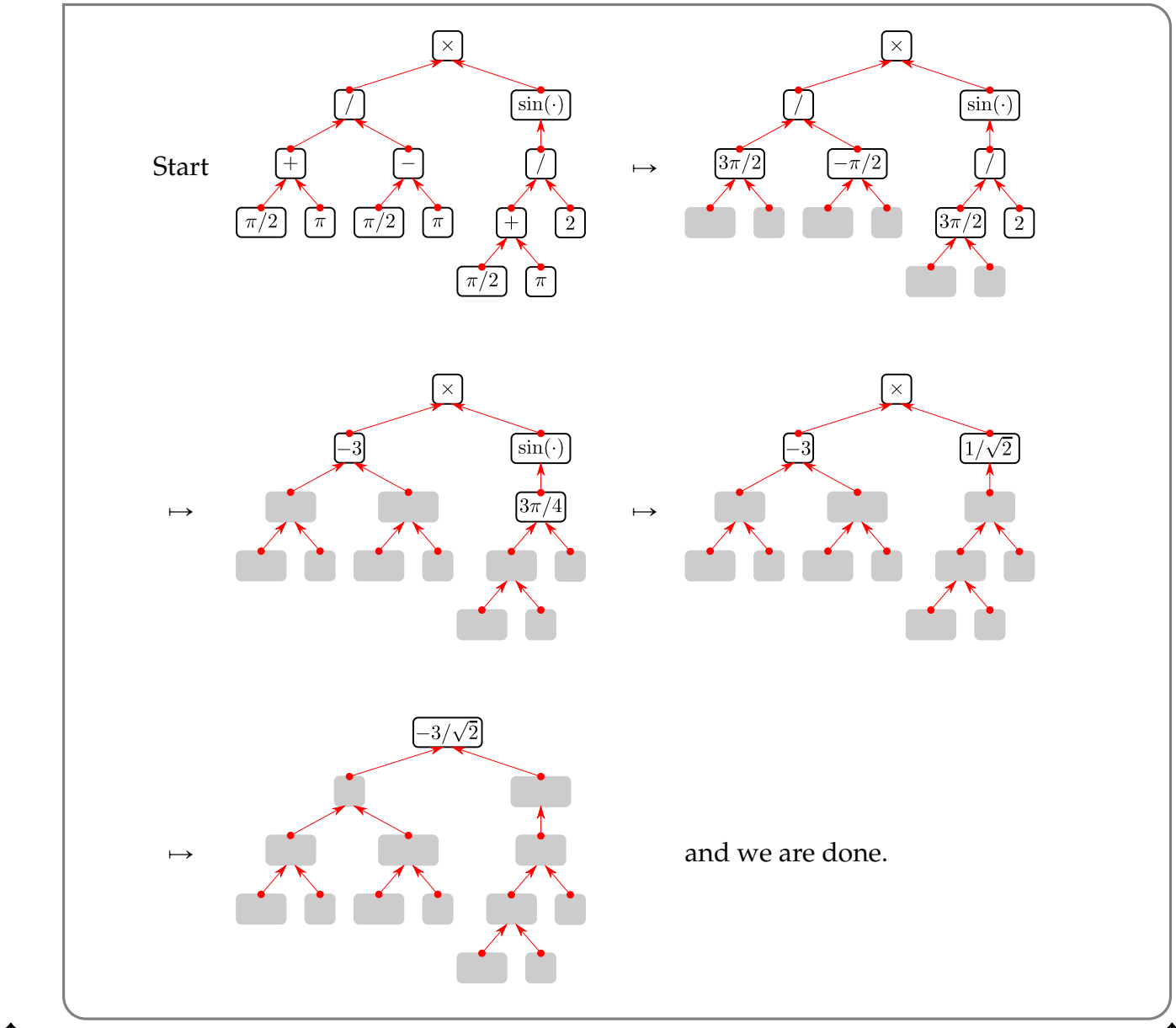
$$g(t) = \left(\frac{t+\pi}{t-\pi}\right) \cdot \sin\left(\frac{t+\pi}{2}\right).$$

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, π 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.

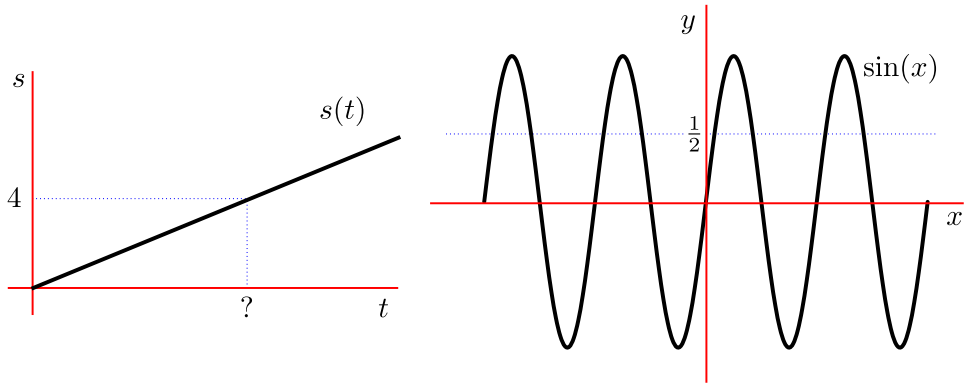
19 Indeed much of encryption exploits the fact that you can find functions that are very quick to do, but very hard to undo. For example — it is very fast to multiply two large prime numbers together, but very hard to take that result and factor it back into the original two primes. The interested reader 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

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

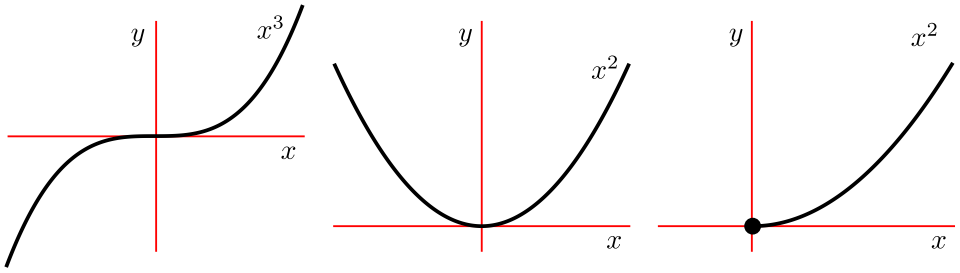
There is an easy way to test this when you have a plot of the function — the horizontal line test.

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.



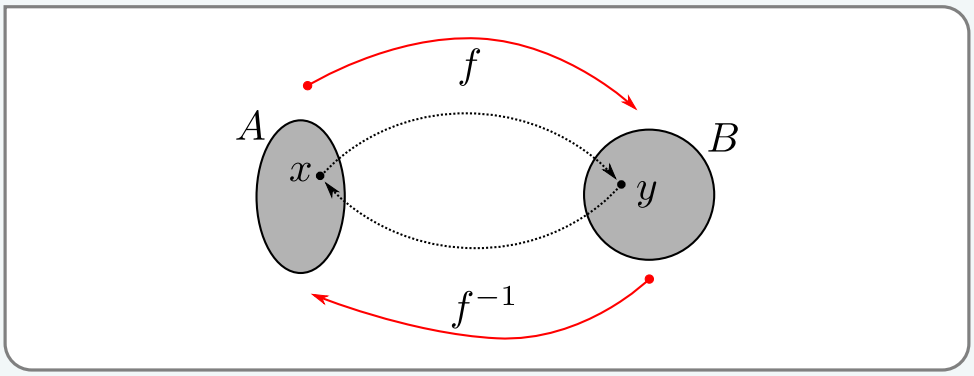
When a function is one-to-one then it has an inverse function.

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

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

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:

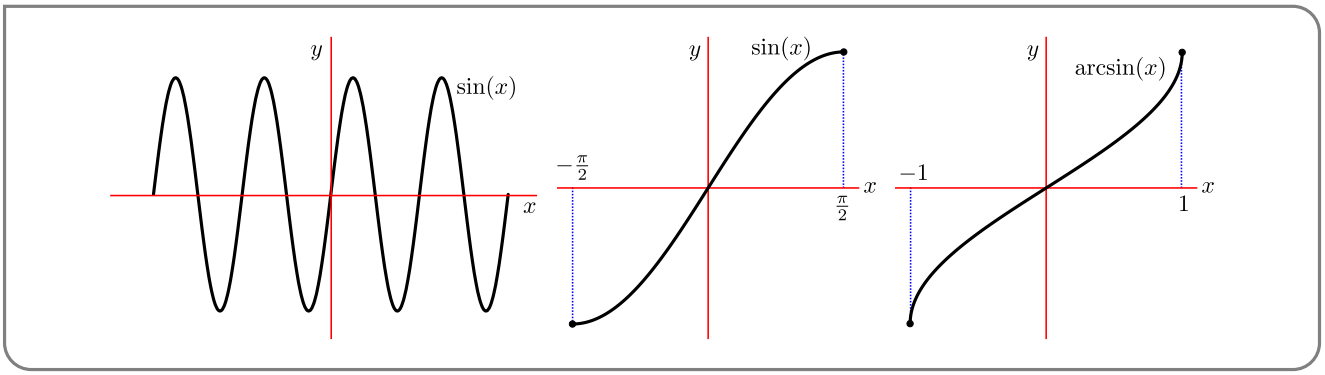
$$\begin{aligned} y &= \sqrt{x-1} \\ y^2 &= x - 1 \\ x &= y^2 + 1 = f^{-1}(y) \qquad \text{or, writing input variable as "}x\text{":} \\ f^{-1}(x) &= x^2 + 1. \end{aligned}$$

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 $\sin(y) = x$.

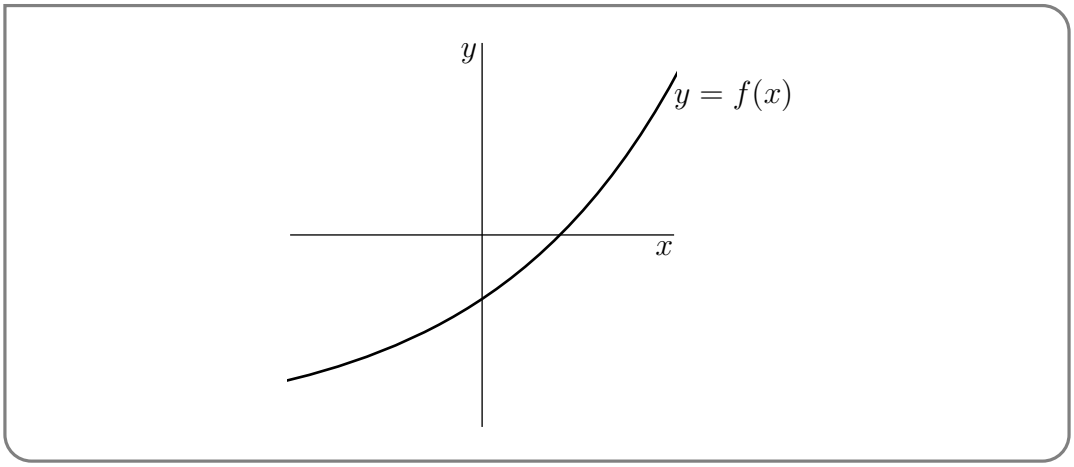
Example 0.6.6

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

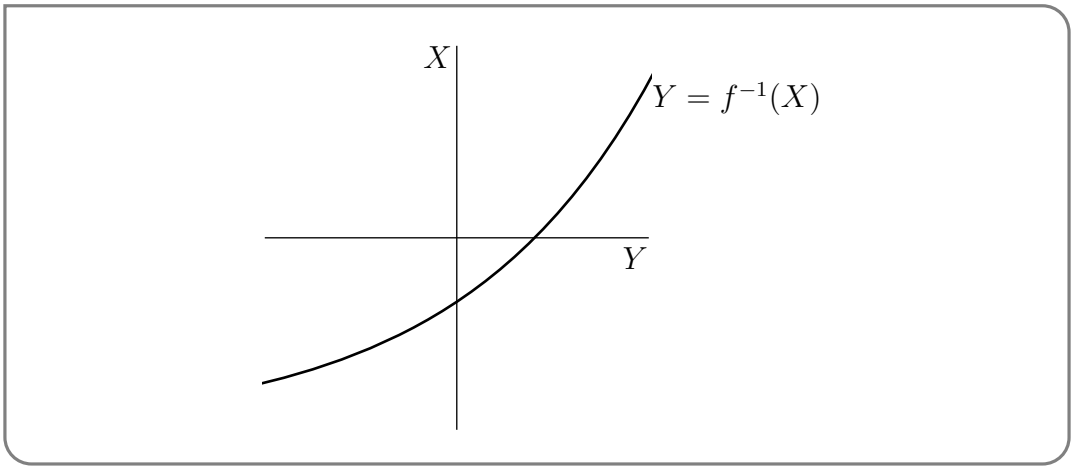
$$Y = f^{-1}(X) \iff f(Y) = X$$

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

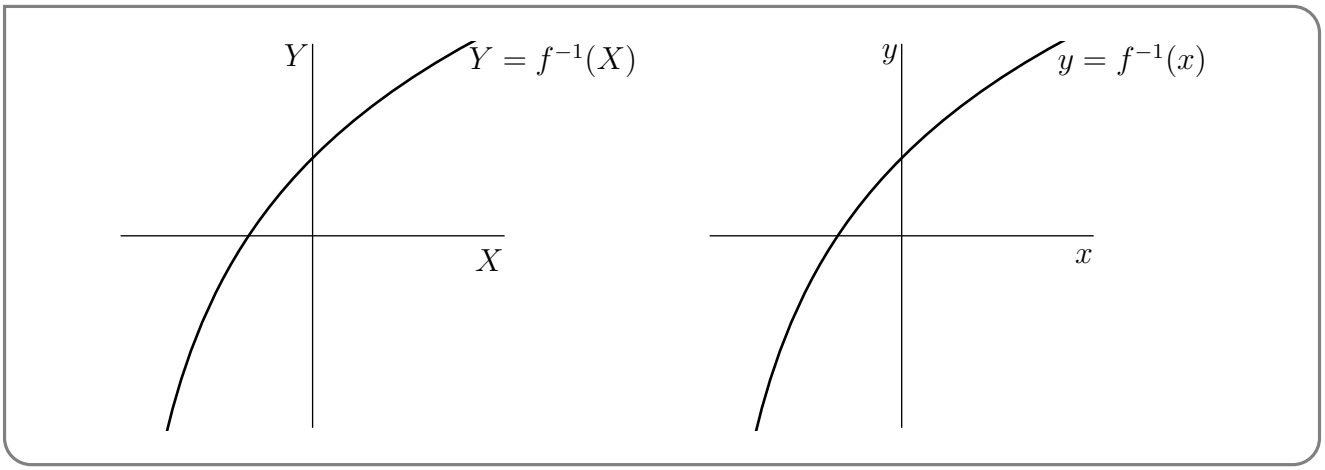
Start by drawing the graph of f, labelling the x– and y–axes and labelling the curve $y = f(x)$.



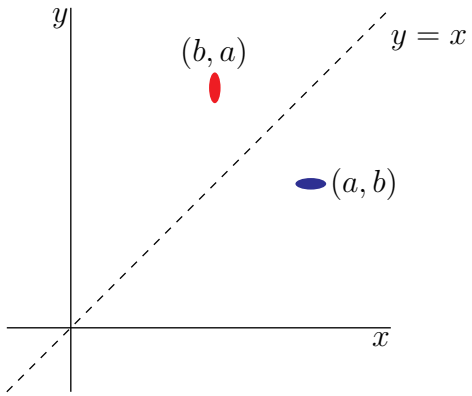
Now replace each x by Y and each y by X and replace the resulting label $X = f(Y)$ on the curve by the equivalent $Y = f^{-1}(X)$.



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).

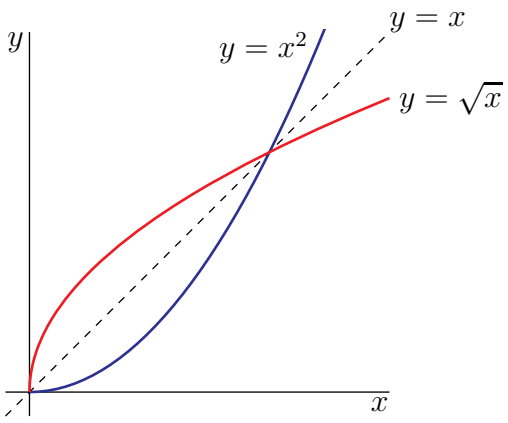


Example 0.6.7

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

Chapter 1

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.

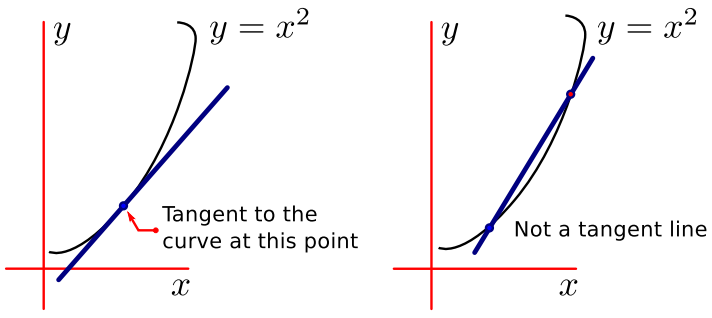
1.1 ▲ Drawing Tangents and a First Limit

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.

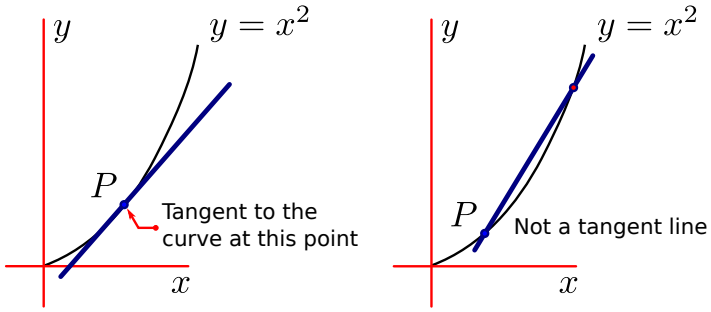


Here we have drawn two very rough sketches of the curve $y = x^2$ for $x \geqslant 0$. These are not very good sketches for a couple of reasons

- 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.

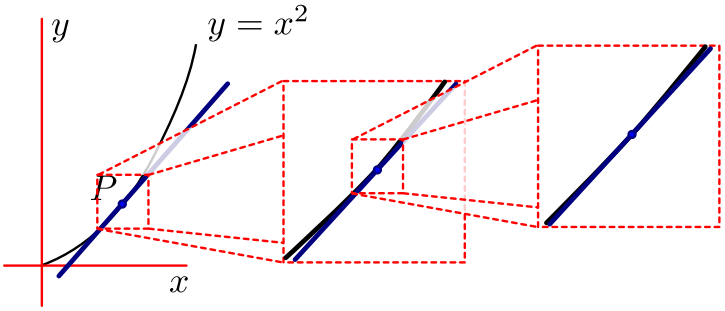
Figure 1.1.2.



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).

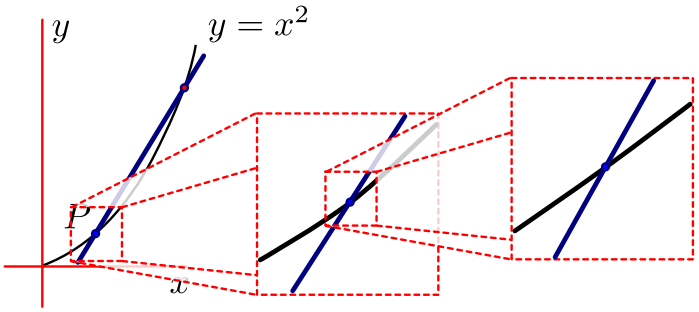
We see that as we zoom in on the point P, the graph of the function looks more and more like a straight line. If we kept on zooming in on P then the graph of the function would be indistinguishable from a straight line. That line is the tangent line (which we

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.

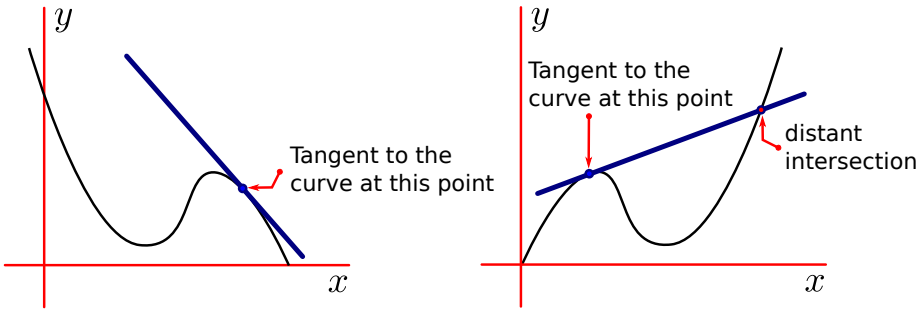
Figure 1.1.4.



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

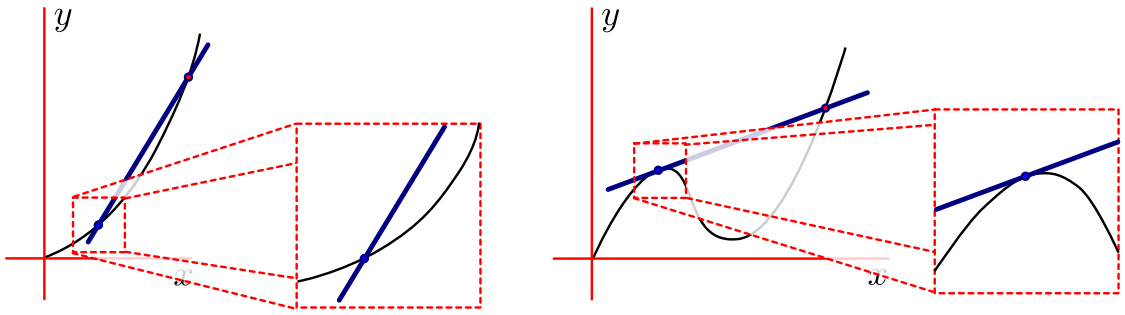
Figure 1.1.5.



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.

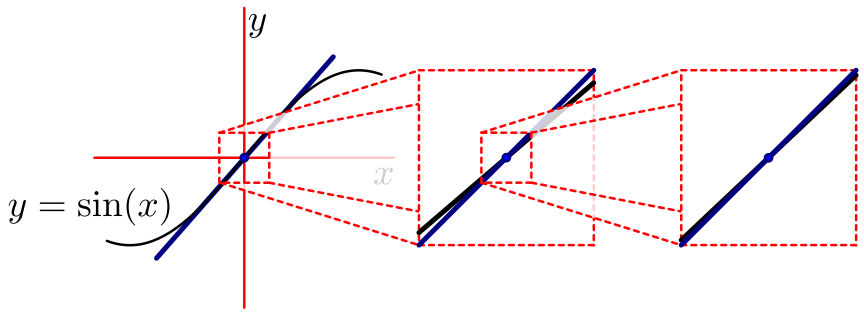


As we saw above in Figure 1.1.4, when we zoom in around our example of "not a tangent line" we see that the straight line looks very different from the curve at the "point of tangency" — i.e. where we are trying to draw the tangent. The line drawn in Figure 1.1.5(right) looks more and more like the function 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 $P = (1, 1)$.

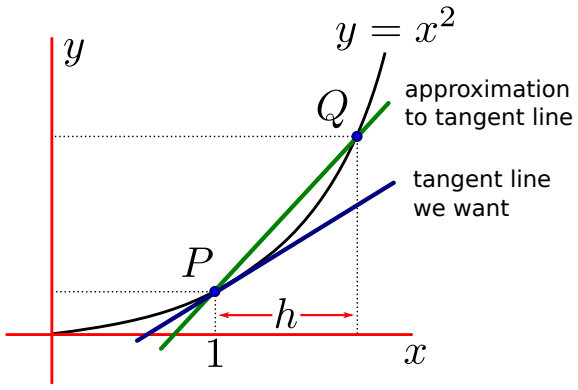
- We don't quite know how to find a line given just 1 point, however we do know how to find a line passing through 2 points. So pick another point on the curves whose coordinates are very close to P. Now rather than picking some actual numbers, I am going to write our second point as $Q = (1 + h, (1 + h)^2)$. 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.



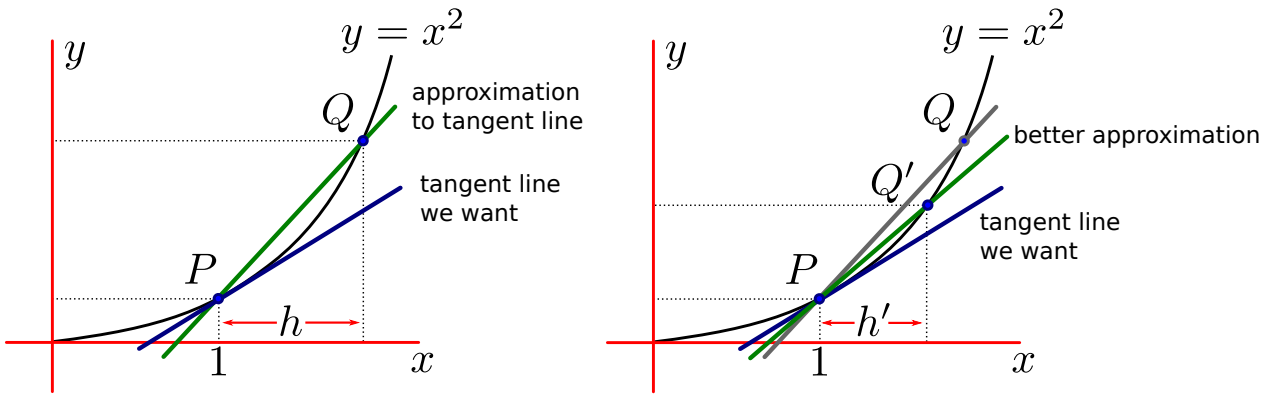
- This line that passes through the curve in two places P and Q is called a "secant line".
- The slope of the line is then

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

  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 $m = 2 + h$:

$$\begin{aligned} h = 0.1 \qquad & \qquad m = 2.1 \\ h = 0.01 \qquad & \qquad m = 2.01 \\ h = 0.001 \qquad & \qquad m = 2.001. \end{aligned}$$

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.

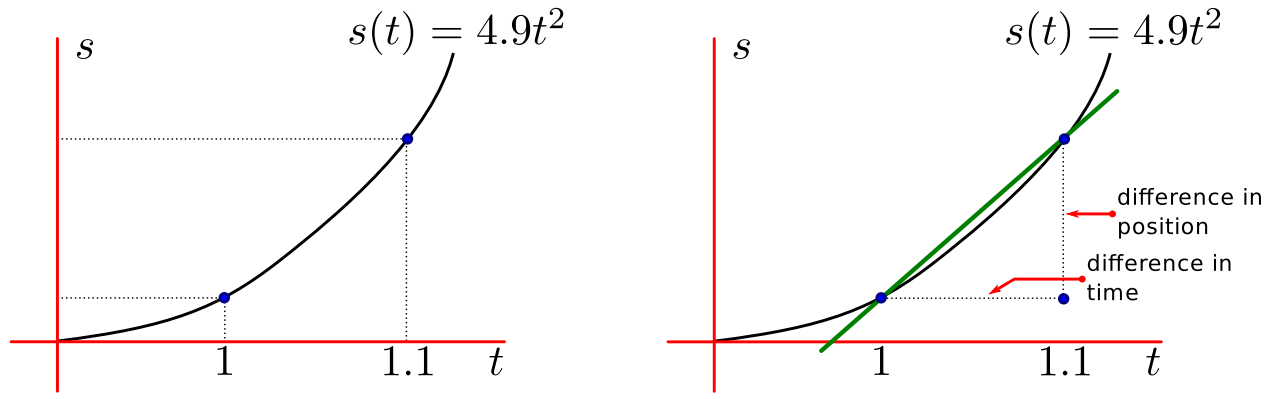
This is the formula for the average velocity of our object between time t_1 and t_2. The denominator is just the difference between these times and the numerator is the difference in position — i.e. position at time t_1 is just $s(t_1)$ and position at time t_2 is just $s(t_2)$.

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

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

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.

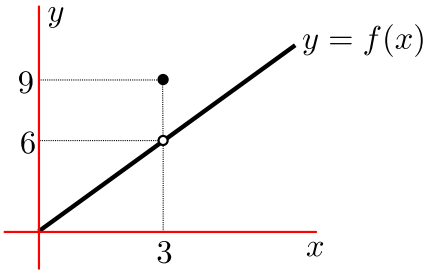
To arrive at the definition of limit, we want to start[6] with a very simple example.

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 $\lim_{x \to a} \frac{f(x) - f(a)}{x - a}$. In this case the function whose limit is being taken, namely $\frac{f(x)-f(a)}{x-a}$, 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	∘	2.001	2.01	2.1
$f(x)$	0.20408	0.20040	0.20004	∘	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

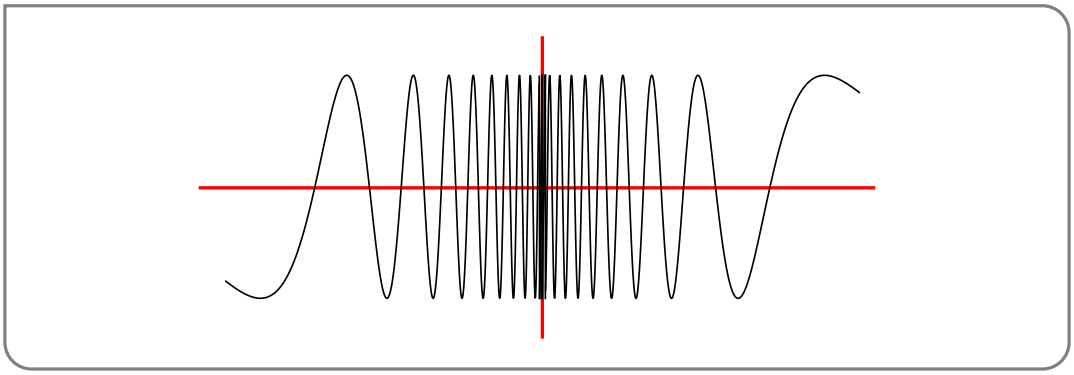
The previous two examples are nicely behaved in that the limits we tried to compute actually exist. We now turn to two nastier examples[9] in which the limits we are interested in do not exist.

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

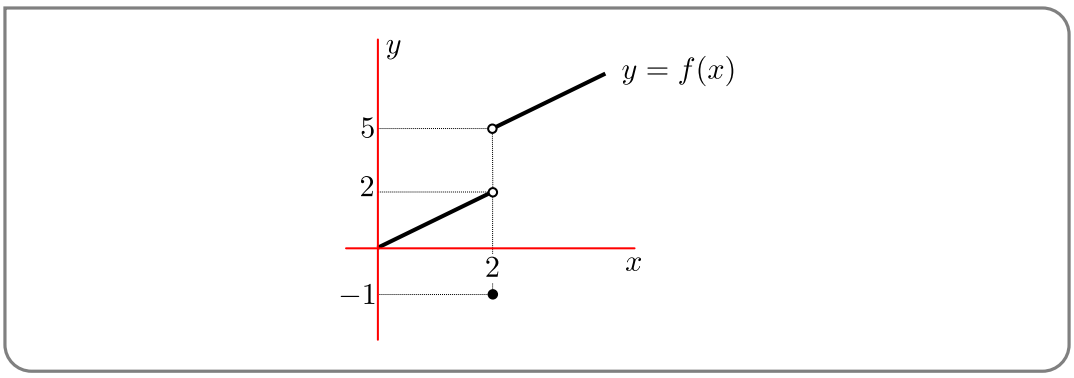
In the following example, the limit we are interested in does not exist. However the way in which things go wrong is quite different from what we just saw.

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	○	2.001	2.01	2.1
$f(x)$	1.9	1.99	1.999	○	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.

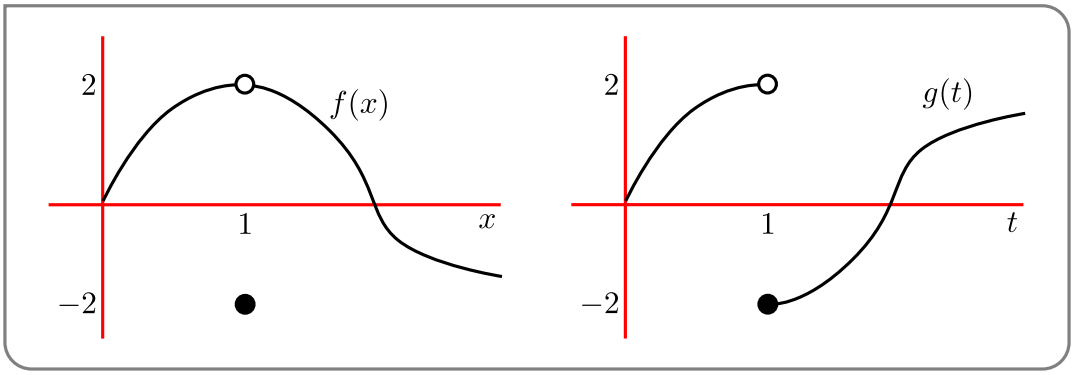
- If either of the left-hand and right-hand limits as x approaches a fail to exist, or if they both exist but are different, then the limit as x approaches a does not exist. AND,
- If the limit as x approaches a does not exist, then the left-hand and right-hand limits are either different or at least 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 — $f(x)$:

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

  so by the previous theorem

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

- Function on the right — $g(t)$:

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

  so by the previous theorem

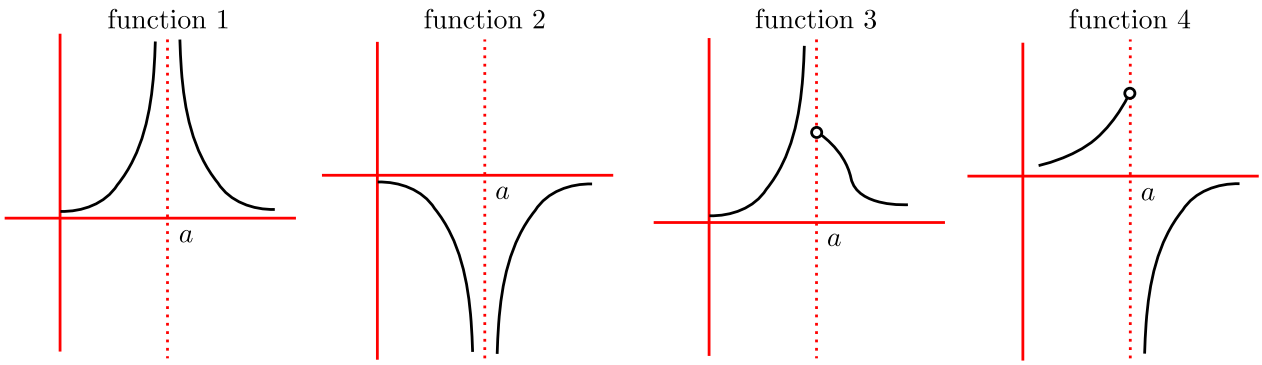
  $$\lim_{t\to 1} g(t) = \text{DNE}$$

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

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

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

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

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

So for function 3 we have

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

and for function 4

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

More examples:

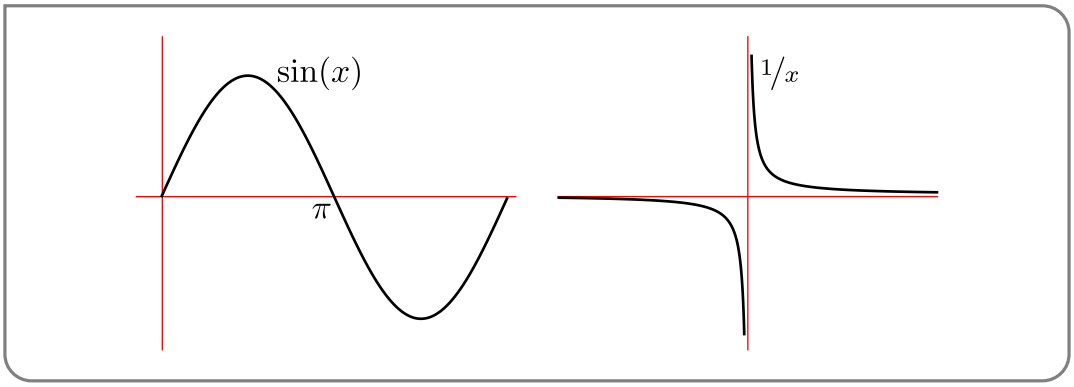
Example 1.3.12

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

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

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.

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

- $\lim\limits_{\mathbf{x\to a}} \mathbf{x} = \mathbf{a}$ — 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

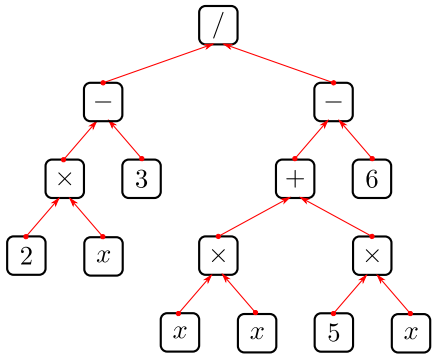
- take x and multiply it by 2
- subtract 3 to the result

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).

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

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.
- $\lim_{x\to a} cf(x) = cF$.
- $\lim_{x\to a}(f(x) \cdot g(x)) = F \cdot G$ — 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

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

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

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:

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

and now the denominator:

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

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

$$\begin{aligned} \lim_{x\to 2} \frac{x}{x-1} &= \frac{\lim_{x\to 2} x}{\lim_{x\to 2}(x-1)} \\ &= \frac{2}{1} \\ &= 2 \end{aligned}$$

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

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

  For example, $\lim_{x\to 0} \frac{1}{x^2} = DNE$.

- 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. $\lim_{x\to 0} \frac{x}{x^2} = \lim_{x\to 0} \frac{1}{x} = DNE$

- the limit is $\pm\infty$, eg. $\lim_{x\to 0} \frac{x^2}{x^4} = \lim_{x\to 0} \frac{1}{x^2} = +\infty$ or $\lim_{x\to 0} \frac{-x^2}{x^4} = \lim_{x\to 0} \frac{-1}{x^2} = -\infty$.
- the limit is zero, eg. $\lim_{x\to 0} \frac{x^2}{x} = 0$
- the limit exists and is non-zero, eg. $\lim_{x\to 0} \frac{x}{x} = 1$

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:

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

Denominator next:

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

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

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

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:

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

(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

$$\begin{aligned}\lim_{x\to 2}(4x^2-3)^{1/3} &= \left((\lim_{x\to 2} 4x^2) - (\lim_{x\to 2} 3)\right)^{1/3} \\ &= \left(4\cdot 2^2 - 3\right)^{1/3} \\ &= (16-3)^{1/3} \\ &= 13^{1/3}\end{aligned}$$

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:

$$\begin{aligned}\lim_{x\to 2}\frac{2x-3}{x^2+5x-6} &= \frac{4-3}{4+10-6} &= \frac{1}{8}\\ \lim_{x\to 2}(4x^2-1) &= 16-1 &= 15\\ \lim_{x\to 2}\frac{x}{x-1} &= \frac{2}{2-1} &= 2\end{aligned}$$

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.

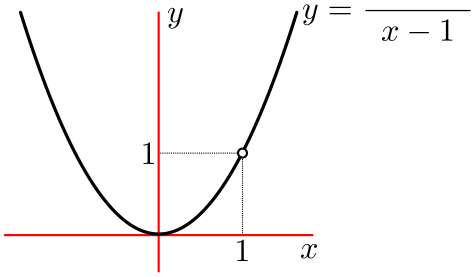
So let us play with the expression a little more before we take the limit:

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

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 $\lim_{x \to 1} x^2$ 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 $\lim_{x \to a} f(x) = \lim_{x \to a} g(x)$ 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:

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

Example 1.4.15

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

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

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

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

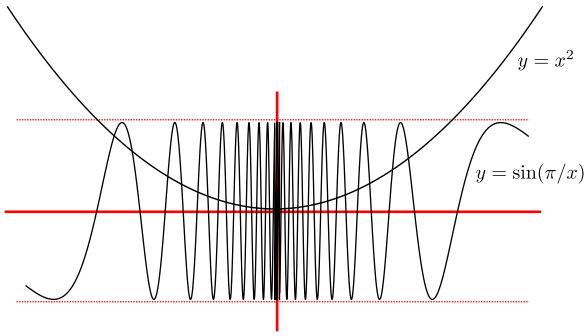
But we just cheated — we cannot use the arithmetic of limits theorem here, because the limit

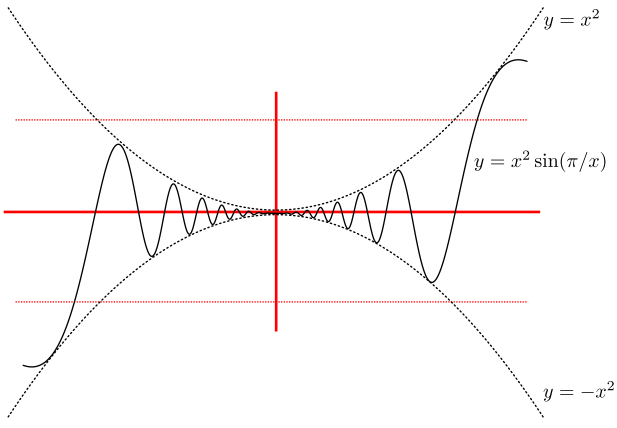
$$\lim_{x\to 0} \sin(\pi/x) = DNE$$

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

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

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

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

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

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

Multiplying the above by x^2 we see that

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

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

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

Example 1.4.18

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

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

Example 1.4.19

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

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

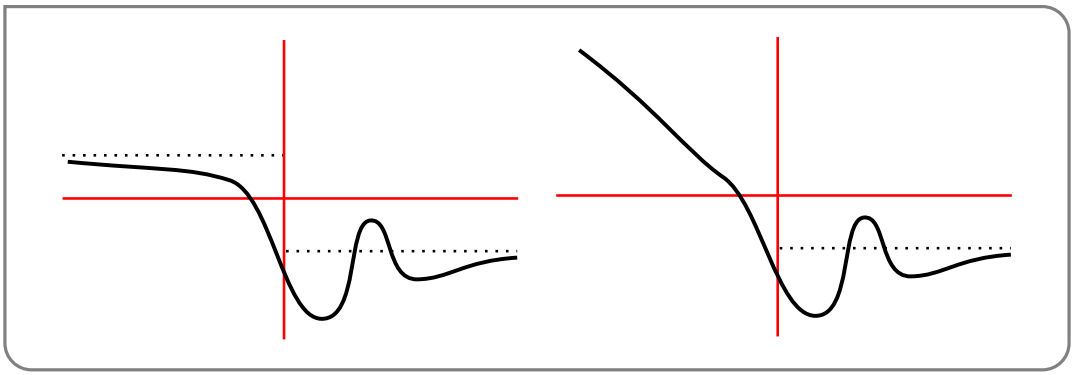
when the value of the function $f(x)$ gets closer and closer to L as we make x larger and larger and positive.
Similarly we write

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

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

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

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

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

exist. Then the following limits hold

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

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

$$\begin{aligned}
\frac{x^2 - 3x + 4}{3x^2 + 8x + 1} &= \frac{x^2\left(1 - \frac{3}{x} + \frac{4}{x^2}\right)}{x^2\left(3 + \frac{8}{x} + \frac{1}{x^2}\right)} \\
&= \frac{1 - \frac{3}{x} + \frac{4}{x^2}}{3 + \frac{8}{x} + \frac{1}{x^2}} && \text{remove the common factors}
\end{aligned}$$

$$\begin{aligned}
\lim_{x\to\infty} \frac{x^2 - 3x + 4}{3x^2 + 8x + 1} &= \lim_{x\to\infty} \frac{1 - \frac{3}{x} + \frac{4}{x^2}}{3 + \frac{8}{x} + \frac{1}{x^2}} \\
&= \frac{\lim_{x\to\infty}\left(1 - \frac{3}{x} + \frac{4}{x^2}\right)}{\lim_{x\to\infty}\left(3 + \frac{8}{x} + \frac{1}{x^2}\right)} && \text{arithmetic of limits} \\
&= \frac{\lim_{x\to\infty} 1 - \lim_{x\to\infty} \frac{3}{x} + \lim_{x\to\infty} \frac{4}{x^2}}{\lim_{x\to\infty} 3 + \lim_{x\to\infty} \frac{8}{x} + \lim_{x\to\infty} \frac{1}{x^2}} && \text{more arithmetic of limits} \\
&= \frac{1 + 0 + 0}{3 + 0 + 0} = \frac{1}{3}
\end{aligned}$$

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

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

Example 1.5.6

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 $f(x)$.

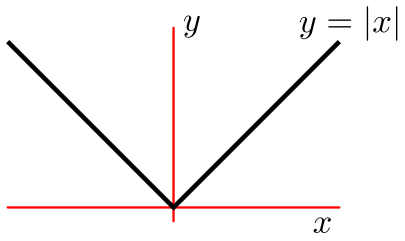
that is the function is returning -1 times the input.

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

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

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

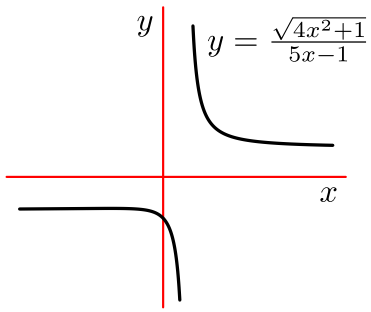
$$\begin{aligned}\lim_{x\to-\infty}\frac{\sqrt{4x^2+1}}{5x-1} &= \lim_{x\to-\infty}\frac{-x\sqrt{4+1/x^2}}{x(5-1/x)} \\ &= \lim_{x\to-\infty}\frac{-\sqrt{4+1/x^2}}{5-1/x} \\ &= -\frac{2}{5}\end{aligned}$$

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.



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

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

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:

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

Their limits as $x \to 0$ are:

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

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,

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

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
- $\lim_{x \to a^-}$ exists and is equal to $f(a)$, and
- $\lim_{x \to a^+}$ 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:

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

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

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

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

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

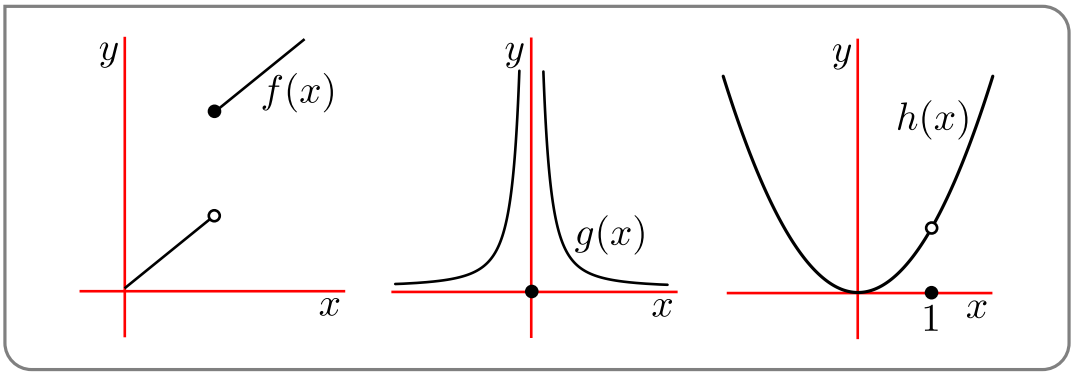
▸▸ Back to the Main Text

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 $\lim_{x\to a} g(x)=b$ then $\lim_{x\to a} f(g(x))=f(b)$. 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?

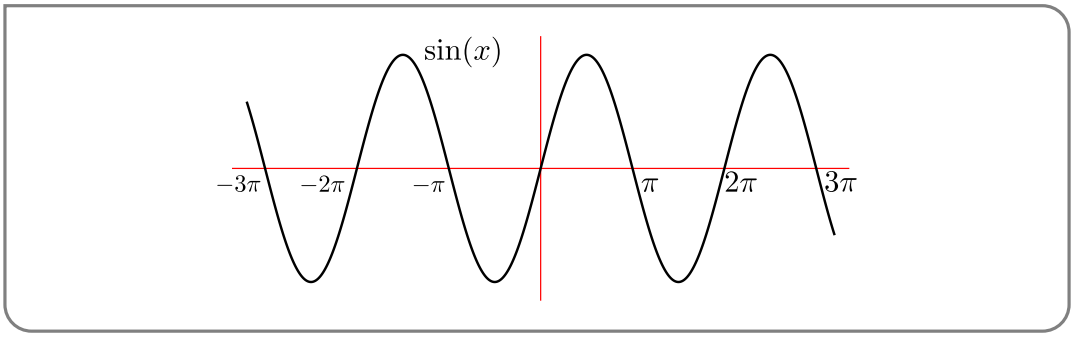
$$f(x)=\sin\left(x^2+\cos(x)\right)$$
$$g(x)=\sqrt{\sin(x)}$$

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.

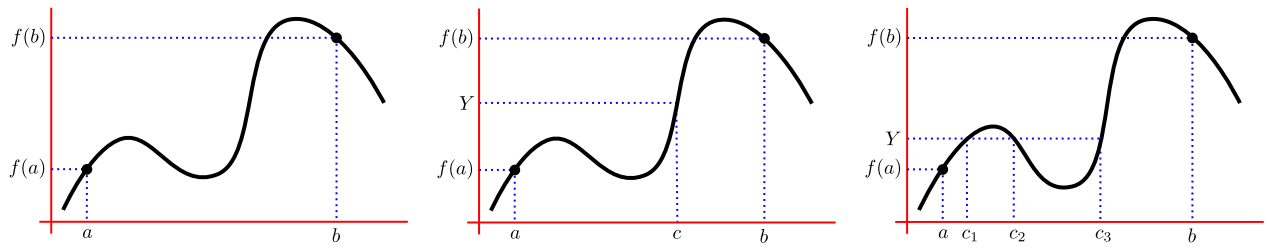
The absence of such jumps leads to the following theorem which, while it can be quite confusing on first glance, actually says something very natural — obvious even. It says, roughly speaking, that, as you draw the graph $y = f(x)$ starting at $x = a$ and ending at $x = b$, y changes continuously from $y = f(a)$ 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

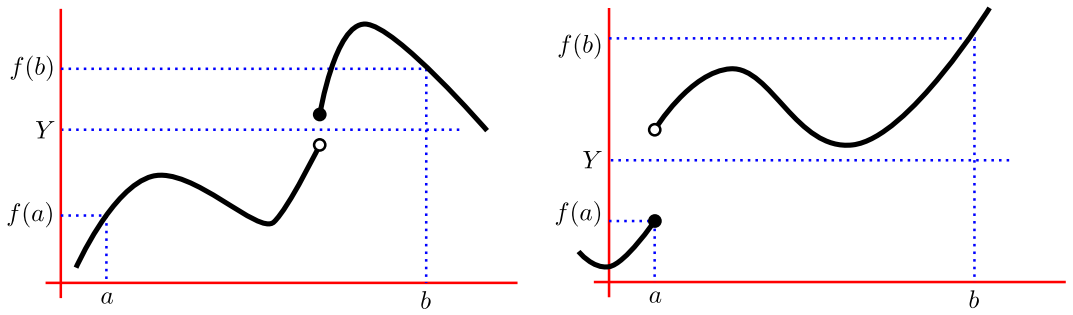
- Now we can pick any Y that lies between $f(a)$ and $f(b)$ — see the middle figure above. The IVT[27] tells us that there must be some x-value that when plugged into the function gives us Y. That is, there is some c between a and b so that $f(c) = Y$. We can also interpret this graphically; the IVT 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 somewhere 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 $a = 10\text{am}$ and $b = 11\text{am}$. 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 applications of the IVT is determining where a function is zero. For quadratics we know (or should know) that

$$ax^2 + bx + c = 0 \qquad \text{when } x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

While the Babylonians could (mostly, but not quite) do the above, the corresponding formula 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.

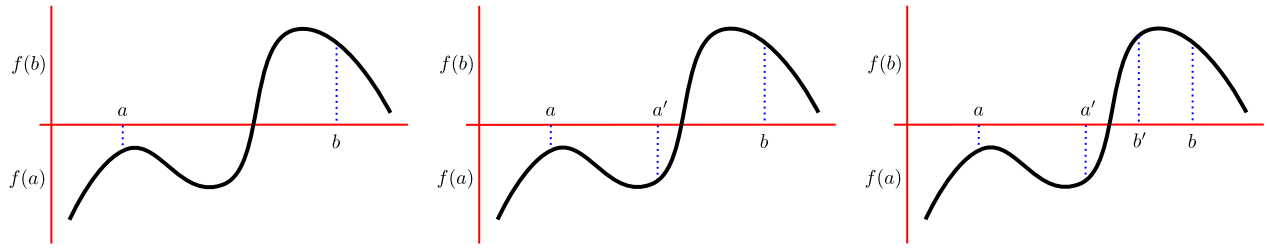
30 It's amazing what facts you can find on Wikipedia.

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 $x = 0.5$

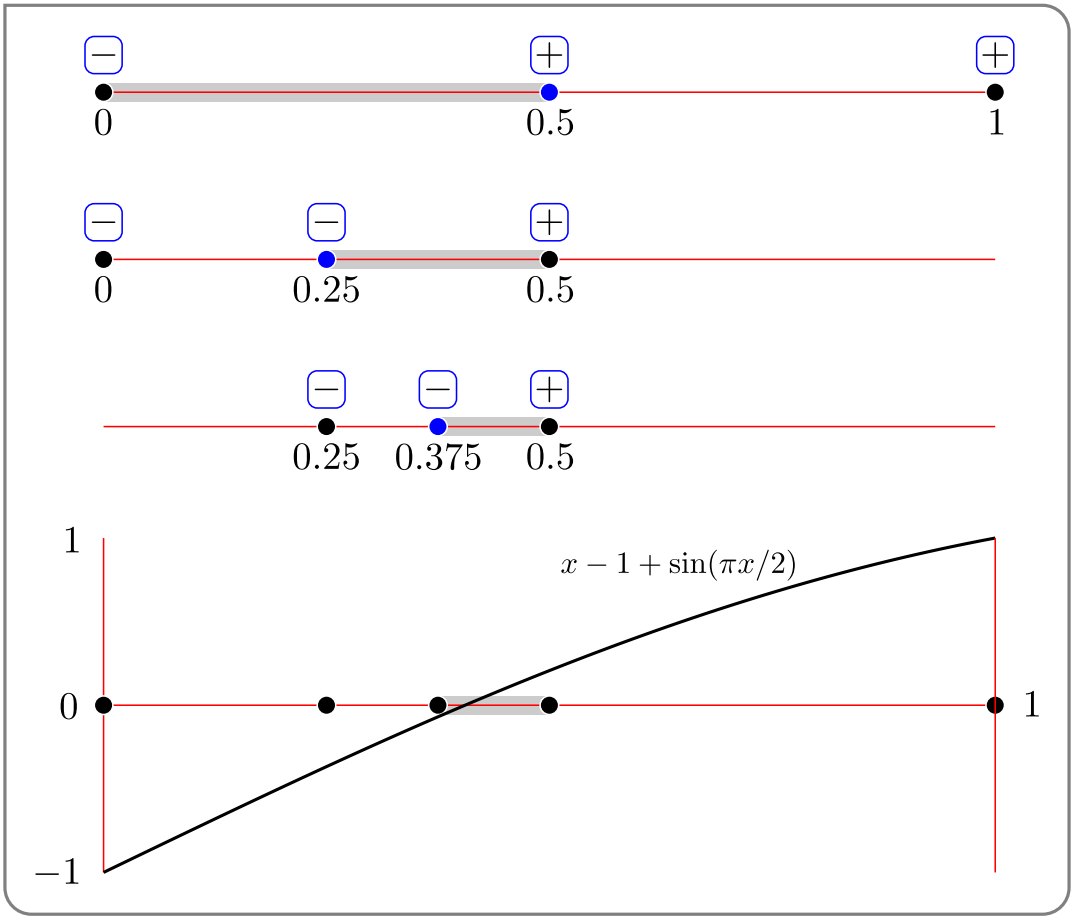
Repeat

- $a = 0.25, b = 0.5$ 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.

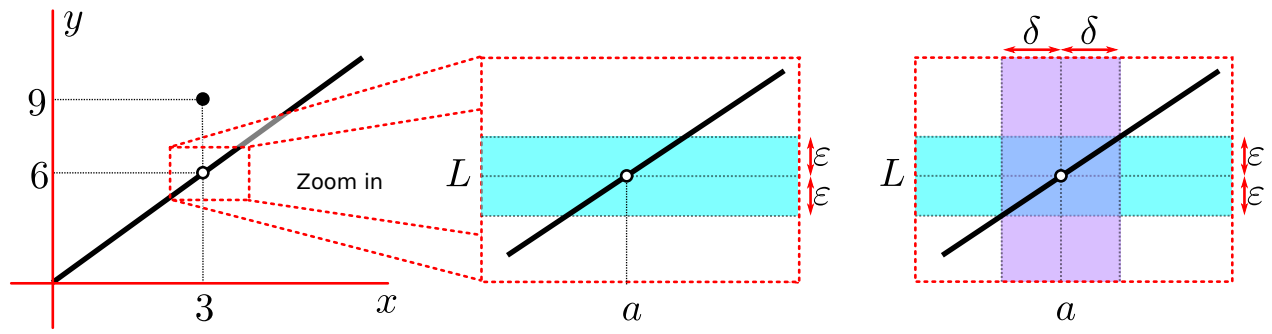
– **if** $0 < |x - a| < \delta$ **then** $|f(x) - L| < \epsilon$ — From the previous point we have our two numbers — any $\epsilon > 0$ then based on that choice of ϵ we have a positive number δ. The current statement says that whenever we have chosen x so that it is very close to a, then $f(x)$ has 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

$$\begin{aligned} 3-\delta < x < 3+\delta \qquad & \text{multiply both sides by 2} \\ 6-2\delta < 2x < 6+2\delta & \end{aligned}$$

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. □

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.

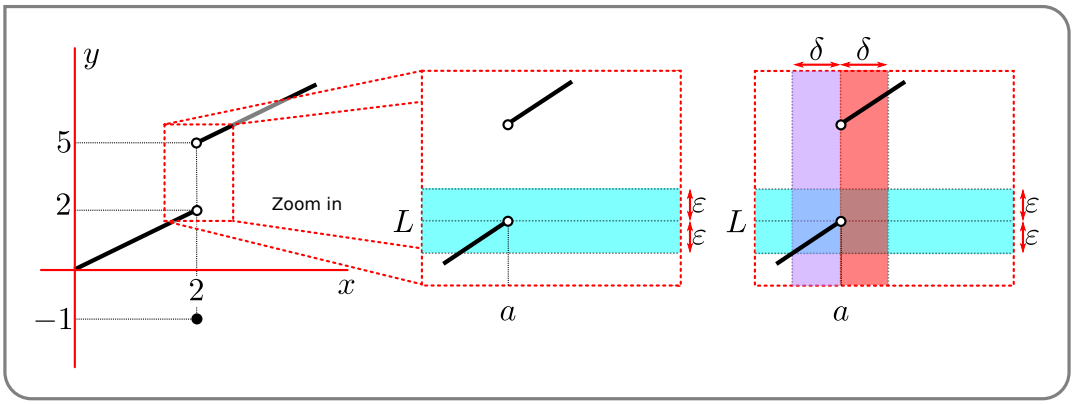
In the above example everything works, but it can be very instructive to see what happens in an example that doesn't work.

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

$$\begin{aligned} a - \delta < x < a + \delta \text{ but } f(x) = x \text{ and } \delta = \epsilon \text{ so} \\ a - \epsilon < f(x) < a + \epsilon \end{aligned}$$

  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: