Calculus

University of York

Autumn term 2022

Contents

Administrative Information

- This module: Calculus (first part, continues in the Spring & Summer)

- Lecturer: Dr Karina Kirkina, karina.kirkina@york.ac.uk, G/N/113.

- Teaching: three one hour lectures each week and in alternating weeks a seminar (small group) or a problem class (large group).

  The lectures will give an overview of the material, while most of the learning will come from doing the problems. The seminars will be more interactive, so I recommend working on the problems before the seminar so that you can ask questions about anything that you do not understand.

- Exercises: released every other week. Try to do all of them, except perhaps those marked as particularly challenging or open-ended. A sample of the questions on each sheet will be compulsory (those marked with stars) and count towards the 5% coursework mark of the module. These questions will be marked and handed back to you during seminars. **Please write your name and your seminar leader's name on your work.**

- Workload: In addition to lectures and seminars you should expect to spend 6–8 hours a week revising lecture notes and working on problems.

- Main course material: these lecture notes, which can be found on Moodle.

- Other sources of information: Any standard university level Calculus textbook such as Thomas's Calculus (any recent edition), Stewart, etc. The website "Paul's Online Math Notes" (`https://tutorial.math.lamar.edu/`) is also a good source of information for Calculus.

- Assessment: 5% homework assignments, 95% final examination in the summer.

1 Introduction, Notation, Terminology

1.1 Notation

In the next section we go over the notation for the number systems we will use. Apart from these, we will also use the symbols shown on the table.

(Note that some authors use the symbol $\subset$ for a subset (not necessarily proper), and some use the symbol $\subsetneq$ for a proper subset, so make sure you check what convention the source you are using is following.)

Symbol	Meaning
$\implies$	"implies"
$\iff$	"if and only if"/"is equivalent to"
$\in$	"is an element of"
$\forall$	"for all"
$\exists$	"there exists"
$\cup$	set union
$\cap$	set intersection
$\setminus$	set difference
$\emptyset$	the empty set
$\subseteq$	subset (which could also be the whole set)
$\subset$	proper subset (i.e. which is *not* equal to the whole set)

Avoid using any other symbols that you might have learned at school or elsewhere unless we use them in this module - doing so can lead to misunderstandings. Agreeing on definitions is crucial in mathematics. Each module you have at university will start by telling you what notation it will use, so try to only use that notation in that module.

1.2 Number Systems

Now we introduce the five number systems we will be using in this module.

$\mathbb{N}$, the set of "natural numbers", i.e. the numbers we use for counting:

$$\mathbb{N} = \{1, 2, 3, \dots\}$$

(Note: some authors also include zero.)

$\mathbb{Z}$, the set of integers, i.e. whole numbers, positive, negative or zero:

$$\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$$

$\mathbb{Q}$, the set of rational numbers, i.e. exact fractions:

$$\mathbb{Q} = \left\{ \frac{p}{q} \mid p \in \mathbb{Z}, q \in \mathbb{Z}, q \neq 0 \right\}$$

$\mathbb{R}$, the set of real numbers. Important: $\mathbb{R}$ does not contain the symbols $-\infty$ or ∞. All real numbers are finite.

$\mathbb{C}$, the set of complex numbers:

$$\mathbb{C} = \{x + \mathrm{i}y \mid x, y \in \mathbb{R}\}$$

where $\mathrm{i}^2 = -1$.

We often use these as convenient abbreviations, for instance "$x \in \mathbb{R}$" means "x is a real number."

The rational numbers have gaps, in the sense that there are physical quantities that cannot be expressed as an exact fraction (such as the diagonal of a square with side length 1). Real numbers that are not rational are called *irrational*. Examples of irrational numbers are $\sqrt{2}$, π, and e.

The real numbers can be viewed as the set of numbers that can be written as decimals. Out of these, the rational numbers are those for which the decimal expansion is eventually repeating, such as $5.3333333... = \frac{16}{3}$, or terminating, such as $0.359372 = 0.3593720000... = \frac{359372}{1000000}$. All irrational numbers have infinite and non-repeating decimal expansions.

Much of this will be discussed in more detail in other modules: sets in *Mathematical Skills 1*, complex numbers in *Algebra*, real numbers in *Real Analysis*.

1.3 Infinity is not a real number and division by zero is "not allowed"

Question: Why do we insist that ∞ is not a real number?

Answer: Because it is not compatible with the usual rules of arithmetic.

Question: Why is division by zero "not allowed"?

Answer: Because it is not compatible with the usual rules of arithmetic.

Naive manipulations with ∞ lead to nonsense!

$$\infty + 1 = \infty \implies 0 = 1$$

(cancelling ∞, additively)

$$2 \cdot \infty = 1 \cdot \infty \implies 2 = 1$$

(cancelling ∞, multiplicatively)

$1/\infty = 0$ and similar expressions are also nonsense:

$$1/\infty = 0 = -0 = -1/\infty \implies \infty = -\infty$$

Equally, if we could make sense of $1/0 = \infty$ then, because $0 = -0$, we would have $\infty = -\infty$.

We want the basic properties of arithmetic to work for all real numbers, so we have to exclude $\pm\infty$.

The upshot of this is: be careful when dealing with infinity, do not write it in calculations as if it were a number.

Concerning zero, we certainly have $1 \cdot 0 = 2 \cdot 0$; if we could divide by zero then we could cancel the zeros to give $1 = 2$.

In every other way, zero works like any other real number. So, we include zero as a real number but with one proviso: we never divide by zero.

1.4 Intervals

An *interval* is a non-trivial subset I of the real line such that if $x, y, z \in \mathbb{R}$ with $x < y < z$ and $x \in I$ and $z \in I$ then $y \in I$. In words:

If I contains x and z then it also contains all points between x and z.

Every interval is of one of the following types (for $a, b \in \mathbb{R}$, $a < b$):

$(a,b) = \{x \in \mathbb{R} : a < x < b\}$	finite, open
$(a,b] = \{x \in \mathbb{R} : a < x \le b\}$	finite, half-open
$[a,b) = \{x \in \mathbb{R} : a \le x < b\}$	finite, half-open
$[a,b] = \{x \in \mathbb{R} : a \le x \le b\}$	finite, closed, compact
$(a,\infty) = \{x \in \mathbb{R} : x > a\}$	infinite, open
$[a,\infty) = \{x \in \mathbb{R} : x \ge a\}$	infinite, closed
$(-\infty,b) = \{x \in \mathbb{R} : x < b\}$	infinite, open
$(-\infty,b] = \{x \in \mathbb{R} : x \le b\}$	infinite, closed
$(-\infty,\infty) = \mathbb{R}$	infinite

The term “non-trivial” here is intended to exclude two degenerate special cases: single points and the empty set. We could express it as “containing at least two different points.” The degenerate cases can actually both be realised using the notation above (e.g. $\{a\} = [a,a]$ and $\emptyset = (a,a)$) but, following Thomas and Stewart, we do not call these sets intervals. Other authors might use a different convention.

The notation is consistent: a square bracket [] means the endpoint is included, a round bracket () means that it is excluded. The endpoints $\pm\infty$ are never included in an infinite interval, because $\pm\infty$ are not real numbers.

The terms *open* and *closed* mean “excluding endpoints” and “including endpoints”; again, note that $\pm\infty$ are excluded even from closed half-lines, because $\pm\infty$ are not real numbers.

The term *compact* is a special case of a more general idea (explored in the *Metric Spaces* module). For our purposes a compact interval is a closed, finite interval.

The terms *bounded* and *unbounded* are sometimes used in place of finite and infinite.

Proving that all intervals have this form is non-trivial and requires a clear understanding of how the real number system works, in particular the ideas of supremum and infimum (covered in the *Real Analysis* module).

1.5 Functions and Formulae

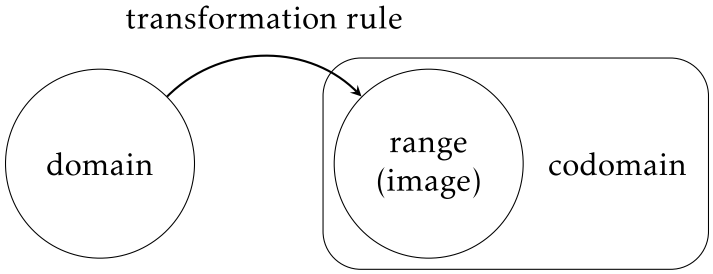
Functions encapsulate mathematical transformations. We put some kind of mathematical object into the function and another one comes out. For example, there is a function called sin; if we put $\pi/2$ in, 1 comes out. We write this as $\sin(\pi/2) = 1$.

A formula describes a sequence of operations or rules that are to be applied to some mathematical object; for example, x^2 is a formula.

Formulae and functions are closely related, but are not the same thing.

To turn a formula into a function, we have to add two extra pieces of information. The first thing we have to do is to specify what the variable (x, in the x^2 example) is allowed to be. Are we thinking about squaring integers, real numbers, complex numbers, matrices? We do this by specifying a set called the *domain* of the function, which contains every object we might want to feed into the function. If we want to think about squaring real numbers, we set the domain be $\mathbb{R}$. If we want to think about squaring 3×3 real matrices, we set the domain to be the set of all real 3×3 matrices, $M_3(\mathbb{R})$.

The second piece of a function is called the *codomain*. It specifies what kind of objects come out of the function. Unlike the domain, it is not an exact description of every value that actually comes out of the function; rather, it is a container large enough to hold every possible value. For example, if our formula is x^2 and our domain is $\mathbb{R}$, we could let the codomain be $\mathbb{R}$, or $[0, \infty)$, or $\mathbb{C}$ - but we could not let it be $\mathbb{Z}$, as squaring a real number does not necessarily give an integer. These three pieces of information together give us a true function, as illustrated below.



The *range* or *image* of the function is the exact set of values that it takes: the set of all $f(x)$ as x varies through the domain of f. We can write

$$f : \mathbb{R} \to \mathbb{R}, \quad f(x) = x^2$$

or

$$f : \mathbb{R} \to \mathbb{R}, \quad f : x \mapsto x^2$$

to denote the function mapping $\mathbb{R}$ to $\mathbb{R}$ by squaring. We could also write

$$f : \mathbb{R} \to [0, \infty), \quad f(x) = x^2$$

which is a slightly different function. It has the same and domain and range, but a different codomain. We could also define

$$f : \mathbb{C} \to \mathbb{C}, \quad f(z) = z^2$$

which is a very different function! Once we have created a function, the letters we use do not matter:

$$\begin{aligned} f : \mathbb{R} \to \mathbb{R}, &\quad f : x \mapsto x^2 \\ f : \mathbb{R} \to \mathbb{R}, &\quad f : y \mapsto y^2 \end{aligned}$$

are exactly the same. A formula like x^2 is about squaring something called x. The function we have defined is about squaring a real number.

(As an aside, while the letters we use do not technically matter, mathematicians prefer to use certain letters for certain objects, for instance x and y for real numbers, z for complex numbers, n and m for natural numbers, etc.; to avoid confusion it is best to stick to these conventions.)

Functions also go beyond formulae: all we need to define a function is an unambiguous rule which, given an element of the domain, describes **exactly one** element of the codomain. This does not have to be a simple algebraic formula; for example

$$|x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0 \end{cases}$$

with domain and codomain $\mathbb{R}$ describes the real absolute value function. Similarly,

$$H(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

with domain and codomain $\mathbb{R}$ is perfectly valid; this function is known as the Heaviside function.

1.6 Maximal Domains

To get from a formula to a function, we need to specify a domain and a codomain. In Calculus, we often do this in the simplest way we can, as follows:

- Let the codomain be $\mathbb{R}$ (or if we are working with complex numbers, $\mathbb{C}$).

- Let the domain be the set of all real numbers for which the formula makes sense. This is called the *maximal domain* or *natural domain* (within $\mathbb{R}$) associated with the formula.

To find the maximal domain, look out for such things as:

- division by zero,
- square roots of negative numbers,
- logarithms of non-positive numbers.

Some examples of maximal domains (within $\mathbb{R}$):

$$\begin{array}{rc}
f_1(x) = x^2 + 1 & \mathbb{R} \\
f_2(x) = \dfrac{1}{x^2+1} & \mathbb{R} \\
f_3(y) = \dfrac{1}{y} & \mathbb{R} \setminus \{0\} = (-\infty, 0) \cup (0, \infty) \\
f_4(u) = \dfrac{u}{u^2-1} & \mathbb{R} \setminus \{-1, 1\} = (-\infty, -1) \cup (-1, 1) \cup (1, \infty) \\
f_5(v) = \log(1+v) & (-1, \infty) \\
f_6(w) = \sqrt{1-w} & (-\infty, 1]
\end{array}$$

Notes:

- There are many ways to describe sets; in examples like this, a unions of intervals is often a good way.
- The variables used in the formulae are arbitrary, and do not form part of the answer. The domain of f_1 is the set $\mathbb{R}$. It is not an expression like "$x \in \mathbb{R}$" because x is part of the formula we used to describe f_1, not part of the function itself. You could also use a variable to describe the maximal domain, e.g. for f_3 you could write

  $$\{y \in \mathbb{R} \mid y \neq 0\}$$

  There is no fundamental reason to use y here, it is simply a placeholder variable, and it would be equally correct to write this set as $\{x \in \mathbb{R} \mid x \neq 0\}$ (but changing variables like this is not good style since it could lead to confusion).
- If we were working with complex numbers, the maximal domain of f_2 (within $\mathbb{C}$) would be $\mathbb{C} \setminus \{\mathrm{i}, -\mathrm{i}\}$

- The logarithm in f_5 with be introduced properly later: it means the real-valued natural logarithm defined on $(0, \infty)$.

- We shall also take a closer look at square roots later: the square root in f_6 is the non-negative square root function defined on $[0, \infty)$.

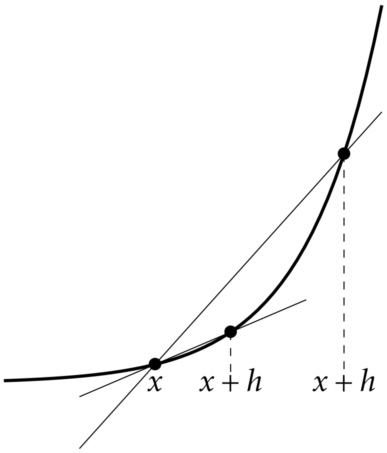
2 Limits

2.1 Difference Quotients

Suppose $f : \mathbb{R} \to \mathbb{R}$, $x, y \in \mathbb{R}$ with $x \neq y$ and let $h = y - x$ (so $h \neq 0$). The ratio

$$\frac{f(y) - f(x)}{y - x} = \frac{f(x+h) - f(x)}{h}$$

is called a *difference quotient*. Geometrically, this is the gradient of the chord joining the points $(x, f(x))$ and $(y, f(y))$.



Fix x and move y closer and closer to x (so h becomes closer and closer to 0); the gradient of the chord becomes closer and closer to the gradient of the tangent line to f at x.

Here are some concrete examples.

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

The difference quotient is

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

As h becomes smaller, $2x+h$ becomes closer to $2x$; we want to conclude that the gradient of the tangent line to f at x is $2x$, but *substituting $h = 0$ makes little sense because the difference quotient, where we started, is not defined when $h = 0$.*

Now try

$$f(x) = x^3$$

The difference quotient is

$$\frac{f(x+h) - f(x)}{h} = \frac{(x+h)^3 - x^3}{h} = 3x^2 + 3xh + h^2$$

Again, *substituting $h = 0$ makes little sense because the difference quotient is not defined when $h = 0$.* We want to argue that if h is small, then $3xh$ is also small (but how small? x could be large) and h^2 is small, so their sum $3xh + h^2$ is also small and hence the difference quotient is close to $3x^2$; we then want to conclude that the gradient of the tangent line to f at x is given by $3x^2$.

Getting arguments like this right requires the idea of a *limit*. This will be thoroughly explored in the *Real Analysis* module, but for now here are the ideas.

2.2 Limits at a Point

Suppose $D \subseteq \mathbb{R}$, $f : D \to \mathbb{R}$, $x_0 \in \mathbb{R}$ and $L \in \mathbb{R}$.

> We say that $f(x)$ *tends to a limit* L as x tends to x_0 if, when $x \in D$ is very close to x_0 BUT NOT EQUAL TO x_0, $f(x)$ is very close to L.

Another way of saying this is that we can get $f(x)$ *arbitrarily* close to L (i.e. as close as we want) by picking an x *sufficiently* close to x_0.

We write this as:

$$f(x) \to L \text{ as } x \to x_0$$

or

$$\lim_{x \to x_0} f(x) = L$$

For example, let $D = \mathbb{R} \setminus \{-1, 1\}$ and $f(x) = (x-1)/(x^2-1) = 1/(x+1)$ (where we were able to cancel by $x-1$ since we assumed that $x = 1$ is not in the domain of f). When x is close to, but not equal to, 1, $f(x) = 1/(x+1)$ is close to $1/2$, so $f(x) \to 1/2$ as $x \to 1$. We cannot evaluate $f(1)$ because the function as given is not defined there, but limits are calculated without using the value of the function at the limit point.

In order for this idea to make sense, there must exist elements of D very close to x_0 but not equal to x_0. Such a point x_0 is called a limit point of D; the exact requirement is that, for every $r > 0$, $((x_0 - r, x_0) \cup (x_0, x_0 + r)) \cap D \neq \emptyset$ but don't worry too much about that — in all our examples, it should work without problems.

Remark: The above definition is not formal. In order to make it precise we would need to be more specific about what we mean by "very close". This is done in detail in *Real Analysis*. For those curious, the formal definition of a limit is

$$\forall\, \varepsilon > 0 \quad \exists\, \delta > 0 \quad \text{such that} \quad |x - x_0| < \delta \implies |f(x) - L| < \varepsilon.$$

Definitions of this form are called "ε-δ (epsilon-delta) definitions". Having a precise definition is essential for being able to prove results about limits. But for the purposes of this module the informal definition will be enough.

To understand the concept of a limit, it helps to consider functions that fail to have a limit at a specific point. Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = \begin{cases} 1 & \text{if } x > 0 \\ x^2 & \text{if } x \leq 0 \end{cases}$$

This function does not have a limit at $x = 0$. This happens because f has a discontinuity there, so, intuitively, near $x = 0$ there are points whose x-values are very close together but whose $f(x)$-values are far apart.

Another example of a function without a limit at $x = 0$ is $g : \mathbb{R} \to \mathbb{R}$, $g(x) = \sin(\frac{1}{x})$. As x approaches 0 this function oscillates between -1 and 1 infinitely often, so there are points near $x = 0$ whose x-values are close together but whose $f(x)$-values are far apart.

2.3 The Algebra of Limits

The material in this section is presented without much consideration of *why* these procedures work. This will be covered in the *Real Analysis* module.

We calculate limits by using a small number of basic examples and a collection of rules that allow us to build up more complicated examples. We start off with some basic ones:

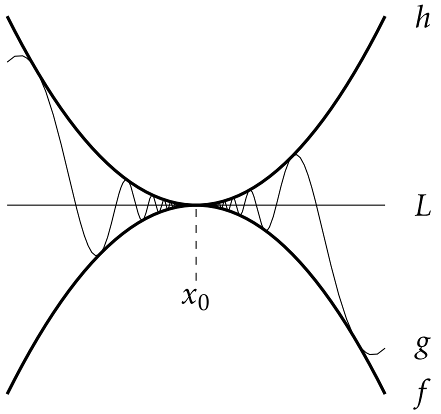
AL1: (localisation principle) For any fixed $r > 0$, $\lim_{x \to x_0} f(x)$ is unaffected by the values, or even existence, of $f(x)$ where $|x_0 - x| > r$.

AL2: If $f(x) \to L$ then, for any constant C, $Cf(x) \to CL$ and $C + f(x) \to C + L$ (all as $x \to x_0$).

AL3: (sandwich theorem) If $f(x) \leq g(x) \leq h(x)$, $f(x) \to L$ and $h(x) \to L$ then $g(x) \to L$ (all as $x \to x_0$).

AL1 says that the limit of a function at a point does not depend on the values of the function away from the point – we could replace the function $f(x)$ with a function $g(x)$ which had the same values near x_0, but different values further away from x_0, and it would have the same limit as $f(x)$ at x_0.

Here is an illustration of AL3, the sandwich theorem: as x approaches x_0, $f(x)$ and $h(x)$ both approach L. Because $f(x) \leq g(x) \leq h(x)$, $g(x)$ must also approach L.



A useful consequence of AL2 and AL3 is:

> If $g(x) \to 0$ as $x \to x_0$ and $|f(x) - L| \le g(x)$ then $f(x) \to L$ as $x \to x_0$.

This is because

$$\begin{aligned} |f(x) - L| \le g(x) &\iff -g(x) \le f(x) - L \le g(x) \\ &\iff L - g(x) \le f(x) \le L + g(x) \end{aligned}$$

and, as $x \to x_0$, $g(x) \to 0$ so $L \mp g(x) \to L$.

We can now look back to our very first example. While we were trying to find the limit of the difference quotient of x^2, we had to conclude that $2x + h \to 2x$ as $h \to 0$; this is due to AL2.

Let us also return to the example of $f(x) = (x-1)/(x^2-1)$ as $x \to 1$. When $x \ne 1$ this is equal to $1/(x+1)$. We think the limit here is $\frac{1}{2}$, so to justify this start off by writing

$$\left|\frac{1}{x+1} - \frac{1}{2}\right| = \frac{|1-x|}{2|1+x|}$$

Now, we can use the localisation principle AL1 to ignore all values of x that are some distance away from 1, say such that $|x-1| > 1$; this leads to $|x-1| \le 1$, so $-1 \le x-1 \le 1$, so $1 \le 1+x \le 3$, and so $|1+x| \ge 1$, which gives us

$$\left|\frac{1}{x+1} - \frac{1}{2}\right| \le \frac{|1-x|}{2} \to 0$$

by AL2, so using AL3 we get $1/(1+x) \to 1/2$ as $x \to 1$.

We can also formulate some properties involving two functions.

Suppose $f, g : D \to \mathbb{R}$ and, as $x \to x_0$, $f(x) \to L$ and $g(x) \to M$ where $L, M \in \mathbb{R}$. Then, as $x \to x_0$,

AL4: $f(x) \pm g(x) \to L \pm M$

AL5: $f(x)g(x) \to LM$

AL6: $f(x)/g(x) \to L/M$ provided $M \neq 0$

When we were trying to find the limit of the difference quotient of x^3, we had to show that $3x^2 + 3xh + h^2 \to 3x^2$ as $h \to 0$. This follows from AL4 and AL5.

Finally, one more sophisticated one:

AL7: (composition property) If $f(x) \to y_0$ as $x \to x_0$ and $g(y) \to z_0$ as $y \to y_0$ then $(g \circ f)(x) = g(f(x)) \to z_0$ as $x \to x_0$.

2.4 A Non-obvious Example

Question Why do we need to be so careful about limits?

Answer Because they can be very non-obvious!

$$\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n = ?$$

Intuition 1 When n is large, $1 + 1/n$ is close to 1, so $(1 + 1/n)^n$ is close to $1^n = 1$; perhaps the limit is 1?

Intuition 2 However large n is, $1+1/n$ is greater than 1. A large power of anything greater than 1 is very large; perhaps the limit is ∞?

Trying out some values shows that both intuitive ideas are wrong:

n	$(1+1/n)^n$	n	$(1+1/n)^n$
1	2.000000	100	2.704814
2	2.250000	1000	2.716924
3	2.370370	10000	2.718146
4	2.441406		
5	2.488320	limit	e
6	2.521626		
7	2.546500		
8	2.565785		
9	2.581175		
10	2.593742		
11	2.604199		
12	2.613035		
13	2.620601		
14	2.627152		
15	2.632879		

This, of course, does not constitute a proof! We shall see later that

$$\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = \mathrm{e}^x$$

for any x.

What goes wrong in both of the intuitive ideas above is that we treated the n's differently - we let one of them tend to infinity while keeping the other fixed. But the n's are identical, so they must either both be treated as finite, or both tend to infinity at the same time.

2.5 Limits of Polynomials and Rational Functions

Here is a fundamental fact about polynomials:

Suppose P is a polynomial and $x_0 \in \mathbb{R}$. Then $P(x) \to P(x_0)$ as $x \to x_0$.

More generally, we can extend this to *rational functions* (ratios of polynomials):

If P and Q are polynomials and $Q(x_0) \neq 0$ then $P(x)/Q(x) \to P(x_0)/Q(x_0)$ as $x \to x_0$.

This follows from the Algebra of Limits: for example, if $P(x) = ax^2 + bx + c$ then we can write

$$\begin{aligned} P(x) &= a.x.x + b.x + c \\ &\to a.x_0.x_0 + b.x_0 + c \text{ as } x \to x_0 \text{ (AL4, AL5)} \\ &= a.x_0^2 + bx_0 + c \\ &= P(x_0) \end{aligned}$$

Polynomials of higher degree work in much the same way. The extension to rational functions follows from AL6.

As a first example, let us find

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

This is the simplest kind of example. The denominator is non-zero at the point 2 (where we are taking the limit), so we can simply calculate

$$\frac{x^2+x+1}{x^2-1} \to \frac{2^2+2+1}{2^2-1} = \frac{7}{3}$$

using the Algebra of Limits. We can also write this as

$$\lim_{x\to 2} \frac{x^2+x+1}{x^2-1} = \frac{7}{3}$$

For a second example, we find

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

This is more complicated, because the numerator and denominator are both zero at the point 1, where we are taking the limit. To proceed, we use the fact that $x-1$ must be a factor of both numerator and denominator. The denominator is easy:

$$x^2-1 = (x-1)(x+1)$$

We also have

$$x^3-x^2+x-1 = (x-1)(x^2+1)$$

so we can write

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

Now, we use the fact that the limit at a point does not depend on the function value at that point to conclude that

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

and we can evaluate the right-hand limit using the Algebra of Limits to be $(1+1)/(1^2+1) = 1$. We thus have

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

2.6 Application: Equating Coefficients and Finding Partial Fractions

Suppose two polynomials are equal as functions, e.g.

$$Ax^2 + Bx + C = ax^2 + bx + c \qquad (x \in \mathbb{R})$$

Why can we "equate coefficients" to give $A = a$, $B = b$, $C = c$?

More generally, if

$$\sum_{n=0}^{N} A_n x^n = \sum_{n=0}^{N} a_n x^n$$

for $x \in \mathbb{R}$, why can we conclude that $A_n = a_n$ for all n?

In the quadratic example:

- Substitute $x = 0$ to give $C = c$. Cancel C against c, to give

  $$Ax^2 + Bx = ax^2 + bx$$

  then divide by x to give

  $$Ax + B = ax + b \qquad (x \neq 0)$$

- Notice we cannot just substitute $x = 0$, because we divided by x.
- Instead, take a limit as $x \to 0$ to give $B = b$, cancel B against b and divide by $x \neq 0$ to give $A = a$.

The same procedure works for polynomials of any degree.

Suppose we want to make a partial fraction decomposition, e.g. find A and B such that

$$\frac{1}{x^2 - 1} = \frac{A}{x-1} + \frac{B}{x+1} \qquad (x \neq \pm 1)$$

Note the exclusion: neither side of the equation makes any sense if $x = \pm 1$. We can multiply through by $x^2 - 1 = (x-1)(x+1)$ to give

$$1 = A(x+1) + B(x-1) \qquad (x \neq \pm 1)$$

but we cannot meaningfully substitute $x = 1$ or $x = -1$ in this formula. However, we can take a limit as $x \to 1$ to give $1 = 2A$ and a limit as $x \to -1$ to give $1 = -2B$ and conclude that $A = 1/2$ and $B = -1/2$.

This leads us to the correct partial fraction decomposition

$$\frac{1}{x^2 - 1} = \frac{1}{2(x+1)} - \frac{1}{2(x-1)} \qquad (x \neq \pm 1)$$

2.7 Finite Limits at Infinity

We can also consider limits at ±∞, by which we mean values approached by $f(x)$ when x is very large. Suppose $D \subseteq \mathbb{R}$, $f : D \to \mathbb{R}$ and $L \in \mathbb{R}$. We say that:

> $f(x)$ tends to a limit L as x tends to ∞ if, when $x \in D$ is very large and positive, $f(x)$ is very close to L.

and that

> $f(x)$ tends to a limit L as x tends to $-\infty$ if, when $x \in D$ is very large and negative, $f(x)$ is very close to L.

We write this as

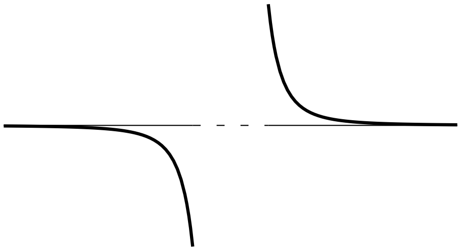
$$f(x) \to L \text{ as } x \to \infty \qquad \text{or} \qquad \lim_{x\to\infty} f(x) = L$$

and similarly for the negative case.

The algebra of limits works in exactly the same way for limits at ±∞ as at finite points except that we replace AL1 by

> **AL1′:** (localisation principle) For any fixed $r > 0$, $\lim_{x\to\infty} f(x)$ is unaffected by the values of $f(x)$ where $x < r$ and $\lim_{x\to-\infty} f(x)$ is unaffected by the values of $f(x)$ where $x > -r$.

Fundamental example (proved in Real Analysis): if $n \in \mathbb{N}$ then $x^{-n} \to 0$ as $x \to \pm\infty$.



Recall (2.5) that a *rational function* is the ratio of two polynomials. We can compute limits of rational functions at ±∞ as follows: given a rational function

$$\frac{x^2 + 3x - 1}{2x^2 + 3}$$

we divide the numerator and denominator by the highest power of x present

$$\frac{x^2+3x-1}{2x^2+3} = \frac{1+3/x-1/x^2}{2+3/x^2} \qquad (x \neq 0)$$

Now, as $x \to \infty$, $1/x \to 0$ and $1/x^2 \to 0$. We can use the algebra of limits to conclude that

$$\frac{x^2+3x-1}{2x^2+3} = \frac{1+3/x-1/x^2}{2+3/x^2} \to \frac{1+0-0}{2+0} = \frac{1}{2}$$

as $x \to 0$. The exclusion $x \neq 0$ is harmless because of localisation (AL1′): the limit at ∞ is not affected by the function's behaviour at the origin. The same technique works for any rational function, e.g.

$$\frac{x}{1+x^2} = \frac{1/x}{1/x^2+1} \to \frac{0}{1+0} = 0$$

as $x \to \infty$.

2.8 Sequences and Series

Sometimes, we have functions that are only, or most conveniently, defined on integers. Our earlier example (2.4) on $(1+1/n)^n$ converging to e as $n \to \infty$ is like this: it can interpreted with $n \in \mathbb{R}$ but is most easily thought of with $n \in \mathbb{N}$.

A function defined on $\mathbb{N}$ is called a *sequence*. We often use subscript notation a_n instead of function notation $a(n)$ for the values of a sequence, and think of it as an infinite list,

$$a_1, a_2, a_3, \ldots$$

Sequences can also be defined on $\mathbb{N} \cup \{0\}$, or with other starting points, e.g.

$$a_0, a_1, a_2, \ldots$$

or

$$a_2, a_3, a_4, \ldots$$

or run towards $-\infty$ instead of $+\infty$:

$$a_0, a_{-1}, a_{-2}, \ldots$$

or

$$a_{-1}, a_{-2}, a_{-3}, \ldots$$

etc.

Recall the definition of the limit of a function at ∞: if $D \subseteq \mathbb{R}$, $f : D \to \mathbb{R}$ and $L \in \mathbb{R}$ then $f(x) \to L$ as $x \to \infty$ if, when $x \in D$ is very large and positive, $f(x)$ is very close to L.

The limit of a sequence is defined by applying this with $D = \mathbb{N}$. In subscript notation:

a_n tends to a limit a as n tends to ∞ if, when $n \in \mathbb{N}$ is very large, a_n is very close to a.

We write this as

$$a_n \to a \text{ as } n \to \infty \qquad \text{or} \qquad \lim_{n\to\infty} a_n = a$$

The case $n \to -\infty$ is a minor and obvious modification.

A *series* is obtained by adding up the terms of a sequence. If $a_n \in \mathbb{R}$ for e.g. $n \in \mathbb{N} \cup \{0\}$, then we consider *partial sums*

$$\sum_{n=0}^{N} a_n = a_0 + a_1 + \cdots + a_N$$

and define the sum of the series by

$$\sum_{n=0}^{\infty} a_n = \lim_{N\to\infty} \sum_{n=0}^{N} a_n$$

We think of this as the sum of all the numbers a_n for $n \in \mathbb{N} \cup \{0\}$, but it can in fact be a bit more complicated than that (rearranging the terms in a series can sometimes change the sum (!) - the Real Analysis module goes into this is more detail, or look up the Riemann Rearrangement Theorem).

Like sequences, series can start at any value of n and sum towards $-\infty$ as well as $+\infty$. Most of our examples later in the course will be Taylor series, which usually start at $n = 0$.

Fundamental example:

If $|x| < 1$ then

$$x^n \to 0 \text{ as } n \to \infty$$

and (geometric sum / series)

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$$

This formula for the infinite geometric sum follows from the finite geometric sum formula, the algebra of limits and the fact that $x^n \to 0$:

$$\sum_{n=0}^{N} x^n = \frac{x^{N+1} - 1}{x-1} \to \frac{-1}{x-1}$$

as $N \to \infty$.

You probably know the finite geometric sum formula, but it will also be explained in Mathematical Skills.

The fact that $x^n \to 0$ as $n \to \infty$ if $|x| < 1$ will be discussed carefully in the Real Analysis module.

2.9 Infinite Limits at a Point

Suppose $D \subseteq \mathbb{R}$, $f : D \to \mathbb{R}$, $x_0 \in \mathbb{R}$. Recall (2.2) the idea of convergence to a limit L:

$f(x) \to L$ as $x \to x_0$ if, when $x \in D$ is very close to x_0 BUT NOT EQUAL TO x_0, $f(x)$ is very close to L.

We can adapt this to "tending to ∞" by replacing "very close to L" with "very large": we say that

> $f(x)$ tends to ∞ as x tends to x_0 if, when $x \in D$ is very close to x_0 BUT NOT EQUAL TO x_0, $f(x)$ is very large and positive.

We write

$$f(x) \to \infty \text{ as } x \to x_0$$

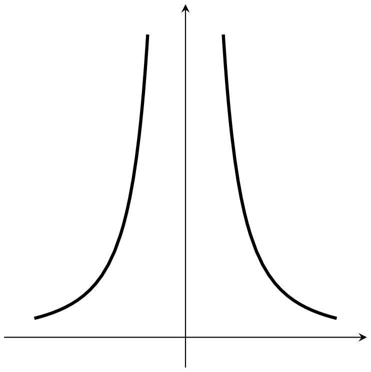
or

$$\lim_{x \to x_0} f(x) = \infty$$

(but remember that $\infty \notin \mathbb{R}$)

Replace "large and positive" with "large and negative" to give the definition of tending to $-\infty$ at x_0.

Fundamental example: if $n \in \mathbb{N}$ then $1/x^{2n} \to \infty$ as $x \to 0$.



Graph of $f(x) = 1/x^2$ near the origin.

2.10 Algebra of Infinite Limits

The algebra of limits does not work for infinite limits! This is for the same reasons that $\pm\infty$ are not real numbers. There is a weaker substitute:

AIL1: (localisation principle) For any fixed $r > 0$, $\lim_{x\to x_0} f(x)$ is unaffected by the values of $f(x)$ where $|x_0 - x| > r$ (exactly the same as AL1)

AIL2: if $f(x) \to \pm\infty$ as $x \to x_0$ and $C \in \mathbb{R}$ with $C \neq 0$ then $Cf(x) \to \pm\infty$ as $x \to x_0$ (the sign on the limit of $Cf(x)$ is determined in the obvious way)

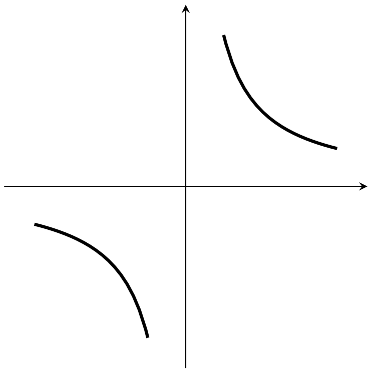
AIL3: if $f(x) \to \infty$ as $x \to x_0$ and $C \in \mathbb{R}$ is such that $g(x) \geq C$ for all x then $f(x)+g(x) \to \infty$ as $x \to x_0$ (modify for $\to -\infty$ in the obvious way).

AIL4: if $f(x) \to \pm\infty$ as $x \to x_0$ then $1/f(x) \to 0$ as $x \to x_0$. If $g(x) \to 0$ as $x \to x_0$ and $g(x) \neq 0$ for x near x_0 then $1/|g(x)| \to \infty$ as $x \to x_0$.

AIL4 is the grain of truth in the "$1/\infty = 0$; $1/0 = \infty$" myth. In essence, it says that if y is large then $1/y$ is small, and if z is small and non-zero then $1/z$ is large in magnitude but of unknown sign (that is why it says $|g(x)|$, not $g(x)$.

2.11 One-Sided Infinite Limits

As noted above (2.9) if $n \in \mathbb{N}$ then $1/x^{2n} \to \infty$ as $x \to 0$. But what about $1/x^{2n-1}$, in particular the simplest example of them all, $1/x$?



Near $x = 0$, $1/x$ does not tend to $+\infty$ or $-\infty$. It approaches both values, depending on the sign of x.

This leads us to the idea of *one-sided limits*.

Suppose $D \subseteq \mathbb{R}$, $f : D \to \mathbb{R}$ and $x_0 \in \mathbb{R}$. Recall (2.9) the idea of tending to ∞ at x_0:

We say that $f(x) \to \infty$ as $x \to x_0$ if, when $x \in D$ is very close to x_0 BUT NOT EQUAL TO x_0, $f(x)$ is very large and positive. We can modify this to describe a left-hand limit: we say that

$f(x)$ tends to ∞ as x tends to x_0 from the left if, when $x \in D$ is very close to x_0 BUT STRICTLY LESS THAN x_0, $f(x)$ is very large and positive.

Notation: "$x \to x_0-$" means "x tends to x_0 from the left." We write

$$\lim_{x \to x_0-} f(x) = \infty; \qquad f(x) \to \infty \text{ as } x \to x_0-$$

Replace "large and positive" with "large and negative" to describe a limit of $-\infty$ instead of $+\infty$.

Of course, the same can be done for right-hand limits: we say that

$f(x)$ tends to ∞ as x tends to x_0 from the right if, when $x \in D$ is very close to x_0 BUT STRICTLY GREATER THAN x_0, $f(x)$ is very large and positive.

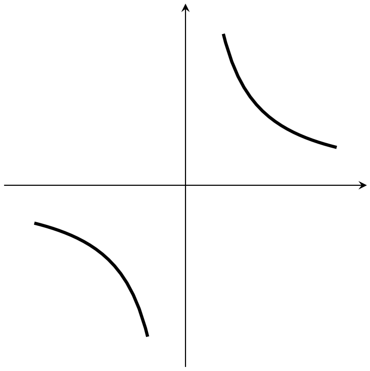
Unsurprisingly, $x \to x_0+$ means "x tends to x_0 from the right." We write

$$\lim_{x \to x_0+} f(x) = \infty; \qquad f(x) \to \infty \text{ as } x \to x_0+$$

Replace "large and positive" with "large and negative" to describe a limit of $-\infty$ instead of $+\infty$.

Fundamental example: if $n \in \mathbb{N}$ then

$$\frac{1}{x^{2n-1}} \to \begin{cases} -\infty & \text{as } x \to 0- \\ +\infty & \text{as } x \to 0+ \end{cases}$$



The algebra of infinite limits (such as it is) works in exactly the same way for one-sided limits as two-sided limits.

2.12 One-Sided Finite Limits

We can also consider finite left and right limits at a point. Suppose $D \subseteq \mathbb{R}$, $f : D \to \mathbb{R}$, $x_0 \in \mathbb{R}$ and $L \in \mathbb{R}$. We say that:

> $f(x)$ tends to L as x approaches x_0 from the left if, when $x \in D$ is very close to x_0 AND STRICTLY LESS THAN x_0, $f(x)$ is very close to L.

We write:

$$\lim_{x \to x_0-} f(x) = L \qquad \text{or} \qquad f(x) \to L \text{ as } x \to x_0-$$

Similarly, we say that

> $f(x)$ tends to L as x approaches x_0 from the right if, when $x \in D$ is very close to x_0 AND STRICTLY GREATER THAN x_0, $f(x)$ is very close to L.

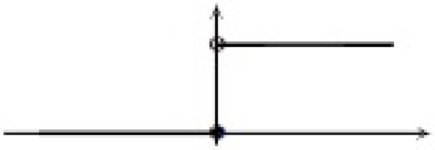
We write this as:

$$\lim_{x \to x_0+} f(x) = L \qquad \text{or} \qquad f(x) \to L \text{ as } x \to x_0+$$

The Heaviside function is defined on $\mathbb{R}$ by

$$H(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

Note that $H(0)$ is sometimes defined as 0, 1, 1/2, or some arbitrary constant c. For this example, it doesn't matter, since the limit at a point does not depend on the value at the point. If x is small and negative then $H(x) = 0$. If x is small and positive, then $H(x) = 1$.



This gives us:

$$H(x) \to 0 \text{ as } x \to 0- \qquad H(x) \to 1 \text{ as } x \to 0+$$

Fact (see Real Analysis): **a function has a limit L at a point x_0 if and only if it has left and right limits at x_0 and those limits both equal L.**

Because H has different left and right limits at the origin, it has no (two-sided) limit there.

2.13 Infinite Limits at Infinity

The last kind of limit we introduce is that of an infinite limit at ±∞. Suppose $D \subseteq \mathbb{R}$ and $f : D \to \mathbb{R}$. (Note: these ideas apply equally to sequences: as for finite limits, just think of a sequence as a function with domain e.g. $\mathbb{N}$.) We say that

> $f(x)$ tends to ∞ as x tends to ∞ if, when $x \in D$ is very large and positive, $f(x)$ is also very large and positive.

We write this as:

$$\lim_{x\to\infty} f(x) = \infty \qquad \text{or} \qquad f(x) \to \infty \text{ as } x \to \infty$$

The three variants $f(x) \to \infty$ as $x \to -\infty$ and $f(x) \to -\infty$ as $x \to \pm\infty$ are made in the obvious way.

Fundamental examples: if $n \in \mathbb{N}$ then

$$\begin{aligned} x^n &\to +\infty \text{ as } x \to +\infty \\ x^{2n} &\to +\infty \text{ as } x \to -\infty \\ x^{2n-1} &\to -\infty \text{ as } x \to -\infty \end{aligned}$$

The algebra of infinite limits (such as it is) works in exactly the same way for limits at ±∞ as it does at finite points, except that we need to modify AIL1 to look like AL1′.

> **AIL1′:** (localisation principle) For any fixed $r > 0$, $\lim_{x\to\infty} f(x)$ is unaffected by the values of $f(x)$ where $x < r$ and $\lim_{x\to-\infty} f(x)$ is unaffected by the values of $f(x)$ where $x > -r$.

Consider a polynomial whose leading coefficient is 1 (a *monic* polynomial):

$$P(x) = x^N + a_{N-1}x^{N-1} + \cdots + a_1 x + a_0$$

Our knowledge of finite limits shows us that

$$\frac{P(x)}{x^N} \to 1 \text{ as } x \to \infty$$

When x is very large, $P(x)/x^N$ is close to 1. In particular, we must have $P(x)/x^N > 1/2$. Rearranging,

$$P(x) > \frac{x^N}{2}$$

for large x. Because $x^N \to \infty$ as $x \to \infty$, the Algebra of Infinite Limits (AIL1′–AIL4) shows that $P(x) \to \infty$ as $x \to \infty$.

A similar argument works at $-\infty$ and we can use AIL2 to consider leading coefficients other than 1.

Important example:

If P is a polynomial with leading term ax^N then

$$P(x) \to \begin{cases} +\infty & \text{if } a > 0 \\ -\infty & \text{if } a < 0 \end{cases}$$

as $x \to \infty$ and

$$P(x) \to \begin{cases} +\infty & \text{if } a > 0 \text{ and } N \text{ is even} \\ -\infty & \text{if } a < 0 \text{ and } N \text{ is even} \\ -\infty & \text{if } a > 0 \text{ and } N \text{ is odd} \\ +\infty & \text{if } a < 0 \text{ and } N \text{ is odd} \end{cases}$$

as $x \to -\infty$.

We can use this to complete our description of limits at infinity of rational functions. The outstanding case is that where the degree of the numerator is greater than that of the denominator. Consider an example:

$$f(x) = \frac{x^2 + 3x - 2}{x - 4} \qquad (x \neq 4)$$

Making the first step in long division of polynomials (see Mathematical Skills) we have

$$f(x) = x + \frac{7x - 2}{x - 4}$$

Now, as $x \to \pm\infty$, the right-hand term tends to 7. This will have no effect on the other term, x, at $\pm\infty$ so we have

$$\frac{x^2 + 3x - 2}{x - 4} \to \begin{cases} +\infty \text{ as } x \to +\infty \\ -\infty \text{ as } x \to -\infty \end{cases}$$

All rational functions where the degree of the numerator is greater than that of the denominator can be handled in this way; the limits at $\pm\infty$ depend only on the signs of the leading coefficients of the numerator and denominator.

This is also another way (2.6) to see that polynomials that are equal as functions have the same coefficients: the difference between two non-equal polynomials is either a non-zero constant or tends to $\pm\infty$ at $\pm\infty$.

3 Continuity and Differentiability

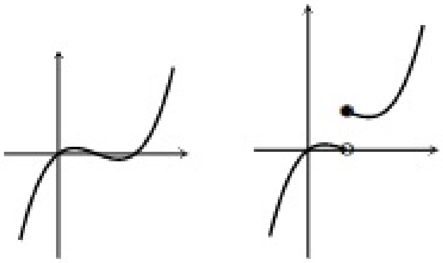
Most ideas in Calculus are based on the idea of a limit. In this section, we investigate two of the most important: continuity and differentiation.

3.1 Continuity

The idea is that a continuous function is one whose graph has no gaps in it. We encode this using limits.

> We say that $f : D \to \mathbb{R}$ is continuous on D if, for every $x_0 \in D$, $f(x) \to f(x_0)$ as $x \to x_0$.

Discontinuous vs. continuous:



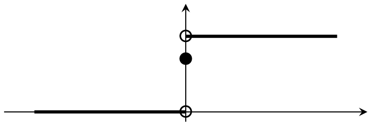
On the left, a continuous function. On the right, a discontinuous (i.e., not continuous) function: at the gap, the function has different left and right limits, so no true limit.

Fundamental example: an earlier result (2.5) can now be expressed in terms of continuity.

> Every polynomial is continuous on $\mathbb{R}$; every rational function is continuous on its domain of definition (i.e. everywhere except the zeros of its denominator).

We can create explicit examples of discontinuous functions using the Heaviside function, which we saw (2.12) earlier. Choose $c \in \mathbb{R}$ and let

$$H(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x < 0 \\ c & \text{if } x = 0 \end{cases}$$



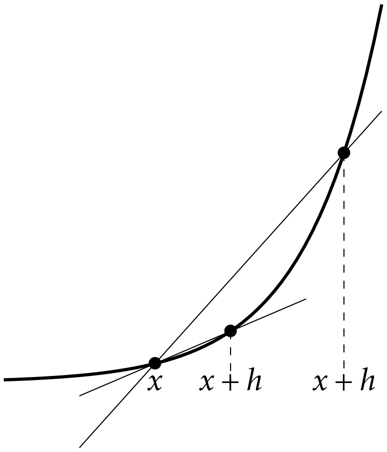
The intuition here is that there is a gap in the graph at 0, so this function is not continuous. More precisely, we saw earlier (2.12) that $H(x) \to 1$ as $x \to 0+$ and $H(x) \to 0$ as $x \to 0-$. Since H has no limit at 0, it is not continuous at 0 — whatever value we give to c.

3.2 Defining the Derivative

Suppose $D \subseteq \mathbb{R}$ and $f : D \to \mathbb{R}$. We define the *derivative* of f at x by

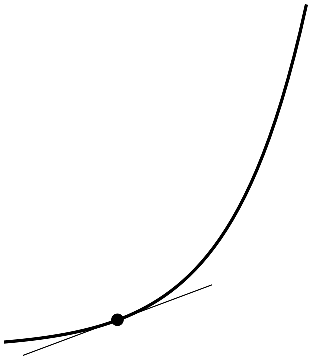
$$\boxed{f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h}}$$

whenever this limit exists. When it does exist, the function is said to be *differentiable* at x.



We describe the line *tangent* to f at x by the formula

$$t(x+h) = f(x) + hf'(x)$$



This function t is also known at the *linearisation* of f about x or the *local linear approximation* to f about x. The difference

$$f(x+h)-t(x+h)$$

is the *error* or *remainder* in the local linear approximation.

3.3 A Few Special Derivatives

We can differentiate some simple functions directly from the difference quotient. Suppose $a, b, c \in \mathbb{R}$ and $n \in \mathbb{N}$. Then we have the following derivatives:

$$\begin{aligned}
&f(x) = a && f'(x) = 0 \\
&f(x) = ax+b && f'(x) = a \\
&f(x) = ax^2+bx+c && f'(x) = 2ax+b \\
&f(x) = x^n && f'(x) = nx^{n-1} \\
&f(x) = 1/x && f'(x) = -1/x^2 \qquad (x \neq 0)
\end{aligned}$$

Constants: if $f(x) = a$ $(x \in \mathbb{R})$ then

$$f'(x) = \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} = \lim_{h\to 0} \frac{a-a}{h} = 0$$

Linear functions: if $f(x) = ax+b$ $(x \in \mathbb{R})$ then

$$f'(x) = \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} = \lim_{h\to 0} \frac{a(x+h)+b-ax-b}{h} = a$$

Quadratics: if $f(x) = ax^2+bx+c$ $(x \in \mathbb{R})$ then

$$\begin{aligned}
f'(x) &= \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} \\
&= \lim_{h\to 0} \frac{a(x+h)^2+b(x+h)+c-ax^2-bx-c}{h} \\
&= \lim_{h\to 0} 2ax+b+h = 2ax+b
\end{aligned}$$

Powers: if $n \in \mathbb{N}$ and $f(x) = x^n$ $(x \in \mathbb{R})$ then

$$\begin{aligned}
f'(x) &= \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} \\
&= \lim_{h\to 0} \frac{1}{h}\left[x^n + nhx^{n-1} + n(n-1)h^2x^{n-2}/2 + \cdots + h^n - x^n\right] \\
&= \lim_{h\to 0} nx^{n-1} + n(n-1)hx^{n-2}/2 + \cdots + h^{n-1}
\end{aligned}$$

All terms apart from the first are of the form $h^k \times$ (something not depending on h) for some $k \in \mathbb{N}$. All such terms tend to zero, so f is differentiable at x and $f'(x) = nx^{n-1}$.

Reciprocal: if $f(x) = 1/x$ ($x \in \mathbb{R}$, $x \neq 0$), then

$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h\to 0} \frac{1}{h}\left[\frac{1}{x+h} - \frac{1}{x}\right] = \lim_{h\to 0} \frac{-1}{x(x+h)} = -\frac{1}{x^2}$$

Example: let $f(x) = 2x^2 - 3x + 1$. Find the line tangent to the graph of f at the point $(2, 3)$.

Start by calculating $f'(x) = 4x - 3$, so $f'(2) = 5$. We need to find the line with gradient 5 passing through the point $(2, 3)$. We can write this as

$$y = 5(x-2) + 3 = 5x - 7$$

3.4 Localisation

The localisation principle for limits (AL1) tells us that

$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h}$$

depends only on values of $f(x+h)$ for h close to zero; equivalently,

$$f'(x) = \lim_{y\to x} \frac{f(y) - f(x)}{y - x}$$

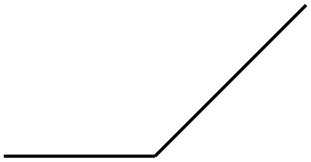
depends only on values of $f(y)$ for y close to x.

> The existence and value of $f'(x)$ depends only on values taken on by f at points close to x.

3.5 An Example of Non-Differentiability

Define $f : \mathbb{R} \to \mathbb{R}$ by

$$f(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } x > 0 \end{cases}$$



Then f is differentiable everywhere except at 0. Its derivative is the Heaviside function:

$$f'(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x > 0 \\ \text{undefined} & \text{if } x = 0 \end{cases}$$

The point of non-differentiability can be seen as a sharp corner in the graph.

The calculations go as follows. If $x_0 > 0$ then, for x close to x_0, we have $f(x) = x$ and hence $f'(x) = 1$ by the localisation principle for derivatives. Similarly,if $x_0 < 0$ then, for x near x_0, we have $f(x) = 0$ so $f'(x) = 0$. But if $x = 0$ then

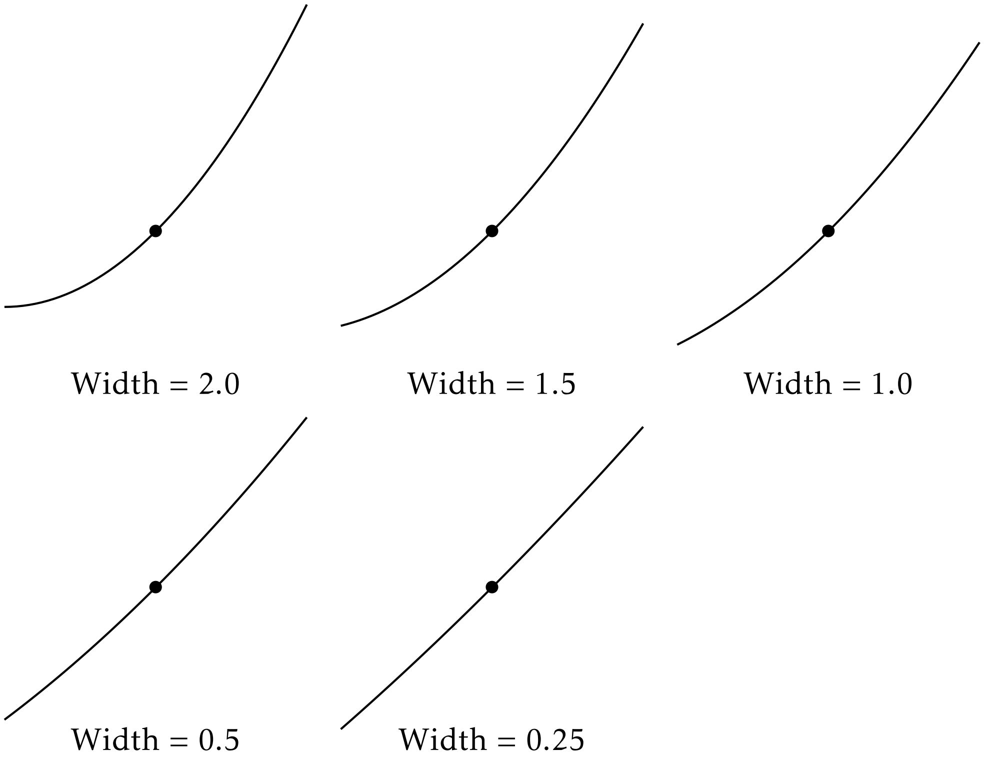
$$\frac{f(x+h) - f(x)}{h} = \begin{cases} 0 & \text{if } h < 0 \\ 1 & \text{if } h > 0 \end{cases}$$

This is the Heaviside function (2.12) which we know has no limit at the origin; f is thus not differentiable at the origin.

Another example of a non-differentiable function is $g : \mathbb{R} \to \mathbb{R}$, $g(x) = |x|$. This function has a slope of -1 on the side of the graph to the left of $x = 0$ and a slope of 1 on the side of the graph to the right of $x = 0$, so there is no derivative at $x = 0$.

3.6 Functions under the Microscope

These are graphs of $f(x) = x^2$, centred around the point (1,1) (marked) at different scales. The more closely we zoom in, the more the function looks like a straight line. That straight line is exactly the tangent line.



Width = 2.0

Width = 1.5

Width = 1.0

Width = 0.5

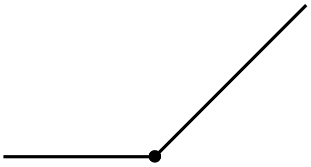
Width = 0.25

All differentiable functions look like this: close to any point, the function is well approximated by the tangent.

In contrast, this is the graph of the non-differentiable function

$$f(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } x > 0 \end{cases}$$

described earlier (3.5) about the point $(0,0)$ (marked)



Width = anything!

This function looks the same at all levels of magnification. No matter how much we zoom in around $(0,0)$, it never looks like a straight line.

3.7 The Remainder Term

As stated earlier, if f is differentiable at x, the difference between $f(x+h)$ and the linear approximation $t(x+h) = f(x) + hf'(x)$ is called the *remainder* or *error* at x and is given by

$$r(h) = f(x+h) - t(x+h) = f(x+h) - hf'(x) - f(x)$$

or, rearranging,

$$\boxed{f(x+h) = f(x) + hf'(x) + r(h)}$$

The remainder has the important property that $r(h)/h \to 0$ as $h \to 0$. Equally important, the process works in reverse: if we can write

$$f(x+h) = f(x) + ah + r(h)$$

where $r(h)/h \to 0$ as $h \to 0$, then f is differentiable at x and $f'(x) = a$.

To see that if f is differentiable then $r(h)/h \to 0$ as $h \to 0$, write

$$\frac{r(h)}{h} = \frac{f(x+h) - hf'(x) - f(x)}{h} = \frac{f(x+h) - f(x)}{h} - f'(x)$$

which, by definition of $f'(x)$, tends to zero as $h \to 0$.

Conversely, if $f(x+h) = f(x) + ah + r(h)$ where $r(h)/h \to 0$ as $h \to 0$, then we can calculate

$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h\to 0} \frac{f(x) + ah + r(h) - f(x)}{h} = \lim_{h\to 0} a + \frac{r(h)}{h} = a$$

so f is differentiable at x and $f'(x) = a$.

The existence of an appropriate remainder term is thus exactly equivalent to the existence of a limit in the difference quotient. In many cases, this is an easier or more illuminating way of using or understanding differentiability.

We can return to an earlier (3.3) example to illustrate this: to differentiate

$$f(x) = ax^2 + bx + c$$

we can write

$$f(x+h) - f(x) = a(x+h)^2 + b(x+h) + c - ax^2 - bx - c = ax^2 + 2axh + ah^2 + bx + bh + c - ax^2 - bx - c$$

Now, we remove cancelling terms and separate any remaining terms that are a constant (w.r.t. h) multiple of h.

$$f(x+h)-f(x) = (2ax+b)h + ah^2$$

We can rearrange this as

$$f(x+h) = \underbrace{f(x) + \underbrace{(2ax+b)}_{\text{derivative}}h}_{\text{local linear approximation}} + \underbrace{ah^2}_{\text{remainder}}$$

Note that the remainder divided by h is ah, which tends to zero as $h \to 0$.

3.8 Differentiability Implies Continuity

If f is differentiable at x then

$$f(x+h) = f(x) + f'(x)h + r(h)$$

As $h \to 0$, the RHS tends to $f(x)$, so f is continuous at x. In short, every differentiable function is continuous. The converse is not true, e.g. the example

$$f(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } x > 0 \end{cases}$$

given above (3.5) is continuous but not differentiable. Most easily-described continuous functions are, like this one, non-differentiable only at a few points.

But, it is possible to construct continuous functions which are nowhere differentiable. More surprisingly, there is a technical sense in which “almost all” continuous functions are nowhere differentiable. Look up the *Weierstrass function*.

3.9 Higher-Order Derivatives

If $f : D \to \mathbb{R}$ is differentiable then its derivative f' is also a function $f' : D \to \mathbb{R}$. If this is also differentiable, we denote its derivative by f'':

$$f'' = (f')'$$

and call this the *second derivative*. We can similarly consider the third derivative f'''. This notation soon becomes unwieldy!

We also use the notation

$$f^{(0)} = f; \qquad f^{(n)} = \left(f^{(n-1)}\right)'$$

so $f^{(n)}$ is the nth derivative of the function f (assuming $f^{(0)}, \ldots, f^{(n-1)}$ are all differentiable on the domain in question).

The term *order* is often used: *second-order derivative*, *higher-order derivative*, etc.

3.10 Variables and Formulae

Instead of using the notation of functions, we often use the notation of *variables*. Each variable is represented by a letter (or some more complicated collection of symbols) and the relationships between the variables are represented by algebraic expressions.

In the simplest setting, we have one variable — called the *dependent variable* — defined in terms of another — called the *independent variable* — for example

$$y = x^2$$

which is roughly the same thing as defining a function by $y(x) = x^2$. With variables, we tend to use *Leibniz notation* for derivatives. Here, we would write

$$\frac{\mathrm{d}y}{\mathrm{d}x} = 2x$$

or

$$\frac{\mathrm{d}}{\mathrm{d}x}x^2 = 2x$$

In this notation, higher-order derivatives are represented by

$$\frac{\mathrm{d}^2y}{\mathrm{d}x^2}, \quad \frac{\mathrm{d}^3y}{\mathrm{d}x^3}, \quad \dots$$

Note that $\mathrm{d}y/\mathrm{d}x$ is simply a shorthand notation for the derivative of y with respect to x - it is not a fraction. The reason for the two notations for derivative (f' and $\mathrm{d}y/\mathrm{d}x$) is largely historic (Newton preferred one and Leibniz preferred the other).

3.11 Combinations of Continuous Functions

The algebra of limits tells us that:

> Suppose f, g are continuous functions on a domain $D \subseteq \mathbb{R}$. Then $f + g$, $f - g$ and fg are continuous on D. Provided g is never zero on D, f/g is also continuous. We can also compose continuous functions: if $g : A \to B$ and $f : B \to C$ are continuous, then $f \circ g : A \to C$ is continuous.

This gives us a slightly different way (3.1) to see why polynomials and rational functions are continuous: starting with the trivial facts that $f(x) = x$ defines a continuous function on $\mathbb{R}$ and that every constant function is continuous, we can build up the polynomials and rational functions by repeated addition, multiplication and division (not by zero), with continuity preserved at every step.

3.12 The Rules of Differentiation

Local linear approximations are a good way to understand how differentiation works. The idea is that any property of linear functions (easy to work with) should correspond to a property of differentiable functions (not so easy to work with), via the derivative.

Suppose f, g are differentiable at x.

Additivity: $f+g$ is differentiable at x and

$$(f+g)'(x) = f'(x) + g'(x)$$

This corresponds to the fact that if we add two linear functions, then their gradients add:

$$(ah+b)+(ch+d) = (a+c)h + (b+d)$$

The product rule: fg is differentiable at x and

$$(fg)'(x) = f(x)g'(x) + f'(x)g(x)$$

This is less obvious: if we multiply two linear functions then we have

$$(ah+b)(ch+d) = (ad+cb)h + bd + ach^2$$

Provided h is small, this is well approximated by the linear part, with gradient $ad+cb$.

The chain rule: if g is differentiable at x and f is differentiable at $g(x)$ then $f \circ g$ is differentiable at x and

$$(f \circ g)'(x) = f'(g(x))g'(x).$$

This corresponds to the fact that if we compose linear functions then their gradients multiply:

$$a(ch+d)+b = (ac)h + ad + b$$

The quotient rule: f/g is differentiable at x and

$$(f/g)'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}$$

(This is most easily derived from the product rule and the chain rule.) If we divide two linear functions then we have:

$$\frac{ah+b}{ch+d} = \frac{ah+b}{ch+d} \cdot \frac{ch-d}{ch-d} = \frac{(ad-cb)h + bd + ach^2}{d^2 - c^2h^2}$$

Provided h is small, this is well approximated by the linear part, with gradient $(ad-cb)/d^2$.

Here are the sum, product and quotient rules in functional and Leibniz notation: f and g are functions and u and v are variables depending on x.

$$\begin{aligned} (f+g)'(x) &= f'(x)+g'(x) & \frac{\mathrm{d}}{\mathrm{d}x}(u+v) &= \frac{\mathrm{d}u}{\mathrm{d}x}+\frac{\mathrm{d}v}{\mathrm{d}x} \\ (fg)'(x) &= f'(x)g(x)+f(x)g'(x) & \frac{\mathrm{d}}{\mathrm{d}x}uv &= u\frac{\mathrm{d}v}{\mathrm{d}x}+v\frac{\mathrm{d}u}{\mathrm{d}x} \\ (f/g)'(x) &= \frac{f'(x)g(x)-f(x)g'(x)}{g(x)^2} & \frac{\mathrm{d}}{\mathrm{d}x}\frac{u}{v} &= \frac{v\mathrm{d}u/\mathrm{d}x-u\mathrm{d}v/\mathrm{d}x}{v^2} \end{aligned}$$

In functional form, the chain rule says

$$(f\circ g)'(x) = f'(g(x))g'(x)$$

In Leibniz form, suppose z depends on y and y depends on x, so z depends on x, via y. Then

$$\frac{\mathrm{d}z}{\mathrm{d}x} = \frac{\mathrm{d}z}{\mathrm{d}y}\frac{\mathrm{d}y}{\mathrm{d}x}$$

This is particularly easy to remember, because it *looks* like a product of fractions, in which the $\mathrm{d}y$ terms have cancelled.

Here are the actual derivations. Suppose f and g are differentiable at x, so $f(x+h) = f(x)+hf'(x)+r(h)$ and $g(x+h) = g(x)+hg'(x)+s(h)$ where, as $h\to 0$, $r(h)/h \to 0$ and $s(h)/h \to 0$.

Additivity:

$$(f+g)(x+h) = f(x)+g(x)+h[f'(x)+g'(x)]+[r(h)+s(h)]$$

Since $[r(h)+s(h)]/h = r(h)/h+s(h)/h \to 0$ as $h\to 0$, we see that $f+g$ is differentiable at x and $(f+g)'(x) = f'(x)+g'(x)$.

Product rule:

$$\begin{aligned} (fg)(x+h) &= f(x+h)g(x+h) \\ &= [f(x)+hf'(x)+r(h)][g(x)+hg'(x)+s(h)] \\ &= f(x)g(x)+h[f'(x)g(x)+f(x)g'(x)]+ \\ &\qquad \underbrace{h^2f'(x)g'(x)+r(h)[g(x)+hg'(x)]+s(h)[f(x)+hf'(x)]+r(h)s(h)}_{=u(h)} \end{aligned}$$

As $h\to 0$, $r(h)/h \to 0$ and $s(h)/h \to 0$; it follows from this that, as $h\to 0$, $u(h)/h\to 0$. We conclude that fg is differentiable at x and $(fg)'(x) = f'(x)g(x)+f(x)g'(x)$. In particular, if c is a constant then $(cf)'(x) = cf'(x)$

Chain rule This is more complicated. A proof will appear in Real Analysis, or you can look at appendix (A.1).

Quotient rule: We know (3.3) that if $q(x) = 1/x$ $(x \neq 0)$ then $q'(x) = -1/x^2$. We can use the chain rule to differentiate $1/g(x) = q(g(x))$ to give $(1/g)'(x) = q'(g(x))g'(x) = -g'(x)/g(x)^2$.

By the product rule,

$$(f/g)'(x) = f'(x)[1/g(x)] + f(x)[-g'(x)/g(x)^2] = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}$$

which is the quotient rule.

In particular, if $f(x) = x^{-n} = 1/x^n$ where $n \in \mathbb{N}$ then $f'(x) = -nx^{n-1}/x^{2n} = -nx^{-n-1}$ $(x \neq 0)$. We now know that $\mathrm{d}/\mathrm{d}x\, x^n = nx^{n-1}$ for any non-zero integer n (of course, $\mathrm{d}/\mathrm{d}x\, x^0 = \mathrm{d}/\mathrm{d}x\, 1 = 0$).

3.13 Differentiating Polynomials and Rational Functions

It follows from these rules and the fundamental facts that $\mathrm{d}/\mathrm{d}x\, x^n = nx^{n-1}$ $(n \in \mathbb{N})$ and that constant functions have derivative zero that

Every polynomial is differentiable on $\mathbb{R}$ and

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(a_N x^N + a_{N-1}x^{N-1} + \cdots + a_1 x + a_0\right) = N a_N x^{N-1} + (N-1)a_{N-1}x^{N-2} + \cdots + a_1$$

So, if we differentiate a polynomial of degree $N \geq 1$ then we obtain a polynomial of degree $N - 1$.

By the quotient rule

Any rational function $f(x) = P(x)/Q(x)$ is differentiable on $\mathbb{R}$, except at the zeros of Q (where it is not even defined).

3.14 Left and Right Derivatives

Just like left and right limits, we can define left and right derivatives. We say that f is *differentiable from the right* at x if

$$\lim_{h \to 0+} \frac{f(x+h) - f(x)}{h}$$

exists; this is denoted by $f'(x+)$ and is called the *right derivative* of f at x. Similarly, the left derivative is defined by

$$f'(x-) = \lim_{h \to 0-} \frac{f(x+h) - f(x)}{h} = \lim_{h \to 0+} \frac{f(x) - f(x-h)}{h}$$

It follows from the corresponding statement for limits (2.12) that f is differentiable at x if and only if $f'(x+)$ and $f'(x-)$ exist and are equal.

For example, if $f(x) = |x|$ then

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

which shows that $f(x) = |x|$ is not differentiable at $x = 0$.

There is a potential conflict of notation here. When we write $f'(x+)$, do we mean the right derivative of f at x, as defined above, or the right limit of f' at x? Happily, this is not a major problem: if f is continuous at x and f' has a right limit at x then it is indeed the right derivative. This follows from the Mean Value Theorem, which we shall see in the next section. However, it is possible for f to be differentiable in an interval I but for f' not to have either a left or right limit at a point of I.

3.15 l'Hôpital's Rule

Suppose f and g are differentiable at x_0, $f(x_0) = g(x_0) = 0$ and $g'(x_0) \neq 0$. Then

$$\frac{f(x)}{g(x)} \to \frac{f'(x_0)}{g'(x_0)}$$

as $x \to x_0$.

This is a basic form of *l'Hôpital's rule*, a useful device for calculating limits of the "0/0" variety (*indeterminate forms*). There are many variants on this theme; all of them will make more sense when we encounter Taylor's Theorem later on.

A more sophisticated version of l'Hôpital's rule is:

$$\lim_{x\to x_0} \frac{f(x)}{g(x)} = \lim_{x\to x_0} \frac{f'(x)}{g'(x)}$$

provided the RHS is defined for x near but not equal to x_0 and the limit on the RHS exists.

To see why the basic version works, we can write

$$\frac{f(x_0+h)}{g(x_0+h)} = \frac{f(x_0)+hf'(x_0)+r(h)}{g(x_0)+hg'(x_0)+s(h)} = \frac{f'(x_0)+r(h)/h}{g'(x_0)+s(h)/h}$$

As $h \to 0$, $r(h)/h$ and $s(h)/h$ tend to zero so, because $g'(x_0) \neq 0$,

$$\frac{f(x_0+h)}{g(x_0+h)} \to \frac{f'(x_0)}{g'(x_0)}$$

As an example of how l'Hôpital's rule is used, consider the limit

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

where the numerator and denominator are both zero when $x = 1$. One way to handle this is to factor it:

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

from which we can see that the limit is 3. Alternatively, we can use l'Hôpital's rule: differentiating the numerator and denominator separately, we get $2x+1$ and 1 so the limit is $(2x+1)/1$ evaluated at $x = 1$, giving us the answer 3. This method can be beneficial when the f and g are polynomials of high degree, when factorisation is not so easy, and is particularly useful when f and g are not polynomials.

Proving the more sophisticated version of l'Hôpital's rule is much harder: see Thomas for the details.

3.16 Parametric curves, velocity and speed

Suppose I is an interval and that $x : I \to \mathbb{R}$ and $y : I \to \mathbb{R}$ are continuous functions. We can define a function $\gamma : I \to \mathbb{R}^2$ by $\gamma(t) = (x(t), y(t))$ (such a function is often called a *path*). The range of γ, i.e. the subset $\Gamma = \{(x(t), y(t)) : t \in I\}$ of $\mathbb{R}^2$ typically represents a curve of some sort. We call it a *parametric curve* and say that the function γ, or equivalently the functions x and y, *parameterise* Γ.

Recall that the graph of a function $f : I \to \mathbb{R}$ is the set

$$\{(t, f(t)) : t \in I\}$$

This is a special kind of parametric curve, with $x(t) = t$ and $y(t) = f(t)$. But parametric curves can describe many shapes that graphs cannot: for example $x(t) = \cos(t)$, $y(t) = \sin(t)$ describes the unit circle.

We can differentiate the path γ by differentiating x and y separately to give $\gamma' : I \to \mathbb{R}^2$,

$$\gamma'(t) = (x'(t), y'(t))$$

(provided $x'(t)$ and $y'(t)$ both exist). If we think of t as time and $\gamma(t)$ as a point (sometimes a "particle") moving in time, then this is the *velocity vector* of the parameterisation: it tells us how fast the point is moving in both the x and y directions. The *speed* is the magnitude (length) of the velocity vector:

$$|\gamma'(t)| = \sqrt{x'(t)^2 + y'(t)^2}$$

For time derivatives, a dot is often used instead of a dash in this notation, and the dependence on t suppressed:

$$\dot{\gamma} = (\dot{x}, \dot{y}) = (x'(t), y'(t)); \qquad |\dot{\gamma}| = \sqrt{\dot{x}^2 + \dot{y}^2}$$

For example, the curve $\gamma(t) = (x(t), y(t))$ with $x(t) = t - t^2$ and $y(t) = 1 - t^2$ has velocity and speed

$$\gamma'(t) = (1 - 2t, -2t); \qquad |\gamma'(t)| = \sqrt{8t^2 - 4t + 1}$$

We can also consider parametric curves in $\mathbb{R}^3$. Here we have an additional continuous function $z : I \to \mathbb{R}$ and consider the curve and (if they exist) velocity and speed

$$\gamma(t) = (x(t), y(t), z(t)); \qquad \gamma'(t) = (x'(t), y'(t), z'(t)); \qquad |\gamma'(t)| = \sqrt{x'(t)^2 + y'(t)^2 + z'(t)^2}$$

or

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

Indeed, we can do it in as many dimensions as we wish — see Vector Calculus.

The second derivative of a path represents the acceleration of the moving point:

$$\gamma''(t) = (x''(t), y''(t), z''(t)); \qquad \ddot{\gamma} = (\ddot{x}, \ddot{y}, \ddot{z})$$

4 Monotonicity, Inverse Functions and Extrema

Two of the most important kinds of problems in Mathematics and its applications are solving equations and maximising or minimising functions. In this section, we look at some of the fundamental tools that Calculus offers for these purposes.

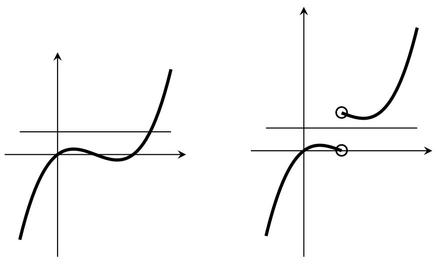
4.1 The Intermediate Value Theorem and the Extreme Value Theorem

Some equations cannot be solved using a simple formula. For quadratic equations we have the quadratic formula which allows us to express the solutions to the equation in terms of the coefficients of the quadratic, but for higher order polynomials this is not always possible. For example $3x^5 - 25x^3 - 60x = 5$ is an example of a quintic equation that cannot be solved using an algebraic formula. However, we might still be able to say that a solution *exists*, and then find an approximate solution using other methods, such as on a computer. The Intermediate Value Theorem gives us a condition for the existence of solutions to certain equations.

The *Intermediate Value Theorem* (IVT) (see Real Analysis for proof) says:

> Suppose $a < b$ and f is continuous on the interval $[a,b]$. Then for every y such that $\min(f(a), f(b)) < y < \max(f(a), f(b))$ there exists $x \in (a,b)$ such that $f(x) = y$.

In other words, if a continuous function attains two values $f(a)$ and $f(b)$ then it also attains every value between $f(a)$ and $f(b)$. This is a statement about solving equations: does the equation $f(x) = y$ have a solution? Yes, provided y lies between $f(a)$ and $f(b)$. Although it does not tell us how to solve the equation, it gives us the vital information that a solution exists.

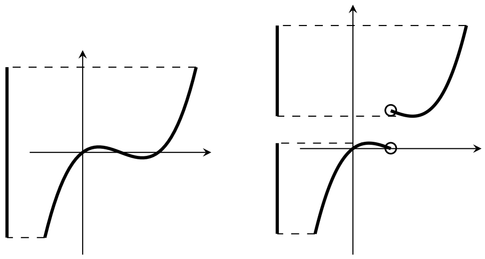


In the left-hand diagram, any horizontal line we draw will meet the graph at some point (unless it goes completely over top or under the bottom: that is the significance of inequality $\min(f(a), f(b)) < y < \max(f(a), f(b))$ in the statement of the theorem). In the right-hand diagram, we can see that there are horizontal lines that separate the top of the graph from the bottom, without touching it. This illustrates the "no gaps in the graph" idea of continuity.

Another way of stating this: if I is an interval then its *range* or *image*

$$f(I) = \{f(x) : x \in I\}$$

is also an interval (or a singleton, in the special case that f is constant): continuous functions map intervals to intervals.

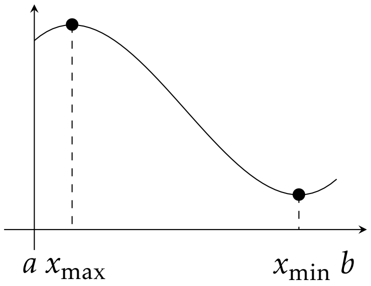


The set of values taken on by each function (the *range* or *image*) is shown by the thick vertical lines. In the continuous example, on the left, the range is a single connected set — an interval. In the discontinuous example, on the right, the range is separated into two pieces: it is a union of two disjoint intervals.

Another important result in Calculus is the Extreme Value Theorem. Suppose $a < b$ and f is continuous on the interval $[a, b]$. The *Extreme Value Theorem* (EVT) (see Real Analysis for proof) says:

> There are points $x_{\text{min}}, x_{\text{max}} \in [a, b]$ such that for all $x \in [a, b]$,
>
> $$f(x_{\text{min}}) \le f(x) \le f(x_{\text{max}}).$$

Another way of saying this is that x_{min} and x_{max} are the points at which f takes on its minimum and maximum values over $[a, b]$. Like the IVT, it does not tell us how to find the extreme values, but it does tell us that they exist.



Continuous functions on other kinds of intervals need not have maxima or minima, e.g. $f(x) = x^2$ on $(0, 1)$ has no maximum or minimum: $f(x)$ approaches 0 and 1 as limits as $x \to 0+$ and $x \to 1-$, but we never have $f(x) = 0$ or $f(x) = 1$ for $x \in (0, 1)$.

For another example, let $g(x) = x - 1/x$ for $x \in (0, \infty)$. As $x \to 0+$, $g(x) \to -\infty$. As $x \to \infty$, $g(x) \to \infty$. This function certainly has no maximum or minimum value.

Discontinuous functions on compact intervals need not have maximum or minimum values, e.g. define $h : [0,1] \to \mathbb{R}$ by

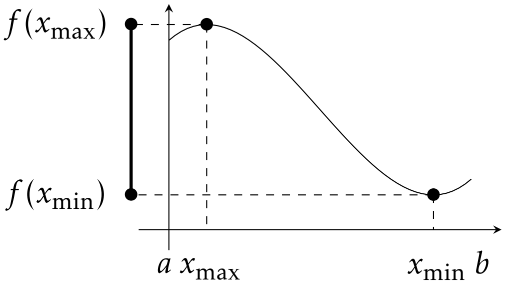
$$h(x) = \begin{cases} x & \text{if } 0 < x < 1 \\ \frac{1}{2} & \text{if } x = 0 \text{ or } x = 1 \end{cases}$$

Here $h([0,1]) = (0,1)$ so, although $[0,1]$ is compact, $h([0,1])$ is open (and, of course, h is not continuous).

The Extreme and Intermediate Value Theorems together tell us that if $f : [a,b] \to \mathbb{R}$ is continuous then (with x_{min} and x_{max} as before)

$$f([a,b]) = [f(x_{\text{min}}), f(x_{\text{max}})]$$

i.e. continuous functions map compact intervals to compact intervals (or singletons, in the special case of a constant function).



We now know that non-constant continuous functions map intervals to intervals, and compact intervals to compact intervals. Other properties of intervals (open, closed, finite, etc.) need not be preserved, e.g. if $I = (-1,1)$ and $f(x) = x^2$ then $f(I) = [0,1)$. Here, f has a minimum value but no maximum value; I is open but $f(I)$ is half-open. Another example is $g(x) = 1/x$ $(x \neq 0)$ which maps the finite interval $(0,1)$ to the infinite interval $(1,\infty)$, and also maps $(1,\infty)$ to $(0,1)$.

4.2 The Inverse Function Theorem for Continuous Functions

Suppose $D \subseteq \mathbb{R}$ and $f : D \to \mathbb{R}$. We say that f is:

increasing if $x < y \implies f(x) \leq f(y)$
strictly increasing if $x < y \implies f(x) < f(y)$
decreasing if $x < y \implies f(x) \geq f(y)$
strictly decreasing if $x < y \implies f(x) > f(y)$

monotonic if f is increasing or decreasing
strictly monotonic if f is strictly increasing or strictly decreasing

for $x, y \in D$.

Every strictly monotonic function is injective (one-to-one).

Injectivity (see Mathematical Skills for more details about this property) is proved as follows: If f is strictly increasing and $x, y \in D$ with $x \neq y$ then either $x < y$, so $f(x) < f(y)$, or $y < x$, so $f(y) < f(x)$; in either case, $f(x) \neq f(y)$. Similarly for strictly decreasing functions.

Fundamental example: if $n \in \mathbb{N}$ then

- $f(x) = x^n$ defines a strictly increasing function on $[0, \infty)$ and $g(x) = x^{-n}$ defines a strictly decreasing function on $(0, \infty)$
- $f(x) = x^{2n}$ defines a strictly decreasing function on $(-\infty, 0]$ and $g(x) = x^{-2n}$ defines a strictly increasing function on $(-\infty, 0)$
- $f(x) = x^{2n-1}$ defines a strictly increasing function on $(-\infty, 0]$ and $g(x) = x^{1-2n}$ defines a strictly decreasing function on $(-\infty, 0)$

One way to see this is to use the fact that if $0 \leq x < y$ and $n \in \mathbb{N}$ then:

$$x^n - y^n = \underbrace{(x-y)}_{<0}\underbrace{(x^{n-1} + x^{n-2}y + \cdots + xy^{n-2} + y^{n-1})}_{>0}$$

All the other results follow e.g. if $0 < x < y$ then $x^n < y^n$ so $x^{-n} > y^{-n}$.

Note that the monotonicity properties could equally have been defined with the $>$ and $\geq$ relations: f is increasing if $x \geq y$ implies $f(x) \geq f(y)$, strictly decreasing if $x > y$ implies $f(x) < f(y)$, etc.

Suppose I is an interval and $f : I \to \mathbb{R}$ is continuous and strictly monotonic. The IVT (4.1) tells us that the range

$$J = f(I) = \{f(x) : x \in I\}$$

is also an interval. If $y \in f(I)$ then $y = f(x)$ for some $x \in I$; moreover, this x is unique, because f is strictly monotonic, hence injective (one-to-one).

We can therefore define another function $f^{-1} : J \to I$ by $f^{-1}(y) = x$; that is $f^{-1}(f(x)) = x$. Moreover, $f(f^{-1}(y)) = y$.

The *Inverse Function Theorem* (IFT) for continuous functions says that:

> Under these conditions ($f : I \to \mathbb{R}$ continuous and strictly monotonic), the inverse function f^{-1} described above is continuous.

Notes:

- Like the IVT (4.1), this is about solving equations. It tells us that we can uniquely solve the equation $f(x) = y$ and that the solution x depends continuously on y.
- if f is strictly increasing, then so is f^{-1} and if f is strictly decreasing, then so is f^{-1}.
- If f is continuous but *not* strictly monotonic, it can be shown that no inverse function exists because, because, for some values of y, the equation $f(x) = y$ has more than one solution; that is, f is not injective (one-one).
- See your Mathematical Skills notes for a full discussion of inverse functions. The property of surjectivity (onto) makes no explicit appearance here because it is implicit in the definition of J as the range (image) of I. The existence of the inverse function is equivalent to the statement that f is a bijection between I and J.
- For a proof of the IFT, see Real Analysis.

4.3 Monotonicity, maxima and minima: first attempt

We now link the ideas of monotonicity to those of differentiability.

A linear function

$$f : \mathbb{R} \to \mathbb{R},\ f(x) = ax + b \qquad (a, b \in \mathbb{R})$$

is increasing if $a > 0$ and decreasing if $a < 0$. We therefore expect a body of knowledge around the idea that “a differentiable function f is increasing if $f'(x) > 0$ and decreasing if $f'(x) < 0$.” First step:

If f is increasing on $[a,b]$ and differentiable at $x \in [a,b]$ then $f'(x) \geq 0$.

If f is decreasing on $[a,b]$ and differentiable at $x \in [a,b]$ then $f'(x) \leq 0$.

Suppose f is increasing, i.e. if $x \leq y$ then $f(x) \leq f(y)$ (4.2). Then

$$f(x+h) - f(x) \begin{cases} \geq 0 & \text{if } h > 0 \\ \leq 0 & \text{if } h < 0 \end{cases}$$

from which it follows that

$$\frac{f(x+h)-f(x)}{h} \geq 0$$

If f is differentiable at x, then we can take a limit as $h \to 0$ to see that $f'(x) \geq 0$. Similarly, if f is decreasing and differentiable then $f'(x) \leq 0$.

Note that the stronger "strictly increasing" property ($x < y \implies f(x) < f(y)$) does not imply that $f'(x) > 0$. For example, if $f(x) = x^3$ for $x \in \mathbb{R}$ then f is strictly increasing (4.2) but $f'(0) = 3x^2|_{x=0} = 0$.

Suppose $D \subseteq \mathbb{R}$ and $f : D \to \mathbb{R}$. We say that

f has a *local maximum* at x_0 if, for some $r > 0$ and all x with $|x - x_0| < r$, $f(x)$ is defined (i.e. $x \in D$) and $f(x) \leq f(x_0)$.

f has a *local minimum* at x_0 if, for some $r > 0$ and all x with $|x - x_0| < r$, $f(x)$ is defined and $f(x) \geq f(x_0)$.

local extremum means "local maximum or local minimum".

Fundamental fact:

If f is differentiable at a local extremum x_0 then $f'(x_0) = 0$. Such a point (x_0, where $f'(x_0) = 0$) is known as a *stationary point*, *turning point* or *critical point* of f.

To see why this works, assume x_0 is a local maximum. Then, for $0 < |h| < r$, $f(x_0 + h) - f(x_0) \leq 0$, so

$$\frac{f(x_0+h)-f(x_0)}{h} \begin{cases} \leq 0 & \text{if } h > 0 \\ \geq 0 & \text{if } h < 0 \end{cases}$$

This tells us that the left limit of the difference quotient, if it exists, is ≥ 0 and the right limit, if it exists, is ≤ 0. If the limit itself exists then it has to be both ≥ 0 and ≤ 0, hence equal to 0. A similar argument works for a local minimum.

Notes:

1. Not all functions have local extrema. Not all functions have stationary points (e.g. $f(x) = x$ on any domain).

2. If $f : [a,b] \to \mathbb{R}$ then the maximum and minimum values of f over $[a,b]$ (4.1) can be attained either at local extrema or at the endpoints a and b (e.g. $f(x) = x$ on any interval $[a,b]$).

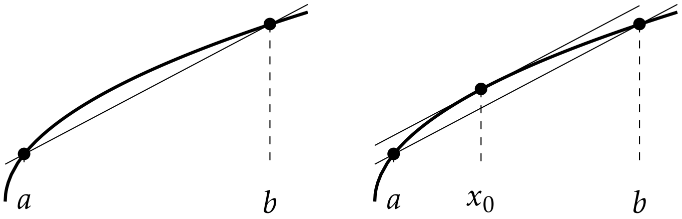
3. If x_0 is a stationary point, i.e. $f'(x_0) = 0$, then x_0 need not be a local extremum (e.g. $f(x) = x^3$, $x_0 = 0$).

4. f can have extrema at which it is not differentiable (e.g. $f(x) = |x|$ on $[-1,1]$).

4.4 The Mean Value Theorem

We saw in the previous section that monotonic differentiable functions have single-signed derivatives. We want to understand the reverse implication: can we say anything about the monotonicity of a function, given information about the sign of the derivative? This turns out to be a good deal more subtle.

Suppose $f : [a,b] \to \mathbb{R}$ is continuous and differentiable on (a,b) and consider the chord between $(a, f(a))$ and $(b, f(b))$.



The Mean Value Theorem (MVT) says that:

> There exists $x_0 \in (a,b)$ such that
>
> $$f'(x_0) = \frac{f(b) - f(a)}{b - a}$$
>
> (the tangent at x_0 is parallel to the chord from $(a, f(a))$ to $(b, f(b))$.)

Geometrically, think of moving the chord line up, preserving its gradient. As we do this, the two points at which it touches the curve become closer together, meeting each other at the point where the line is just about to separate from the curve. At this point, the line is tangent to the curve.

A more formal argument can be given, using the Extreme Value Theorem to find the mean value x_0 — see (A.2) for the details. Alternatively, at the expense of requiring f' to be continuous, a much easier derivation is possible once we have the basics of integration theory working; this will be presented later.

Compare the MVT to the local linear approximation:

$$f(x+h) = f(x) + hf'(x_0); \qquad f(x+h) = f(x) + hf'(x) + r(h).$$

The MVT eliminates the remainder term, at the expense of changing $f'(x)$ to $f'(x_0)$ for some unknown x_0 between x and $x+h$.

4.5 Monotonicity, maxima and minima: second attempt

The MVT allows us to understand the sign of the derivative much more clearly.

Suppose f is continuous on $[a,b]$ and differentiable on (a,b). If $a \le x < y \le b$ then we can apply the MVT to f on $[x,y]$ and rearrange to give

$$f(y) = f(x) + (y-x)f'(z)$$

for some $z \in (x,y)$. This gives several important conclusions (remember $x < y$)

- If $f'(z) \ge 0$ for all $z \in (a,b)$ then $f(y) \ge f(x)$, so f is increasing.
- If $f'(z) > 0$ for all $z \in (a,b)$ then $f(y) > f(x)$, so f is strictly increasing.
- If $f'(z) \le 0$ for all $z \in (a,b)$ then $f(y) \le f(x)$, so f is decreasing.
- If $f'(z) < 0$ for all $z \in (a,b)$ then $f(y) < f(x)$, so f is strictly decreasing.
- If $f'(z) = 0$ for all $z \in (a,b)$ then $f(y) = f(x)$, so f is constant.

All these apply on the interval $[a,b]$.

In combination with earlier results (4.3) we now know that, for a differentiable function on an interval, the increasing property is equivalent to the derivative being non-negative. Also, if the derivative is strictly positive then the function is strictly increasing. This implication is not reversible, though: $f(x) = x^3$ describes a strictly increasing function (see (4.7)) with $f'(0) = 0$.

For example, consider the cubic polynomial

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

If we differentiate this, we get

$$P'(x) = x^2 - 3x + 2 = (x-1)(x-2)$$

By considering the signs of the factors $x-1$ and $x-2$, we can see that

$$P'(x)\begin{cases} >0 & \text{if } x<1 \\ =0 & \text{if } x=1 \\ <0 & \text{if } 1<x<2 \\ =0 & \text{if } x=2 \\ >0 & \text{if } x>2 \end{cases}$$

We can now find the *intervals of monotonicity*: P is strictly increasing on $(-\infty,1]$, strictly decreasing on $[1,2]$ and strictly increasing on $[2,\infty)$. It follows that 1 is a local maximum and 2 is a local minimum.

This allows us to find the maximum and minimum of P over any interval, by considering the values at the endpoints and included critical points. For example, to find the extrema over $[0,2]$ we need to consider the values of P at 0 (endpoint), 1 (included critical point) and 2 (endpoint and included critical point). These are given by

$$P(0)=-1; \qquad P(1)=-\frac{1}{6}; \qquad P(2)=-\frac{1}{3}$$

so the minimum is -1 and the maximum $-1/6$.

For another example, consider $f:\mathbb{R}\to\mathbb{R}$

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

Using the quotient rule, we see that

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

and we can again find the intervals of monotonicity: the denominator is always positive and we see from the numerator that $f'(\pm1)=0$, $f'<0$ on $(-\infty,-1)$ and on $(1,\infty)$ and $f'>0$ on $(-1,1)$. The function is thus strictly decreasing on $(-\infty,-1]$, has a local minimum at -1, is strictly increasing on $[-1,1]$, has a local maximum at 1 and is then decreasing on $[1,\infty)$. The values of f at the critical points are $f(-1)=-1/2$ and $f(1)=1/2$.

On a finite interval, we would also have to consider endpoint values. Working over the whole line, the relevant information is the limits at $\pm\infty$, both of which are 0. We can now conclude that the minimum of f over $\mathbb{R}$ is $-1/2$, attained at -1, and the maximum is $1/2$, attained at 1.

Note, though, that functions on $\mathbb{R}$ need not have maxima or minima. For example,

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

has a maximum value of 1, attained at 0, but no minimum value: as $x \to \pm\infty$, $g(x)$ becomes closer and closer to 0, but never actually reaches 0. In Real Analysis, this will be called an *infimum*, as opposed to a minimum.

The "second derivative test" for classifying critical points as maxima or minima will come later in the course, but as the examples above show the same information can be found by understanding the intervals of monotonicity.

4.6 The Inverse Function Theorem for Differentiable Functions

Suppose f is continuous on $[a,b]$ and differentiable on (a,b) and $f'(x) > 0$ for all $z \in (a,b)$. Then f is strictly increasing on $[a,b]$, so (by the Inverse Function Theorem for Continuous Functions (4.2)), it has a continuous inverse $f^{-1} : [f(a), f(b)] \to [a,b]$. Similarly, if $f'(x) < 0$ for all $x \in (a,b)$, f has a continuous inverse. It turns out that f^{-1} is differentiable and we have the *Inverse Function Theorem for Differentiable Functions*:

- Suppose $f : [a,b] \to \mathbb{R}$ is continuous on $[a,b]$ and differentiable on (a,b).
- If $f'(x) > 0$ for all $x \in (a,b)$ then $f^{-1} : [f(a), f(b)] \to [a,b]$ is differentiable.
- If $f'(x) < 0$ for all $x \in (a,b)$ then $f^{-1} : [f(b), f(a)] \to [a,b]$ is differentiable.
- In either case,
$$(f^{-1})'(y) = \frac{1}{f'(f^{-1}(y))}$$

Recall the chain rule in Leibniz notation: suppose z depends on y and y depends on x, so z depends on x, via y. Then
$$\frac{\mathrm{d}z}{\mathrm{d}x} = \frac{\mathrm{d}z}{\mathrm{d}y}\frac{\mathrm{d}y}{\mathrm{d}x}$$

Now suppose that x and y are related in such a way that either variable can be expressed in terms of the other (equivalently, one variable depends on the other in a strictly monotonic way). Then this formula makes sense with $z = x$ and we have
$$\frac{\mathrm{d}x}{\mathrm{d}x} = 1 = \frac{\mathrm{d}x}{\mathrm{d}y}\frac{\mathrm{d}y}{\mathrm{d}x}$$

or
$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{\mathrm{d}x/\mathrm{d}y}$$

Like the chain rule, this is easy to remember because it looks like a fraction being inverted.

Assuming that f^{-1} is differentiable, we can find $(f^{-1})'$ using the chain rule: because $f(f^{-1}(y)) = y$, differentiating both sides we get $f'(f^{-1}(y))(f^{-1})'(y) = 1$ and hence $(f^{-1})'(y) = 1/f'(f^{-1}(y))$.

Actually *proving* that f^{-1} is differentiable is a bit harder. This will be covered in Real Analysis but, if you'd like to see the details, see (A.3).

4.7 Application: Rational Powers

If $x \in \mathbb{R}$ and $p \in \mathbb{N}$, we define x^p by repeated multiplication and note that $x^p x^q = x^{p+q}$. This motivates our definition of $x^0 = 1$ and (for $x \neq 0$) $x^{-p} = 1/x^p$. With these in place, we have (for any $x \in \mathbb{R}$ and $p, q \in \mathbb{Z}$):

- $x^p x^q = x^{p+q}$ for any integers p and q
- $(xy)^p = x^p y^p$
- $(x^p)^q = x^{pq}$

unless any part of the formula contains a negative power of 0.

Let $I = [0, \infty)$ and $f(x) = x^2$. This is continuous (see (3.1)) and strictly increasing (see (4.2)). It maps 0 to 0 and tends to ∞ at ∞ (see (2.13)).

By the IFT for continuous functions, we have a continuous, strictly increasing inverse function $f^{-1} : [0, \infty) \to [0, \infty)$. This function is the (non-negative) square root function, which we denote by $\sqrt{\cdot}$ as usual. The same principle works for any even power.

Now let $I = \mathbb{R}$ and $f(x) = x^3$. This is continuous and strictly increasing and $f(x) \to \infty$ as $x \to \infty$ and $f(x) \to -\infty$ as $x \to -\infty$.

By the IFT for continuous functions, we have a continuous, strictly increasing inverse function $f^{-1} : \mathbb{R} \to \mathbb{R}$. This function is the (real) cube root function. The same principle works for any odd power.

Suppose $x > 0$. The identity $(x^p)^q = x^{pq}$ motivates our definition of $x^{p/q}$ ($p \in \mathbb{Z}$, $q \in \mathbb{N}$) to be the qth root of x^p (where the qth root of x is the unique positive real number such that $(x^{1/q})^q = x$ (the existence and uniqueness of this number follows from the IFT, as sketched above)). With this in place, we have (for $x, y > 0$, $p, r \in \mathbb{Z}$, $q, s \in \mathbb{N}$)

- $(xy)^{p/q} = x^{p/q}y^{p/q}$
- $x^{p/q}x^{r/s} = x^{p/q+r/s}$
- $(x^{p/q})^{r/s} = x^{(pr)/(qs)}$

We have now described $x^{p/q}$ ($p \in \mathbb{Z}$, $q \in \mathbb{N}$) in terms of x, p and q but:

- Some care needs to be taken to ensure that the definition even makes sense: for example, we should really check that $2^{1/2}$ and $2^{2/4}$ give the same answer (they do!).
- Actually verifying the three properties stated is a tedious and fiddly exercise.
- We still cannot interpret e.g. $2^{\sqrt{2}}$, because $\sqrt{2}$ is irrational. For expressions like this, we need the exponential function (described in the next chapter).

We defined rational powers by using the Inverse Function Theorem for continuous functions. We can differentiate them using the Inverse Function Theorem for differentiable functions. Suppose $p \in \mathbb{Z} \setminus \{0\}$ and $q \in \mathbb{N}$.

If $y = x^{p/q}$ then

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{p}{q}x^{p/q-1}$$

Domains: this holds for $x \in (0, \infty)$. If q is odd, it also holds for $x \in (-\infty, 0)$. It applies for $x = 0$ if and only if $p/q > 1$. This leads to a useful approximation:

If h is small and $u \in \mathbb{Q}$ then $(1+h)^u \approx 1 + uh$.

Remark: it is best to avoid the $\approx$ symbol (meaning "is approximately equal to") in formal writing (such as your assignments), unless you are asked to find a numerical approximation for a value. This is because this symbol is imprecise - what exactly do we mean by "approximately"? Do we mean that the difference between the two tends to 0, or the ratio tends tends to 1, or something else entirely? Throughout your degree you will see several different precise ways of expressing this intuitive idea.

To see how we get this approximation, start by setting $x = y^q$, so $y = x^{1/q}$. We know that

$$\frac{\mathrm{d}x}{\mathrm{d}y} = qy^{q-1}$$

so by the Inverse Function Theorem

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{qy^{q-1}} = \frac{1}{q(x^{1/q})^{q-1}} = \frac{1}{q}x^{1/q\,-1}$$

so the usual formula for differentiating a power works for qth roots.

Now, we can use the chain rule:

$$\frac{\mathrm{d}}{\mathrm{d}x}x^{p/q} = \frac{\mathrm{d}}{\mathrm{d}x}(x^p)^{1/q} = \frac{1}{q}(x^p)^{1/q\,-1}px^{p-1} = \frac{p}{q}x^{p/q\,-1}$$

so the usual formula for differentiating a power works for all non-zero rational powers. Note that we haven't yet described what irrational powers mean.

We can use this to find the approximation above: if $u \in \mathbb{Q}$ then

$$\frac{\mathrm{d}}{\mathrm{d}h}(1+h)^u = u(1+h)^{u-1}$$

which evaluates to u if $h = 0$. Using the local linear approximation we can write

$$(1+h)^u = 1 + uh + s(h)$$

where $s(h)/h \to 0$ as $h \to 0$. Informally, if h is small, then $(1+h)^u \approx 1 + uh$.

4.8 Application: polynomials of odd degree

Suppose

$$P(x) = a_0 + a_1x + a_2x^2 + \cdots + a_Nx^N$$

with N odd and $a_N \neq 0$. Then $P(x_0) = 0$ for some $x_0 \in \mathbb{R}$. Phrased slightly differently,

Every real polynomial of odd degree has a real zero (root).

This is because (see (2.13))

- If $a_N > 0$ then $P(x) \to \infty$ as $x \to \infty$ and $P(x) \to -\infty$ as $x \to -\infty$. In particular, $P(x) > 0$ when x is large and positive and $P(x) < 0$ when x is large and negative.

- If $a_N < 0$ then $P(x) \to -\infty$ as $x \to \infty$ and $P(x) \to \infty$ as $x \to -\infty$. In particular, $P(x) < 0$ when x is large and positive and $P(x) > 0$ when x is large and negative.

Because $P(x)$ takes on both positive and negative values, the IVT (4.1) tells us that it must be zero at some point.

If P has even degree, this need not be true: $P(x) = x^2 + 1$, with the property $P(x) \geq 1$ for all $x \in \mathbb{R}$, is a simple example. Examples like this lead us into *complex numbers* and the *Fundamental Theorem of Algebra*. This is all discussed in the companion Algebra module.

5 Complex Numbers

Let us now briefly consider what happens for complex-valued functions (where *complex-valued* means that the codomain is $\mathbb{C}$; the domain remains a subset of $\mathbb{R}$). The complex numbers cannot be ordered (there is no concept of which of two complex numbers is greater or smaller) so we have the following.

Everything we have done so far works for complex-valued functions as well as for real-valued functions, except in situations where the order relation is involved.

- Limits: copy the "real" definition and interpret "$f(t)$ is close to L" in the complex plane, i.e. $|f(t)-L|$ is small, where $|\cdot|$ is the complex modulus.
- Or use real and imaginary parts: if $z(t)=x(t)+\mathrm{i}y(t)$ and $L=x_0+\mathrm{i}y_0$ (x and y real-valued functions, $x_0, y_0 \in \mathbb{R}$) then

$$z(t)\to L \text{ as } t\to t_0 \iff x(t)\to x_0 \text{ and } y(t)\to y_0 \text{ as } t\to t_0$$

- Derivatives: use the complex limit to define the derivative.

$$z'(t)=\lim_{h\to 0}\frac{z(t+h)-z(t)}{h}$$

- Or use real and imaginary parts:

$$z'(t)=(\operatorname{Re} z)'(t)+\mathrm{i}(\operatorname{Im} z)'(t)$$

Example:

$$\lim_{t\to\infty}\frac{t^2+(\mathrm{i}-2)t+1}{2\mathrm{i}t^2-\mathrm{i}t+3}=\lim_{t\to\infty}\frac{1+(\mathrm{i}-2)/t+1/t^2}{2\mathrm{i}-\mathrm{i}/t+3/t^2}=\frac{1}{2\mathrm{i}}=-\frac{\mathrm{i}}{2}$$

For another example, we can define a square root function on the whole of $\mathbb{R}$ by

$$f(x)=\begin{cases} x^{1/2} & \text{if } x\geq 0\\ (-x)^{1/2}\mathrm{i} & \text{if } x<0\end{cases}$$

(this formula involves no square roots of negative numbers except for i itself; interpret the half-powers as non-negative as usual). Because differentiation is localised (3.4) and we know how to differentiate rational powers, we have

$$f'(x)=\begin{cases} \dfrac{x^{-1/2}}{2} & \text{if } x>0\\ \dfrac{-(-x)^{-1/2}\mathrm{i}}{2} & \text{if } x<0\end{cases}$$

f is not differentiable at 0 because

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

so the limit does not exist.

Some concepts do not carry over to complex numbers, because they involve the order relation and the complex numbers cannot be ordered.

Some examples of concepts that do not carry over to complex numbers are:

- Tending to $\pm\infty$ (but we can say $|f(t)| \to \infty$ as $t \to t_0$ if f is complex-valued).
- The Intermediate Value Theorem.
- The Mean Value Theorem.
- Monotonicity (increasing and decreasing properties).
- Extrema (maxima and minima) of any kind.

The theory of limits and differentiability of functions defined on domains in $\mathbb{C}$ is a different story: much of the theory works but, in some cases, in a significantly different way. This is covered in the second-year Functions of a Complex Variable module.

6 Elementary Functions

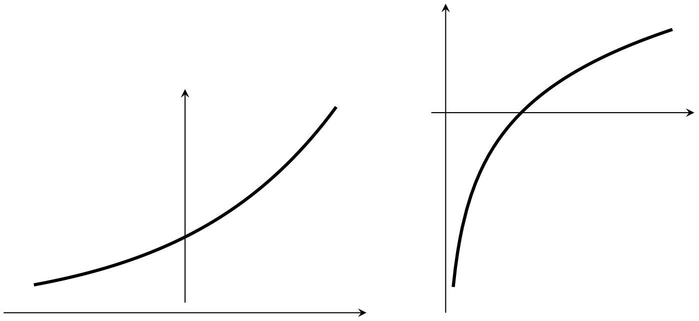
6.1 The exponential function and the logarithm

Fact (proof omitted): there is a unique function $f : \mathbb{R} \to \mathbb{R}$ such that $f' = f$ and $f(0) = 1$. We define exp, *the exponential function*, to be this unique function. So $\exp : \mathbb{R} \to \mathbb{R}$ is defined by the properties that $\exp' = \exp$ and $\exp(0) = 1$

Any function $f : \mathbb{R} \to \mathbb{C}$ with $f' = f$ is a constant multiple of exp. We can use this to derive some properties of the exponential function:

- $\exp(x+y) = \exp(x)\exp(y)$ for any $x, y \in \mathbb{R}$.
- $\exp(x) > 0$ for all $x \in \mathbb{R}$ and exp is strictly increasing, tends to 0 at $-\infty$ and tends to ∞ at $+\infty$.
- If we define $\mathrm{e} = \exp(1)$ then we have $\exp(q) = \mathrm{e}^q$ for any $q \in \mathbb{Q}$.
- There is an inverse function $\log : (0, \infty) \to \mathbb{R}$ such that $\log(\exp(x)) = x$ and $\exp(\log(y)) = y$ for all $x \in \mathbb{R}$ and $y \in (0, \infty)$.
- log is differentiable on $(0, \infty)$ and $\log'(y) = 1/y$.
- $\log(1) = 0$ and $\log(ab) = \log(a) + \log(b)$ for $a, b \in (0, \infty)$.

Graphs of the exponential and logarithm functions:



Let us look at how these properties are derived.

- If f satisfies $f' = f$ and $f(0) \neq 0$ then $f(x) = A\exp(x)$ for some $A \in \mathbb{R}$, so, evaluating at $x = 0$, we get $A = f(0)$, i.e. $f(x) = f(0)\exp(x)$.

- Fix $y \in \mathbb{R}$ and let $f(x) = \exp(x+y)$. Then $f'(x) = \exp(x+y) = f(x)$ so $f(x) = f(0)\exp(x)$, i.e. $\exp(x+y) = \exp(x)\exp(y)$.
- Because $\exp(x)\exp(-x) = \exp(0) = 1$, $\exp(x) \neq 0$. By the IVT and $\exp(0) = 1 > 0$, $\exp(x) > 0$ for all x. Because $\exp'(x) = \exp(x) > 0$ for all $x \in \mathbb{R}$, exp is strictly increasing.
- Because $\exp(0) = 1$ and exp is increasing, $\exp(x) > \exp(0) = 1$ for all $x > 0$. For $x > 0$, the MVT gives $\exp(x) = \exp(0) + x\exp'(x_0) > 1 + x$ (for some $x_0 \in (0, x)$) so $\exp(x) \to \infty$ as $x \to \infty$. Because $\exp(-x) = 1/\exp(x)$, $\exp(x) \to 0$ as $x \to -\infty$.
- If $q \in \mathbb{N}$ then we have

  $$\exp(\underbrace{1 + 1 + \cdots + 1}_{q \text{ terms}}) = \underbrace{\exp(1)\exp(1)\ldots\exp(1)}_{q \text{ terms}} = \mathrm{e}^q$$

  so $\exp(q) = \mathrm{e}^q$ for $q \in \mathbb{N}$ (where e^q means repeated multiplication and $\exp(q)$ is a value of the function described above). Similar arguments show that $\exp(q) = \mathrm{e}^q$ for $q \in \mathbb{Z}$ and finally $q \in \mathbb{Q}$.
- By the Inverse Function Theorem for differentiable functions (4.6)

  $$(\log')(y) = \frac{1}{\exp'(\log(y))} = \frac{1}{y}$$

- Because $\exp(0) = 1$, $\log(1) = 0$. If $a, b > 0$ then

  $$\exp[\log(a) + \log(b)] = \exp(\log(a))\exp(\log(b)) = ab.$$

  Take logarithms to give $\log(a) + \log(b) = \log(ab)$.

Note that, in Mathematics, we usually use log to denote the “natural” logarithm described above. The symbol ln is also commonly used. In other subject areas, log might be used for the logarithm to another base, most commonly 10.

6.2 Arbitrary powers

We previously saw how to define rational powers of a real number. Now we can use the exponential function to define arbitrary powers of a real number.

For $x \in (0, \infty)$ and $y \in \mathbb{R}$, we define

$$x^y = \exp(y \log x)$$

We also adopt the conventions that $0^y = 0$ for $y > 0$ and $0^0 = 1$.

This agrees with, and extends, the earlier definition of x^q for $q \in \mathbb{Q}$. We have the identities

$$\log(x^y) = y\log(x) \quad (x > 0) \qquad (x^y)^z = x^{yz} \qquad x^{y+z} = x^y x^z$$

The derivatives with respect to x and y of x^y are

$$\frac{\mathrm{d}}{\mathrm{d}x}x^y = yx^{y-1} \qquad\qquad \frac{\mathrm{d}}{\mathrm{d}y}x^y = \log(x)x^y$$

To see why this works, note first that the identity $\log(x^y) = y\log(x)$ is immediate from the definition of x^y. We can now calculate:

$$(x^y)^z = \exp(z\log(x^y)) = \exp(zy\log(x)) = x^{yz}$$

and:

$$x^{y+z} = \exp((y+z)\log(x)) = \exp(y\log(x) + z\log(x)) = \exp(y\log(x))\exp(z\log(x)) = x^y x^z$$

using the functional equation for exp.

To find the derivatives:

$$\frac{\mathrm{d}}{\mathrm{d}x}x^y = \frac{\mathrm{d}}{\mathrm{d}x}\exp(y\log x) = \exp(y\log x)\frac{y}{x} = x^y\frac{y}{x} = yx^{y-1}$$

$$\frac{\mathrm{d}}{\mathrm{d}y}x^y = \frac{\mathrm{d}}{\mathrm{d}y}\exp(y\log x) = \log(x)\exp(y\log x) = \log(x)x^y$$

Note that there is a tension between the identities $0^y = 0$ $(y > 0)$ and $x^0 = 1$ $(x > 0)$. The convention $0^0 = 1$ is the most common resolution, but some authors prefer to leave 0^0 undefined.

6.3 An important formula for $\exp(x)$

Earlier (2.4) we saw some numerical evidence suggesting that

$$\left(1 + \frac{1}{n}\right)^n \to \mathrm{e}$$

as $n \to \infty$ $(n \in \mathbb{N})$. We can now show that for any $x \in \mathbb{R}$

$$\boxed{\left(1 + \frac{x}{y}\right)^y \to \exp(x) \text{ as } y \to \infty}$$

Notice that $(1+x/y)^y$ is only defined if $1+x/y>0$, but this is not a problem because we are letting y tend to infinity. The calculation goes as follows (using the local linear approximation at 1):

$$\log\left[\left(1+\frac{x}{y}\right)^y\right]=y\log\left(1+\frac{x}{y}\right)=y\left(\log(1)+\Big(\frac{x}{y}\Big)\log'(1)+r\Big(\frac{x}{y}\Big)\right)$$

where $r(h)/h\to 0$ as $h\to 0$. We have $\log(1)=0$ and $\log'(1)=1$ so

$$\log\left[\left(1+\frac{x}{y}\right)^y\right]=y\Big(\frac{x}{y}+r\Big(\frac{x}{y}\Big)\Big)=x+yr\Big(\frac{x}{y}\Big)=x+x\cdot\frac{y}{x}\cdot r\Big(\frac{x}{y}\Big)=x+x\frac{r(x/y)}{x/y}\to x$$

as $y\to\infty$ (since as $y\to\infty$, $x/y\to 0$, so $(r(x/y))/(x/y)\to 0$). Now, use the fact that exp is continuous to move from

$$\log\left[\left(1+\frac{x}{y}\right)^y\right]\to x \text{ as } y\to\infty$$

to

$$\left(1+\frac{x}{y}\right)^y\to\exp(x) \text{ as } y\to\infty$$

6.4 The circular functions

Fact (proof omitted): there is a unique function $f:\mathbb{R}\to R$ such that $f''=-f$ and $f(0)=0$ and $f'(0)=1$. We define the *sine function*, sin, to be this unique function. So $\sin:\mathbb{R}\to\mathbb{R}$ is defined by the properties that $\sin''=-\sin$, $\sin(0)=0$ and $\sin'(0)=1$. Next we define the *cosine function*, $\cos:\mathbb{R}\to R$, as the derivative of the sine function, so $\cos=\sin'$, and $\cos'=-\sin$.

Any solution of $u''=-u$ satisfies $u(x)=A\cos(x)+B\sin(x)$.

Some consequences of this are:

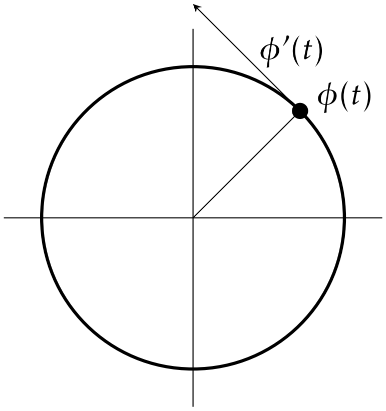
- sin is an odd function and cos is an even function.
- If $x,y\in\mathbb{R}$ then

$$\sin(x+y)=\sin(x)\cos(y)+\cos(x)\sin(y)$$
$$\cos(x+y)=\cos(x)\cos(y)-\sin(x)\sin(y)$$

- If $x\in\mathbb{R}$ then $\cos^2(x)+\sin^2(x)=1$
- Define π to be the smallest number such that $\pi>0$ and $\sin(\pi)=0$. Then sin and cos are 2π-periodic (and 2π is the minimal period).
- Other trigonometric functions can now be defined: $\tan=\sin/\cos$, etc.

Why are these the trigonometric functions we know so well? Because they parameterise the unit circle at constant speed and with period 2π.

If we let $\phi(t) = (\cos(t), \sin(t))$, then $\phi'(t) = (-\sin(t), \cos(t))$. Because $\cos^2 + \sin^2 = 1$, we can see that $|\phi(t)| = 1$, so $\phi(t)$ lies on the unit circle, and $|\phi'(t)| = 1$, so ϕ has constant speed. We can also note that $\phi'(t) \cdot \phi(t) = 0$ (where $\cdot$ denotes the dot or scalar product of vectors), so the velocity vector at t is at right angles to the position vector (relative to the origin).



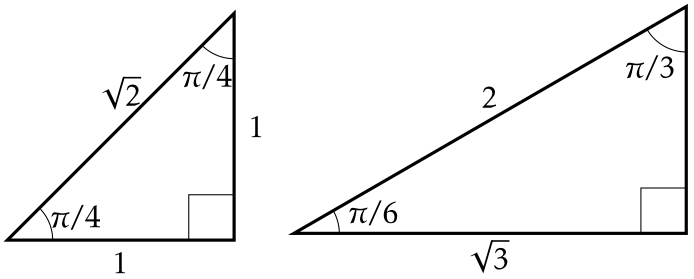
These facts can all be derived from the differential equations, in a similar way to the basic facts about the exponential function (6.1). The exercise is much more fiddly, though, so the details are an optional extra (A.4).

6.5 Special values of circular functions

We can deduce that:

θ	$\sin(\theta)$	$\cos(\theta)$	$\tan(\theta)$
0	0	1	0
$\frac{\pi}{6}$	$\frac{1}{2}$	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{3}}{3}$
$\frac{\pi}{4}$	$\frac{\sqrt{2}}{2}$	$\frac{\sqrt{2}}{2}$	1
$\frac{\pi}{3}$	$\frac{\sqrt{3}}{2}$	$\frac{1}{2}$	$\sqrt{3}$
$\frac{\pi}{2}$	1	0	undefined

These can be visualised on two standard triangles:



Other values can be deduced from the translation rules (which follow from the addition formulae):

$$\sin(x+\pi/2) = \cos(x); \qquad \cos(x+\pi/2) = -\sin(x)$$
$$\sin(x+\pi) = -\sin(x); \qquad \cos(x+\pi) = -\cos(x)$$

We also have the sign diagram

S	A
T	C

which can be remembered by the mnemonic “all stations to Crewe”. It shows in which quadrants All, Sin, Tan and Cosine are positive. Apart from in the All quadrant, the two functions not mentioned are negative (with various functions zero at the boundaries).

Graphs of sin, cos and tan: Note that tan is undefined at $(k+1/2)\pi$ $(k \in \mathbb{Z})$, where it has left and right limits $+\infty$ and $-\infty$, respectively.

None of this is really difficult to derive, but it is rather fiddly. The details for the interested reader are in (A.5).

6.6 Useful approximations for sin and cos near the origin

Because $\sin(0) = 0$ and $\sin'(0) = \cos(0) = 1$, we can write

$$\sin(0+h) = \sin(0) + h \cdot 1 + r(h) \implies \sin(h) = h + r(h)$$

where $r(h)/h \to 0$ as $h \to 0$. Dividing by h, we see that

$$\boxed{\frac{\sin(h)}{h} \to 1 \text{ as } h \to 0}$$

More crudely, $\sin(h) \approx h$ when h is small.

We have already seen (in (4.7)) that $(1-h)^{1/2} \approx 1 - h/2$ when h is small; since $\cos(h) = (1-\sin^2(h))^{1/2}$, we also have, informally, $\cos(h) \approx 1 - h^2/2$ if h is small.

6.7 Inverse Trigonometric Functions

We have $\sin(-\pi/2) = -1$, $\sin(\pi/2) = 1$ and $\sin'(x) = \cos(x) > 0$ on $(-\pi/2, \pi/2)$. So sin is strictly increasing on $[-\pi/2, \pi/2]$ and the Inverse Function Theorem allows us to define an inverse function $\sin^{-1} : [-1, 1] \to [-\pi/2, \pi/2]$. This function is also commonly called arcsin.

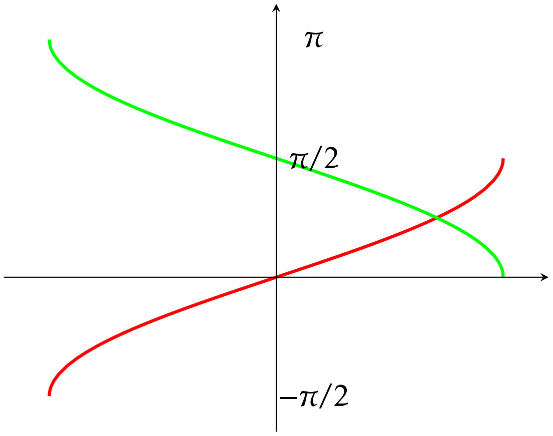
A word of warning: when we use the $\sin^{-1}$ notation, we mean the inverse of the sine function, not the reciprocal of it: $\sin^{-1}(x) \neq (\sin(x))^{-1} = 1/(\sin(x)) = \operatorname{cosec}(x)$.

For any $y \in [-1, 1]$, $x = \sin^{-1}(y)$ is a solution to the equation $\sin(x) = y$, but not the only solution: the others are $\sin^{-1}(y) + 2k\pi$ and $(2k-1)\pi - \sin^{-1}(y) \quad (k \in \mathbb{Z})$. We call this particular solution the *principal value* of the inverse sine and the function assigning this value the *principal branch* of the inverse sine function.

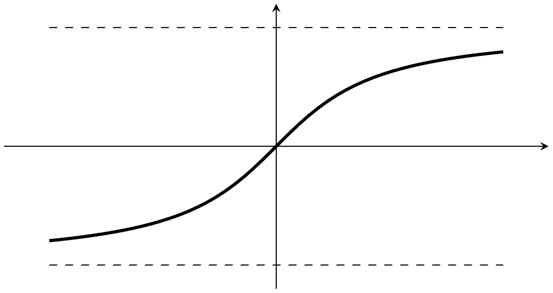
Similarly, $\cos(0) = 1$, $\cos(\pi) = -1$ and $\cos'(x) = -\sin(x) < 0$ on $(0, \pi)$. We can therefore define an inverse function $\cos^{-1} : [-1, 1] \to [0, \pi]$ called the *principal branch* of the inverse cosine. This function is also commonly called arccos.

tan is strictly increasing on $(-\pi/2, \pi/2)$ and tends to $\pm\infty$ at $\pm\pi/2$. The principal branch of the inverse tangent is the inverse mapping $\tan^{-1} : \mathbb{R} \to (-\pi/2, \pi/2)$, also commonly called arctan. The other inverse tangents of y are $\tan^{-1}(y) + k\pi$ $(k \in \mathbb{Z})$.

Graphs of $\sin^{-1}$, $\cos^{-1}$ on $[-1, 1]$:



Graph of $\tan^{-1}$ on $\mathbb{R}$:



As $x \to \pm\infty$, $\tan^{-1}(x) \to \pm\pi/2$.

The derivatives of the principal branches of the inverse trigonometric functions are as follows:

$$\boxed{\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x}\sin^{-1}(x) &= \frac{1}{\sqrt{1-x^2}} \\ \frac{\mathrm{d}}{\mathrm{d}x}\cos^{-1}(x) &= \frac{-1}{\sqrt{1-x^2}} \\ \frac{\mathrm{d}}{\mathrm{d}x}\tan^{-1}(x) &= \frac{1}{1+x^2} \end{aligned}}$$

We calculate these using the Inverse Function Theorem for differentiable functions (4.6)

$$\begin{aligned} & y = \sin^{-1}(x) \\ \Longrightarrow\quad & \sin(y) = x \\ \Longrightarrow\quad & \cos(y) = \frac{\mathrm{d}x}{\mathrm{d}y} \\ \Longrightarrow\quad & \frac{\mathrm{d}x}{\mathrm{d}y} = \pm\sqrt{1-(\sin(y))^2} = \pm\sqrt{1-x^2} \\ \Longrightarrow\quad & \frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{\pm\sqrt{1-x^2}} \qquad \text{by the IFT} \end{aligned}$$

The sign on the square root depends on which branch of the inverse sine we use. The principal branch of the inverse sine, $\sin^{-1}$, increases from $-\pi/2$ to $\pi/2$ on $[-1,1]$, so the derivative is non-negative and we should use the non-negative square root.

The calculation for $\cos^{-1}$ is almost identical. The principal branch of the inverse cosine, $\cos^{-1}$, decreases from π to 0 on $[-1,1]$, so the derivative is negative and we should use the negative square root.

If we use a different branch, we might have a different sign. For example, we could define $f(y)$ to be the unique $x \in [\pi/2, 3\pi/2]$ such that $\sin(x) = y$; this branch of the inverse sine function is decreasing, so the derivative is $-1/\sqrt{1-x^2}$.

Is it surprising that $\sin^{-1}$ and $\cos^{-1}$ have the same derivative? Not really: $\sin(x+\pi/2) = \cos(x)$, so $\sin^{-1}$ and $\cos^{-1}$ differ by a constant.

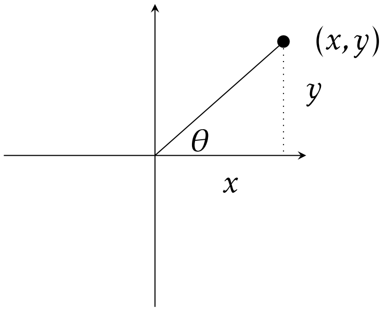
The inverse tangent calculation is similar.

6.8 Radian Measure

We have now, in effect, described the radian measure of an angle. Suppose $(x,y) \in \mathbb{R}^2$ and $(x,y) \neq (0,0)$. Then there is a unique solution to the equations

$$x = r\cos(\theta), \qquad y = r\sin(\theta)$$

for $\theta \in (-\pi, \pi]$ and $r > 0$.



We can describe this explicitly by $r = \sqrt{x^2+y^2} > 0$ and

$$\theta = \begin{cases} \tan^{-1}(y/x) & \text{if } x > 0 \\ \pi/2 & \text{if } x = 0, y > 0 \\ \tan^{-1}(y/x) + \pi & \text{if } x < 0, y \geq 0 \\ \tan^{-1}(y/x) - \pi & \text{if } x < 0, y < 0 \\ -\pi/2 & \text{if } x = 0, y < 0 \end{cases}$$

where $\tan^{-1} : \mathbb{R} \to (-\pi/2, \pi/2)$ is the principal branch. This number θ is the *radian measure* of the angle from the positive x axis anticlockwise to the line joining the origin to (x,y).

Where does this come from? Essentially, it is because the equations

$$x = r\cos\theta, \quad y = r\sin\theta \tag{A}$$

are “almost” equivalent to the equations

$$x^2 + y^2 = r^2, \quad \tan(\theta) = y/x \tag{B}$$

These statements are not quite equivalent, because in writing down $\tan\theta = y/x$ we have lost the individual signs of x and y (this is why many computer programming languages have a two-argument inverse tangent of the form $\tan_2^{-1}(y,x)$, implementing something similar to the table above). The details are worked though in the appendix (A.6).

It is more conventional to describe radian measure in terms of arc length or sector area, but this is equivalent. See below for the connections.

6.9 Complex exponentials and trigonometric functions

Question: Can we put complex numbers into the exponential and trigonometric functions, in such a way that formulae like $\exp(x+y) = \exp(x)\exp(y)$ continue to work?

Answer: Try it and see if it works!

- If $x, y \in \mathbb{R}$, we would like to have $\exp(x+\mathrm{i}y) = \exp(x)\exp(\mathrm{i}y)$.
- We would like to have
$$\frac{\mathrm{d}}{\mathrm{d}y}\exp(\mathrm{i}y) = \mathrm{i}\exp(\mathrm{i}y)$$
- If this works, we should also have
$$\frac{\mathrm{d}^2}{\mathrm{d}y^2}\exp(\mathrm{i}y) = -\exp(\mathrm{i}y)$$
so $u(y) = \mathrm{e}^{\mathrm{i}y}$ should be a solution to the differential equation $u'' = -u$. This is the equation that defines the circular functions, so we should have
$$\mathrm{e}^{\mathrm{i}y} = A\cos(y) + B\sin(y)$$
for some $A, B \in \mathbb{C}$
- We can find A and B as follows: substitute $y = 0$ to give $1 = A$; differentiate and substitute $y = 0$ to give $\mathrm{i} = B$.

Conclusion: If exp is to work for complex arguments in the same way as real arguments, we must have
$$\mathrm{e}^{x+\mathrm{i}y} = \exp(x)(\cos(y) + \mathrm{i}\sin(y))$$

Now we make the definition:

For $x, y \in \mathbb{R}$, *define*
$$\exp(x+\mathrm{i}y) = \exp(x)\,(\cos(y) + \mathrm{i}\sin(y))$$
and write e^z for $\exp(z)$ where convenient.

We need to check that this definition works properly. The main point is to check that $\exp(z+w) = \exp(z)\exp(w)$.

If $z = x + \mathrm{i}y$ and $w = u + \mathrm{i}v$ $(u, v, x, y \in \mathbb{R})$ then

$$\begin{aligned}\exp(z)\exp(w) &= \mathrm{e}^{x}\,[\cos(y) + \mathrm{i}\sin(y)]\,\mathrm{e}^{u}[\cos(v) + \mathrm{i}\sin(v)] \\ &= \mathrm{e}^{x+u}\,[\cos(y)\cos(v) - \sin(y)\sin(v) + \mathrm{i}(\cos(y)\sin(v) + \sin(y)\cos(v))] \\ &= \mathrm{e}^{x+u}\,[\cos(v + y) + \mathrm{i}\sin(v + y)] \\ &= \exp(z + w)\end{aligned}$$

so this complex function really does work as an exponential function should.

Note that there are many other ways to describe $\exp(z)$, e.g. Taylor series or

$$\lim_{n\to\infty}\left(1 + \frac{z}{n}\right)^{n}$$

Which one counts as the definition is a matter of choice and convenience.

We often encounter e^{it} for $t \in \mathbb{R}$. The following are worth remembering:

$$\mathrm{e}^{\mathrm{i}\pi} = -1 \qquad \mathrm{e}^{2\pi\mathrm{i}} = 1 \qquad \mathrm{e}^{z+2k\pi\mathrm{i}} = \mathrm{e}^{z} \quad (k \in \mathbb{Z})$$

$$1/\mathrm{e}^{it} = \mathrm{e}^{-\mathrm{i}t} = \overline{\mathrm{e}^{\mathrm{i}t}} \qquad |\mathrm{e}^{\mathrm{i}t}| = 1$$

Next, we observe that for $\theta \in \mathbb{R}$ we have

$$\begin{aligned}\mathrm{e}^{\mathrm{i}\theta} &= \cos(\theta) + \mathrm{i}\sin(\theta) \\ \mathrm{e}^{-\mathrm{i}\theta} &= \cos(\theta) - \mathrm{i}\sin(\theta)\end{aligned}$$

and we can solve these to give

$$\begin{aligned}\cos(\theta) &= \frac{1}{2}(\mathrm{e}^{\mathrm{i}\theta} + \mathrm{e}^{-\mathrm{i}\theta}) \\ \sin(\theta) &= \frac{1}{2\mathrm{i}}(\mathrm{e}^{\mathrm{i}\theta} - \mathrm{e}^{-\mathrm{i}\theta})\end{aligned}$$

Now *define* for $z \in \mathbb{C}$

$$\begin{aligned}\cos(z) &= \frac{1}{2}(\mathrm{e}^{\mathrm{i}z} + \mathrm{e}^{-\mathrm{i}z}) \\ \sin(z) &= \frac{1}{2\mathrm{i}}(\mathrm{e}^{\mathrm{i}z} - \mathrm{e}^{-\mathrm{i}z})\end{aligned}$$

From sin and cos, we define all the other trigonometric functions on $\mathbb{C}$: $\tan(z) = \sin(z)/\cos(z)$, $\cot(z) = 1/\tan(z)$, $\sec(z) = 1/\cos(z)$, $\mathrm{cosec}(z) = 1/\sin(z)$.

All the basic identities for trigonometric functions still work with complex arguments: $\cos^2(z) + \sin^2(z) = 1$, $\sin(z + w) = \sin(z)\cos(w) + \cos(z)\sin(w)$, $\cos(z + w) = \cos(z)\cos(w) -$

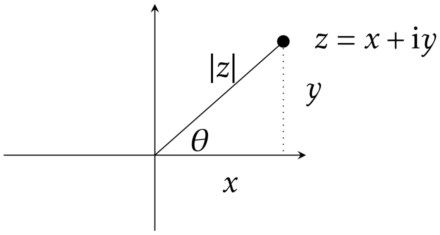
$\sin(z)\sin(w)$, etc. The easiest way to prove this, at this stage, is just to use the definitions. For example,

$$\cos^2(z)+\sin^2(z)=\frac{(e^{iz}+e^{-iz})^2}{4}-\frac{(e^{iz}-e^{-iz})^2}{4}=1$$

You will see in the second-year Functions of a Complex Variable module that any relationship of this type that holds for real numbers must also hold for complex numbers (the Identity Theorem).

For real x, $-1\le\sin(x),\cos(x)\le 1$. For complex z, though, $\sin(z)$ and $\cos(z)$ can take on any complex value.

6.10 Modulus, argument and logarithms of complex numbers



Given $z=x+iy$ with $z\neq 0$, we let $|z|=\sqrt{x^2+y^2}>0$ — this is the modulus of z. We can then (see (6.8)) find θ such that $x=|z|\cos(\theta)$ and $y=|z|\sin(\theta)$; we then have

$$z=|z|(\cos(\theta)+i\sin(\theta))=|z|e^{i\theta}=|z|e^{i(\theta+2k\pi)}$$

for $k\in\mathbb{Z}$. These numbers $\theta+2k\pi$ are called *arguments* of z; the *principal argument* is the one with $-\pi<\theta+2k\pi\le\pi$.

If $z=0$, we can still write $z=|z|e^{i\theta}$ but in this case the argument is completely undefined because this holds for any complex number θ. Numbers written in the form $re^{i\theta}$ are said to be in *polar form*, as opposed to the *Cartesian form* $x+iy$.

If $z\in\mathbb{C}$ with $z\neq 0$, the equation

$$\exp(w)=z$$

has solutions

$$\boxed{w=\log|z|+i(\theta+2k\pi)}$$

where θ is an argument of z and $k\in\mathbb{Z}$. These are the *logarithms* of z. The *principal value* of the logarithm is the one using the principal value of the argument, i.e. $-\pi<\theta+2k\pi\le\pi$.

The equation $\exp(w)=0$ has no solutions in $\mathbb{C}$.

To solve the equation $\exp(w) = z$, we write z and w in Cartesian form as $z = x + \mathrm{i}y$ and $w = u + \mathrm{i}v$, so our equation becomes

$$\mathrm{e}^u(\cos(v) + \mathrm{i}\sin(v)) = x + \mathrm{i}y$$

For this to be true, both sides must have the same modulus, i.e. $\mathrm{e}^u = |z| = \sqrt{x^2 + y^2}$. Because $z \neq 0$, we can uniquely solve this to give $u = \log|z|$ (an ordinary real logarithm). We then have to solve $\cos(v)+\mathrm{i}\sin(v) = z/|z|$, which we have already done: v is an argument of z.

The reason that 0 has no logarithms is that $|\mathrm{e}^w| = \mathrm{e}^u > 0$, so the exponential of a complex number is never zero.

Be very careful with complex logarithms! The identity $\exp(\log(z)) = z$ is true for all $z \neq 0$. However, the reverse relation does not necessarily work: the best we can say is $\log(\exp(z)) = z + 2k\pi\mathrm{i}$ for some $k \in \mathbb{Z}$. Similarly, if $w, z \neq 0$ then $\log(zw) = \log(z) + \log(w) + 2k\pi\mathrm{i}$.

6.11 Hyperbolic Functions

We have now defined trigonometric functions on the whole complex plane. On the real axis, these are the functions we are familiar with. What about on the imaginary axis?

If $x \in \mathbb{R}$, then we have

$$\cos(\mathrm{i}x) = \frac{\mathrm{e}^{-x} + \mathrm{e}^{x}}{2} \qquad \sin(\mathrm{i}x) = \frac{\mathrm{e}^{-x} - \mathrm{e}^{x}}{2\mathrm{i}} = \mathrm{i}\frac{\mathrm{e}^{x} - \mathrm{e}^{-x}}{2}$$

We can see here two functions mapping $\mathbb{R}$ to $\mathbb{R}$, which we call the *hyperbolic sine* and the *hyperbolic cosine*:

$$\cosh(x) = \frac{\mathrm{e}^{x} + \mathrm{e}^{-x}}{2} \qquad \sinh(x) = \frac{\mathrm{e}^{x} - \mathrm{e}^{-x}}{2}$$

so $\cosh(x) = \cos(\mathrm{i}x)$ and $\sinh(x) = -\mathrm{i}\sin(\mathrm{i}x)$.

As well as mapping $\mathbb{R}$ to $\mathbb{R}$, we can think of these as mapping $\mathbb{C}$ to $\mathbb{C}$: for $z \in \mathbb{C}$ define

$$\boxed{\cosh(z) = \frac{\mathrm{e}^{z} + \mathrm{e}^{-z}}{2} \qquad \sinh(z) = \frac{\mathrm{e}^{z} - \mathrm{e}^{-z}}{2}}$$

From these, we define the hyperbolic analogues of the other trigonometric functions

$$\tanh(z) = \frac{\sinh(z)}{\cosh(z)} \qquad \operatorname{sech}(z) = \frac{1}{\cosh(z)}$$
$$\operatorname{cosech}(z) = \frac{1}{\sinh(z)} \qquad \coth(z) = \frac{1}{\tanh(z)}$$

These have some points of non-definition, because $\cosh(z) = 0$ when $z = (k+1/2)\pi\mathrm{i} \quad (k \in \mathbb{Z})$ and $\sinh(z) = 0$ when $z = k\pi\mathrm{i} \quad (k \in \mathbb{Z})$.

Properties of trigonometric functions lead automatically to properties of hyperbolic functions. For example, from $\cos^2 + \sin^2 = 1$ we substitute

$$\begin{aligned} \cos^2(\mathrm{i}z) + \sin^2(\mathrm{i}z) &= 1 \\ \implies \cosh^2(z) + (\mathrm{i}\sinh(z))^2 &= 1 \\ \implies \cosh^2(z) - \sinh^2(z) &= 1 \end{aligned}$$

This is why they are called hyperbolic functions: the curve $(\cosh(t), \sinh(t))$ for $t \in \mathbb{R}$ describes one branch of the hyperbola $\{(x,y) \in \mathbb{R} : x^2 - y^2 = 1\}$. Compare with the circular functions: the curve $(\cos(t), \sin(t))$ with $t \in \mathbb{R}$ describes the circle $\{(x,y) \in \mathbb{R} : x^2 + y^2 = 1\}$.

We could also have proved the previous identity directly from the definition:

$$\cosh^2(z) - \sinh^2(z) = \frac{(\mathrm{e}^z + \mathrm{e}^{-z})^2}{4} - \frac{(\mathrm{e}^z - \mathrm{e}^{-z})^2}{4} = 1$$

Either from the definition, or from the corresponding trigonometric identities, we can see that e.g.

$$\begin{aligned} \sinh(x+y) &= \sinh(x)\cosh(y) + \cosh(x)\sinh(y) \\ \cosh(x+y) &= \cosh(x)\cosh(y) + \sinh(x)\sinh(y) \\ \tanh(x+y) &= \frac{\tanh(x) + \tanh(y)}{1 + \tanh(x)\tanh(y)} \end{aligned}$$

We also have

$$\sinh' = \cosh; \qquad \cosh' = \sinh; \qquad \tanh' = \operatorname{sech}^2$$

etc.

If we compare trigonometric and hyperbolic identities, we can see a pattern.

$$\cos^2(z)+\sin^2(z)=1 \qquad \cos(2z)=\cos^2(z)-\sin^2(z)$$
$$\cosh^2(z)-\sinh^2(z)=1 \qquad \cosh(2z)=\cosh^2(z)+\sinh^2(z)$$

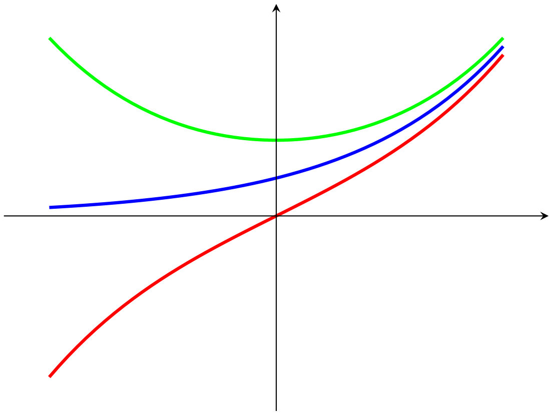
Osborn's rule states that we can convert a valid trigonometric identity to a valid hyperbolic identity by changing all the functions to their hyperbolic equivalents and changing the sign of the product of two (hyperbolic) sine terms. Here we should interpret $\tan = \sin/\cos$, $\cot = \cos/\sin$ and $\operatorname{cosec} = 1/\sin$ as sine terms, because they are defined in terms of a sine.

Don't stretch this rule too far, but it works for the Pythagoras-type identities (sums and differences of squares) and the addition formula for sine, cosine and tangent up to triple-angle. Osborn himself remarks "This rule would fail for terms of the 4th degree, but it covers everything that is likely to be required, and is very convenient for teaching purposes" [Osborn, G. "Mnemonic for Hyperbolic Formulae." Math. Gaz. 2, 189, 1902.]

The reason that Osborn's rule works is that $\cos(\mathrm{i}z)=\cosh(z)$ and $\sin(\mathrm{i}z)=\mathrm{i}\sinh(z)$, so when we replace z by $\mathrm{i}z$ we convert the circular functions to hyperbolic ones, with extra factors of i on the sine terms; these combine to give -1 when two sine terms are multiplied.

6.12 Inverse Hyperbolic Functions

Graphs of sinh, cosh and exp/2 on $\mathbb{R}$ ($\sinh(0)=0$, $\cosh(0)=1$):

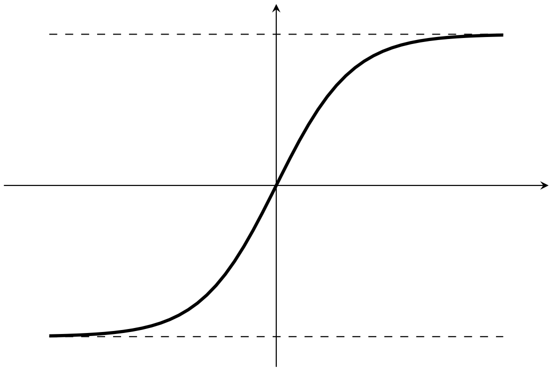


Note that sinh is odd and cosh is even. For all $x \in \mathbb{R}$,

$$\sinh(x) < \frac{\exp(x)}{2} < \cosh(x)$$

For large positive x, the three quantities are very close.

Graph of tanh on $\mathbb{R}$:



As $x \to \pm\infty$, $\tanh(x) \to \pm 1$.

We can check that:

- sinh is a strictly increasing function from $\mathbb{R}$ to $\mathbb{R}$
- cosh is a strictly increasing function from $[0, \infty)$ to $[1, \infty)$
- tanh is a strictly increasing function from $\mathbb{R}$ to $(-1, 1)$

We can therefore define inverse functions:

- $\sinh^{-1} : \mathbb{R} \to \mathbb{R}$
- $\cosh^{-1} : [1, \infty) \to [0, \infty)$
- $\tanh^{-1} : (-1, 1) \to \mathbb{R}$

For sinh and tanh, we find the unique real inverse, but for $y > 1$ there are two real solutions to $\cosh(x) = y$; the inverse function returns the positive solution. The other solution is $-\cosh^{-1}(y)$.

Inverse hyperbolic functions can be expressed in terms of logarithms. For example,

$$\begin{aligned}
\sinh(x) = y &\iff \mathrm{e}^x - \mathrm{e}^{-x} = 2y \\
&\iff \mathrm{e}^{2x} - 2y\mathrm{e}^x - 1 = 0 \\
&\iff \mathrm{e}^x = \frac{2y \pm \sqrt{4y^2 + 4}}{2} \quad \text{(quadratic in } \mathrm{e}^x\text{)} \\
&\iff \mathrm{e}^x = y + \sqrt{1 + y^2} \quad \text{(negative root can't equal } \mathrm{e}^x \text{ for } x \text{ real)} \\
&\iff x = \log(y + \sqrt{1 + y^2})
\end{aligned}$$

We thus have

$$\sinh^{-1}(y) = \log(y + \sqrt{1 + y^2})$$

This formula also works for complex inversion, provided the complex logarithms are properly tracked. All the other inverse hyperbolic functions can be described in terms of logarithms in a similar way. Indeed, complex inverse trigonometric functions can be described in terms of complex logarithms.

We can differentiate inverse hyperbolic functions in much the same way as (6.7) inverse trigonometric functions. The answers are very similar, but have different sign patterns.

$$\boxed{\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\sinh^{-1}(x) &= \frac{1}{\sqrt{1+x^2}}\\ \frac{\mathrm{d}}{\mathrm{d}x}\cosh^{-1}(x) &= \frac{1}{\sqrt{x^2-1}}\\ \frac{\mathrm{d}}{\mathrm{d}x}\tanh^{-1}(x) &= \frac{1}{1-x^2}\end{aligned}}$$

For example,

$$\begin{aligned} y &= \sinh^{-1}(x)\\ \implies \sinh(y) &= x\\ \implies \cosh(y) &= \frac{\mathrm{d}x}{\mathrm{d}y}\\ \implies \sqrt{1+x^2} &= \frac{\mathrm{d}x}{\mathrm{d}y}\\ \implies \frac{\mathrm{d}y}{\mathrm{d}x} &= \frac{1}{\sqrt{1+x^2}}\end{aligned}$$

so

$$\frac{\mathrm{d}}{\mathrm{d}x}\sinh^{-1}(x) = \frac{1}{\sqrt{1+x^2}}$$

For $\sinh^{-1}$ and $\cosh^{-1}$ defined as above, the square roots are non-negative. For the negative branch of $\cosh^{-1}$, the negative square root should be used.

6.13 Hyperbolic functions from their differential equations

We introduced sin and cos as solutions to the differential equation $u'' = -u$. We remark that hyperbolic functions can be thought of as solutions to the equation $u'' = u$. Any solution to that equation has the form

$$u(x) = A\cosh(x) + B\sinh(x)$$

or equivalently

$$u(x) = C\mathrm{e}^{x} + D\mathrm{e}^{-x}$$

6.14 Three Differential Equations

Here are three fundamental differential equations and their general solutions. In the equation, c is a non-zero real number. In the solution, A and B are arbitrary constants.

$$\begin{aligned} u'(x) &= \quad cu(x) & \qquad u(x) &= Ae^{cx} \\ u''(x) &= -c^2u(x) & u(x) &= A\cos(cx) + B\sin(cx) \\ u''(x) &= \quad c^2u(x) & u(x) &= Ae^{cx} + Be^{-cx} \\ & & &= C\cosh(cx) + D\sinh(cx) \end{aligned}$$

Given the corresponding results for $c = 1$ (see (6.1), (6.4), (6.13)), these are easy to deduce. For example, if we have $u'(x) = cu(x)$ then by the chain rule

$$\frac{\mathrm{d}}{\mathrm{d}x}u(\frac{x}{c}) = c\frac{u(x/c)}{c} = u(\frac{x}{c})$$

The function $x \mapsto u(x/c)$ thus satisfies the equation defining the exponential function, so we we have (6.1) $u(x/c) = Ae^x$ for all x and some $A \in \mathbb{R}$. Changing notation to $y = x/c$, we have $u(y) = Ae^{cy}$, as claimed (changing notation like this makes no difference to the function u: the symbols x and y are just part of how we choose to describe the function, not part of the function itself).

Similar calculations work for the other differential equations.

Example: suppose $u''(x) = -4u(x)$, $u(0) = 1$ and $u(\pi/4) = 2$ (this is known as a *boundary value problem* or BVP). From the differential equation, we have

$$u(x) = A\cos(2x) + B\sin(2x)$$

Substituting $x = 0$ and $x = \pi$, we have

$$1 = A; \qquad 2 = B$$

so the solution to the BVP is

$$u(x) = \cos(2x) + 2\sin(2x)$$

7 Integration

In this section, we describe a basic theory of integration for complex-valued functions on real intervals. The fact that they are complex valued makes little difference, except at one point. If you prefer to think of the functions as real-valued, you lose very little.

7.1 Riemann Sums

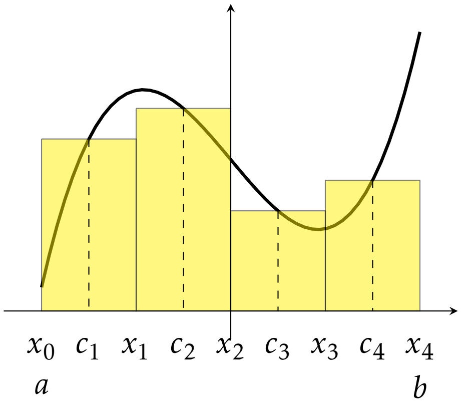
Suppose $f : [a,b] \to \mathbb{C}$. A *Riemann sum* is constructed out of numbers

$$a = x_0 < c_1 < x_1 < c_2 < \cdots < x_{n-1} < c_n < x_n = b$$

(called a *partition* of $[a,b]$) by the formula

$$R = \sum_{j=1}^{n} (x_j - x_{j-1}) f(c_j)$$

If f is real-valued, the Riemann sum can be viewed as the sum of the areas of the rectangles in this diagram:



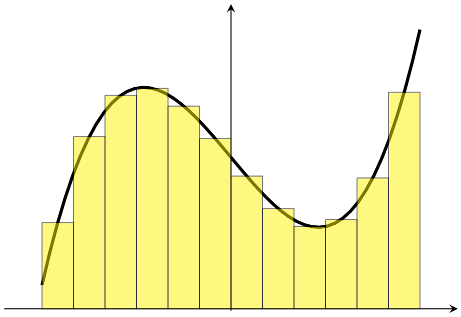
Note that the rectangles do not need to be of same width, and c_i can be any point between x_{i-1} and x_i (not necessarily the midpoint).

An individual Riemann sum can be thought of as an approximation to the "area under the curve."

The *definite integral* $\int_a^b f$ is defined as a limit of Riemann sums, as the *mesh*

$$\max_{j=1}^{n} (x_j - x_{j-1})$$

tends to zero. It represents the exact "area under the curve."



The idea is that, as the rectangles get thinner, the exact choice of c_i in the interval (x_{i-1}, x_i) becomes less significant. So different ways of choosing c_i and different types of partitions lead to the same limit. The exact details of this are quite complicated, so we leave them until the Real Analysis module.

Notation: if $f : [a,b] \to \mathbb{R}$, write

$$\int_a^b f$$

for the definite integral. If y is a variable depending on x, write

$$\int_a^b y \,\mathrm{d}x$$

Or combine them:

$$\int_a^b f(x)\,\mathrm{d}x$$

But don't write

$$\int_a^b f\,\mathrm{d}x \quad \text{✗}$$

for a function f or

$$\int_a^b y \quad \text{✗}$$

for y depending on x.

7.2 Basic Properties of the Definite Integral

The definite integral has the following basic properties, which follow from corresponding properties of the Riemann sums (in the Real Analysis module, these will be proved from the precise definition of the Riemann integral).

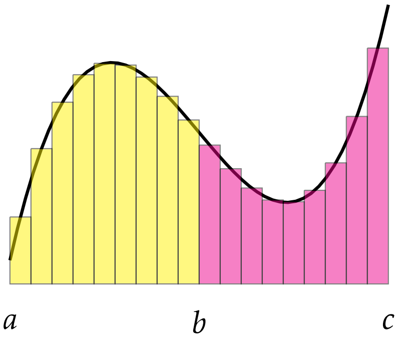
- Scope: if $f : [a,b] \to \mathbb{C}$ is bounded (i.e. for some $M \geq 0$ we have $|f(x)| \leq M$ for all $x \in [a,b]$) and has only finitely many points of discontinuity then f is integrable (i.e. $\int_a^b f$ is defined).

- Linearity: if $\lambda, \mu \in \mathbb{R}$ then $\displaystyle\int_a^b (\lambda f + \mu g) = \lambda \int_a^b f + \mu \int_a^b g$
- Positivity: if $f \geq 0$ (i.e. takes real, non-negative values) on $[a,b]$ then $\int_a^b f \geq 0$
- Additivity: if $a < b < c$ then $\displaystyle\int_a^c f = \int_a^b f + \int_b^c f$
- Normalisation: $\displaystyle\int_a^b c = (b-a)c$
- Triangle inequality: $|\int_a^b f| \leq \int_a^b |f|$

There are some implicit hypotheses here. For positivity, f must be an integrable function. The identity in "linearity" should be taken to mean that if the RHS is defined then so is the LHS and the two sides are equal. The additivity property means that if *either* side of the identity is defined, then so is the other and the two sides are equal.

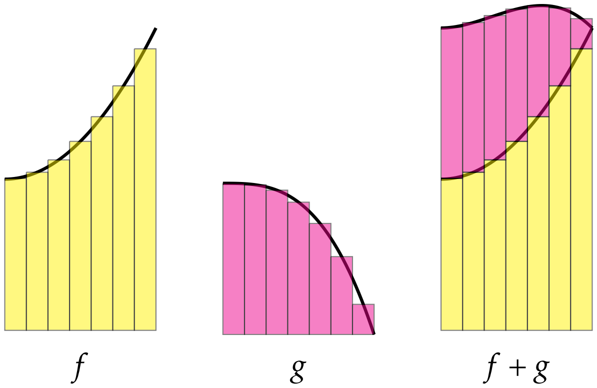
An illustration of the additivity property for Riemann sums: in the limit, this gives the identity

$$\int_a^b f + \int_b^c f = \int_a^c f$$



An illustration of the linearity property for Riemann sums: in the limit, this gives the identity

$$\int_a^b f + \int_a^b g = \int_a^b (f+g)$$



7.3 Generalised additivity property

If $a > b$, we define

$$\int_a^b f = -\int_b^a f$$

We also define

$$\int_a^a f = 0$$

This leads to the identity

$$\int_a^c f = \int_a^b f + \int_b^c f$$

working for *all* $a, b, c \in \mathbb{R}$, not just those with $a < b < c$. It also follows that the value of f at one point (and, by induction, at finitely many points) has no effect on the integral. This sometimes allows us to integrate a function that is not defined at finitely many points in $[a, b]$: set it to be, say, zero on those points and integrate the resulting function.

Verifying that the additivity identity works for all a, b, c is rather tedious. Firstly, we should check that the original identity

$$\int_a^c f = \int_a^b f + \int_b^c f$$

works in the cases $a = b < c$ or $a < b = c$ and $a = b = c$. For example, if $a = b < c$ then the LHS is $\int_a^c f$ and the RHS is $0 + \int_a^c f$; the other cases are similar. Then, we need to consider six possible cases:

$$a \leq b \leq c \qquad a \leq c \leq b \qquad b \leq a \leq c \qquad b \leq c \leq a \qquad c \leq a \leq b \qquad c \leq b \leq a$$

The first of these we already understand. To establish the identity in the second case, $a \leq c \leq b$, start with

$$\int_a^b f = \int_a^c f + \int_c^b f$$

and rearrange to give

$$\int_a^c f = \int_a^b f - \int_c^b f$$

then exchange the order of limits in the last integral to give

$$\int_a^c f = \int_a^b f + \int_b^c f$$

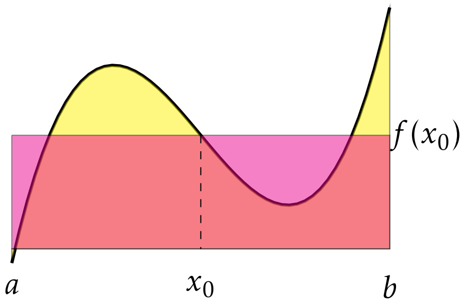
The four other cases are similar.

7.4 The Mean Value Theorem for Integrals (MVTI)

The Mean Value Theorem for Integrals (MVTI) says:

> Suppose f and g are continuous, real-valued functions on $[a,b]$, $g \geq 0$ and g is not identically zero. Then, for some $x_0 \in [a,b]$,
>
> $$\int_a^b fg = f(x_0)\int_a^b g$$
>
> If g is the constant function 1, this reduces to
>
> $$f(x_0) = \frac{1}{b-a}\int_a^b f$$

Illustration of the MVTI with $g = 1$: the area under the straight line, $(b-a)f(x_0)$, is equal to the area under the curve.



By analogy with the formula

$$\frac{1}{N}\sum_{j=1}^{N} a_j$$

we can think of

$$f(x_0) = \frac{1}{b-a}\int_a^b f$$

as the mean, or average, value of f over the interval $[a,b]$.

To see why this works, let $m = \min_{[a,b]} f$ and $M = \max_{[a,b]} f$, so

$$mg(x) \le f(x)g(x) \le Mg(x)$$

for $x \in [a,b]$. Integrating and rearranging, we see that

$$m\int_a^b g \le \int_a^b fg \le M\int_a^b g; \qquad m \le \frac{\int_a^b fg}{\int_a^b g} \le M$$

By the Intermediate Value Theorem (4.1)

$$\frac{\int_a^b fg}{\int_a^b g} = f(x_0)$$

for some $x_0 \in [a,b]$.

If we exchange a and b, then both sides of the MVTI change sign: the result therefore still holds with the limits in reverse order, except that we would have $x_0 \in [b,a]$ instead of $x_0 \in [a,b]$.

7.5 The Fundamental Theorem of Calculus (FTC)

The (first) *Fundamental Theorem of Calculus* (FTC) says that if $f : [a,b] \to \mathbb{C}$ is continuous then:

In functional notation: if we define $F : [a,b] \to \mathbb{R}$ by

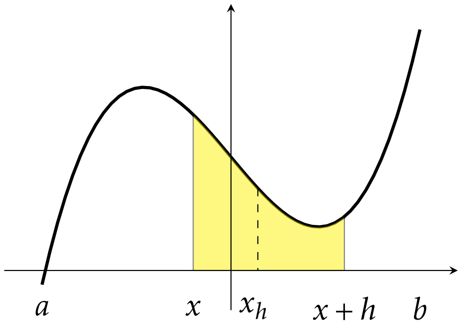
$$F(x) = \int_a^x f$$

then $F' = f$.

In Leibniz notation:

$$\frac{\mathrm{d}}{\mathrm{d}x}\int_a^x f(y)\,\mathrm{d}y = f(x)$$

Illustration of the proof of the (first) FTC:



The area of the shaded region is $F(x+h)-F(x)$, the area under f between x and $x+h$, and we have

$$f(x_h) = \frac{F(x+h)-F(x)}{h}$$

To see why the first FTC holds for real-valued functions, consider the difference quotient:

$$\frac{F(x+h)-F(x)}{h} = \frac{1}{h}\int_x^{x+h} f = f(x_h)$$

for some x_h between x and $x+h$ (by the MVTI). As $h \to 0$, $x_h \to x$ so (because f is continuous) $f(x_h) \to f(x)$. We can now conclude that $F'(x) = f(x)$.

For complex functions, write $f = g + \mathrm{i}h$ where g and h are real-valued and define G and H analogously to F. Then, by linearity of the integral, we have $F = G + \mathrm{i}H$. By linearity of the derivative, $F' = G' + \mathrm{i}H'$. Finally, by the real FTC, $F' = g + \mathrm{i}h = f$.

The (second) Fundamental Theorem of Calculus (FTC) says that

If f is differentiable on $[a,b]$ and f' is continuous on $[a,b]$ then

$$\int_a^b f' = f(b) - f(a)$$

Such a function f is said to be *continuously differentiable* on $[a,b]$. Note that the derivatives at a and b are respectively (3.14) right and left derivatives.

To see why the second FTC works, let

$$g(x) = \int_a^x f'$$

Then $g' = f'$ by the first FTC, so $g = f + c$ for some constant c. Also, $g(a) = 0$ so $c = -f(a)$; this gives $g(x) = f(x) - f(a)$, i.e. $\int_a^x f' = f(x) - f(a)$. Putting $x = b$ gives the second FTC.

The term *Fundamental Theorem of Calculus* (FTC) is often used to refer to either or both of these facts, or to the general idea that integration and differentiation are inverse operations.

The (second) FTC allows us to calculate integrals by using *antiderivatives*: an antiderivative of f is a function F such that $F' = f$. For example, to find

$$\int_0^1 (\sin(x) + \exp(x) + x^3)\,\mathrm{d}x$$

we use our knowledge of derivatives to observe that

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(-\cos(x) + \exp(x) + \frac{x^4}{4}\right) = \sin(x) + \exp(x) + x^3$$

so by the FTC

$$\int_0^1 (\sin(x) + \exp(x) + x^3)\,\mathrm{d}x = \left[-\cos(x) + \exp(x) + \frac{x^4}{4}\right]_{x=0}^{1} = -\cos(1) + \mathrm{e} + \frac{1}{4}$$

Combining the FTC with the Chain Rule allows us to differentiate integrals with variable limits of integration, without necessarily working out the integral. For example,

$$\frac{\mathrm{d}}{\mathrm{d}x}\int_0^x \frac{1}{1+y^4}\,\mathrm{d}y = \frac{1}{1+x^4}$$

is the FTC, whereas

$$\frac{\mathrm{d}}{\mathrm{d}x}\int_0^{x^3} \frac{1}{1+y^4}\,\mathrm{d}y = \frac{3x^2}{1+x^{12}}$$

is a combination of the chain rule and the FTC.

7.6 Indefinite Integrals

So far, we have mostly considered definite integrals: $\int_a^b f$ for specific a and b. An *indefinite integral* of f is a factory for calculating definite integrals of f for arbitrary a and b. Precisely:

An *indefinite integral* F of a function f is a function with the property that

$$\int_a^b f = F(b) - F(a)$$

for all a, b in the domain of f.

Adding a constant C to F does not affect this property, so if F is an indefinite integral of f then so is $F+C$. Less obviously, if F and G are both indefinite integrals of f, then $G = F + C$ for some C.

We sometimes write

$$\int f = F \quad \text{or} \quad \int f = F + C$$

to represent the statement that F is an indefinite integral of f.

We can see why any two indefinite integrals differ by a constant without any actual calculus: if $F(b) - F(a) = G(b) - G(a)$ for all a and b then $F(b) - G(b) = F(a) - G(a)$ for all a and b. Since the LHS depends only on b, the RHS must remain constant as a varies; equally, the LHS must remain constant as b varies. We thus have $F - G = C$ where C is a constant. Given any indefinite integral F of f, we can therefore describe all the indefinite integrals by considering $F + C$, for all constants C. This is the "constant of integration."

Indefinite integrals can be constructed from definite integrals by fixing one endpoint and varying the other.

> If we fix c and let
>
> $$F(x) = \int_c^x f(t)\,\mathrm{d}t$$
>
> then F is an indefinite integral of f.

More usefully for calculating integrals, the FTC says that, for a continuous function f, an indefinite integral is the same thing as an antiderivative:

> If $F' = f$ and f is continuous then
>
> $$\int_a^b f = F(b) - F(a)$$

To see that

$$F(x) = \int_c^x f(t)\,\mathrm{d}t$$

defines an indefinite integral, note that (by the generalised additivity property (7.3))

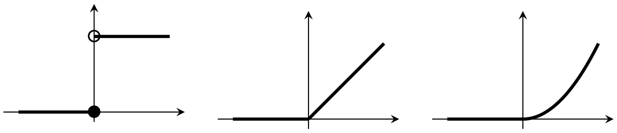
$$\int_a^b f = \int_a^c f + \int_c^b f = \int_c^b f - \int_c^a f = F(b) - F(a)$$

Note that, in this formula, the arbitrary constant of integration has been replaced by an arbitrary starting point c for the definite integral.

Example: For $n \in \mathbb{Z}$, $n \geq 0$, and $x \in \mathbb{R}$ let

$$f_n(x) = \begin{cases} x^n & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

Graphs of f_0, f_1 and f_2 look like:



f_n for larger n looks similar to f_2.

These are clearly continuous except possibly at 0. If $n \geq 1$, then the right and left limits at 0 are both 0, so f is continuous at 0. If $n = 0$, then the left limit is 0 and the right limit is 1. We can also differentiate f away from the origin to give:

$$f_n'(x) = \begin{cases} nx^{n-1} & \text{if } x > 0 \text{ and } n > 0 \\ 0 & \text{if } x < 0 \text{ or } (n = 0 \text{ and } x > 0) \end{cases}$$

At 0, f_0 is not continuous and hence not differentiable. f_1' has a left limit of 0 and a right limit of 1, so f_1 has different left and right derivatives at 0 and is therefore not differentiable at 0. For $n \geq 2$, f_n' has left and right limits at 0 sat 0 so is differentiable with derivative 0 at 0 and f_n' is continuous.

We shall integrate f_n' from -1 to 1. If $n \geq 2$ then, because f_n' is continuous, we satisfy the hypotheses of the FTC and can integrate directly:

$$\int_{-1}^{1} f_n' = f_n(1) - f_n(-1) = 1$$

Now consider f_1. Because f_1' does not exist at 0, we cannot apply the FTC on $[-1,1]$. But, we can break the integral up into two pieces using the additivity property:

$$\int_{-1}^{1} f_1' = \int_{-1}^{0} 0 + \int_{0}^{1} 1$$

The non-existent $f_1'(0)$ has been replaced on the left by $f_1'(0-)$ and on the right by $f_1'(0+)$. The FTC applies to both of these separately so we can integrate to give $\int_{-1}^{1} f_1' = 0 + 1 = 1$.

Finally, we look at f_0. This is not continuous, so the FTC certainly does not apply. But, we can break it into two pieces as before and write to give $\int_{-1}^{1} f_0' = \int_{-1}^{0} 0 + \int_{0}^{1} 0 = 0$

What would have happened if we hadn't noticed the problems at the origin, and incorrectly applied the FTC? For $n = 1$, we would have written down $f_1(1) - f_1(-1) = 1$, which is the right answer. For $n = 0$, we would written down $f_0(1) - f_0(-1) = 1$, which is the wrong answer.

We have three cases here.

1. If $n \geq 2$, the FTC applies and gives the correct answer.
2. If $n = 0$, the FTC does not apply and, if used, gives the wrong answer.
3. If $n = 1$, we're in an awkward middle ground. The FTC as we know it does not apply, but it still gives the right answer. This is not by chance: there are more powerful versions of the FTC, which are applicable to this case.

In this course, we are not going to pursue the best and most powerful FTC. To illustrate the complexity, the easiest version which would cope with f_1 says that if f is continuous and f' has left and right limits at every point and in any bounded interval there are only finitely many points at which f is not differentiable then f is an indefinite integral of f'. The best version of the FTC involves the Lebesgue integral and an idea called "absolute continuity" — developing these ideas would take almost an entire lecture course in its own right.

Moral: if a function behaves in an odd way at a point, it's safest to split the domain at that point and and integrate the pieces separately.

7.7 Integration using multiple-angle formulae

We know that

$$\cos^2(x) + \sin^2(x) = 1 \qquad \cos^2(x) - \sin^2(x) = \cos(2x)$$

From here, we can write

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

leading to

$$\begin{aligned} \int \cos^2(x)\,\mathrm{d}x &= \frac{x}{2} + \frac{\sin(2x)}{4} + C = \frac{x + \sin(x)\cos(x)}{2} + C \\ \int \sin^2(x)\,\mathrm{d}x &= \frac{x}{2} - \frac{\sin(2x)}{4} + C = \frac{x - \sin(x)\cos(x)}{2} + C \end{aligned}$$

Larger powers of sin and cos can be integrated using similar techniques with higher-order multiple-angle formulae, or by the use of (7.10) reduction formulae. The same techniques work for hyperbolic functions.

7.8 The logarithmic derivative trick

By the chain rule,

$$\frac{\mathrm{d}}{\mathrm{d}x}\log(f(x)) = \frac{f'(x)}{f(x)}$$

This is known as the *logarithmic derivative* of f. Used in reverse, it can be a handy shortcut to evaluate some integrals, where the numerator of a fraction is (almost) the derivative of the denominator.

$$\int \frac{f'(x)}{f(x)}\,\mathrm{d}x = \log(f(x)) \quad + C$$

Example:

$$\int \frac{x}{1+x^2}\,\mathrm{d}x = \frac{1}{2}\log(1+x^2) \quad + C$$

Here, we notice that the numerator is (apart from a factor of 2) the derivative of the denominator.

$$\int \tan(x)\,\mathrm{d}x = \int \frac{\sin(x)}{\cos(x)}\,\mathrm{d}x = -\log(\cos(x)) \ + C = \log(\sec(x)) \ + C$$

Here, we notice that the numerator is (apart from a sign) the derivative of the denominator. Similarly, $\int \cot(x)\,\mathrm{d}x = \log(\sin(x)) \ + C$.

The logarithmic derivative trick is equivalent to the substitution (7.11) $y = f(x)$ where $f(x)$ is the denominator, but is quicker and easier to use.

7.9 Integration by Parts

Integration by parts is a powerful technique based on the FTC and the product rule for derivatives. Suppose f is continuously differentiable, g is continuous and $G' = g$. Then $(fG)' = fg + f'G$, Rearranging, $fg = (fG)' - f'G$, and integrating gives

$$\boxed{\int fg = fG - \int f'G}$$

Integrating fg is thus equivalent to integrating $f'G$, where G is an antiderivative of g. If $f'G$ is easier to integrate that fg, this is useful. Any antiderivative of g can be used — any constants of integration can be added on at the end of the process.

Note that the formula is highly asymmetrical between f and g — given an integrand, choosing an appropriate f and g is part of the skill of using the method. Example:

$$\int x\cos(x)\,\mathrm{d}x$$

To integrate by parts, we need to choose which factor to differentiate and which to integrate. If we differentiate x and integrate $\cos(x)$, we have 1 and $\sin(x)$; if we do it the other way round, we have $x^2/2$ and $\cos(x)$. On the basis that there's nothing much to choose between $\sin(x)$ and $\cos(x)$, but 1 is simpler to handle than $x^2/2$, we differentiate x and integrate $\sin(x)$, i.e. apply the formula with $f(x) = x$ and $g(x) = \cos(x)$, so $G(x) = \sin(x)$.

$$\int x\cos(x)\,\mathrm{d}x = x\sin(x) - \int 1 \times \sin(x)\,\mathrm{d}x$$

The remaining integral is easy to evaluate, so we have

$$\int x\cos(x)\,\mathrm{d}x = x\sin(x) - \int 1 \cdot \sin(x)\,\mathrm{d}x = x\sin(x) + \cos(x) \ + C$$

If we had applied the formula the other way round, we would have

$$\int x\cos(x)\,\mathrm{d}x = \cos(x)\frac{x^2}{2} + \int \sin(x)\frac{x^2}{2}\,\mathrm{d}x$$

which is true, but not helpful — the remaining integral is no easier than the one we started with.

Sometimes, we have to apply the procedure more than once.

$$\int x^2\mathrm{e}^x\,\mathrm{d}x = x^2\mathrm{e}^x - \int (2x)\mathrm{e}^x\,\mathrm{d}x = x^2\mathrm{e}^x - 2\int x\mathrm{e}^x\,\mathrm{d}x$$

(integrating e^x, differentiating x^2). We evaluate the remaining integral by using the parts formula again:

$$\int x\mathrm{e}^x\,\mathrm{d}x = x\mathrm{e}^x - \int \mathrm{e}^x\,\mathrm{d}x = x\mathrm{e}^x - \mathrm{e}^x \ + C = (x-1)\mathrm{e}^x \ + C$$

Now we can assemble the integral:

$$\int x^2\mathrm{e}^x\,\mathrm{d}x = (x^2 - 2x + 2)\mathrm{e}^x \ + C$$

In general, if P is a polynomial then any formula of the form

$$P(x)\mathrm{e}^{ax}; \qquad P(x)\sin(ax); \qquad P(x)\cos(ax)$$

or any linear combination of these can be tackled by parts, requiring n applications if the degree of P is n.

We can also integrate terms like $\mathrm{e}^x \sin(x)$ like this.

$$\begin{aligned}
I &= \int \mathrm{e}^x \sin(x) \\
&= -\mathrm{e}^x \cos(x) + \int \mathrm{e}^x \cos(x)\,\mathrm{d}x \\
&= -\mathrm{e}^x \cos(x) + \mathrm{e}^x \sin(x) - \int \mathrm{e}^x \sin(x)\,\mathrm{d}x \\
&= -\mathrm{e}^x \cos(x) + \mathrm{e}^x \sin(x) - I
\end{aligned}$$

After two stages, the calculation comes around almost full circle, and produces an answer in terms of the integral we started with. We can then solve for I to give

$$I = \int \mathrm{e}^x \sin(x) = \frac{\mathrm{e}^x}{2}(\sin(x) - \cos(x)) \ + C$$

We can also read off from the first step and the final answer

$$\int \mathrm{e}^x \cos(x)\,\mathrm{d}x = I + \mathrm{e}^x \cos(x) = \frac{\mathrm{e}^x}{2}(\sin(x) + \cos(x)) \ + C$$

We could also have differentiated the trig term and integrated the exponential one. Be careful to be consistent, though: having started off by differentiating the exponential term and integrating the trig term, we have to do the same in the next stage otherwise we just end up with the equation $I = I$.

For definite integrals we can, of course, work out the indefinite integral and then substitute the limits of integration. Alternatively, we can use a definite version of the formula:

$$\int_a^b fg = [fG]_a^b - \int_a^b f'G$$

For example,

$$\int_0^\pi x\cos(x)\,\mathrm{d}x = [x\sin(x)]_{x=0}^\pi - \int_0^\pi \sin(x)\,\mathrm{d}x = [\cos(x)]_{x=0}^\pi = -2$$

7.10 Reduction Formulae

Suppose f_n is a function depending on a parameter $n \in \mathbb{N}$ or $n \in \mathbb{N} \cup \{0\}$. A *reduction formula* is a formula for $\int f_n$ in terms of $\int f_m$ where m is smaller than n (most commonly $m = n-1$ or $m = n-2$). Repeated use of the formula eventually expresses $\int f_n$ in terms of $\int f_1$ or $\int f_0$, which we hope we can calculate explicitly.

$$I_n = \int x^n \mathrm{e}^x\,\mathrm{d}x \qquad I_n = x^n \mathrm{e}^x - nI_{n-1} \qquad I_0 = e^x$$

$$J_n = \int \sin^n(x)\,\mathrm{d}x \qquad J_0 = x \qquad J_1 = -\cos(x)$$

$$J_n = -\frac{1}{n}\sin^{n-1}(x)\cos(x) + \left(1 - \frac{1}{n}\right)J_{n-2}$$

The calculation for I_n is a direct use of integration by parts, integrating e^x and differentiating x^n. For J_n it is a little more complicated.

$$\begin{aligned} J_n &= \int \sin^n(x)\,\mathrm{d}x \\ &= \int \sin^{n-1}(x)\sin(x)\,\mathrm{d}x \\ &= -\sin^{n-1}(x)\cos(x) + \int (n-1)\sin^{n-2}(x)\cos(x)\cos(x)\,\mathrm{d}x \\ &= -\sin^{n-1}(x)\cos(x) + (n-1)\int \sin^{n-2}(x)(1-\sin^2(x))\,\mathrm{d}x \\ &= -\sin^{n-1}(x)\cos(x) + (n-1)[J_{n-2} - J_n] \end{aligned}$$

Now, we can solve this for J_n to give

$$J_n = -\frac{1}{n}\sin^{n-1}(x)\cos(x) + \left(1 - \frac{1}{n}\right)J_{n-2}$$

7.11 Integration by substitution

Integration by substitution is another powerful technique for integration, based on the chain rule. Suppose $F' = f$, then

$$\frac{\mathrm{d}}{\mathrm{d}y}F(g(y)) = f(g(y))g'(y)$$

Expressed as an integral,

$$\int_\alpha^\beta f(g(y))g'(y)\,\mathrm{d}y = F(g(\beta)) - F(g(\alpha)) = \int_{g(\alpha)}^{g(\beta)} f(x)\,\mathrm{d}x$$

More usefully, letting $a = g(\alpha)$ and $b = g(\beta)$,

$$\int_a^b f(x)\,\mathrm{d}x = \int_{g^{-1}(a)}^{g^{-1}(b)} f(g(y))g'(y)\,\mathrm{d}y$$

In applications, we start with the integral on the left and hope to think of a useful function g such that the integral on the right is easier to evaluate.

Think of this as a procedure:

1. Replace x by $g(y)$.

2. Replace $\mathrm{d}x$ by $g'(y)\,\mathrm{d}y$. In Leibniz notation, if $x = g(y)$ then $\mathrm{d}x/\mathrm{d}y = g'(y)$ so " $\mathrm{d}x = g'(y)\,\mathrm{d}y$".

3. Replace the limits by solving $g(y) = a$ and $g(y) = b$ for y. Ask the question "what value does y have when $x = a$ and $x = b$"?

For example, to evaluate

$$\int_0^1 \frac{1}{\sqrt{1+x^2}}\,\mathrm{d}x$$

we seek a substitution for x that makes the integrand simpler. Among the trigonometric and hyperbolic identities we find

$$\cosh^2 = 1 + \sinh^2$$

If we let $x = \sinh(u)$, then $\mathrm{d}x/\mathrm{d}u = \cosh(u)$ so " $\mathrm{d}x = \cosh(u)\,\mathrm{d}u$". When $x = 0$, $u = 0$ and when $x = 1$, $u = \sinh^{-1}(1)$. The integral therefore becomes

$$\int_0^{\sinh^{-1}(1)} \frac{\cosh(u)}{\cosh(u)}\,\mathrm{d}u = \sinh^{-1}(1) = -\log(\sqrt{2} - 1)$$

We could also try to use the identity $\sec^2 = 1 + \tan^2$ by substituting $u = \tan(x)$. This would give the integral

$$\int_0^{\pi/4} \frac{\sec^2(u)}{\sec(u)}\,\mathrm{d}u = \int_0^{\pi/4} \sec(u)\,\mathrm{d}u$$

This is correct but, unless you know the integral of sec, not helpful!

The corresponding formula for indefinite integrals is

$$\int f(x)\,\mathrm{d}x = \int f(g(y))g'(y)\,\mathrm{d}y$$

where $x = g(y)$. For definite integrals, we have two choices:

1. Use the definite integral formula, which involves solving the equations $g(y) = a, b$.

2. Use the indefinite integral formula, then change y back to x. If the answer happens to be obviously expressed in terms of $g(y)$, this might be easier.

Example:

$$\int \frac{1}{1+x^2}\,\mathrm{d}x$$

We seek a substitution for x that makes $1+x^2$ simpler. In our trigonometric formulae, we find $\sec^2 y = 1+\tan^2 y$ which suggests that $x = \tan(y)$ might be useful. Differentiating, we see that $\mathrm{d}x/\mathrm{d}y = \sec^2(y)$, so we formally put "$\mathrm{d}x = \sec^2(y)\,\mathrm{d}y$." Now assemble the pieces:

$$\int \frac{1}{1+x^2}\,\mathrm{d}x = \int \frac{1}{1+\tan^2(y)}\sec^2(y)\,\mathrm{d}y = \int dy = y = \tan^{-1}(x)\ +C$$

In this case, the new integral turned out to be (much) easier than the old one, and we have our answer. But we can never be sure that a substitution will work. We could also have noted that $1+\sinh^2(y) = \cosh^2(y)$ and tried the substitution $x = \sinh(y)$, so "$\mathrm{d}x = \cosh(y)\,\mathrm{d}y$" and hence

$$\int \frac{1}{1+x^2}\,\mathrm{d}x = \int \frac{1}{1+\sinh^2(y)}\cosh(y)\,\mathrm{d}y = \int \frac{\mathrm{d}y}{\cosh(y)}$$

which, unless you know how to integrate a hyperbolic secant, is not much help! As a spinoff, though, we can use the earlier answer to observe that

$$\int \frac{\mathrm{d}y}{\cosh(y)} = \tan^{-1}(x) = \tan^{-1}(\sinh(y))\ +C$$

Compare this example with the earlier one, integrating $1/\sqrt{1+x^2}$. Here, the trigonometric substitution worked much better than the hyperbolic one; there, the hyperbolic substitution worked better. There is no way to predict this: try a substitution and see what happens. In the same way that this example had a side-effect of integrating sech, the previous example can be used to show that $\int \sec(x)\,\mathrm{d}x = \sinh^{-1}(\tan(x))$. The more familiar formula $\log(\sec(x)+\tan(x))$ can be derived from the formula (6.12) for $\sinh^{-1}$ in terms of logarithms.

Example:

$$\int \sqrt{1-x^2}\,\mathrm{d}x$$

We look for a trigonometric or hyperbolic substitution which will simplify $1-x^2$. The most obvious ones are $1-\cos^2 = \sin^2$ and $1-\sin^2 = \cos^2$; there are also some hyperbolic ones that we could try, but let's try $x = \cos(u)$, so $\mathrm{d}x/\mathrm{d}u = -\sin(u)$ and $\sqrt{1-x^2} = \sin(u)$. Our integral now takes the form

$$-\int \sin^2(u)\,\mathrm{d}u$$

which we can evaluate using the trigonometric identity $\sin^2(u) = (1-\cos(2u))/2$:

$$\frac{1}{2}\int \cos(2u)-1\,\mathrm{d}u = \frac{1}{4}\sin(2u) - \frac{u}{2}$$

Now we put the answer back in terms of x, remembering that $\sin(2u) = 2\sin(u)\cos(u) = 2x\sqrt{1-x^2}$:

$$\frac{1}{2}x\sqrt{1-x^2} - \frac{\cos^{-1}(x)}{2}$$

We could also have substituted $\sin(u) = x$, which would have given

$$\frac{1}{2}x\sqrt{1-x^2} + \frac{\sin^{-1}(x)}{2}$$

These superficially different answers actually differ by a constant, so they are different, but correct, antiderivatives.

7.12 The "multiplication by 1" trick

A few functions can be integrated by the following trick, using integration by parts with a factor of 1.

$$\int f(x)\,\mathrm{d}x = \int f(x)\cdot 1\,\mathrm{d}x = f(x)x - \int f'(x)x\,\mathrm{d}x$$

When does this help? Usually, when $f'(x)$ looks very different from $f(x)$, e.g.

$$\int \log(x)\,\mathrm{d}x = x\log(x) - x \ + C$$
$$\int \cos^{-1}(x)\,\mathrm{d}x = x\cos^{-1}(x) - \sqrt{1-x^2} \ + C$$

All other inverse trigonometric and hyperbolic functions can be handled using this technique.

The calculation for the logarithm is as follows: if $f(x) = \log(x)$ then $f'(x) = 1/x$ and $xf'(x) = 1$. This gives us

$$\int \log(x)dx = x\log(x) - x \ + C$$

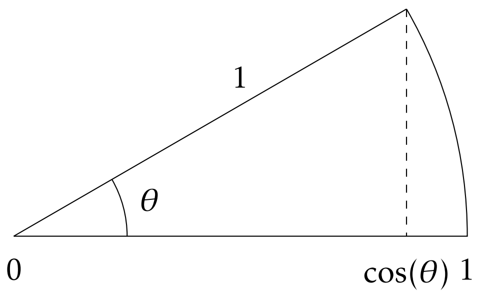
And for $\cos^{-1}$:

$$\begin{aligned}\int \cos^{-1}(x)\,\mathrm{d}x &= x\cos^{-1}(x) + \int \frac{x}{\sqrt{1-x^2}}\,\mathrm{d}x \\ &= x\cos^{-1}(x) - \int \frac{\mathrm{d}u}{2\sqrt{u}} \\ &= x\cos^{-1}(x) - u^{1/2} \\ &= x\cos^{-1}(x) - \sqrt{1-x^2} \ + C\end{aligned}$$

making the substitution $1 - x^2 = u$, "$\mathrm{d}u = 2x\,\mathrm{d}x$".

7.13 Radian Measure, second attempt

What is the area of a sector of a circle? Suppose $0 < \theta < \pi/2$.



Then the area in the sector of the unit circle between angle 0 (x-axis) and angle θ is (additivity property)

$$\int_0^{\cos\theta} \tan(\theta)x\,\mathrm{d}x + \int_{\cos(\theta)}^1 \sqrt{1-x^2}\,\mathrm{d}x = \cdots = \frac{\theta}{2}$$

The first integral is easy to calculate: $\tan(\theta)$ comes out as a factor and the integral is

$$\tan(\theta)\left[\frac{x^2}{2}\right]_{x=0}^{\cos(\theta)} = \frac{1}{2}\tan(\theta)\cos^2(\theta) = \frac{1}{2}\sin(\theta)\cos(\theta)$$

The second one is actually an example we have seen before (7.11):

$$\int \sqrt{1-x^2}\,\mathrm{d}x = \frac{1}{2}x\sqrt{1-x^2} - \frac{\cos^{-1}(x)}{2}$$

At $x = 1$, this gives 0. At $x = \cos(\theta)$, it gives

$$\frac{1}{2}\cos(\theta)\sin(\theta) - \frac{\theta}{2}$$

The whole integral can therefore be assembled as:

$$\frac{1}{2}\sin(\theta)\cos(\theta) - \frac{1}{2}\cos(\theta)\sin(\theta) + \frac{\theta}{2} = \frac{\theta}{2}$$

which confirms that the idea of radian measure we introduced earlier (6.8) is indeed the familiar one: at least for $0 < \theta < \pi/2$, the area of a sector of angle θ in a circle is half the radian measure of its angle, and (taking a limit as $\theta \to \pi/2$) a quarter-circle has area $\pi/4$. For angles larger than $\pi/2$, decompose the area into a sum of one or more quarter-circles and a segment of angle less than $\pi/2$.

It is possible to start here, by defining the radian measure of an angle in terms of the area of a sector of a circle and the trigonometric functions in terms of this. See Spivak's *Calculus* Chapter 15 for details.

7.14 Integration using partial fractions

Rational functions are often best integrated using partial fractions. For example, to calculate

$$\int \frac{x}{x^2 - 3x + 2}\,\mathrm{d}x$$

we write

$$\frac{x}{x^2 - 3x + 2} = \frac{A}{x-2} + \frac{B}{x-1} \qquad (x \neq 1, 2)$$

then multiply through to give

$$x = A(x-1) + B(x-2) \qquad (x \neq 1, 2)$$

and takes limits as $x \to 2$ and as $x \to 1$ to give $A = 2$ and $B = -1$. Now, we can integrate:

$$\int \frac{x}{x^2 - 3x + 2}\,\mathrm{d}x = \int \frac{2}{x-2} - \frac{1}{x-1}\,\mathrm{d}x = 2\log(x-2) - \log(x-1) \ + C$$

More complicated examples can give rise to more complicated partial fractions, but the following should cover all eventualities:

$$\int \frac{1}{x-a}\,\mathrm{d}x = \log(x-a) \quad + C$$
$$\int \frac{x}{x^2+a^2}\,\mathrm{d}x = \frac{1}{2}\log(x^2+a^2) \quad + C$$
$$\int \frac{1}{x^2+a^2}\,\mathrm{d}x = \frac{1}{a}\tan^{-1}\left(\frac{x}{a}\right) \quad + C$$

Here, a is intended to be real: complex partial fractions work in principle but in practice it is often difficult to track the correct branches of the complex logarithms and inverse trigonometric functions.

7.15 Improper Integrals

Begin with an example:

$$\int_0^1 \frac{1}{\sqrt{x}}\,\mathrm{d}x$$

Here we have a singularity at 0. We cannot apply the FTC directly because the integrand $1/\sqrt{x} \to +\infty$ as $x \to 0+$. We can, however, integrate from r to 1 for any $r \in (0,1)$

$$\int_r^1 \frac{1}{\sqrt{x}}\,\mathrm{d}x = [2\sqrt{x}]_{x=r}^{x=1} = 2 - 2\sqrt{r}$$

As $r \to 0+$

$$\int_r^1 \frac{1}{\sqrt{x}}\,\mathrm{d}x \to 2$$

This is an *improper integral*, but we still write

$$\int_0^1 \frac{1}{\sqrt{x}}\,\mathrm{d}x = 2$$

Note that the antiderivative $2\sqrt{x}$ is not differentiable at 0.

Another example:

$$\int_0^\infty \mathrm{e}^{-x}\,\mathrm{d}x$$

Here the range of integration is infinite, so the FTC does not apply. We proceed in a similar way, by evaluating

$$\int_0^R \mathrm{e}^{-x}\,\mathrm{d}x = [-\mathrm{e}^{-x}]_{x=0}^{x=R} = 1 - \mathrm{e}^{-R}$$

and let $R \to \infty$, so $1 - \mathrm{e}^{-R} \to 1$. We write

$$\int_0^\infty \mathrm{e}^{-x}\,\mathrm{d}x = 1$$

This is also known as an improper integral. In each case, we integrate over a finite range then take a limit, either to approach a singularity or to stretch the region of integration to an infinite length. Sometimes, we want to integrate over the whole line:

$$\int_{-\infty}^\infty \frac{1}{1+x^2}\,\mathrm{d}x$$

We evaluate this by integrating from $-R$ to $+S$:

$$\int_{-R}^S \frac{1}{1+x^2}\,\mathrm{d}x = [\tan^{-1}(x)]_{x=-R}^{x=S} = \tan^{-1}(S) + \tan^{-1}(R) \to \frac{\pi}{2} + \frac{\pi}{2} = \pi$$

as $R \to \infty$ and $S \to \infty$. We write

$$\int_{-\infty}^\infty \frac{1}{1+x^2}\,\mathrm{d}x = \pi$$

There are many ways to send R and S tend to ∞, e.g.

1. First let $R \to \infty$, then let $S \to \infty$.

2. First let $S \to \infty$, then let $R \to \infty$.

3. Let $S = R$, then let $R \to \infty$.
4. Let $S = 2R$, then let $R \to \infty$.
5. etc.

In this example, these all give the same answer. Sometimes, however, they can give different answers:

$$\int_{-R}^{S} \frac{x}{1+x^2}\,\mathrm{d}x = \left[\frac{\log(1+x^2)}{2}\right]_{x=-R}^{x=S} = \frac{1}{2}\log\left(\frac{1+S^2}{1+R^2}\right)$$

Now:

1. If we fix R and let $S \to \infty$, we get $+\infty$
2. If we fix S and let $R \to \infty$, we get $-\infty$
3. If we let $S = R$ and then let $R \to \infty$, we get 0
4. Let $S = 2R$ and let $R \to \infty$, we get $\log(2)$
5. etc.

The integral

$$\int_{-\infty}^{\infty} \frac{x}{1+x^2}\,\mathrm{d}x$$

is not really well defined. One of these alternatives is more useful than the others, though, and is known as the *principal value integral*. It is based on symmetric integrals around the origin:

$$\mathrm{PV}\int_{-\infty}^{\infty} \frac{x}{1+x^2}\,\mathrm{d}x = \lim_{R\to\infty}\int_{-R}^{R} \frac{x}{1+x^2}\,\mathrm{d}x = 0$$

In general, we write

$$\mathrm{PV}\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \lim_{R\to\infty}\int_{-R}^{R} f(x)\,\mathrm{d}x$$

if this limit exists.

7.16 Examples of Improper Integrals

When we calculate improper integrals, we do not usually write down the limiting steps, but present the calculation as if we're using the FTC, replacing function evaluation with limit evaluation whenever necessary.

$$\int_{0}^{\infty} \mathrm{e}^{-x}\,\mathrm{d}x = [-\mathrm{e}^{-x}]_{x=0}^{\infty} = 0-(-1) = 1$$

Here the evaluation of $-\mathrm{e}^{-x}$ "at $x = \infty$" is really a limit as $x \to \infty$.

One way to evaluate

$$\int_0^\infty \mathrm{e}^{-x} \sin(x)\,\mathrm{d}x$$

is to integrate by parts. Differentiating $\sin(x)$ and integrating e^{-x}, we have

$$\begin{aligned} I &= \int_0^\infty \mathrm{e}^{-x} \sin(x)\,\mathrm{d}x \\ &= [-\sin(x)\mathrm{e}^{-x}]_{x=0}^\infty + \int_0^\infty \cos(x)\mathrm{e}^{-x}\,\mathrm{d}x \\ &= \int_0^\infty \cos(x)\mathrm{e}^{-x}\,\mathrm{d}x \\ &= [-\cos(x)\mathrm{e}^{-x}]_{x=0}^\infty - \int_0^\infty \sin(x)\mathrm{e}^{-x}\,\mathrm{d}x \\ &= 1 - I \end{aligned}$$

which we solve to give $I = 1/2$. We can also read off from the third line that

$$\int_0^\infty \cos(x)\mathrm{e}^{-x}\,\mathrm{d}x = \int_0^\infty \mathrm{e}^{-x} \sin(x)\,\mathrm{d}x = \frac{1}{2}$$

Again, all of the evaluations "at ∞" are really limits as $x \to \infty$.

Another way to do the same calculation is to use the fact that $\sin(x) = \mathrm{Im}(\mathrm{e}^{\mathrm{i}x})$. It follows that

$$\begin{aligned} \int_0^\infty \mathrm{e}^{-x} \sin(x)\,\mathrm{d}x &= \mathrm{Im} \int_0^\infty \mathrm{e}^{-x}\mathrm{e}^{\mathrm{i}x}\,\mathrm{d}x \\ &= \mathrm{Im} \int_0^\infty \mathrm{e}^{(-1+\mathrm{i})x}\,\mathrm{d}x \\ &= \mathrm{Im} \left[\frac{\mathrm{e}^{(-1+\mathrm{i})x}}{-1+\mathrm{i}} \right]_{x=0}^\infty \\ &= \mathrm{Im} - \frac{1}{-1+\mathrm{i}} \qquad \text{(taking a limit at } \infty\text{)} \\ &= \mathrm{Im}\, \frac{1+\mathrm{i}}{2} \\ &= \frac{1}{2} \end{aligned}$$

Again, this calculation gives the corresponding cosine integral for free: replacing Im with Re throughout gives

$$\int_0^\infty \cos(x)\mathrm{e}^{-x}\,\mathrm{d}x = \frac{1}{2}$$

7.17 The Gamma Function

For $\alpha \in \mathbb{R}$ with $\alpha > 0$, define:

$$\boxed{\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \mathrm{e}^{-x} \, \mathrm{d}x}$$

This is called the *Gamma Function*. Its most important features are the functional equation and consequent identity:

$$\boxed{\begin{aligned} \Gamma(\alpha+1) &= \alpha\Gamma(\alpha) && (\alpha > 0) \\ \Gamma(n) &= (n-1)! && (n \in \mathbb{N}) \end{aligned}}$$

By extension, we sometimes write

$$\alpha! = \Gamma(\alpha+1)$$

for $\alpha \in \mathbb{R}$, $\alpha > -1$.

We establish the functional equation by integrating by parts:

$$\Gamma(\alpha+1) = \int_0^\infty x^\alpha \mathrm{e}^{-x} \, \mathrm{d}x = [-x^\alpha \mathrm{e}^{-x}]_{x=0}^{x=\infty} - \int_0^\infty \alpha x^{\alpha-1} \mathrm{e}^{-x} \, \mathrm{d}x = \alpha\Gamma(\alpha)$$

We also note that

$$\Gamma(1) = \int_0^\infty \mathrm{e}^{-x} \, \mathrm{d}x = [-\mathrm{e}^{-x}]_{x=0}^\infty = 1$$

We now have

$$\Gamma(2) = 1\Gamma(1) = 1; \qquad \Gamma(3) = 2\Gamma(2) = 2; \qquad \Gamma(4) = 3\Gamma(3) = 6; \qquad \Gamma(5) = 4\Gamma(4) = 24$$

etc. By induction, for any $n \in \mathbb{N}$, we have

$$\Gamma(n) = (n-1)!$$

7.18 The Gaussian Integral

We mention in passing the Gaussian or Normal integral:

$$\int_{-\infty}^\infty \mathrm{e}^{-x^2} \, \mathrm{d}x = \sqrt{\pi}$$

This is most easily proved using integrals of functions of two variables, which will be explained in the second part of the course.

We can use this to show that

$$\Gamma(1/2) = \sqrt{\pi}$$

and hence that

$$(1/2)! = \frac{\sqrt{\pi}}{2}$$

To reduce $\Gamma(1/2)$ to the Gaussian integral, write down the integral defining it and substitute $u^2 = x$:

$$\begin{aligned}\Gamma(1/2) &= \int_0^\infty x^{-1/2} \mathrm{e}^{-x}\,\mathrm{d}x \\ &= \int_0^\infty u^{-1} \mathrm{e}^{-u^2} 2u\,\mathrm{d}u \\ &= 2\int_0^\infty \mathrm{e}^{-u^2}\,\mathrm{d}u \\ &= \sqrt{\pi}\end{aligned}$$

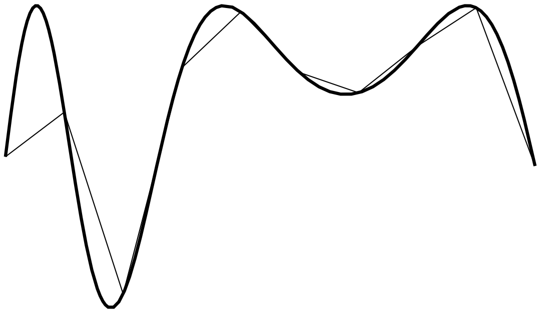
and hence

$$(1/2)! = \Gamma(3/2) = (1/2)\Gamma(1/2) = \frac{\sqrt{\pi}}{2}$$

7.19 Arc lengths

Integration can do more than compute areas. Suppose we have a parameterised curve in the plane, say $\gamma : [a,b] \to \mathbb{R}^2$ where $\gamma(t) = (x(t), y(t))$ and x, y are continuously differentiable. Recall that $\gamma'(t)$ denotes the velocity of the parametric curve, given by $\gamma(t) = (x'(t), y'(t))$. Then the length of the curve is given by

$$\boxed{L(\gamma) = \int_a^b \sqrt{(x'(t))^2 + (y'(t))^2}\,\mathrm{d}t = \int_a^b |\gamma'(t)|\,\mathrm{d}t}$$



To see how to get this formula, let us try computing the length of the curve by cutting the interval $[a,b]$ into a number of small intervals, say

$$a = t_0 < t_1 < \cdots < t_n = b$$

which also cuts the curve into small arcs. We approximate the curve on each arc by a straight line joining the endpoints. On the interval $[t_{j-1}, t_j]$, the length of this straight line is

$$\sqrt{(x(t_j) - x(t_{j-1}))^2 + (y(t_j) - y(t_{j-1}))^2}$$

But, by the MVT,

$$x(t_j) - x(t_{j-1}) = (t_j - t_{j-1})x'(c_j); \qquad y(t_j) - y(t_{j-1}) = (t_j - t_{j-1})y'(d_j)$$

for some $c_j, d_j \in (t_{j-1}, t_j)$ so we have

$$\sqrt{(x(t_j) - x(t_{j-1}))^2 + (y(t_j) - y(t_{j-1}))^2} = (t_j - t_{j-1})\sqrt{(x'(c_j))^2 + (y'(d_j))^2}$$

and the total length of the approximating lines is

$$\sum_{j=1}^{n} \sqrt{(x(t_j) - x(t_{j-1}))^2 + (y(t_j) - y(t_{j-1}))^2} = \sum_{j=1}^{n} (t_j - t_{j-1})\sqrt{(x'(c_j))^2 + (y'(d_j))^2}$$

This is very similar to a Riemann sum for

$$\int_a^b \sqrt{(x'(t))^2 + (y'(t))^2}\,\mathrm{d}t$$

(the difference being that c_j is not necessarily equal to d_j) and in fact it can be shown that, as the length of the approximating intervals tends to zero, the sum of their lengths tends to this integral. We can therefore describe the length of the curve by

$$L(\gamma) = \int_a^b \sqrt{(x'(t))^2 + (y'(t))^2}\,\mathrm{d}t = \int_a^b |\gamma'(t)|\,\mathrm{d}t$$

Actually proving that this works requires an idea called "uniform continuity" which will be introduced in the Real Analysis module.

If we think of t as representing time, then the integrand is the speed of motion along the path at time t (recall from (3.16) that the speed is the magnitude of the velocity vector, that is $|\gamma'(t)| = \sqrt{(x'(t))^2 + (y'(t))^2)}$. Each term in the sum is the product of the length of time represented by the interval and a sample speed within the interval, which is an approximation to the total distance travelled in that interval; the sum itself is thus an approximation to the total distance travelled, i.e. the length of the curve.

For example, consider the curve

$$\gamma(t) = (t, t^{3/2})$$

for $t \in [0,1]$. The velocity and speed are given by

$$\gamma'(t) = (1, (3/2)t^{1/2}); \qquad |\gamma'(t)| = \sqrt{1 + (9/4)t}$$

To find the length of the curve, we integrate the speed

$$L(\gamma) = \int_0^1 \sqrt{1 + (9/4)t}\,\mathrm{d}t = \left[\frac{1}{27}(4 + 9t)^{3/2}\right]_{t=0}^1 = \frac{13\sqrt{13} - 8}{27}$$

The same idea works in three dimensions: if $\gamma : [a,b] \to \mathbb{R}^3$ is given by

$$\gamma(t) = (x(t), y(t), z(t))$$

$$L(\gamma) = \int_a^b \sqrt{(x'(t))^2 + (y'(t))^2 + (z'(t))^2}\,\mathrm{d}t = \int_a^b |\gamma'(t)|\,\mathrm{d}t.$$

In fact, this works in any number of dimensions — see Vector Calculus.

7.20 Radian Measure, third attempt

An arc of curve between the angles θ and ϕ ($\theta < \phi$) on the circle centred at 0 with radius $r > 0$ can be parameterised by

$$\gamma(t) = (r\cos(t), r\sin(t))$$

for $\theta \le t \le \phi$. We find its length by evaluating the integral

$$\int_\theta^\phi \sqrt{r^2\sin^2(t) + r^2\cos^2(t)}\,\mathrm{d}t = r(\phi - \theta)$$

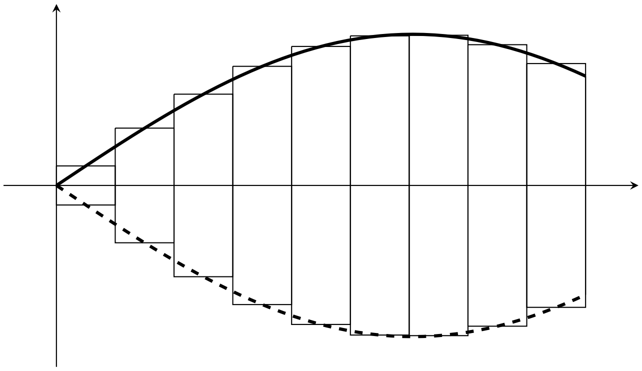
so the length of the arc is r times the angle traversed.

This is another common way of describing radian measure, and again confirms that our definition is consistent with other ways of describing angles.

In principle, one could start here and define radian measure in terms of arc length and trigonometrical functions in terms of that (this is a difficult place to start, though).

7.21 Surfaces and Volumes of Revolution

Suppose $r : [a,b] \to \mathbb{R}^+$ is continuous, so r describes a curve above the axis. Imagine rotating the curve about the x axis to describe the surface of a solid figure.



The volume of this figure is

$$V = \pi \int_a^b (r(x))^2 \, \mathrm{d}x$$

We establish this formula by dividing the interval $[a,b]$ into small subintervals, with endpoints $(x_j)_{j=0}^n$ and choosing sample points $c_j \in (x_{j-1}, x_j)$. We approximate the solid figure by a collection of discs of thickness $x_j - x_{j-1}$ and radius $r(c_j)$. Each disc has volume $\pi(r(c_j))^2(x_j - x_{j-1})$, so the total volume is

$$\sum_{j=1}^n \pi(r(c_j))^2(x_j - x_{j-1})$$

which is a Riemann sum for the integral

$$\pi \int_a^b (r(x))^2 \, \mathrm{d}x$$

so, letting the mesh of the partition tend to zero, we see that we can describe the volume of the solid by

$$V = \pi \int_a^b (r(x))^2 \, \mathrm{d}x$$

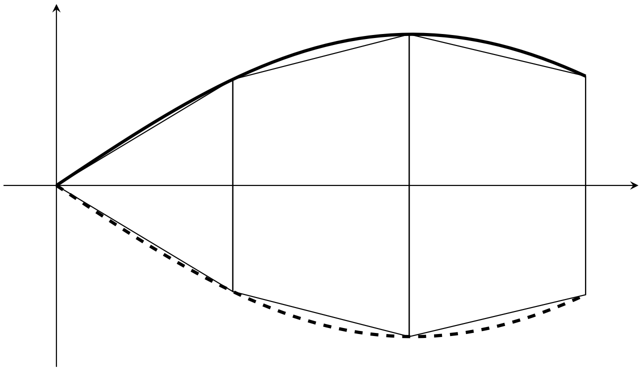
For example, what is volume of the figure formed by rotating one arc of a sine wave about the axis? We can calculate this as

$$V = \pi \int_0^\pi \sin^2(x) \, \mathrm{d}x = \frac{\pi}{2} \int_0^\pi (1 - \cos(2x)) \, \mathrm{d}x = \frac{\pi^2}{2}$$

(integrating a sine or cosine over one or more full periods gives zero).

We can also use the same decomposition to find the surface area, but this is much more delicate.

Suppose $r : [a,b] \to \mathbb{R}^+$ is continuously differentiable, so r describes a curve above the axis. Imagine rotating the curve about the x axis to describe the surface of a solid figure.



The surface area of this figure (excluding the flat ends) is

$$A = 2\pi \int_a^b r(x)\sqrt{1 + r'(x)^2}\,\mathrm{d}x$$

To see this, approximate the surface area associated with the interval $[x_{j-1}, x_j]$ by the surface area of the conical frustum with radii $r(x_{j-1})$ and $r(x_j)$ and height $x_j - x_{j-1}$, i.e.

$$\pi(r(x_{j-1}) + r(x_j))\sqrt{(r(x_j) - r(x_{j-1}))^2 + (x_j - x_{j-1})^2}$$

We can draw out a factor of $x_j - x_{j-1}$ from the square root to give

$$\pi(r(x_{j-1}) + r(x_j))(x_j - x_{j-1})\sqrt{\left(\frac{r(x_j) - r(x_{j-1})}{x_j - x_{j-1}}\right)^2 + 1}$$

and use the MVT to see that this is equal to

$$\pi(r(x_{j-1}) + r(x_j))(x_j - x_{j-1})\sqrt{r'(c_j)^2 + 1}$$

for some $c_j \in (x_{j-1}, x_j)$. Now, we can add up the contributions to give

$$\sum_{j=1}^{n} \pi r(x_{j-1})(x_j - x_{j-1})\sqrt{r'(c_j)^2 + 1} + \sum_{j=1}^{n} \pi r(x_j)(x_j - x_{j-1})\sqrt{r'(c_j)^2 + 1}$$

Just as for arc length (7.19) these are almost, but not exactly, Riemann sums, this time for the integral

$$\pi \int_a^b r(x)\sqrt{1 + r'(x)^2}\,\mathrm{d}x$$

Again, it can be shown that, as the mesh of the partition tends to zero, they both converge to this integral. We can therefore describe the surface area by

$$A = 2\pi \int_a^b r(x)\sqrt{1 + r'(x)^2}\,\mathrm{d}x$$

For example, let us find the area of the surface formed by rotating the cosh curve about the origin, between −1 and 1. Our formula is

$$\begin{aligned} A &= 2\pi \int_{-1}^{1} \cosh(x)\sqrt{1+\sinh^2(x)}\,\mathrm{d}x \\ &= 2\pi \int_{-1}^{1} \cosh(x)^2\,\mathrm{d}x \\ &= \pi \int_{-1}^{1} \cosh(2x)+1\,\mathrm{d}x \\ &= \pi[\sinh(2x)/2+x]_{x=-1}^{1} \\ &= \pi(\sinh(2)+2) \end{aligned}$$

8 Taylor's Theorem

8.1 Introduction

Recall from section (3.7) that we can approximate a function near a point by the local linear approximation at the point, which was the equation of the tangent line at the point. For the point x_0 and the function f the local linear approximation can be written as $f(x_0)+f'(x_0)(x-x_0)$. We saw that we can then write f as

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + r(x - x_0)$$

where r is the remainder function, which has the property that $\frac{r(h)}{h} \to 0$ as $h \to 0$.

The local linear approximation of f at the point x_0 is a polynomial which has the property that it has the same function value as f at x_0, and the same first derivative as f at x_0.

Taylor's Theorem allows us to extend this notion, by approximating a function near a point by higher order polynomials. These polynomials, known as *Taylor polynomials*, have the property that they have the same higher order derivatives as f at the point x_0.

Recall the notation $f^{(n)}$ for the nth derivative of f. We also follow the convention that $f^{(0)} = f$.

If I is an interval, $I \subseteq \mathbb{R}$, denote by $C^n(I)$ the set of all functions $f : I \to \mathbb{R}$ such that all the derivatives $f', f'', \dots f^{(n)}$ exist and are continuous on I. $C^n(I)$ is in fact also a vector space over $\mathbb{R}$, so it is often referred to as a space of functions. We also write $C^\infty(I)$ for the space of functions $f : I \to \mathbb{R}$ that can be differentiated infinitely many times. These sets are nested, in the sense that we have $C^\infty(I) \subseteq \ldots \subseteq C^3(I) \subseteq C^2(I) \subseteq C^1(I) \subseteq C^0(I)$, where $C^0(I)$ is the space of continuous functions $f : I \to \mathbb{R}$.

8.2 Taylor's Theorem with Integral Remainder

Let us now look at how to construct Taylor polynomials, using the Fundamental Theorem of Calculus and integration by parts. Suppose I is an interval and $x_0 \in I$. The FTC tells us that if $f \in C^1(I)$ (that is if f is differentiable and f' is continuous) and $x \in I$ then

$$\int_{x_0}^{x} f'(t)\,\mathrm{d}t = f(x) - f(x_0)$$

which we can rewrite as

$$f(x) = f(x_0) + \int_{x_0}^{x} f'(t)\,\mathrm{d}t$$

Now suppose moreover that $f \in C^2(I)$. Use the "multiplication by 1" trick to write $f'(t) = 1 \cdot f'(t)$ and integrate by parts, differentiating f' and integrating 1. Instead of the standard antiderivative t, we use the antiderivative $t-x$ (since $-x$ is a constant with respect to t).

$$\begin{aligned} f(x) &= f(x_0) + [(t-x)f'(t)]_{t=x_0}^{x} - \int_{x_0}^{x} (t-x)f''(t)\,\mathrm{d}t \\ &= f(x_0) + (x-x_0)f'(x_0) - \int_{x_0}^{x} (t-x)f''(t)\,\mathrm{d}t \\ &= f(x_0) + f'(x_0)(x-x_0) + \int_{x_0}^{x} (x-t)f''(t)\,\mathrm{d}t \end{aligned}$$

Note the connection with the first order approximation: $f(x_0)+f'(x_0)(x-x_0)$ is the local linear approximation for f near x_0 and $\int_{x_0}^{x}(x-t)f''(t)\,\mathrm{d}t$ is an explicit formula for the remainder term $r(x-x_0)$ in terms of f''.

Similarly, assuming $f \in C^3(I)$, integrate by parts, integrating $x-t$ to give $-\frac{(x-t)^2}{2}$ and differentiating f'', to give

$$\begin{aligned} \int_{x_0}^{x} (x-t)f''(t)\,\mathrm{d}t &= \left[-\frac{(x-t)^2}{2}f''(t)\right]_{t=x_0}^{x} + \int_{x_0}^{x} \frac{(x-t)^2}{2}f'''(t)\,\mathrm{d}t \\ &= \frac{f''(x-x_0)}{2}(x-x_0)^2 + \frac{1}{2}\int_{x_0}^{x}(x-t)^2 f'''(t)\,\mathrm{d}t \end{aligned}$$

Substitute this back into our expression for $f(x)$ to give:

$$f(x) = f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x-x_0)}{2}(x-x_0)^2 + \frac{1}{2}\int_{x_0}^{x}(x-t)^2 f'''(t)\,\mathrm{d}t$$

Carrying on like this, we integrate $\frac{(x-t)^2}{2}$ to give (up to a sign) $\frac{(x-t)^3}{6}$, then $\frac{(x-t)^4}{24}$, etc. This leads to:

Taylor's Theorem (with the integral form of the remainder): if $f \in C^{N+1}(I)$ and $x \in I$ then

$$f(x) = \sum_{n=0}^{N} \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n + \frac{1}{N!}\int_{x_0}^{x}(x-t)^N f^{(N+1)}(t)\,\mathrm{d}t$$

The terms

$$\sum_{n=0}^{N} \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n$$

are called the *Taylor polynomial* (of f, at x_0) of order N. The term

$$R_N(x) = \frac{1}{N!}\int_{x_0}^{x} (x-t)^N f^{(N+1)}(t)\,\mathrm{d}t$$

is called the *error* or *remainder*, in *integral form*. We shall soon see other ways to express the same quantity.

When $N = 1$, the Taylor polynomial is the local linear approximation, i.e. the straight line that touches the graph of f at x_0 and has the same gradient as f at x_0.

More generally,

The Taylor polynomial P of order N at x_0 is the unique polynomial such that

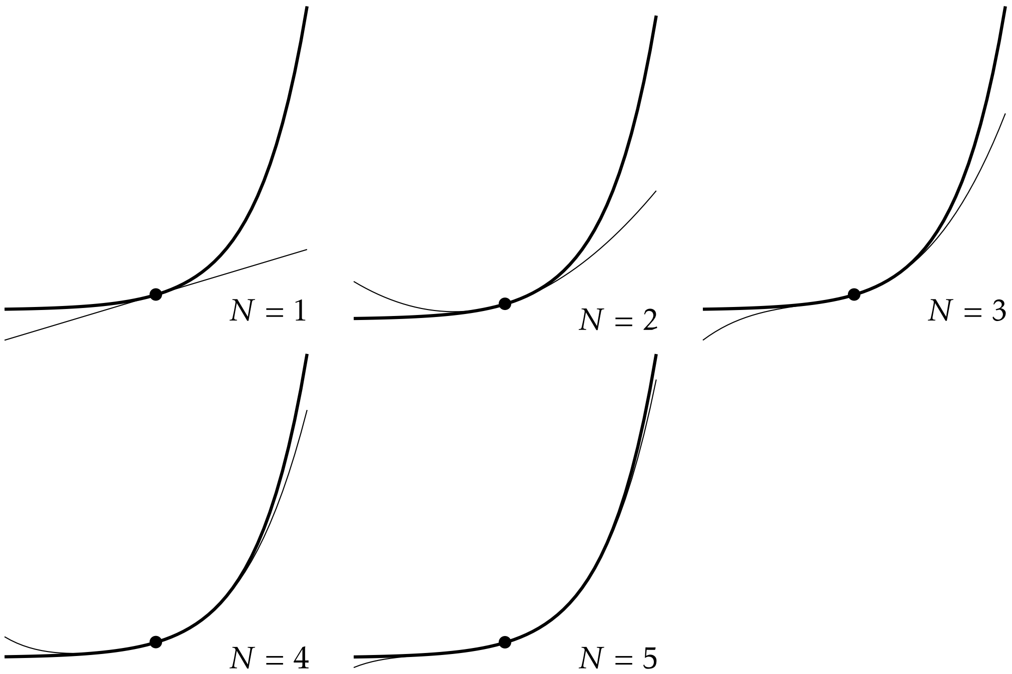
$$P^{(n)}(x_0) = f^{(n)}(x_0)$$

for all $0 \leq n \leq N$.

In other words, the Taylor polynomial P of order N at x_0 is the unique polynomial whose graph touches that of f at x_0 (since $P(x_0) = P^{(0)}(x_0) = f^{(0)}(x_0) = f(x_0)$) and whose first n derivatives at x_0 agree with those of f. Intuitively, the Taylor polynomial at x_0 has the "same shape" as f near x_0.

This means that $P(x)$ is a good approximation for $f(x)$ when x is close to x_0 and that the larger we make N, the better the approximation becomes.

The first few Taylor polynomials for the exponential function at the origin look like this:



These are easily calculated: because $\exp' = \exp$, we have $\exp^{(n)}(0) = 1$ for all n. The coefficient of x^n is thus $\frac{1}{n!}$ and the Taylor polynomial is

$$\sum_{n=0}^{N} \frac{x^n}{n!}$$

We could also expand around any other point. For example about $x_0 = 1$: we have $\exp^{(n)}(1) = \mathrm{e}$ for all n, so the Taylor polynomial of order N is

$$\sum_{n=0}^{N} \frac{\mathrm{e}}{n!}(x-1)^n$$

This is a good approximation for $\exp(x)$ for x near 1.

8.3 The Lagrange form of the remainder

If $I \subseteq \mathbb{R}$ and $f \in C^{N+1}(I)$ then we saw above that the integral form of the remainder is

$$R_N(x) = \frac{1}{N!}\int_{x_0}^{x} (x-t)^N f^{(N+1)}(t)\,\mathrm{d}t$$

We can (because $f^{(N+1)}$ is continuous) apply the MVTI (7.4) on the interval between x_0 and x, with $f^{(N+1)}$ playing the role of f and $g(t) = (x-t)^N$, to the remainder term to see that it can also be written as

$$R_N(x) = \frac{1}{N!} f^{(N+1)}(c) \int_{x_0}^{x} (x-t)^N\,\mathrm{d}t$$

for some c between x_0 and x. Computing the integral we get

$$R_N(x) = \frac{1}{N!} f^{(N+1)}(c)\left[-\frac{(x-t)^{N+1}}{(N+1)}\right]_{t=x_0}^{x} = \frac{1}{N!} f^{(N+1)}(c)\frac{(x-x_0)^{N+1}}{N+1}$$

so we end up with

$$R_N(x) = \frac{(x-x_0)^{N+1}}{(N+1)!} f^{(N+1)}(c)$$

for some c between x_0 and x.

This is the *Lagrange form* or *derivative form* of the remainder.

It can be shown that Taylor's Theorem and the various forms of the error term are true even if $f^{(N+1)}$ is not continuous, as long as $f^{(N)}$ is differentiable on the interior of I. The usual argument is rather more complicated, and is commonly based on a generalised MVT for derivatives. See Spivak's *Calculus* Chapter 19.

8.4 Second Derivative Tests

The remainder term in Taylor's Theorem allows us to interpret the second derivative at a critical point.

Suppose

- $f : [a,b] \to \mathbb{R}$ and $f \in C^2([a,b])$
- $x_0 \in (a,b)$ and $f'(x_0) = 0$

Then:

- If $f''(x_0) > 0$ then x_0 is a local minimum.
- If $f''(x_0) < 0$ then x_0 is a local maximum.
- If $f''(x_0) = 0$, no conclusion of this type can be drawn.

To see why this works, write down the Taylor expansion of f about x_0 to order 1 with the Lagrange form of the remainder.

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(c)(x - x_0)^2$$

for some c between x and x_0. Because we are assuming $f'(x_0) = 0$, this reduces to

$$f(x) = f(x_0) + \frac{1}{2} f''(c)(x - x_0)^2$$

Now, if x is close to, but not equal to, x_0 then $f''(c)$ is close to $f''(x_0)$. It follows that if $f''(x_0) < 0$ then $f''(c) < 0$ and if $f''(x_0) > 0$ then $f''(c) > 0$. Because $(x - x_0)^2 > 0$, we have $f(x) > f(x_0)$ if $f''(x_0) > 0$ and $f(x) < f(x_0)$ if $f''(x_0) < 0$. That is, x_0 is a local minimum if $f''(x_0) > 0$ and a local maximum if $f''(x_0) < 0$.

If $f''(x_0) = 0$, we can see by example that any behaviour is possible. The following all satisfy $f'(0) = f''(0) = 0$:

- $f(x) = x^3$ is strictly increasing on $\mathbb{R}$
- $f(x) = x^4$ has a local minimum at the origin
- $f(x) = -x^4$ has a local maximum at the origin

There is a more detailed version of the test (A.7) but it is rarely useful.

8.5 Taylor Series of Some Elementary Functions

We saw above that the Taylor polynomial of order N of f at x_0 is given by

$$P(x) = \sum_{n=0}^{N} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$

The infinite series

$$\sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$

is known as the *Taylor series* of f at x_0. Truncating the series after a finite number of terms gives us a Taylor polynomial. At $x = x_0$ this series clearly converges to $f(x_0)$ (since all the other terms are 0), but away from x_0 this series may or may not converge to $f(x)$.

To see whether the Taylor series of f at x_0 converges to $f(x)$, we look at the remainder term, and consider whether it tends to 0 as $N \to \infty$.

Let us look at some examples of Taylor series.

The standard formula (2.8)

$$\sum_{n=0}^{N} x^n = \frac{1 - x^{N+1}}{1 - x} \qquad (x \neq 1)$$

for a geometric series can be written as

$$\frac{1}{1 - x} = \sum_{n=0}^{N} x^n + \frac{x^{N+1}}{1 - x}$$

This is exactly the Taylor polynomial and remainder term of $f(x) = \frac{1}{1-x}$ at the origin. The remainder tends to 0 as $N \to \infty$ if $|x| < 1$. So the Taylor series of $f(x) = \frac{1}{1-x}$ at $x = 0$, $\sum_{n=0}^{\infty} x^n$, converges to the function value $f(x)$ when $|x| < 1$. This also works if x is complex.

Consider the exponential function $\exp : \mathbb{R} \to \mathbb{R}$. We have already seen that the Taylor polynomial of order N at $x_0 = 0$ is

$$\sum_{n=0}^{N} \frac{x^n}{n!}$$

and, from Taylor's Theorem, the remainder term is

$$R_N(x) = \frac{1}{N!} \int_0^x (x - t)^N \exp(t)\,\mathrm{d}t$$

or, using the Lagrange form of the remainder, we have

$$R_N(x) = \frac{x^{N+1}}{(N + 1)!} \exp(c)$$

for some c between 0 and x. Consider the absolute value of this expression, $|R_N(x)| = \frac{|x|^{N+1}}{(N+1)!}\exp(c)$. For fixed x and c, as N increases, the factorial denominator $(N+1)!$ grows faster than the exponential numerator $|x|^{N+1}$, so we get $R_N(x) \to 0$ as $N \to \infty$.

So the Taylor series of $\exp(x)$ at 0 converges to $\exp(x)$ for any x, so we can write

$$\boxed{\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!} \qquad (x \in \mathbb{R})}$$

A Taylor series at 0 is sometimes also called a *Maclaurin series*.

The idea here is that, as N becomes larger, the Taylor polynomials become better and better approximations to exp and, in the limit, they give an exact formula for exp.

An infinite sum of the form

$$\boxed{\sum_{n=0}^{\infty} a_n (x - x_0)^n}$$

is called *power series* (about the point x_0). Taylor series are examples of power series.

To find the series for sin, note that the derivatives of sin (starting from sin itself) are sin, cos, sin, −cos, after which they repeat cyclically. The derivatives at 0 are therefore $0, 1, 0, -1, \ldots$. Putting these into Taylor's Theorem gives the coefficients in the sine series. The remainder is

$$R_N(x) = \frac{1}{N!}\int_0^x (x-t)^N \sin(t)\,\mathrm{d}t$$

As before, we can write this in the Lagrange form, as

$$R_N(x) = \frac{x^{N+1}}{(N+1)!}\sin^{(N+1)}(c)$$

for some c between 0 and x. The $N+1$-st derivative of sin, regardless of what exactly it is, satisfies $|\sin^{(N+1)}(c)| \leq 1$. So we have $|R_N(x)| \leq \frac{|x|^{N+1}}{(N+1)!}$, which as we saw above tends to 0 as $N \to \infty$.

The calculation for cos is near-identical; the derivatives at the origin start with $(1, 0, -1, 0)$ and repeat cyclically.

The calculations for sinh and cosh are similar, but their derivatives cycle with period 2, not period 4. To show that the remainder tends to 0 we can either follow the method above, or use the exponential definition of sinh and cosh and the fact that we already showed that the Taylor series of $\exp(x)$ at 0 converges to $\exp(x)$ for any x.

In conclusion, for any $x \in \mathbb{R}$, the Taylor series for sin, cos, sinh and cosh at 0 converge to their respective functions, so we can write

$$\boxed{\begin{aligned} \sin(x) &= \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!} & \cos(x) &= \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!} \\ \sinh(x) &= \sum_{n=0}^{\infty} \frac{x^{2n+1}}{(2n+1)!} & \cosh(x) &= \sum_{n=0}^{\infty} \frac{x^{2n}}{(2n)!} \end{aligned}}$$

The first term of the series for $\sin(x)$ is x; the first two terms of the series for $\cos(x)$ are $1-x^2/2$. These are the approximations for $\sin(x)$ and $\cos(x)$ for small x that we saw earlier (6.6).

Consider the approximation "$\sin(x) \approx x$" for small x. We can analyse this more carefully using the Lagrange form of the remainder. Because the quadratic term is absent in the Taylor series, we can write

$$\sin(x) = x - \frac{c^3}{6}$$

where $|c| \leq |x|$. The error in the approximation is thus no larger than $|x|^3/6$. Suppose we want an error of less than 0.01 (roughly two decimal places of accuracy). Then we want $|x|^3/6 < 0.01$, or $|x| < 0.39$. If we work out the error at $x = 0.39$, we actually have

$$|\sin(0.39) - 0.39| \approx 0.0098 < 0.01$$

As a final remark for this section, all the examples we have seen so far involve Taylor series that converge to the function, but this is not always the case. An example of this is $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = \begin{cases} e^{-\frac{1}{x^2}} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

It can be shown that this function is continuous and infinitely differentiable. At $x_0 = 0$, $f(0) = 0$, and all the derivatives of this function at 0 are 0, so the Taylor series of f at 0 is 0. But the function is not equal to 0 away from 0, so the Taylor series at 0 does not converge to $f(x)$ for $x \neq 0$.

8.6 Complex Taylor Series

The power series for exp, sin, cos, sinh and cosh also work for complex arguments.

Euler's formula

$$\mathrm{e}^{\mathrm{i}\theta} = \cos(\theta) + \mathrm{i}\sin(\theta)$$

is clearly visible in the Taylor series.

$$\begin{aligned}\exp(\mathrm{i}\theta) &= \sum_{n=0}^{\infty} \frac{(\mathrm{i}\theta)^n}{n!} \\ &= \sum_{\substack{n=0 \\ n \text{ even}}}^{\infty} \frac{(-1)^{n/2}\theta^n}{n!} + \mathrm{i} \sum_{\substack{n=0 \\ n \text{ odd}}}^{\infty} \frac{(-1)^{(n-1)/2}\theta^n}{n!} \\ &= \cos(\theta) + \mathrm{i}\sin(\theta)\end{aligned}$$

The details are a bit fiddly, and are relegated to the appendix (A.8). This is another place we could have started the construction of the elementary functions: by *defining* them as the sums of their Taylor series.

8.7 The Binomial Theorem

Suppose $\alpha \in \mathbb{R}$ and let $f(x) = (1+x)^\alpha$ for $x \in (-1,1)$. Differentiate this:

$$\begin{aligned}f'(x) &= \alpha(1+x)^{\alpha-1} \\ f''(x) &= \alpha(\alpha-1)(1+x)^{\alpha-2} \\ f'''(x) &= \alpha(\alpha-1)(\alpha-2)(1+x)^{\alpha-3} \\ &\vdots \\ f^{(n)}(x) &= \alpha(\alpha-1)\ldots(\alpha-n+1)(1+x)^{\alpha-n}\end{aligned}$$

At the origin, this gives us

$$f^{(n)}(0) = \alpha(\alpha-1)\ldots(\alpha-n+1)$$

so the Taylor polynomial of f at the origin is given by

$$\sum_{n=0}^{N} \binom{\alpha}{n} x^n$$

where

$$\binom{\alpha}{n} = \frac{1}{n!}\alpha(\alpha-1)\dots(\alpha-n+1)$$

If $\alpha \in \mathbb{N}$ and $0 \leq \alpha \leq n$, this is an ordinary binomial coefficient.

We can show that, if $|x| < 1$, then the error term tends to zero giving

$$(1+x)^\alpha = \sum_{n=0}^{\infty}\binom{\alpha}{n}x^n$$

for any $x \in \mathbb{R}$ with $|x| < 1$.

If $\alpha \in \mathbb{N}$, then $\binom{\alpha}{n} = 0$ for $n > \alpha$, so this is exactly the usual binomial theorem (which holds for all $x \in \mathbb{R}$).

The first two terms of this expansion are $1 + \alpha x$; this is the approximation for $(1+x)^\alpha$ for small x that we saw earlier (4.7) from the linear approximation for the derivative.

We can now see (Lagrange form (8.3) of the remainder) that

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

where $|c| < |x|$, so the error in this approximation is no larger than

$$\frac{1}{2}|\alpha(\alpha-1)|x^2$$

Analysing the error term in the binomial series is more difficult than for the elementary functions above (because, for large N, $f^{(N)}$ has a singularity at -1) so the details are consigned to the appendix (A.9).

8.8 Radius of Convergence

Consider a general power series

$$\sum_{n=0}^{\infty} a_n(z-z_0)^n$$

It can be shown (Real Analysis, Complex Analysis and Integral Transforms) that there are three possibilities:

1. The series converges for all $z \in \mathbb{C}$ (like exp, sin, cos, sinh, cosh)

2. The series converges only for $z = z_0$ (useless!)
3. There exists $R > 0$ such that the series converges if $|z - z_0| < R$ and diverges if $|z - z_0| > R$ (like the geometric series)

R is called the *radius of convergence*. We write $R = \infty$ and $R = 0$ for the first two cases. In $\mathbb{C}$, the inequality $|z - z_0| < R$ describes a disc; in $\mathbb{R}$, it describes an interval.

Suppose $R > 0$ and

$$f(x) = \sum_{n=0}^{\infty} a_n x^n$$

for $x \in (-R, R)$. It can be shown that

$$f'(x) = \sum_{n=1}^{\infty} n a_n x^{n-1}$$

for $x \in (-R, R)$, that

$$F(x) = \sum_{n=0}^{\infty} \frac{a_n}{n+1} x^{n+1}$$

converges for $x \in (-R, R)$ and that $F' = f$.

Within the radius of convergence, a power series can be differentiated and integrated term-by-term just like a polynomial.

This will be explained in the second-year Functions of a Complex Variable module.

Some inverse function Taylor series, all valid for $|x| < 1$:

$$\log(1+x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{n+1}}{n+1}$$

$$\tan^{-1}(x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{2n+1}$$

$$\tanh^{-1}(x) = \sum_{n=0}^{N} \frac{x^{2n+1}}{2n+1}$$

To establish these, start with

$$\begin{aligned}
\frac{1}{1+x} &= \sum_{n=0}^{\infty}(-1)^n x^n \\
\frac{1}{1+x^2} &= \sum_{m=0}^{\infty}(-1)^n x^{2n} \\
\frac{1}{1-x^2} &= \sum_{m=0}^{\infty} x^{2n}
\end{aligned}$$

and integrate term-by-term; the constant of integration is zero in each case because each of these functions is zero at the origin.

9 First-Order Differential Equations

9.1 Differential Equations

An *Ordinary Differential Equation* (ODE) is an equation involving an unknown function of one variable and one or more of its derivatives. We seek to solve the equation, i.e. find the unknown function.

The *order* of the equation is the highest order of derivative present.

We have seen (6.14) three differential equations and their solutions already:

$$\begin{aligned}
\frac{\mathrm{d}y}{\mathrm{d}x} &= cy & y &= A\exp(cx) \\
\frac{\mathrm{d}^2y}{\mathrm{d}x^2} &= -c^2y & y &= A\cos(cx) + B\sin(cx) \\
\frac{\mathrm{d}^2y}{\mathrm{d}x^2} &= c^2y & y &= A\cosh(cx) + B\sinh(cx) \\
& & &= C\exp(cx) + D\exp(-cx)
\end{aligned}$$

The first of these is first-order, the other two are second-order.

We expect the solution of a first-order problem to have one unknown constant, and the solution to a second-order problem to have two unknown constants, but this is not always the case. This term, we look at first-order equations only. Study of second-order equations begins in the second part of the Calculus module, in January, and also in the Introduction to Applied Mathematics module.

9.2 Simple Differential Equations

A *simple* differential equation is one of the form

$$\frac{\mathrm{d}y}{\mathrm{d}x} = f(x)$$

The solutions are

$$y = \int f(x)\,\mathrm{d}x + C$$

i.e. an indefinite integral, plus an arbitrary constant.

Whenever we integrate a function, we are solving a simple differential equation.

When we integrate to solve a simple differential equation, we *must* supply a constant of integration, otherwise we lose many solutions.

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \sin(x) \qquad \Longrightarrow \qquad y = -\cos(x) + C$$

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \exp(x) + 2\cos(x) \qquad \Longrightarrow \qquad y = \exp(x) + 2\sin(x) + C$$

etc.

9.3 Separable Differential Equations

A *separable* first-order equation has the form

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{f(x)}{g(y)}$$

where f and g are given functions and we want to express y in terms of x (or x in terms of y). We solve this by formally cross-multiplying to give

$$\text{“}g(y)\,\mathrm{d}y = f(x)\,\mathrm{d}x\text{”}$$

and integrating

$$\int g(y)\,\mathrm{d}y = \int f(x)\,\mathrm{d}x + C$$

then solving this algebraic equation for x or y, as required.

Why does this work? Let F and G be antiderivatives of f and g, so $F' = f$ and $G' = g$. Rearrange the equation to

$$g(y)\frac{\mathrm{d}y}{\mathrm{d}x} = f(x)$$

and notice (by the chain rule) that the left hand side is the derivative with respect to x of $G(y)$, whereas the RHS is the derivative w.r.t. x of $F(x)$. We thus have

$$\frac{\mathrm{d}}{\mathrm{d}x}G(y) = \frac{\mathrm{d}}{\mathrm{d}x}F(x)$$

from which we conclude that $G(y) = F(x) + C$ for some constant C.

For example, to solve

$$\frac{\mathrm{d}y}{\mathrm{d}x} = 1 + y^2$$

we rewrite the equation in the form

$$\int \frac{\mathrm{d}y}{1+y^2} = \int \mathrm{d}x + C$$

and integrate to give

$$\tan^{-1} y = x + C$$

or

$$y = \tan(x + C)$$

Note that it is important to put the constant of integration in immediately after integrating.

In an *Initial Value Problem* (IVP), we are also given the value taken on by the unknown function at one point. This allows us to calculate the otherwise unknown constant.

For example,

$$\frac{\mathrm{d}y}{\mathrm{d}x} = -xy; \qquad y = 1 \text{ when } x = 0$$

Separate the variables:

$$\int \frac{\mathrm{d}y}{y} = -\int x\,\mathrm{d}x$$

and integrate:

$$\log(y) = -\frac{x^2}{2} + C$$

Now substitute $x = 0$ and $y = 1$ to give $0 = 0 + C$, so $C = 0$. The solution is

$$\log(y) = \frac{-x^2}{2} \qquad \Longrightarrow \qquad y = \exp(-x^2/2)$$

9.4 An IVP with many solutions

Consider the IVP

$$\frac{\mathrm{d}y}{\mathrm{d}x} = y^{1/3}; \qquad y = 0 \text{ when } x = 0$$

As well as the trivial solution $y = 0$, this has infinitely many other solutions on $\mathbb{R}$. To find some of them, separate variables to obtain

$$\int y^{-1/3}\,\mathrm{d}y = \int x\,\mathrm{d}x$$

$$\Longrightarrow \qquad \frac{3}{2}y^{2/3} = x + C$$

Substituting $x = 0, y = 0$ gives $C = 0$ so we have

$$y = \left(\frac{2}{3}x\right)^{3/2}$$

This solution is very different from the trivial $y = 0$ solution already mentioned. It is real for $x \geq 0$ and imaginary for $x < 0$, so it raises the spectre of real ODEs having complex solutions. It is also really four different solutions in one: we can choose either branch of the real square root function for $x > 0$ and, independently, either branch of the imaginary square root function for $x < 0$.

We can go beyond this: if we let

$$y = \begin{cases} 0 & \text{if } x \le 0 \\ \left(\frac{2}{3}x\right)^{3/2} & \text{if } x > 0 \end{cases}$$

(a combination of the trivial and non-trivial solutions), we obtain another, purely real, solution. In fact, two solutions: we can use either branch of the square root. Finally, if $k > 0$ then we can define

$$y = \begin{cases} 0 & \text{if } x \le k \\ \left(\frac{2}{3}(x-k)\right)^{3/2} & \text{if } x > k \end{cases}$$

This simply moves the hybrid solution k to the right; either branch of the square root can be used. This is also a solution to the IVP, giving us infinitely many solutions.

There are theorems guaranteeing that, under certain hypotheses, IVPs like this do have unique solutions. This particular equation, of course, cannot satisfy the hypotheses of such a theorem: the underlying problem is that the RHS, $y^{1/3}$, is not differentiable at 0.

9.5 Integrating Factors

A first-order ODE is said to be *linear* if it has the form

$$a(x)\frac{\mathrm{d}y}{\mathrm{d}x} + b(x)y + c(x) = 0$$

The important thing is that y and $\mathrm{d}y/\mathrm{d}x$ appear only in this way: multiplied by functions of x only, and added. Such an equation can be put into *standard form* by dividing by $a(x)$ to give an equivalent equation

$$\frac{\mathrm{d}y}{\mathrm{d}x} + P(x)y + Q(x) = 0$$

and solved to give

$$y = -\frac{\int Q(x)F(x)\,\mathrm{d}x + C}{F(x)}$$

where

$$F(x) = \exp\left(\int P(x)\,\mathrm{d}x\right)$$

is called the *integrating factor*.

The idea is that $\mathrm{d}y/\mathrm{d}x + P(x)y$ looks like the result of a product rule and, if we multiply by a suitable factor, we can integrate it exactly. We seek an *integrating factor* $F(x)$ with the property that

$$F(x)\frac{\mathrm{d}y}{\mathrm{d}x} + F(x)P(x)y$$

is the derivative, via the product rule, of some simple expression. The first term, $F(x)\mathrm{d}y/\mathrm{d}x$, appears if we differentiate $F(x)y$:

$$\frac{\mathrm{d}}{\mathrm{d}x}F(x)y = F(x)\frac{\mathrm{d}y}{\mathrm{d}x} + F'(x)y$$

so we seek $F(x)$ such that

$$F'(x) = F(x)P(x)$$

Dividing by $F(x)$ and integrating, we see that $\log F(x) = \int P(x)\,\mathrm{d}x$, or equivalently $F(x) = \exp(\int P(x)\,\mathrm{d}x)$. So, we multiply by $F(x)$ to give

$$\frac{\mathrm{d}y}{\mathrm{d}x}F(x) + P(x)F(x)y = -Q(x)F(x)$$

where the LHS is exactly

$$\frac{\mathrm{d}}{\mathrm{d}x}yF(x)$$

so we can integrate to give

$$yF(x) = -\int Q(x)F(x)\,\mathrm{d}x + C$$

where C is an unknown constant, and solve this equation for y.

Example:

$$\frac{\mathrm{d}y}{\mathrm{d}x} + xy = x$$

This is already in standard form, so the integrating factor is $\exp(\int x\,\mathrm{d}x) = \mathrm{e}^{x^2/2}$. There is no need for a constant of integration at this point; if you add one, it will cancel out later. Multiplying by the IF we have

$$\mathrm{e}^{x^2/2}\frac{\mathrm{d}y}{\mathrm{d}x} + x\mathrm{e}^{x^2/2}y = x\mathrm{e}^{x^2/2}$$

The LHS is exactly the derivative of y times the integrating factor (always check this, so you pick up any mistake in calculating the IF)

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathrm{e}^{x^2/2}y = x\mathrm{e}^{x^2/2}$$

and we can integrate the RHS to give $\mathrm{e}^{x^2/2}$, so we have

$$\mathrm{e}^{x^2/2}y = \mathrm{e}^{x^2/2} + C$$

$$\implies \qquad y = 1 + C\mathrm{e}^{-x^2/2}$$

Note that this equation is also separable:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = x(1-y)$$

$$\implies \qquad \int \frac{\mathrm{d}y}{1-y} = \int x\,\mathrm{d}x$$

$$\implies \qquad -\log(1-y) = x^2/2 + C$$

$$\implies \qquad \frac{1}{1-y} = \mathrm{e}^C \mathrm{e}^{x^2/2}$$

$$\implies \qquad y = 1 + \mathrm{e}^{-C}\mathrm{e}^{-x^2/2}$$

Just as for separable equations, the constant of integration can be determined by an initial value. Example:

$$x^2\frac{\mathrm{d}y}{\mathrm{d}x} + xy = \frac{1}{x}; \qquad y = 1 \text{ when } x = 1$$

To put this in standard form, we divide by x^2:

$$\frac{\mathrm{d}y}{\mathrm{d}x} + \frac{y}{x} = \frac{1}{x^3}$$

Now we can read off the integrating factor:

$$\exp\left(\int 1/x\,\mathrm{d}x\right) = \exp(\log(x)) = x$$

and multiply through

$$x\frac{\mathrm{d}y}{\mathrm{d}x} + y = \frac{1}{x^2}$$

The LHS integrates to xy and the RHS to $-1/x$. The solution is thus

$$xy = -\frac{1}{x} + C$$

We fix C by substituting $y = x = 1$:

$$1 = -1 + C$$

so $C = 2$ and the solution is

$$xy = -\frac{1}{x} + 2 \qquad \implies \qquad y = \frac{2}{x} - \frac{1}{x^2}$$

9.6 Some longer examples

A body sits at temperature T_0, in a room whose air temperature is also T_0. If we start to heat it with power P, its temperature starts to rise and at temperature T it loses energy at the rate $hA(T - T_0)$ where A is the surface area and h is a constant called the heat transfer coefficient (Newton's law of cooling). Its temperature is governed by a differential equation:

$$\frac{\mathrm{d}T}{\mathrm{d}t} = \frac{P - hA(T - T_0)}{C}$$

where t is the time since heating began and C is another constant, called the heat capacity. We can solve this equation to find that the temperature at time T is

$$T = T_0 + \frac{P(1 - \exp(-hAt/C))}{hA}$$

To solve this equation, we separate variables:

$$\int \frac{C\,\mathrm{d}T}{P - hA(T - T_0)} = \int \mathrm{d}t$$

$$\frac{C\log(P - hA(T - T_0))}{-hA} = t + K$$

and we determine the constant K by substituting $t = 0$ and $T = T_0$:

$$\frac{C\log(P)}{-hA} = K$$

Moving K to the left and combining logarithms gives

$$-\frac{C}{hA}\log(1 - hA(T - T_0)/P) = t$$

and hence

$$1 - hA(T - T_0)/P = \exp(-hAt/C)$$

$$T = T_0 + \frac{P(1 - \exp(-hAt/C))}{hA}$$

As $t \to \infty$, the exponential term tends to zero and the temperature converges to

$$T = T_0 + P/(hA)$$

As expected, this is the point where the loss of heat, $(T - T_0)hA$ exactly balances the power of the heater, P.

Some plausible values:

A saucepan of water of diameter 18 cm and height 10 cm holds about 2.5 kg of water and has a surface area of about $0.1\mathrm{m}^2$. Internet searches suggest that h should be about

$10\,\mathrm{Wm^{-2}K^{-1}}$ and the heat capacity of water is $4200\,\mathrm{Jkg^{-1}K^{-1}}$, giving $C = 4200 \times 2.5 \approx 10000\,\mathrm{JK^{-1}}$. To represent room temperature, take $T_0 = 20°\mathrm{C}$. Water boils at $100°\mathrm{C}$. As far as the difference $T - T_0$ goes, °C and K are identical.

This allows us to answer some questions:

- How long would a 1 kW heater take to boil the water?
- How long would a 2 kW heater take to boil the water?
- How powerful a heater is needed to make the water boil?

Answers:

- Here $T = 100$, $P = 1000$ and the equation is

  $$100 = 20 + 1000(1 - \mathrm{e}^{-t/10000})$$

  Answer: 833 s.

- Here $T = 200$, $P = 1000$ and the equation is

  $$100 = 20 + 1000(1 - \mathrm{e}^{-t/10000})$$

  Answer: 408 s. This is less than half the time taken by the 1 kW heater: less heat is lost because the pan spends less time heating up.

- The limiting temperature is $T = T_0 + P/(hA)$, so we solve $100 = 20 + P/(10 \times 0.1)$, i.e. $P = 80$. An element of power less than 80 W will never boil the water, an element of higher power will eventually boil it. Note that the model breaks down when the water boils because of the latent heat of vaporisation.

At time t, a voltage $V\sin(\omega t)$ (for some constant $\omega > 0$) is applied to an inductor of inductance L and a resistor of resistance R, in series. The current I in the circuit satisfies the ODE

$$L\frac{\mathrm{d}I}{\mathrm{d}t} + RI = V\sin(\omega t)$$

We can solve this equation to find that the current at time t is

$$\frac{V}{\sqrt{R^2 + L^2\omega^2}}\sin(\omega t - \beta) + C\exp(-(R/L)t)$$

where C is an unknown constant and

$$\beta = \tan^{-1}\frac{L\omega}{R} \in (0, \pi/2)$$

To solve this equation, we use an integrating factor. In standard form, the ODE is

$$\frac{\mathrm{d}I}{\mathrm{d}t} + (R/L)I = (V/L)\sin(\omega t)$$

so the integrating factor is $\exp((R/L)t)$. Multiplying through,

$$\exp((R/L)t)\frac{\mathrm{d}I}{\mathrm{d}t} + (R/L)\exp((R/L)t)I = (V/L)\exp((R/L)t)\sin(\omega t)$$

and the LHS integrates to

$$I\exp((R/L)t)$$

The RHS can be integrated (7.9) by parts or (7.16) by converting to complex exponentials. Whichever way you do it, the answer is

$$\frac{V}{L^2\omega^2 + R^2}\exp((R/L)t)(R\sin(\omega t) - L\omega\cos(\omega t))$$

Now we can assemble the solution to the ODE:

$$I = \frac{V}{L^2\omega^2 + R^2}(R\sin(\omega t) - L\omega\cos(\omega t)) + C\exp(-(R/L)t)$$

where C is some unknown constant. Another way to write this is with an auxiliary angle: if we suppose

$$R\sin(\omega t) - L\omega\cos(\omega t) = A\sin(\omega t + \alpha)$$

and expand the RHS and equate coefficient of $\sin(\omega t)$ and $\cos(\omega t)$, we get the equations

$$R = A\cos(\alpha); \qquad L\omega = A\sin(\alpha)$$

These are exactly the equations we looked at (6.8) in the context of radian measure. The solutions are of the form

$$A = \sqrt{R^2 + L^2\omega^2} \qquad \alpha = -\tan^{-1}(L\omega/R)$$

Here, L and R are certainly positive and we assumed that $\omega > 0$. If we take A to be the positive square root then $\tan^{-1}$ is the principal branch, between 0 and $-\pi/2$. Our solution now looks like

$$\frac{V}{\sqrt{R^2 + L^2\omega^2}}\sin(\omega t - \beta) + C\exp(-(R/L)t)$$

where $\beta = -\alpha \in (0, \pi/2)$.

As t becomes larger, the second term tends to zero (whatever the value of the unknown constant C). In the limit, the current has the same frequency as the voltage, but is out of phase: it is retarded by $\tan^{-1}(L\omega/R)$.

A Appendix, for enthusiasts

This section contains some details omitted from the main text. Working through any or all of this is entirely optional, so don't worry too much if you try but don't quite get it.

A.1 The Chain Rule

Here is a proof of the Chain Rule. Hypotheses are as stated in (3.12).

Suppose g is differentiable at x, so $g(x+h) = g(x) + hg'(x) + r(h)$ where $r(h)/h \to 0$ as $h \to 0$ and f is differentiable at $g(x)$, so $f(g(x)+k) = f(g(x)) + kf'(g(x)) + s(k)$ where $s(k)/k \to 0$ as $k \to 0$. We have:

$$\begin{aligned} f(g(x+h)) &= f(g(x)+hg'(x)+r(h)) \\ &= f(g(x)) + [hg'(x)+r(h)]f'(g(x)) + s(hg'(x)+r(h)) \\ &= f(g(x)) + hf'(g(x))g'(x) + r(h)f'(g(x)) + s(hg'(x)+r(h)) \end{aligned}$$

Because $r(h)/h \to 0$ as $h \to 0$, we have $r(h)f'(g(x)) \to 0$ as $h \to 0$. It remains to show that $s(hg'(x)+r(h))/h \to 0$ as $h \to 0$. At this point, it becomes convenient to introduce a function

$$\sigma(k) = \begin{cases} s(k)/k & \text{if } k \neq 0 \\ 0 & \text{if } k = 0 \end{cases}$$

so $s(k) = k\sigma(k)$ for all k and $\sigma(k) \to 0$ as $k \to 0$. Now,

$$\frac{s(hg'(x)+r(h))}{h} = \frac{(hg'(x)+r(h))\sigma(hg'(x)+r(h))}{h} = (g'(x)+r(h)/h)\sigma(hg'(x)+r(h))$$

As $h \to 0$, $hg'(x)+r(h) \to 0$, so $\sigma(hg'(x)+r(h)) \to 0$. The other factor tends to $g'(x)$, so the product tends to zero. This proves the chain rule: $(f \circ g)'(x) = f'(g(x))g'(x)$.

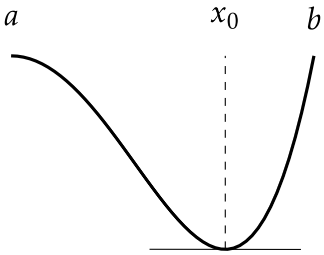
The point of introducing $\sigma(k)$, instead of just working with $s(k)/k$, is that $\sigma(0)$ is well defined, and indeed a point of continuity of σ, whereas $s(0)/0$ makes no sense. This means that we don't have to worry about the possibility that $hg'(x)+r(h)$ could be zero.

A.2 The Mean Value Theorem

The Mean Value Theorem (MVT) was introduced (4.4) without proof. We shall see a derivation later in the course, once we have developed some of the theory of integration, but here is a direct proof (something similar to this will appear in Real Analysis). We begin with a special case that has its own name.

If $f : [a,b] \to \mathbb{R}$ is continuous and f is differentiable on (a,b) and $f(a) = f(b)$ then, for some $x_0 \in (a,b)$, we have $f'(x_0) = 0$

This is *Rolle's Theorem*.

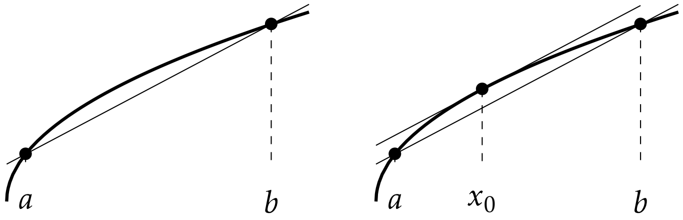


To see this, note firstly that Rolle's Theorem is trivially true if f is constant (because then $f'(x) = 0$ for all x).

Now suppose f is not constant. We know from the Extreme Value Theorem (4.1), that f takes on its minimum and maximum values; these are not the same, because f is not constant. Because $f(a) = f(b)$, either the minimum or the maximum must be attained at a point $x_0 \in (a,b)$. At this point, $f'(x_0) = 0$.

Note that there could be many x_0 for which $f'(x_0) = 0$; Rolle's Theorem guarantees the existence of at least one.

Now suppose $f : [a,b] \to \mathbb{R}$ is continuous and differentiable on (a,b) and consider the chord between $(a, f(a))$ and $(b, f(b))$.



The Mean Value Theorem (MVT) says that:

There exists $x_0 \in (a,b)$ such that

$$f'(x_0) = \frac{f(b) - f(a)}{b - a}$$

(the tangent at x_0 is parallel to the chord from $(a, f(a))$ to $(b, f(b))$.)

To see this, note that the equation of the chord is

$$g(x) = f(a) + \frac{f(b) - f(a)}{b - a}(x - a)$$

and the difference $f - g$ has $(f - g)(a) = (f - g)(b) = 0$, so satisfies the hypotheses of Rolle's Theorem: there must therefore exist $x_0 \in (a, b)$ such that $(f - g)'(x_0) = 0$, i.e. $f'(x_0) = g'(x_0)$. But $g'(x_0) = (f(b) - f(a))/(b - a)$.

If we change the notation so $a = x$, $b = x + h$ $(h > 0)$ then we have

$$f'(x_0) = \frac{f(x+h) - f(x)}{h} \qquad \Longleftrightarrow \qquad f(x+h) = f(x) + hf'(x_0)$$

for some $y \in (x, x + h)$. Similarly, if we change the notation so $a = x + h$, $b = x$ $(h < 0)$ then we have

$$f'(x_0) = \frac{f(x) - f(x+h)}{-h} \qquad \Longleftrightarrow \qquad f(x+h) = f(x) + hf'(x_0)$$

for some $x_0 \in (x + h, x)$. Regardless of sign, we have

$$f(x+h) = f(x) + hf'(x_0)$$

for some x_0 between x and $x + h$.

A.3 The Inverse Function Theorem for Differentiable Functions

We saw (4.6) that if $f : [a.b] \to \mathbb{R}$ has a single-signed derivative then the inverse function f^{-1} exists and, assuming it is differentiable, satisfies $(f^{-1})'(y) = 1/f'(f^{-1}(y))$ (this is rather informally stated; see the main body of the notes for the exact statement). As remarked, proving that f^{-1} is differentiable is a bit tricky. Here is one way to do it:

Fix $y \in (f(a), f(b))$, so $y = f(x)$ where $x \in (a, b)$ is uniquely determined. For any small k, there is a unique h (depending on k) such that $f(x+h) = y+k$ (we should really write $h(k)$, but abbreviate this to h to simplify notation). We can write h explicitly as $h = f^{-1}(y+k)-x$; we can see from this that, because f^{-1} is continuous, $h \to 0$ as $k \to 0$. Now, the definition of differentiability gives:

$$y + k = f(x + h) = f(x) + f'(x)h + r(h)$$

and we can cancel $y = f(x)$ to give

$$k = f'(x)h + r(h)$$

Dividing by h, we have $k/h = f'(x) + r(h)/h$; as $k \to 0$, $h \to 0$ and so $r(h)/h \to 0$; that is, $k/h \to f'(x)$ so $h/k \to 1/f'(x)$. Now, we solve for h:

$$h = \frac{k}{f'(x)} - \frac{r(h)}{f'(x)}$$

and go back to $f(x+h) = y+k$ to give

$$f^{-1}(y+k) = x+h = f^{-1}(y) + \frac{k}{f'(x)} - \frac{r(h)}{f'(x)}$$

To finish, we need to show that the remainder term tends to zero when divided by k:

$$\frac{r(h)}{f'(x)k} = \frac{r(h)}{h}\frac{h}{k}\frac{1}{f'(x)}$$

As $k \to 0$, $h \to 0$ so $r(h)/h \to 0$ and also $h/k \to 1/f'(x)$. The limit is therefore zero, and we can see that f^{-1} is differentiable with $(f^{-1})(y) = 1/f'(x)$.

A.4 The circular functions

Here is a derivation of the fundamental properties of the circular (trigonometric) functions, as stated in (6.4), from their defining differential equations, The most difficult part is the construction of π.

- Let $u(x) = \sin(-x)$. Then $u'' = -u$, $u(0) = 0$ and $u'(0) = -1$. We must have $u(x) = A\cos(x) + B\sin(x)$ and hence $u'(x) = -A\sin(x) + B\cos(x)$. Since $u(0) = 0$, $0 = A$. Since $u'(0) = -1$, $B = -1$, so $\sin(-x) = -\sin(x)$. Similarly, $\cos(-x) = \cos(x)$, so sin is odd and cos is even.

- Fix y and let $u(x) = \sin(x+y)$. Then $u'' = -u$, so we must have $u(x) = A\cos(x) + B\sin(x)$, i.e. $\sin(x+y) = A\cos(x) + B\sin(x)$ for all x. Putting $x = 0$ we see that $\sin(y) = A$. Differentiating gives $\cos(x+y) = -A\sin(x) + B\cos(x)$. Putting $x = 0$ gives $\cos(y) = B$. Assembling these gives

  $$\sin(x+y) = \sin(x)\cos(y) + \cos(x)\sin(y); \qquad \cos(x+y) = \cos(x)\cos(y) - \sin(x)\sin(y).$$

- Differentiate $\cos^2(x) + \sin^2(x)$ to give $-2\sin(x)\cos(x) + 2\sin(x)\cos(x) = 0$, so $\cos^2 + \sin^2$ is constant. Since $\cos^2(0) + \sin^2(0) = 1$, $\cos^2(x) + \sin^2(x) = 1$ for all x.

- We have $\sin(0) = 0$ and $\sin'(0) = 1$, so $\sin(x)$ is initially increasing from 0 as x increases from 0. We claim that the equation $\sin(x) = 0$ has a solution $x > 0$. If not then, by the IVT, we have $\sin(x) > 0$ for all x. If for some x we have $\sin(x) > 1/\sqrt{2}$ then (because $\cos^2 x + \sin^2 x = 1$), we have $\cos(x) < 1/\sqrt{2}$. We then have $\sin(0) - \cos(0) < 0$ and $\sin(x) - \cos(x) > 0$, so $\sin(x) = \sin(y)$ for some $y \in (0,x)$. Now, $\sin(2y) = \cos^2(y) - \sin^2(y) = 0$ and $\sin(4y) = 2\sin(2y)\cos(2y) = 0$. We are left with the remaining case where $0 < \sin(x) \leq 1/\sqrt{2}$ for all $x > 0$, so $\cos(x) \geq 1/\sqrt{2}$ for all $x > 0$. By the MVT, $\sin(x) = x\cos(x_0)$ for some $x_0 \in (0,x)$, so $\sin(x) \geq x/\sqrt{2} \to \infty$ which is impossible because $|\sin(x)| \leq 1$.

- Let π be the smallest positive solution to $\sin(\pi) = 0$. Because π is the minimal positive solution and sin is increasing at 0, we must have $\sin > 0$ on $(0, \pi)$; since $\cos' = -\sin$, cos is decreasing on $(0, \pi)$. We have $\cos^2(\pi) = 1$, so we must have $\cos(\pi) = -1$. Multiple angle formulae now give $\sin(x + \pi) = -\sin(x)$ and $\cos(x + \pi) = -\cos(x)$ and hence $\sin(x + 2\pi) = \sin(x)$ and $\cos(x + 2\pi) = \cos(x)$ so sin and cos are 2π-periodic. The sign pattern shows that there is no smaller period.

A.5 Special values of circular functions

Here is a careful derivation of the sign properties of the circular functions and some special values they take on.

- Triple etc. angle formulae follow from the addition rules in the usual way.
- We have already seen that $\sin(0) = 0$, $\cos(0) = 1$, $\sin(\pi) = 0$, $\cos(\pi) = -1$, that sin and cos are 2π-periodic, that $\sin(x + \pi) = -\sin(x)$ and that $\cos(x + \pi) = -\cos(x)$. We also have $\sin > 0$ on $(0, \pi)$, so we must have $\sin < 0$ on $(\pi, 2\pi)$.
- Let $s = \sin(\pi/2)$ and $c = \cos(\pi/2)$. Then $-1 = \cos(\pi) = c^2 - s^2$, and $1 = c^2 + s^2$. Add to give $0 = 2c^2$, so $c = 0$, and hence $s^2 = 1$. Because $\sin > 0$ on $(0, \pi)$, we have $\sin(\pi/2) = 1$. We know that cos is decreasing on $(0, \pi)$, so we have $\cos > 0$ on $(0, \pi/2)$ and $\cos < 0$ on $(\pi/2, \pi)$. Since $\cos(x + \pi) = -\cos(x)$, we must have $\cos > 0$ on $(\pi/2, 3\pi/2)$ and $\cos < 0$ on $(3\pi/2, 2\pi)$. We can now draw the ASTC diagram.
- Let $s = \sin(\pi/4)$ and $c = \cos(\pi/4)$. Then $c^2 - s^2 = \cos(\pi/2) = 0$ and $c^2 + s^2 = 1$. Adding and subtracting give $2c^2 = 2s^2 = 1$. From the sign diagram, $s > 0$ and $c > 0$; we therefore have $\cos(\pi/4) = \sin \pi/4) = \sqrt{1/2} = \sqrt{2}/2$.
- Let $s = \sin(\pi/6)$ and $c = \cos(\pi/6)$. Then (cosine triple angle formula) $0 = \cos(3\pi/6) = 4c^3 - 3c$. We know from the sign diagram that $c > 0$, so $4c^2 = 3$ and hence $\cos(\pi/6) = c = \sqrt{3}/2$. We also have $c^2+s^2 = 1$ and $s > 0$, so $\sin(\pi/6) = 1/2$. From the double-angle formulae, $\sin(\pi/3) = \sqrt{3}/2$ and $\cos(\pi/3) = 1/2$.

A.6 Radian Measure

Here are the details from (6.8) about radian measure (equivalently, plane polar coordinates, which will play a role in the second part of the course).

- Any solution to (A) must, because $\cos^2(\theta) + \sin^2(\theta) = 1$, satisfy $x^2 + y^2 = r^2$.
- Any solution to (A) with $x \neq 0$ must have $\tan(\theta) = \sin(\theta)/\cos(\theta) = y/x$.

- Any solution to (B) with $x \neq 0$ must, because $\sec^2(\theta) = 1 + \tan^2(\theta)$, satisfy $\sec^2(\theta) = 1 + y^2/x^2 = r^2/x^2$, so $\cos^2(\theta) = x^2/r^2$ and $\sin^2(\theta) = 1 - \cos^2(\theta) = y^2/r^2$. We thus have $x = \pm r\cos\theta$ and $y = \pm r\sin\theta$.

The ambiguities in signs are associated with the fact that tan is π-periodic, so the equation $\tan\theta = y/x$ has two solutions in $(-\pi, \pi]$, differing by π. To completely solve the problem, we consider the four quadrants separately (also the case $x = 0$ and, for convenience, $y = 0$). Note that any interval $(\alpha, \alpha + \pi/2)$ cannot contain two different inverse tangents to y/x.

- If $x > 0$ and $y > 0$, we must have $\sin\theta > 0$ and $\cos(\theta) > 0$, so $\theta \in (0, \pi/2)$. Because $y/x > 0$, $\tan^{-1}(y/x) \in (0, \pi/2)$, so we must have $\theta = \tan^{-1}(y/x)$.
- If $x > 0$ and $y < 0$, we must have $\sin\theta < 0$ and $\cos(\theta) > 0$, so $\theta \in (-\pi/2, 0)$. Because $y/x < 0$, $\tan^{-1}(y/x) \in (0, \pi/2)$, so we must have $\theta = \tan^{-1}(y/x)$.
- If $x < 0$ and $y > 0$, we must have $\sin\theta > 0$ and $\cos(\theta) < 0$, so $\theta \in (\pi/2, \pi)$. Because $y/x < 0$, $\tan^{-1}(y/x) \in (-\pi/2, 0)$, so we have to add π to find the solution we want: $\theta = \pi + \tan^{-1}(y/x)$.
- If $x < 0$ and $y < 0$, we must have $\sin\theta < 0$ and $\cos(\theta) < 0$, so $\theta \in (-\pi, -\pi/2)$. Because $y/x > 0$, $\tan^{-1}(y/x) \in (0, \pi/2)$, so we have to subtract π to find the solution we want: $\theta = \tan^{-1}(y/x) - \pi$.
- If $y = 0$, then $\sin(\theta) = 0$ so $\theta = 0$ or $\theta = \pi$, corresponding to $\cos(\theta) = 1$ and $\cos\theta = -1$. We must therefore have $\theta = 0$ if $x > 0$ and $\theta = \pi$ if $x < 0$.
- If $x = 0$ then $\cos(\theta) = 0$, so $\theta = \pi/2$ or $\theta = -\pi/2$, corresponding to $\sin\theta = 1$ and $\sin(\theta) = -1$. We must therefore have $\theta = \pi/2$ if $y > 0$ and $\theta = -\pi/2$ if $y < 0$.

A.7 Higher Derivative Tests

Here is the ultimate version of the second derivative test (8.4). If $f \in C^{N+1}$ and $f^{(n)}(x_0) = 0$ for $1 \le n \le N$ but $f^{(N+1)}(x_0) \neq 0$ then the Taylor polynomial of f at x_0 to order N is just $f(x_0)$, so we have

$$f(x) = f(x_0) + \frac{(x - x_0)^{N+1}}{(N+1)!} f^{(N+1)}(c)$$

using the Lagrange form of the remainder. We can then conclude:

- If N is odd and $f^{(N+1)}(x_0) > 0$ then x_0 is a local minimum
- If N is odd and $f^{(N+1)}(x_0) < 0$ then x_0 is a local maximum
- If N is even then x_0 is neither a local maximum nor a local minimum

The first two work exactly like the $N = 1$ case because $(x-x_0)^{N+1} > 0$. In the third case, because $(x-x_0)^{N+1}$ changes sign according to whether $x < x_0$ or $x > x_0$, if x is close to x_0 then $f(x)-f(x_0)$ also changes sign in a similar way: the exact sign pattern depends on the sign of $f^{(N+1)}(x_0)$, but it is always different in the $x < x_0$ and $x > x_0$ cases.

A.8 Complex Taylor Series

This is why the Taylor series for the exponential function converges for complex numbers, as well as real numbers (8.6).

Fix $\theta \in \mathbb{R}$ and considering $f : \mathbb{R} \to \mathbb{C}$ defined by

$$f(x) = \exp(x\mathrm{e}^{\mathrm{i}\theta})$$

By the chain rule,

$$f'(x) = \mathrm{e}^{\mathrm{i}\theta}\exp(x\mathrm{e}^{\mathrm{i}\theta}); \qquad f''(x) = \mathrm{e}^{2\mathrm{i}\theta}\exp(x\mathrm{e}^{\mathrm{i}\theta})$$

and by induction

$$f^{(n)}(x) = \mathrm{e}^{\mathrm{i}n\theta}\exp(x\mathrm{e}^{\mathrm{i}\theta})$$

In particular, $f^{(n)}(0) = \mathrm{e}^{\mathrm{i}n\theta}$. This gives us the Taylor series at the origin:

$$\sum_{n=0}^{N}\frac{\mathrm{e}^{\mathrm{i}n\theta}x^n}{n!} + \frac{1}{N!}\int_0^x (x-t)^N\mathrm{e}^{\mathrm{i}(N+1)\theta}\exp(t\mathrm{e}^{\mathrm{i}\theta})\,\mathrm{d}t$$

Looking at the remainder term, the width of the integration region is $|x|$, $|(x-t)^N| \le |x|^N$, $|\mathrm{e}^{\mathrm{i}(N+1)\theta}| = 1$ and

$$|\exp(t\mathrm{e}^{\mathrm{i}\theta})| = \exp(\mathrm{Re}(t\mathrm{e}^{\mathrm{i}\theta})) = \exp(t\cos(\theta)) \le \exp(|x|)$$

Combining these gives an upper bound for the remainder:

$$\frac{|x|^{N+1}\mathrm{e}^{|x|}}{N!}$$

which tends to zero as $N \to \infty$. If $z = r\mathrm{e}^{\mathrm{i}\theta}$, we can let $x = r$ to see that

$$\exp(z) = \sum_{n=0}^{\infty}\frac{z^n}{n!}$$

For the other trigonometric and hyperbolic functions, express the function in terms of exponentials and use this result.

A.9 The Binomial Theorem

The error term in the Taylor series for $(1+x)^\alpha$ (A.9) is

$$\frac{1}{N!}\int_0^x (x-t)^N f^{(N+1)}(t)\,\mathrm{d}t = \binom{\alpha}{N}(N-\alpha)\int_0^x (x-t)^N (1+t)^{\alpha-N-1}$$

This is quite tricky to control. We can isolate the N dependency in the integral by rewriting this as

$$\binom{\alpha}{N}(\alpha-N)\int_0^x (1+t)^{\alpha-1}\left(\frac{x-t}{1+t}\right)^N \mathrm{d}t$$

Now, if $x>0$ then $0<t<x$ and $(x-t)/(1+t)$ is positive and decreasing as a function of t, so it attains its maximum magnitude of x at $t=0$. If $x<0$ then $x<t<0$ and $(x-t)/(1+t)$ is negative and decreasing as a function of t, so it attains its maximum magnitude magnitude of $-x$ at $t=0$. It follows that, for the whole of the region of integration, $|[(x-t)/(1+t)]^N| \leq |x|^N$. We can therefore estimate the error term above by

$$R_N = \binom{\alpha}{N}(\alpha-N)|x|^N\int_0^x (1+t)^{\alpha-1}\,\mathrm{d}t$$

We could work out the remaining integral easily enough, but we don't need to (because it doesn't depend on N). We now show that this estimate converges to zero. We do this by looking at the ratio of two consecutive terms:

$$\frac{R_{N+1}}{R_N} = \frac{\alpha-N}{N+1}\frac{\alpha-N-1}{\alpha-N}|x|$$

which tends to $|x|$ as $n\to\infty$. Because $(1+|x|)/2>|x|$, we eventually have $R_{N+1}/R_N < (1+|x|)/2$; say this holds for $N>M$. We then have $R_N < R_M[(1+|x|)/2]^{N-M}$; this tends to zero as $N\to\infty$ because $|x|<1$, so $(1+|x|)/2<1$.