4CCM111A Calculus I Lecture Notes

Dr. Aled Walker
Dr. David Sheard
Department of Mathematics, King's College London.

Based on previous notes by Dr. Paul P. Cook, Prof. Ton Coolen, Prof. Peter Sollich, and Prof. Gerard Watts.

26th September 2023

"And what are these Fluxions? The Velocities of evanescent Increments? And what are these same evanescent Increments? They are neither finite Quantities nor Quantities infinitely small, nor yet nothing. May we not call them the ghosts of departed quantities?"

George Berkeley, in The Analyst, 1734.

0. Contents

1 Introduction **7**

1.1 Introduction 7

1.2 Course organisation 14

1.2.1 Programming 14

1.2.2 Skills sessions 14

1.2.3 Tutorials 14

1.2.4 Assessment 15

2 Functions of One Variable **19**

2.1 Mathematical Definitions 19

2.2 The Definition of a Function. 21

2.2.1 Domain and Range 24

2.2.2 Injections, Surjections and Bijections 27

2.3 Inverse functions 33

2.4 Monotonically increasing and monotonically decreasing functions 37

2.5 Graph sketching 37

2.6 Some standard functions 43

2.6.1 Exponents, Logarithms and Polynomial Functions 43

2.6.2 Exponential functions 46

2.6.3 The Logarithm 49

2.6.4 Polynomial Functions 52

2.6.5 A look ahead at derivatives... 53

2.6.6 The Exponential Function: power series 55

2.6.7 The Trigonometric Functions. 58

2.6.8 The Hyperbolic Functions. 71

2.6.9 Trigonometric and Hyperbolic Functions with Complex Variables 79

2.6.10 Functions of Functions 85

2.7 A Zoo of Functions 86

3 The Limit **89**
3.1 Motivation 89
3.2 The Existence of the Limit. 92
3.2.1 Continuous Functions. 101
3.2.2 Limits Involving Infinity. 106
3.3 Working with Limits 107
3.3.1 Rules for Limits of Composite Expressions. 108
3.3.2 Multiple Limits. 112
3.3.3 Evaluating standard limits 114

4 The Derivative **125**
4.1 Differentiation from first principles 128
4.2 Differentiable Functions 136
4.3 Properties of the derivative 139
4.3.1 The Chain Rule 139
4.3.2 The Sum Rule 141
4.3.3 The Product Rule 142
4.3.4 The Quotient Rule 143
4.4 Derivatives of Implicit Functions 143
4.4.1 Inverse Functions 144
4.4.2 Curves 146
4.4.3 Parametric Functions 150
4.5 The Mean Value Theorem 152

5 Integration **157**
5.0.1 Motion at constant speed: the area under a straight line 158
5.1 The Riemann Integral 159
5.2 The Fundamental Theorem of Calculus 171
5.2.1 Indefinite and Definite Integrals. 176
5.3 Properties of the Integral and some Techniques for Integration 177
5.3.1 Solving Integrals by Substitution 179
5.3.2 Integration by Parts 186
5.3.3 Partial Fractions 189
5.3.4 Recursion Relations 193
5.4 Some Applications: Length, Area and Volume. 195

5.4.1 Areas of Circles and Ellipses 195
5.4.2 Volumes of Revolution 198
5.4.3 The Length of a Curve 200

6 Power Series **203**
6.1 Infinite sums 204
6.2 Convergence 204
6.2.1 Series Convergence Criteria. 207
6.3 Series as Functions of x. 209
6.4 Taylor's Theorem 214
6.4.1 Short-cuts for finding a Taylor Expansion 222
6.5 l'Hôpital's Rule 223

Appendix **227**
6.A The Ratio Test Implies the Root Test. 227

Acknowledgements

These lecture notes were first written in August-September 2015, with major revisions undertaken in August-September 2018 and September 2023. They built upon the previous excellent set of notes written by Prof. Ton Coolen. Each overhaul of a standard set of lecture notes builds its foundations on the previous courses and the work of Prof. Peter Sollich and Dr Gerard Watts has had a profound impact on previous iterations of the course. All typographical errors and errors of any other type are, of course, my fault and I will be grateful for all comments of any kind of mistake. Thanks are due to Jane Bennett-Rees, Robert Evans, Asuka Kumon, (Anthony) Peter Young, Michael Yiasemides, Haodong Sun, Shuo Huang, Pablo de Castro, Rishi Mouland, Alessio Sarti, Senan Sekhon, Keith Glennon, Veno Mramor, Véronique Fischer and many others for pointing out typos in previous versions and making other comments on the text.

Aled Walker, 13^{th} September, 2023.

1. Introduction

In which we introduce and motivate the study of calculus; describe the course in outline; explain the organisation of the course and how it will be examined.

This will be covered during week 1 of lectures.

1.1 Introduction

Calculus is the mathematical study of change. It concerns the study of functions and their derivatives, integrals, and limits – all terms that we will define in the coming weeks. First developed in the 17^{th} century, by the English mathematician and physicist Isaac Newton and the German polymath Gottfried Leibniz, calculus is sometimes viewed as the 'starting point' of modern mathematics. It is apt to begin your university studies with it.

Depending on the particular education system in which you were raised, you may or may not have encountered calculus before. As it happens, those students who were educated in the UK, and who took both Maths and Further Maths A-Level, will already be well-acquainted with the topic – though perhaps to the detriment of their exposure to alternative mathematical subjects. Other students will have had a different experience, and others different again. Thus, part of the purpose of this course is to create a uniform starting point of 'pre-university level' mathematics, giving a stable platform from which you may tackle the rest of a mathematics degree.

Yet, such a description might suggest that there will be little for the well-prepared A-Level student to learn here. This is not true. The study of calculus is suffused with subtlety, and students who think they know the subject may find themselves surprised, when the same ground that they covered at school is reconsidered at university. Even if the basics are already familiar to you, they are but the beginning of your journey.

The questions that calculus addresses are related to old philosophical quandaries – thousands of years old, in fact. These concern the notion of *divisibility*. Such discussion will not directly help us to calculate the answers to the problems on the calculus tutorial sheets in the coming weeks! Yet, these historical points will serve to explain why the mathematical theory developed in the way that it did, and will provide some extra context that may be illuminating in other ways.

A certain school of Ancient Greek philosophers (the Atomists, Leucippus and Democritus of the 5^{th} century BCE) contended that physical space is made up fundamental objects which are physically indivisible. They called these objects atoms; this is not the same use of the word atom as in modern physics, of course – we now know physical atoms to be highly divisible – but more of a philosophical construct. For the later philosopher Plato, who was an atomist and a geometer, there arose the question of whether the theory of atomism applied to *mathematical* objects: this is where the trouble begins.

With the assumption of indivisible geometric atoms, many paradoxes readily arise. For example, given a continuous line, we might ask: how many atoms does it contain? Suppose we posit that a continuous line is made of N atomic parts. A problem now arises, as we can split a continuous line in two, forming two new continuous lines. If each continuous line contains N atoms, then the pair of lines must therefore contain $2N$ atoms together – contradicting the assumption that the original line contained N atoms.

There is another possibility: perhaps each continuous line contains an infinite number of indivisible parts? This is indeed the modern mathematical interpretation, but it raises further questions. Of particular relevance for calculus – when performing a process called 'integration', involving summing many small lengths together–: how can a sum of infinitely many points create a continuous line of finite length?

The most famous poser of such paradoxical problems was Zeno of Elea (another Ancient Greek philosopher, about whom we know very little except that which was communicated by Plato and Aristotle). His paradox of 'Achilles and the tortoise' encapsulates the difficulties of considering space as infinitely divisible. It goes like this. Achilles is chasing a tortoise. Achilles runs much faster than the tortoise, so you would expect him to easily catch the tortoise. However, to do so he first has to catch up with where the tortoise currently is. When he has done that, the tortoise will have walked a little way forward. So Achilles much catch up with where the tortoise has walked to. But by the time he does that, the tortoise will have walked a bit further on... and so on and so forth, continuing forever, as one considers space to be infinitely divisible. To give the succinct summary of Aristotle (in translation):

> "In a race, the quickest runner can never overtake the slowest, since the pursuer

> must first reach the point whence the pursued started, so that the slower must always hold a lead."

Whether a line comprises finitely many 'indivisibles' or infinitely many 'infinitesimals', a paradox seems to arise.

Parts of the ancient world did make use of infinitesimal mathematics, although such methods were not trusted. For example, Archimedes was able to find the volume of a sphere, a cone and a cylinder using his "method of indivisibles", but he verified his results using geometrical derivations as well – just to be safe. Zu Chongzhi and his son Zu Gengzhi – Chinese mathematicians of the 5^{th} and 6^{th} centuries – were able to make similar computations. They also produced an estimate for π that was not surpassed for nearly a thousand years.

In Europe, the collapse of the Western Roman Empire greatly interrupted the transmission of knowledge from the Ancient Greeks; the question of the indivisibles or infinitesimals remained unprogressed. Geometric understanding only returned to Europe in the Renaissance, when texts of Euclid, Archimedes, Apollonius, Diophantus and others arrived again in Italy, often via Arabic translations (where the tradition had been preserved and developed), or from the academy at Trebizond prior to the collapse of the Byzantine Empire.

By the 17^{th} century, much was changed. Many mathematicians in Europe were experimenting with infinitesimals and infinite quantities, such as Evangelista Torricelli, Pierre de Fermat, and John Wallis (an English contemporary of Newton). It is Wallis to whom we owe the symbol ∞ representing infinity. Alongside this, one has René Descartes' groundbreaking (if impenetrable) text *La Géométrie* of 1637, in which he first introduced the coordinate plane as a way of relating algebraic and geometric mathematics. It is from this background that Newton (in the late 1660s) and Leibniz (in the 1670s) constructed their systematised theories, giving to them the credit of 'discovering calculus'. Though Newton's work slightly predated Leibniz's, Newton didn't publish until much later, and it was Leibniz's notation (such as $\frac{dy}{dx}$) that became widely used. This led to a famously bitter dispute over precedence, with Newton accusing Leibniz of plagiarism.

The word calculus means 'pebble' in Latin. Latin was the language of the Roman Empire, and a common language of intellectual enquiry in Europe until the 19^{th} century; thus, as the Romans used pebbles to perform simple arithmetic, so the word calculus came to be associated with computation. Initially it had a broad meaning – Descartes referred to his innovation as *calcul géometrique*, for example – but it came to be used solely for the mathematics around Newton and Leibniz's work. Sometimes the subject was called 'the infinitesimal calculus', to distinguish it from other calculi; at other times you might see it called, rather grandly, '*the*

calculus'. The word now refers primarily to the study of functions, limits, differentials (studying the gradients of curves) and integrals (studying the areas under curves).

However, it should always be remembered that the word has its root in computation and calculation. Newton and Leibniz were not just proposing an abstract mathematical theory of change: they were proposing a theory that could be used for quantitative calculations. We should not be surprised, therefore, that contemporaries of these mathematicians struggled to reckon with how ephemeral infinitesimal quantities could be used to calculate answers to concrete questions – what is the gradient of the tangent to this curve, how fast is the projectile travelling, what distance did the boat travel, etc.

Furthermore, in that era it was highly philosophically disturbing to contemplate the idea that the mathematical world was underpinned by infinitesimal quantities rather than indivisible atoms. The solidity and surety of atoms of geometry sat easily with the idea of a stable universe, in which stable kingdoms are governed by unquestionable systems of law and religion. The idea of infinitesimals which were not mathematically well-understood was not welcome[1] and there are famous examples of respectable and historic figures who rejected the idea of an infinitesimal quantity, notably Descartes (who first embraced and then turned away from infinitesimals) and George Berkeley, who said that infinitesimals – which Newton called 'fluxions' – were

> "the ghosts of departed quantities."

in his famous mid-18^{th}-century polemic.

A great part of mathematics was created to resolve such challenges. This is not just the mathematics of the early-modern era: although certain introductory calculus textbooks make out that the issue of infinitesimals was entirely resolved hundreds of years ago, there were fundamental disagreements well into the 20^{th} century . Look up the Hilbert–Brouwer controversy, or the debates around the so-called *axiom of choice*. In the third-year Fundamentals of Probability course, you will meet another 20^{th} century attempt at taming infinitesimals: Lebesgue's 'measure theory'. In short, there is a wild and difficult mathematical landscape out there, surrounding notions of infinity. We need to start somewhere, but what you'll learn in this course is but a small part of the overall story.

To get a purchase on these questions, let us restrict our attention to calculus itself. Broadly speaking, there are now two theories:

- **Analysis**. This theory was created by the schools of Cauchy (in Paris) and Weierstrass (in Berlin) during the 19^{th} century. They avoided the infinitesimal quantities entirely: all

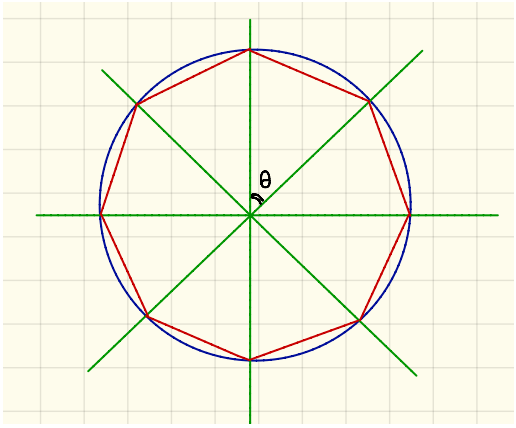
[1] There is an intriguing book called *Infinitesimal* by Amir Alexander which discusses how dangerous an idea the infinitesimal calculus was at the time.

computation was achieved with finite quantities, before introducing an appropriate 'limit' where the finite quantities became smaller and smaller. This is the standard theory: you will study it here, during the Sequences and Series course, and in the 2nd year Real Analysis course. Analysis, alongside *algebra* and *geometry*, is one of the three pillars of modern mathematics.

- **Non-standard analysis**. This is a highly mathematically sophisticated theory, built on the rigorous development of mathematical logic that was undertaken in the first half of the 20^{th} century. Through this, it was understood how to enrich the usual arithmetic of the number line with extra genuinely infinitesimal quantities, smaller than any positive real number and yet bigger than 0. Though it is remarkable that Newton and Leibniz's original ideas (such as fluxions) can be put on a rigorous footing, non-standard analysis is exactly that: non-standard. If you study for a PhD in a related discipline, you may learn about these ideas properly (involving objects called *ultrafilters*). We will not mention non-standard analysis again in these notes, and you should forget that you ever heard about it.

The aim in this course, therefore, is to give you an introduction to the techniques of mathematical *analysis*, in particular to the notion of a *limit*. Informally, taking a limit is simply manipulating a sequence of mathematical objects until they get closer and closer to some other mathematical object (which is called the 'limit' of the sequence). The objects themselves could be numbers, sums, functions, curves, etc. However, it turns out that an informal treatment can also lead to troubling paradoxes, and we must only consider limits when we have carefully justified their use. We close this introduction with some examples, specifically designed to make you really scared.

The first two examples consider taking limits of lines. Firstly, consider the simple slicing up of a disc into similar segments each of which subtends an angle θ at the centre of the circle:



The straight lines (in red above) can be summed, in an attempt to find the length of the circumference of the circle. As the angle θ is decreased the better the approximation of the circumference will be; in the 'limit' where $\theta \to 0$ the approximation of the circumference can be expected to become exact. However, in that limit, the length of the red lines goes to zero and we see we are arguing that an infinite sum of infinitesimal lengths is neither zero nor infinity but is finite. The central question is to understand under which circumstances this is the case.

The second example highlights the problem. Consider the square $ABCD$ whose edges are all of 1 metre long.

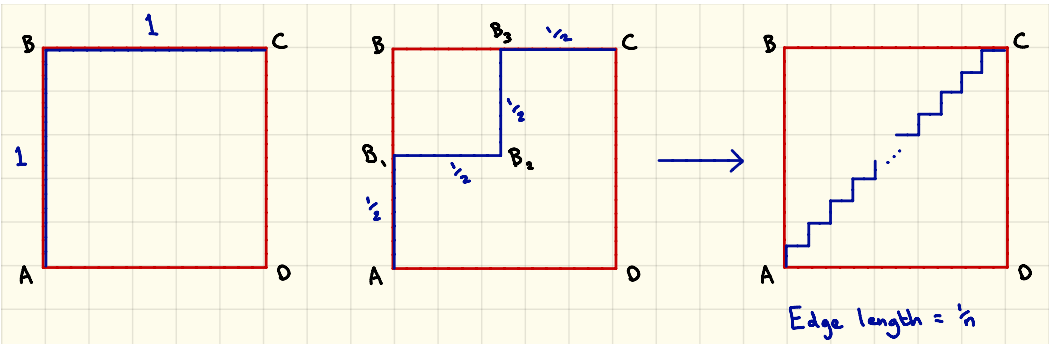


Figure 1.1: An ant travels along a staircase shaped path, the path always has length two metres, regardless of the number of steps.

Imagine an ant travels the path along the edges AB and then BC, denote the path ABC. The distance travelled by the ant is 2 metres. Now consider a new path within the square $AB_1B_2B_3C$ whose edges all measure 0.5 metres as shown by the middle square in figure 1.1, i.e. $|AB_1| = |B_1B_2| = |B_2B_3| = |B_3C| = 0.5m$. The ant travels the path $AB_1B_2B_3C$ which has length $4 \times \frac{1}{2} = 2$ metres. Now repeat the process: halve the length of each edge and form the staircase path whose edges all of length $\frac{1}{n}$ metres. There are now $2n$ edges of length $\frac{1}{n}$ giving a total path length of, again, 2 metres. No matter how small the edge length $\frac{1}{n}$ is made, the total path length remains 2 metres. However we recognise intuitively that as n gets larger the path travelled by the ant seems to approach the diagonal of the unit square, which we know from Pythagoras' theorem has length $\sqrt{2}$ metres. So there is some fundamental difference between the finite n case and the limit case, which seemed not to be the case when approximating the circumference of the circle in the previous example. But what is the difference? It is clear that the process of taking the limit to infinity needs to be carefully constructed!

Consider a final example, which was the subject of significant debate in the 18^{th} century: the series

$$1 - 1 + 1 - 1 + 1 - 1 + 1 - 1 + \dots .$$

The modern view is that this infinite sum does not have an answer: we call it *divergent*, as the

sequence of partial sums does not tend to a limit. Summing from the left, you get the answer 1, then 0, then 1, then 0, and so on, which doesn't approach any fixed number. You will study these kind of series in more depth in the Sequences and Series course. However you would be in the good company of Euler if you tried to argue instead that the series is an example of a geometric series of the form

$$1 + x + x^2 + x^3 + ...$$

(with $x = -1$), which we are used to evaluating as the sum $\frac{1}{1-x} = \frac{1}{2}$. You would also be forgiven if, without a rigorous system of analysis, you argued that you could sum the series to zero or one.

A final confusion that we would like to emphasise is a little more abstract. It also relies on the definition of a 'function', which we have not yet formally introduced. However, it is probable that most students will have been taught about the concept of a function before – even if only in a rudimentary way – so hopefully this discussion will nonetheless be meaningful.

The confusion was apparent already within the foundations of calculus. Suppose that for every function f we had a transformation $f \mapsto f + df$. Here df is meant to represent some very small quantity related to f (which we leave intentionally vague for the purpose of this general discussion). The question: given two functions f and g with a product fg, can you write the map $fg \mapsto fg + d(fg)$ in terms of the quantities df and dg, and if so how? One might try to write something like

$$fg + d(fg) = (f + df)(g + dg) = fg + f\,dg + g\,df + df\,dg$$

after multiplying out the brackets. Subtracting fg from both sides of the equation, one would end up with the formula

$$d(fg) = f\,dg + g\,df + df\,dg.$$

However Leibniz[2] argued that the infinitesimal variation of a product of the two functions is instead

$$d(fg) = f\,dg + g\,df,$$

that is $df\,dg$ is treated as zero, whereas df or dg is infinitesimally small but not treated as zero.

Why is it valid to treat $df\,dg$ as zero? The answer to this question requires us to consider what we mean by a function, and to develop a mathematical method that will allow us to work with finite but very small quantities rather than to attempt an algebra of zeros. It will turn out that there are many mysteries which will need to be vanquished, and even the most beguilingly simple expressions will require a clear definition.

[2]This is Leibniz famous product rule, albeit written in a slightly unfamiliar notation.

1.2 Course organisation

An approximate indication of the programme and the material to be covered is shown in table 1.1.

1.2.1 Programming

This year we are including a small amount of Python programming in Calculus I. The idea is to develop a basic familiarity and competence with Python so that, at the very least, lectures in future modules can demonstrate examples with Python code while reasonably expecting that the majority of students will be able to follow. Many students may already have some knowledge of Python, and we do not intend to teach any advanced skills in this module. Later modules in your degree, particularly courses such as Numerical and Computational Methods with Python, will develop these.

This programming material will mostly be taught asynchronously, with short videos prepared by Dr. Lassina Dembele produced each week (usually released on the Friday of each week). These videos will have accompanying exercises, on an interface called CoCalc. There will also be a few in-person lectures (included as part of the usual Calculus I lecture timetable) given by Dr. Dembele and devoted to programming issues. The schedule for these lectures will be given on the KEATS page. Although some of the mathematics that you will learn to program will be directly related to the Calculus I course, the programming element of the course may also include other mathematical topics, e.g. ideas from linear algebra.

These lecture notes will not include programming examples or exercises. Dr. Dembele will provide all the programming material on KEATS and CoCalc.

1.2.2 Skills sessions

You will have a weekly 'Skills session', in a group of around 80 students. The purpose of this session is to help you turn theoretical knowledge (which you will learn in lectures) into mathematical problem-solving strategies (which you will need for tutorials and for the exams). In addition, general study techniques – sometimes called 'soft skills' – will be discussed. These sessions will be highly interactive.

1.2.3 Tutorials

Each week you will receive a sheet of mathematical problems to complete. Tutorials are 50 minutes sessions, either led by a Graduate Teaching Assistant (GTA) or a member of the

academic faculty, in which you are guided to work on these problems collaboratively. Tutorials will vary in size – sometimes comprising 10 students , sometimes 20 – but will follow a similar structure. Initially, you will be presented with the solution to one of the problems by the tutorial leader. Then, you will be encouraged to work on the rest of the problems in groups. If you have questions or are stuck, the tutorial leader will help you. But they are unlikely to tell you the answer, as being able to work through your own difficulties is a key component of active learning.

You should continue to work on the questions in your own time after the tutorial. The solutions will be released on KEATS the following week.

1.2.4 Assessment

The distribution of marks is as follows:

- Participation Marks, 10%
- Class Tests, 10% combined
- Final examination, 80%.

Participation marks. Three times during the term, you will have an assessed homework (usually based on some questions from the exercise sheets, on which you will have worked during tutorials). You will need to submit your solutions via GradeScope, at the KEATS link provided under the relevant week. This work will then be marked by a Graduate Teaching Assistant and you will get written feedback. You will gain full credit if you submit a decent attempt at each homework, even if there are some mathematical mistakes. These submissions will open during weeks 3, 6, and 9 of the course, and will be due on **20th October**, **17th November**, and **8th December**.

Class tests. There are two class tests. Each test will take place in person on campus, so please check your timetables carefully for where to go. They will take place at **10am, Wednesday 25th October** and **10am, Wednesday 29th November**. The tests will be conducted under exam conditions with invigilators, and full KCL exam regulations apply. The tests are in the form of timed moodle quizzes; at the start of each test, you will be directed to the KEATS page to access the quiz.

Final exam. The course 4CCM111A Calculus I is examined in the January exam window at KCL. This is called 'Exam Period 1' or 'EP1'. This year the dates of this period are January

5th – January 11th 2024, though we don't yet know when the exam itself will be. The final exam timetable is usually produced at some point in mid-November, by the central university. Make sure that you arrange any travel over the Christmas and New Year period to enable you to be in London to sit this exam.

The exam will be 2 hours long, and will take place in-person under invigilated exam conditions. You will have to write your answers by hand.

The exam will comprise two sections. Section A will be short questions (possibly multiple choice), whereas Section B may require some longer written responses and mathematical arguments. To help you practice, on each tutorial sheet there will be (at least) one 'exam-style question', which is a longer form Section B-style question. There are also many past papers available on the KEATS page.

	Topics
Week 1	Historical introduction, course organisation, functions, domain, range, injections, surjections and bijections.
Week 2	Inverse functions, exponents, logarithms, polynomial functions and trigonometric functions.
Week 3	The inverse trigonometric functions, the hyperbolic functions, trigonometric functions with complex arguments, double-angle formulae and function composition.
Week 4	The limit, the $\epsilon - \delta$ definition of the limit, continuous functions, the intermediate value theorem, limits involving infinity and rules for working with sums and products of limits.
Week 5	More work with limits: limits of composite functions, multiple limits standard limits, the sandwich theorem and logarithmic vs polynomial vs exponential limits.
Week 6	**Reading week.**
Week 7	Differentiation from first principles, differentiable functions, properties of the derivative, the chain rule, the sum rule, the product rule and the quotient rule.
Week 8	Derivatives of implicit functions, derivatives of inverse functions, derivatives of parametric functions and the mean value theorem.
Week 9	Integration, the Riemann integral and the fundamental theorems of calculus.
Week 10	Indefinite and definite integrals, integration by substitution, integration by parts, recursion relations, partial fractions and surface areas.
Week 11	Volumes of revolution, the length of a curve, infinite sums, convergence criteria and power series.
Week 12	Taylor's theorem, l'Hôpital's rule and analytic functions.

Table 1.1: The approximate organisation of the course lectures.

2. Functions of One Variable

In which we introduce the most important character in the course: the function. We will define functions of one variable, build up a catalogue of examples of functions, and study the properties of these examples. In our thinking we will come across the infinite sum, which will foreshadow the idea of an analytic function. We will chase this idea until we meet it formally at the culmination of the course.

This will be covered during weeks 1-3 of lectures.

2.1 Mathematical Definitions

During your time studying mathematics at university, you will see many *definitions*. They are important, and frequently the product of decades – or even centuries – of mathematical research. It is not always easy to understand a mathematical definition: much of these lecture notes will be devoted to explaining, developing, and contextualising the mathematical definitions of a function, continuity, a derivative, and so on. However, there is in fact a more fundamental difficulty. Certain educational researchers[1] have found that first-year students arrive at university with some confusion around the very concept of defining an object mathematically. It's as well that we begin the course by addressing this matter.

A definition is a statement of the exact meaning of a word. However, when defining words from ordinary language – as in a casual conversation, or even in a dictionary – we often allow some room for interpretation. After all, words rarely appear without any associated context, and one of the key features of everyday language is that it can be used flexibly. For example, an entry in the Oxford English Dictionary definition for the word *table* reads:

> "A flat and comparatively thin piece of wood, stone, metal, or other solid material."

[1] I learnt about this from the book *Ideas from mathematics education* by Lara Alcock and Adrian Simpson

You might contend that this definition is incomplete: after all, if you built a large solid block of wood, 1-metre by 1-metre by 1-metre – and then put it in your house and ate dinner on it – you would probably still call it a table, despite it not being 'comparatively thin'! However, the given definition still captures an essence of 'table-ness' in most situations; we are usually happy to leave it at that.

Mathematical definitions are different. A mathematical definition is the founding statement of an object. The definition is precisely what the object *is*. Mathematical definitions are not intended to be the subject of any interpretation or vagueness. When answering a question about a mathematical object, it is almost always inadequate to have merely a hazy or diffuse sense of the properties of the object: you must refer to the precise definition.

School mathematics does not always make this notion clear. To answer most A-Level questions, for example, it is enough to have a general intuitive sense of the mathematical concept involved – this is something that is called the *concept image* in mathematics education –, and to have practiced a huge number of example problems. This is not the case at university. Deep understanding requires the precise *concept definition*.

We will soon meet the definition of a function, and then quickly the definition of an injective function, a surjective function, a bijective function, and so on. Seeing examples of different definitions remains highly important, and it is only through the exposure to many examples of mathematical definitions that you will gain a deep familiarity with how to use them. However, we have to start somewhere. As a first illustration, we'll make up a mathematical definition with entirely fictitious words – so we don't get confused by using any mathematical words you may have already met – and discuss what this definition means.

Here is the definition of a made-up mathematical noun called a *jaberwocky*. It uses another made-up mathematical noun called a *tove*.

Definition 2.1.1. A *jaberwocky* is a *tove* with properties A, B and C.

This definition means three things:

1. any time you encounter a *jaberwocky* in a mathematical problem, you may assume that it is a *tove* and that is satisfies properties A, B, and C;

2. anything mathematical object that is a *tove* and satisfies properties A, B, and C is also a *jaberwocky*;

3. suppose that there is another property D, that may be deduced from the property of being a *tove* and properties A, B, C. Then every *jaberwocky* satisfies property D.

When trying to answer a mathematical question about *jaberwockies*, all three of these consequences may need to be considered.

Do not be concerned if this diversion to *jaberwockies* seemed a little bizarre to you. It was the best example that could be given before we have actually covered any mathematics. In some sense, the entire rest of the course will be an explanation and elaboration of this illustration. For example, here is one of the most complicated definitions we will encounter in this course (certainly you should not expect to understand this immediately!):

Definition 2.1.2 (Analytic function)**.** *An analytic function $f : \mathbb{R} \to \mathbb{R}$ is any function $f : \mathbb{R} \to \mathbb{R}$ for which for all $x_0 \in \mathbb{R}$ there exists $\delta > 0$ and a convergent power series $\sum_{n=0}^{\infty} a_n x^n$ with coefficients $a_n \in \mathbb{R}$ for which*

$$f(x) = \sum_{n=0}^{\infty} a_n (x - x_0)^n$$

for all x that satisfy $|x - x_0| < \delta$.

There's a lot going on here! It is easy to get confused. Yet, this definition follows the same structure as the definition of a *jaberwocky* above. The role of *jaberwocky* is played by 'analytic function', the role of *tove* is played by 'function', and the remainder of the definition are the various properties A, B and C.

2.2 The Definition of a Function.

Functions are the basic object for most of modern mathematics. As such, there is no predetermined reason why it is the Calculus I course in which they are introduced at KCL: any of your four 1^{st} semester courses could begin this way. However, at some point we decided that Calculus I would include the introductory material on functions, and that is as good a choice as any.

There is an immediate problem, alas: to understand the definition of a function, it is necessary to know the definition of a *set*. Yet, sets are covered in the Sequences and Series course – in great detail – as complicated manipulation of sets will play a greater role in that module. As such, we will provide a short basic definition here, but leave the broader explanation and other associated words (like *element*, *member*, *subset*, *union*, *intersection*, $\in$, $\notin$, $\cap$, $\cup$, $\subset$, and the sets $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$) to the Sequences and Series course.

Definition 2.2.1 (Set)**.** A *set* is any collection of objects.

For example, $\{1, 2, 3, 4, 5\}$ denotes the collection of whole numbers 1, 2, 3, 4, 5 (and these are the five elements of the set). Note that, to make the notes easier to revise quickly, we have put a short sub-heading for the definition straight after the number.

Definition 2.2.2 (Function)**.** Let A and B be two sets. A *function*, denoted $f : A \to B$, associates a unique element in B to each element in A.

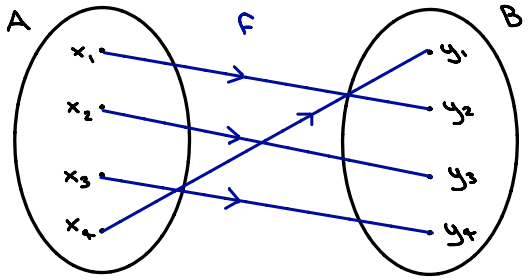
We'll give plenty of examples soon. However, as part of the point of the definition is its generality (i.e. it applies to any two sets A and B and any way of assigning a unique element in B to each element in A), we will hold off for now, and give some general guidance first.

Comment(s). (Functions)

1. We often think of a function as representing a sort of *transformation*, or even a *machine*, taking as inputs the elements from the set A and giving as its output the unique element in B that is associated to the input element. The set A indicates the possible inputs to this machine, while the set B specifies what kinds of outputs the machine can produce. This is a useful perspective. In time, however, you will need to also consider a function as an object in its own right, and not just as a process that is applied to other objects.

2. It is useful to have some notation for how the function $f : A \to B$ associates an element $y \in B$ to a specific element $x \in A$. This is $f : x \mapsto y$. Note the arrow we use: the symbol $\mapsto$ is used when a function is acting on a single element in a set, while the symbol $\to$ is used (as in the definition of a function) to indicate a function between sets of elements.

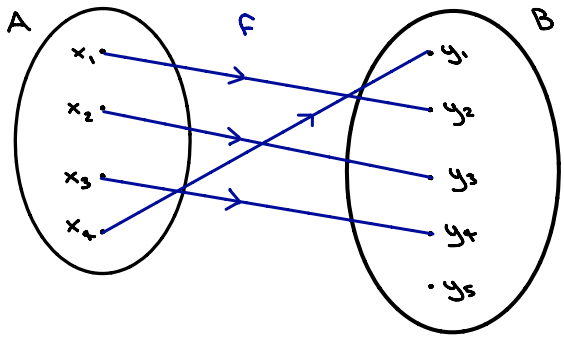
3. We might define a function $f : A \to B$ pictorially by drawing arrows between elements of A and elements of B indicating the action of f, e.g.



describes the function $f : A \to B$ where $A = \{x_1, x_2, x_3, x_4\}$ and $B = \{y_1, y_2, y_3, y_4\}$ given by $f : x_1 \mapsto y_2$, $f : x_2 \mapsto y_3$, $f : x_3 \mapsto y_4$ and $f : x_4 \mapsto y_1$. Defining functions in this way is only convenient when A and B are sets of small order, of course, as the pictures get increasingly complicated when the sets A and B are large!

4. If $f : A \to B$ is a function, for $x \in A$ we write $f(x)$ for the unique element of B that is associated to x by the function f. Hence we might sometimes write a function $f : A \to B$ as $f : x \mapsto f(x)$ for all $x \in A$.

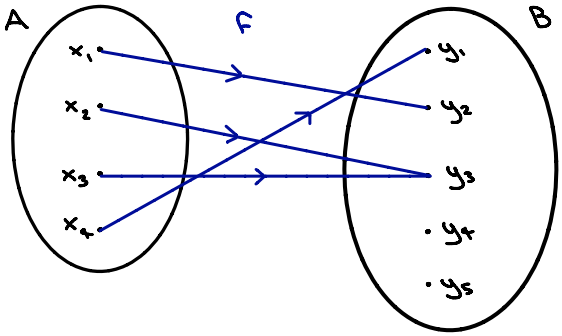
5. A function $f : A \to B$ acts on all elements in A to give an element in B (by definition), but it is not necessary that every element in B is associated with an element in A by the function. For example, the function indicated by



,

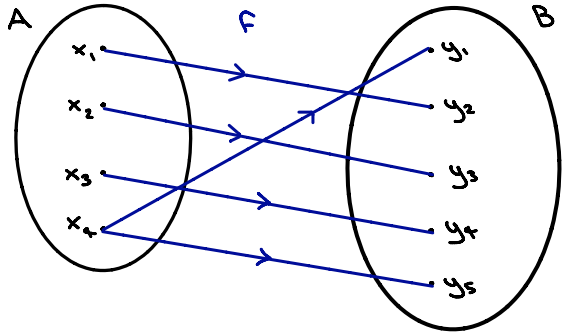
where now $B = \{y_1, y_2, y_3, y_4, y_5\}$, is a well-defined[2] function $f : A \to B$ even though there is no element in A which is mapped to y_5 in B.

6. The definition of a function $f : A \to B$ requires that the output of f is a unique element in B. For example, the map defined by



[2]An object in mathematics is *well-defined* if it satisfies all the conditions of its definition without any logical inconsistencies.

is a well-defined function, whereas the map defined by



is not a well-defined function, as it does not map x_4 to a unique element in B. The definition implies that functions (though they can be 'many-to-one' maps) cannot be 'one-to-many' maps.

2.2.1 Domain and Range

Definition 2.2.3 (Domain and Range)**.** For a function $f : A \to B$, the set A is called the *domain* of f and the set B is called the *range* of f.

Comment(s). (On the domain and range)

1. There are different names for the domain and range of a function, which you may find are used sometimes instead – though not in this course. The most common alternatives are the words *co-domain* or *image* to describe what we are calling the range, and correspondingly the word *pre-image* for what we are calling the domain. For us, the word *image* has the following more-precise meaning:

   Definition 2.2.4 (Image)**.** Let $f : A \to B$ be a function. Then the *image* of f is the set of all elements $f(x)$, i.e. $\{f(x) : x \in A\}$.

   Note that the image of a function must always be contained within its range, but needn't be the same as the range[3].

[3]As a warning, conventions also differ in the use of the word 'range'. In this course we have identified the range with the co-domain; however, another common convention is to identify the range with the image. It is not uncommon for fundamental definitions to differ in mathematics. What is important is that you have a working knowledge of the different conventions, and understand which convention is being used by an author at a certain moment.

2. Where the domain and range are known, we might swiftly refer to a function as $f(x)$ where x lies in the domain. This is a useful algebraic notation that will enable us to define functions whose domain and ranges are infinite sets, such as $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ or $\mathbb{C}$.

3. The domain and range of a function might be unions of sets, for example

$$f : \mathbb{R} \to (-\infty, -3] \cup [7, \infty), \qquad f(x) = \begin{cases} x^2 + 7 & x \geq 7 \\ x - 3 & x \leq -3. \end{cases}$$

   Here we use the notation convention $(-\infty, a]$ to refer to the set of all real numbers that are less than or equal to a; similarly $[b, \infty)$ denotes the set of all real numbers that are greater than or equal to b. For more detail on this notation, see the Sequences and Series course.

4. The range B of a function $f : A \to B$ contains the elements $f(x)$ for each element $x \in A$, i.e. the range contains the image. However, the range could be a substantially larger set than the image. In other words, the range B may contain elements which are not mapped to by f. For example,

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(x) = \begin{cases} x^2 + 7 & x > 0 \\ x - 3 & x < 0 \end{cases}$$

   is well-defined, since every x in the domain is associated to a unique y in the range. However, there is no $x \in \mathbb{R}$ for which $f(x) = 0$, so 0 is not in the image of f.

When defining a function it is necessary to specify the domain and range. The same operation applied over different ranges and domains gives a different function. As we will see in some of the following examples the same function operation defined over different domains and ranges may not even give a well-defined function in all cases

---

Example 2.1. State whether the following function is well-defined:

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(x) = x + 2.$$

The function defined above is very simple: it translates the real number line by +2. We are immediately content that the output is unique, so this is a well-defined function. The specification of the domain tells us that $x \in \mathbb{R}$, and the definition of the range that $f(x) \in \mathbb{R}$. Evidently various parts of the definition of the function are independent, and it is up to us as mathematicians to check that the domain and range are suitable for any explicitly defined function $f(x)$.

There are many different ways we could have pick the domain and range so that this function was not well-defined, e.g. if we had changed the range so that $f : \mathbb{R} \to \mathbb{Z}$, for $f(x) = x + 2$ the range would simply be incorrect.

Example 2.2. Check that the following function is well-defined:

$$f : [0, \infty) \to [0, 12] \cup (14, 16) \cup \{9\}, \qquad f(x) = \begin{cases} 3x & \text{for } x \in [0, 4] \\ 2x + 6 & \text{for } x \in (4, 5) \\ 9 & \text{for } x \geq 5. \end{cases}$$

Here we have a complicated definition for a function, which is perfectly valid (as you should check). Notice again the use of the notation $[0, \infty)$ to denote the set of all non-negative numbers.

Example 2.3. Is

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(x) = \pm\sqrt{x}$$

a well-defined function?

It is not well-defined for two reasons. First $f(x)$ gives two outputs for every input. So let us correct this and redefine the function as

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(x) = \sqrt{x}.$$

We emphasise that the notation $\sqrt{x}$ denotes the principal square root of a number, i.e. its positive square root. The function is still not well-defined for if $x < 0$ then $f(x) \notin \mathbb{R}$. However we may again amend the definition to

$$f : [0, \infty) \to \mathbb{R}, \qquad f(x) = \sqrt{x}.$$

This is well-defined, even though there is some redundancy in the range; we could have tightened the definition to

$$f : [0, \infty) \to [0, \infty), \qquad f(x) = \sqrt{x}$$

but it is not strictly necessary.

We have restricted our examples to simple algebraic functions, but we could certainly have constructed well-defined functions of a more abstract and curious nature e.g.

$$f : \mathbb{R} \to \{0, 1\}, \qquad f(x) = \begin{cases} 1, & \text{if } x \in \mathbb{Q} \\ 0 & \text{if } x \notin \mathbb{Q} \end{cases}.$$

This function, although obeying none of the 'regularity' properties that we will study late in the course – continuity, differentiability –, is well-known in the development of mathematical analysis[4] as it is in fact discontinuous at every point x. We will meet the definition of continuity later in the course, which will help us to better appreciate the horror of the above function.

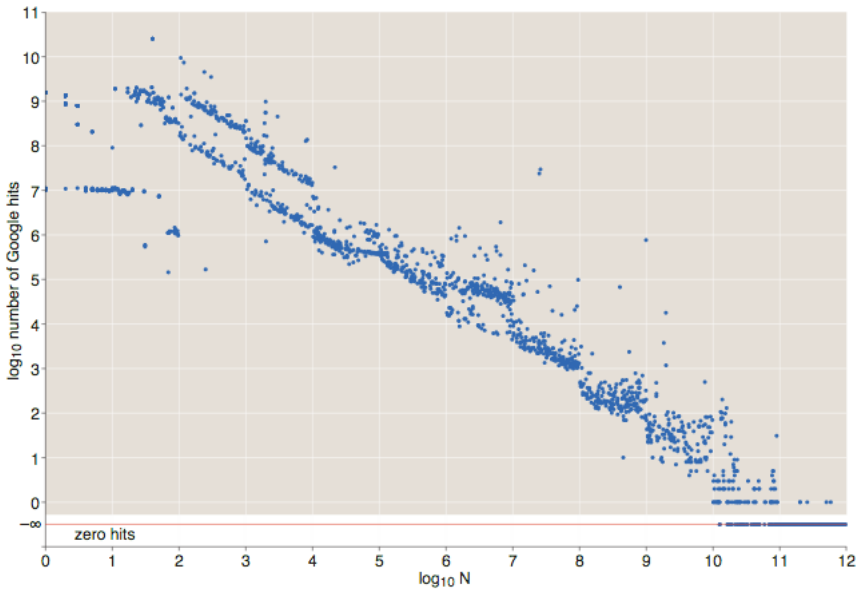


Figure 2.1: A log plot of the number of google hits of the integers from 0 to 10^{12}.

We could also have defined interesting functions in less mathematical language, for example the function which outputs the number of results when one googles an integer. The outputs of this function is shown in a log graph in figure 2.1. This graph was made in 2014 by Brian Hayes and comes from his blog where there is a detailed and interesting discussion of the patterns in the output, see `http://bit-player.org/2014/600613`.

By this point in our discussion, we may begin to feel that there are an overwhelming variety of possibilities for functions – and we would be right. Intuitively we have sense that certain functions are "nicer" than others – those that we can sketch a smooth graph of, for example – while there are many others that are less palatable, such as the google hits function above. The desire to put functions into classes based on their mathematical behaviour will motivate much of our work in these lectures.

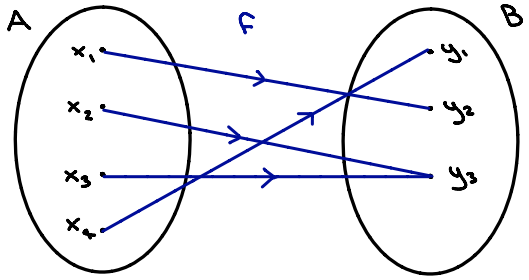
2.2.2 Injections, Surjections and Bijections

We will begin our classification of functions with the following definitions.

[4]It is called the Dirichlet function after 19^{th} century mathematician Peter Gustav Lejeune Dirichlet.

Definition 2.2.5 (Surjective)**.** A function $f : A \to B$ is *surjective* (or *onto*) if every element in B is mapped to by at least one element of A. In other words, f is surjective if its range and image are the same. A surjective function is called a *surjection*.

For example the following picture defines a surjective function, as every element of B is mapped to from A:

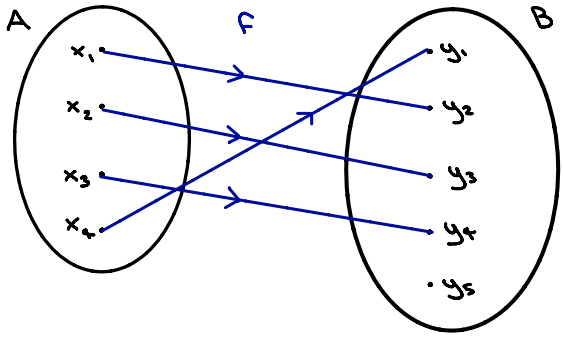


We can write the definition of a surjective function in notation as follows.

> Let $f : A \to B$ be a function. If $\forall y \in B \, \exists x \in A$ such that $y = f(x)$, then we say that f is a surjection.

Definition 2.2.6 (Injective)**.** A function $f : A \to B$ is *injective* (or *one-to-one*) if every element $f(x) \in B$ is mapped to by at most one element in the domain A. An injective function is called an *injection*.

For example the following picutre defines an injective function, as each element of A is associated with a unique element of B:



We can write the definition of an injection in mathematical notation as follows.

> Let $f : A \to B$ be a function. If, $\forall \, x_1, x_2 \in A$, $(f(x_1) = f(x_2)) \implies (x_1 = x_2)$, then we say that $f(x)$ is an injective function.

Example 2.4. State which, if any, of the following functions are injective, surjective or both?

(a) $f_1 : \mathbb{R} \to \mathbb{R}$ *given by* $f_1(x) = x^2$

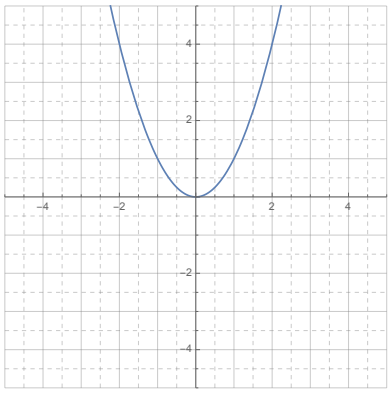
(b) $f_2 : \mathbb{R} \to \mathbb{R}$ *given by* $f_2(x) = x^3$

(c) $f_3 : \mathbb{R}^+ \to \mathbb{R}^+$ *given by* $f_3(x) = \sqrt{x}$

(d) $f_4 : \mathbb{R} \to \mathbb{R}$ *given by* $f_4(x) = 2^x$

Let's consider each function in turn.

(a) The graph of f_1 is



The graph considers the function as a map from a point on the x-axis to a point on the y-axis, and these points are paired together to give the coordinates of the curve we have sketched. If you wanted to draw an arrow from a real number on the x-axis to find its image in the real numbers on the y-axis you would draw a vertical line from the number on the x-axis until it meets the curve at a point, and then draw a horizontal line from that point until it meets the y-axis. You've known how to do this for many years, of course. However, it is worth reflecting that this extraordinarily useful mathematical device – linking the algebraic definition of a function with a geometric object – was only introduced in 1637, by Descartes.

Let us first ask whether f_1 is surjective, i.e. is every number on the y-axis in the image of f_1? We can see from the graph, which is always positive, that none of the negative numbers on the y-axis are in the image of the function. This immediately means that the function is ***not surjective***.

Once you are more familiar with these notions, you will be happy to move on after such an observation. However, since this is our first example, it is worth proceeding a little more slowly. To properly prove that the function is not surjective, we need only show

that one element in $\mathbb{R}$ (which is the specified range of f_1) is not in the image of f_1. We will show that -1 is not in the image of f_1.

We will give two 'proofs' that -1 is not in the image of f_1. The first is of a kind that you may from time to time be expected to produce in this course.

Proof that there does not exist $x \in \mathbb{R}$ for which $x^2 = -1$. Let $x \in \mathbb{R}$. Since the square of every real number is positive, $x^2 \geqslant 0$. Therefore, $x^2 > -1$. Since x was arbitrary, we conclude that $x^2 > -1$ for all $x \in \mathbb{R}$ and therefore there does not exist $x \in \mathbb{R}$ such that $x^2 = -1$. □

Diversion: non-examinable. That seemed all very straightforward: just a formal write-up of the intuitive visual notion that the graph of $y = x^2$ is always positive. However, you may be nervous: how do we *know* that the square of every real number is non-negative? Has anyone ever proved it to you? Certainly you cannot have checked it for every example, even with a calculator, as there are infinitely many real numbers! Such concerns are not part of this module, and in fact they are barely part of the Sequences and Series module either. To properly handle these questions, we would need to delve into the axioms of the real numbers as a *complete totally-ordered field* in which the rationals are dense, and to discuss the constructions of Dedekind and Cauchy. We will not do this, as it is too advanced for our main theme. However, I would be short-changing you if I didn't show you what such a proof might look like, written in this more detailed language. Similar ideas will appear in the Sequences and Series course.

Proof that there does not exist $x \in \mathbb{R}$ for which $x^2 = -1$. We begin by noting that $(-1)^2 = 1$. Indeed, by the distributive property of multiplication in a field and the multiplicative identity property of 1, $(-1)^2 - 1 = (-1)^2 - 1^2 = (-1-1)(-1+1) = -2 \times 0 = 0$.

Now, let $x \in \mathbb{R}$. We claim that $x^2 \geqslant 0$. There are three cases: either $x > 0$, $x = 0$, or $x < 0$ (and these are exhaustive, as the order relation is a total ordering on $\mathbb{R}$).

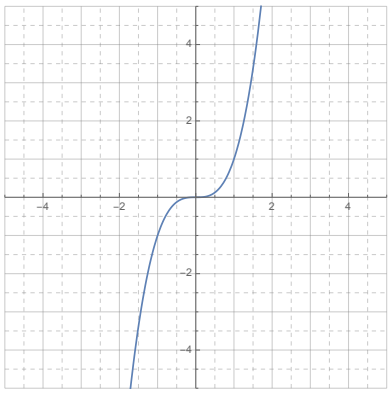
- If $x > 0$ then, multiplying both sides of the inequality by positive x, we have $x^2 > 0$ and thus the claim is satisfied in this case.
- If $x = 0$ then $x^2 = 0$ and thus the claim is satisfied in this case.
- If $x < 0$, then adding $(-x)$ to both sides we conclude that $(-x) > 0$. Repeating the argument from the first case we conclude that $(-x)^2 > 0$. However, $(-x)^2 = (-1)^2 x^2 = 1 \cdot x^2 = x^2$, so $x^2 > 0$ too.

Since x was arbitrary, we conclude that $x^2 \geqslant 0$ for all $x \in \mathbb{R}$. Since $0 > -1$, by transitivity of the ordering we conclude that $x^2 > -1$ for all $x \in \mathbb{R}$, and hence there is no $x \in \mathbb{R}$ for which $x^2 = -1$. □

This is the end of the diversion and the return to examinable material.

Examinable material again. From the graph we can also quickly deduce that f_1 is not injective either, as for all x both x and $-x$ are mapped to the same point on the y-axis (i.e. the curve is symmetric about the line $x = 0$). As two different numbers are mapped to the same number on the y-axis, we see that the function is not one-to-one. If we had wanted to prove that f_1 is not injective algebraically, we would have needed only to give one example where the map is not one-to-one (e.g. $f_1(2) = 4 = f_1(-2)$).

(b) The graph of f_2 is



which indicates that f_2 is both surjective and injective. Again, for our purposes in this course we will be happy to move on at this point. However, it is worth reflecting again on how one would formally prove that f_2 is both surjective and injective.

We have it within our power to prove that f_2 is injective. There are a few different ways of handling the details: here is one using algebraic factorisation.

Indeed, suppose that $x_1^3 = x_2^3$. We have to show that $x_1 = x_2$. You may be tempted to 'apply the cube-root' to both sides of the equation $x_1^3 = x_2^3$, but how do you know that the cube-root exists as a well-defined operation? When we cover 'inverse functions' in the next section, you will see that proving that the cube-root is well-defined is precisely the same task as showing that f_2 is injective and surjective. Therefore, it would be logically circular to invoke the existence of the cube-root operation when trying to show that f_2 is injective and surjective.

Rather, by algebraic factorisation, we have that $0 = x_1^3 - x_2^3 = (x_1 - x_2)(x_1^2 + x_1x_2 + x_2^2)$. So either $x_1 = x_2$ or $x_1^2 + x_1x_2 + x_2^2 = 0$. However this second possibility can almost never

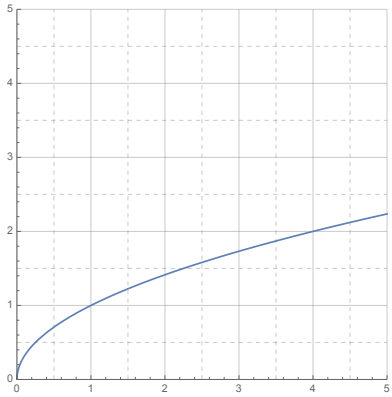
happen, since

$$x_1^2 + x_1x_2 + x_2^2 = \frac{(x_1 + x_2)^2 + x_1^2 + x_2^2}{2} > 0$$

unless $x_1 = x_2 = 0$. So in either case we conclude that $x_1 = x_2$, and hence f_2 is injective.

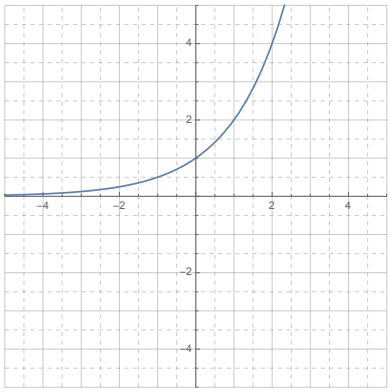
To prove that f_2 is surjective, we need to show that for any value $y \in \mathbb{R}$ we can identify a value $x \in \mathbb{R}$ for which $f_2(x) = y$, i.e. we must solve the equation $x^3 = y$ for x in terms of y. You might be already very comfortable with rules of exponents and writing $x = y^{1/3}$, and therefore claiming that you have identified the value of x as the cube-root of y. But again, how do you even know that $y^{1/3}$ exists in $\mathbb{R}$? Certainly, if $y \in \mathbb{Q}$ it is not always the case that $y^{1/3} \in \mathbb{Q}$ too, e.g. $2^{1/3} \notin \mathbb{Q}$. The existence of cube roots in $\mathbb{R}$ is a subtle thing, which uses the fact that $\mathbb{R}$ is *complete*. This is to do with the existence of objects called supremums and infimums, which will be covered towards the end of the Sequences and Series course.

(c) The graph of f_3 is



The square-root function is both injective and surjective (defined as a map from $[0\,\infty) \to [0, \infty)$). Again, we leave unsaid the question of proving that square-roots of all positive numbers actually exist in $\mathbb{R}$.

(d) The graph of f_4 is



The function f_4 is injective but not surjective (defined as a map from $\mathbb{R} \to \mathbb{R}$). We also remark that we have not formally defined exponential functions of the form 2^x yet, though many students will have come across them before. This issue will be discussed further in the next section.

Definition 2.2.7 (Bijective)**.** A function $f : A \to B$ is *bijective* if it is both surjective and injective. A bijective function is called a *bijection*.

In the examples above, both f_2 and f_3 are bijections. Bijections are very useful functions because by being both injective and surjective between two sets A and B it means that every element in A is uniquely paired with an element in B and vice versa. This implies that a bijective function can be *inverted*.

2.3 Inverse functions

Definition 2.3.1 (Inverse function)**.** The *inverse function* of a bijective function $f : A \to B$ is a function denoted by f^{-1} and defined by

$$f^{-1} : B \to A, \qquad f^{-1}(f(a)) = a \qquad \forall\, a \in A.$$

Observe that f^{-1} is really a 'two-sided inverse', in the following sense:

- $\forall a \in A$, $f^{-1}(f(a)) = a$ by definition;
- $\forall b \in B$, $f(f^{-1}(b)) = b$. We can show this by the following short argument. Since f is bijective, for all $b \in B$ there is a unique $a \in A$ such that $f(a) = b$. Then $f(f^{-1}(b)) = f(f^{-1}(f(a))) = f(a) = b$, where we used the definition of the inverse function f^{-1} to replace $f^{-1}(f(a))$ by a.

Comment(s). (On inverse functions)

1. The notation is potentially confusing: f^{-1} is a special exception to the exponent notation and one should be careful to make sure its meaning is clear. When f is a bijective function, f^{-1} does not signify the reciprocal $\frac{1}{f}$ but the inverse function, e.g. if $f : [0, \infty) \to [0, \infty)$ is given by $f(x) = x^2$ then $f^{-1}(x) = \sqrt{x}$.
2. For a function f to be invertible it must be both surjective and injective. Otherwise we say that the inverse function f^{-1} does not exist.

- Consider the example of the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^2$, which is not injective as, for $x \neq 0$, the distinct values x and $-x$ are both mapped to x^2. Consequently if we attempted to invert the map we would not know whether x^2 originated from x or $-x$. Written in general language, if f is not injective then there does not exists a function $f^{-1} : B \to A$ for which $f^{-1}(f(a)) = a$ for all $a \in A$.
- If $f : A \to B$ is injective, however, one can still construct a so-called *one-sided inverse* function $g : B \to A$. This is a function for which $g(f(a)) = a$ for all $a \in A$, and it is well-defined since the map f is one-to-one. In other words, injectivity allows the inverse map to be constructed for those elements in the image of f. However, if f is not also surjective, then there will exist some $b \in B$ for which $f(g(b)) \neq b$. So g is not a two-sided inverse in this case.
- If $f : A \to B$ is surjective, one can also construct a different *one-side inverse* function $g : B \to A$ by choosing $g(b)$ to be some element $a \in A$ for which $g(a) = b$. (The element a always exists by the surjectivity assumption). Therefore, $f(g(b)) = b$ for all $b \in B$. However, if f is not injective then there exists some $a \in A$ for which $g(f(a)) \neq a$. So g is not a two-sided inverse in this case.
- Let us look at a concrete example. The following function is injective but not surjective, $f : [0, \infty) \to \mathbb{R}$ given by $f : x \mapsto \sqrt{x}$. It is not surjective since $\sqrt{x}$ is always non-negative. What problems does this cause when we attempt to define a two-sided inverse function? To construct the inverse function f^{-1} we invert the map from elements of A to elements of B. Therefore, if $y \geqslant 0$ we must choose $f^{-1}(y) = y^2$. In this case $f^{-1}(f(a)) = a$ for all $a \in [0, \infty)$, so we have a one-sided inverse. Next, we are free to choose the value of $f^{-1}(y)$ when $y < 0$, as this doesn't affect the one-sided formula.

  However, no matter what we choose, if $y < 0$ it will never be the case that $f(f^{-1}(y)) = y$ (since $f(x) \geqslant 0$ for all x). This means $\exists y \in \mathbb{R}$ such that $f(f^{-1}(y)) \neq y$. Hence there does not exist a two-sided inverse.

3. The inverse function of a bijective function is also a bijection. In the pictorial representation of a bijective function $f : A \to B$ in terms of arrows, the inverse function $f^{-1} : B \to A$ is just constructed by reversing the direction of all the arrows.

Now if we are given a bijection how do we find the inverse function? If the function is given by an algebraic formula, we have the following three-step recipe:

(i) Write $y = f(x)$.

(ii) Rearrange the equation to obtain $x = g(y)$.

(iii) As $y = f(x) = f(g(y))$, we have $g(y) = f^{-1}(y)$. Changing variable, write $g(x) = f^{-1}(x)$.

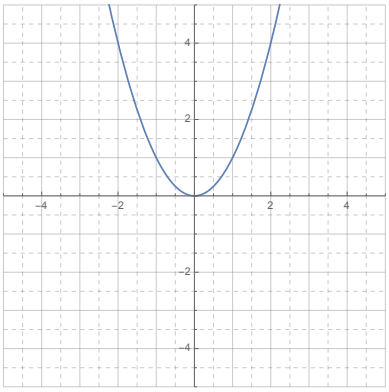
This procedure is sometimes straightforward. For example, if we were trying to find the inverse function of the function $f : \mathbb{R} \to \mathbb{R}$ given by the algebraic formula $f(x) = x + 4$, we could write $y = x+4$, and so $x = y-4$. Therefore $g(y) = y-4$, and so $f^{-1}(x) = x-4$ as expected. However it is worth practising further for yourself so you are happy that all the steps are sensible.

Exercise 2.1. Find the inverse function $f^{-1}(x)$ for

1. $f : \mathbb{R} \to \mathbb{R}$ *given by* $f(x) = 6x + 4$
2. $f : [0, \infty) \to [5, \infty)$ *given by* $f(x) = \sqrt{x} + 5$.

Many common functions are not bijections. However, we can restrict the domain and range of their definition so as to construct a bijection, which can then be inverted on the restricted domain and range. We will consider a simple example in this section, and then in later sections we will look at more complicated inverse functions.

Let us consider again the quadratic function $f : \mathbb{R} \to \mathbb{R}$ given by $f : x \mapsto x^2$, whose graph is

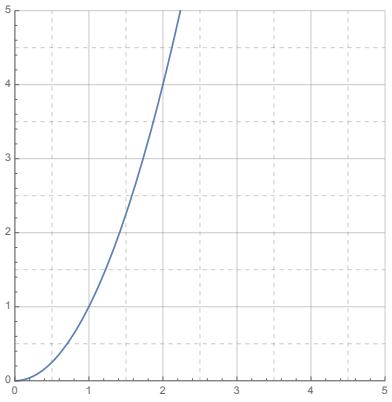


.

The function f fails to be a bijection because it is not injective, nor is it surjective onto $\mathbb{R}$. We can see both these observations from the graph. The image of f is $[0, \infty)$, so that if we were to change the definition of f and restrict its range as $f : \mathbb{R} \to [0, \infty)$ by $f : x \mapsto x^2$, then the new function would be surjective. Furthermore we can see from the graph that f is generally a two-to-one function (apart from at zero): if we drew a horizontal line across the graph for any positive value on the y-axis, that line would meet the curve at exactly two points. Therefore, if we restricted the domain of the function's definition to, for example, just $[0, \infty)$, it would become an injective function. In other words $f : [0, \infty) \to [0, \infty)$ given by $f : x \mapsto x^2$ is both an injective and a surjective function. Therefore it is a bijection and its inverse function exists,

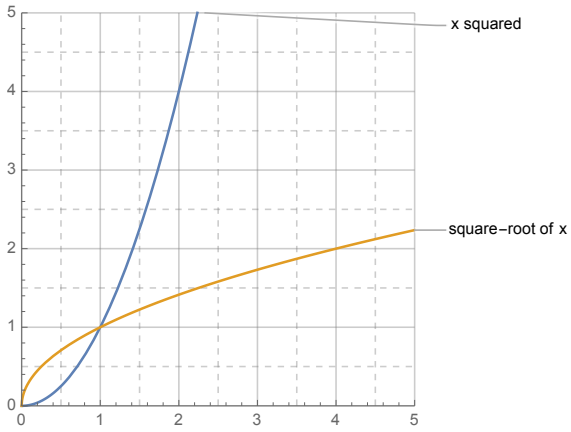
given by the unique map $f^{-1} : [0, \infty) \to [0, \infty)$ such that $f^{-1}(f(x)) = x$ for all $x \in [0, \infty)$. This is a function we have a short-hand for, namely it is given by $f^{-1}(x) = \sqrt{x}$ (and this is the way the square-root function is actually defined).

Let's draw the graph of the bijective function $f : [0, \infty) \to [0, \infty)$ given by $f : x \mapsto x^2$:



This is the restriction of the parabola $f(x) = x^2$ constructed over $\mathbb{R}$. Compared to the graph above, we see that by restricting the domain we have effectively cut the function at its turning point, and by 'stopping it turning' we have ensured that what remains is a one-to-one function. This idea – of restricting a function's domain to lie between its turning points so as to make it injective on the new domain – is how we will set about constructing more complicated inverse functions.

We will make one final remark about inverse functions here: if $A, B \subset \mathbb{R}$ and $f : A \to B$ is a bijection, the graph of the inverse function $f^{-1} : B \to A$ is given by the graph of the original function but with the x-axis and y-axis interchanged. Hence if we were to plot both f and f^{-1} on the same set of axes we would find that f^{-1} is the reflection of f in the line $y = x$. For the example we have considered here we plot both f and f^{-1} below so we can see that one is a reflection of the other.



This is a beautiful geometric idea: inverse functions are defined purely algebraically, but when plotted obey a close geometric relationship to the original function.

2.4 Monotonically increasing and monotonically decreasing functions

There are a final couple of basic definitions which we will need throughout the course.

Definition 2.4.1 (Monotonically increasing)**.** Let $A \subset \mathbb{R}$. We say that a function $f : A \to \mathbb{R}$ is *monotonically increasing* if $f(x) \leqslant f(y)$ whenever $x, y \in A$ and $x \leqslant y$.

For example, the function $f : \mathbb{R} \longrightarrow \mathbb{R}$ given by $f(x) = x^3$ is monotonically increasing.

There is one subtlety in this definition. Consider the example of $f : [0, \infty) \to \mathbb{R}$ given by

$$f(x) = \begin{cases} x & \text{if } x \in [0, 5) \\ 5 & \text{if } x \geqslant 5. \end{cases}$$

This is a well-defined function, and it is monotonically increasing, because the definition permits f to have sections of constant value.

Definition 2.4.2 (Monotonically decreasing)**.** Let $A \subset \mathbb{R}$. We say that a function $f : A \to \mathbb{R}$ is *monotonically decreasing* if $f(x) \leqslant f(y)$ whenever $x, y \in A$ and $x \geqslant y$.

For example, the function $f : \mathbb{R} \to \mathbb{R}$ given by $x \mapsto -x + 7$ monotonically decreasing.

2.5 Graph sketching

The ability to sketch the graph of a function $f : \mathbb{R} \to \mathbb{R}$ is extremely important to learn, even in the presence of graphing software (such as WolframAlpha). This is because a sketch of a graph demonstrates all the most important aspects of a function; by sketching the graph by hand, you yourself are making the intellectual effort to identify these aspects, thus improving your understanding of the function itself. It is hard to develop the same understanding by looking uncritically at a computer-generated plot.

In teaching graph sketching, we reach a tricky pedagogical point. To illustrate the techniques we must provide several examples and determine the key parts of a graph to consider. However, we have yet to talk rigorously about concepts such as 'asymptotes', 'limits at infinity', and

'derivatives', nor have we even formally defined polynomial functions, exponential functions, trigonometric functions, etc: so how could we give any examples?

We will proceed as follows. In this short section, we will explain some of the key 'rules-of-thumb' for graph sketching. They will be illustrated with examples, albeit with the proviso that a rigorous analysis of these examples has yet to be undertaken. Later on in the course we will cover more advanced techniques in graph sketching, related to understanding the gradients of curves.

Here are some of the key principles to bear in mind when sketching a graph:

1. first principles;
2. links to known examples;
3. axis intercepts;
4. asymptotes;
5. when is the function positive/negative?;
6. how does the function behave as $x \to \infty$ and $x \to -\infty$ (i.e. as x gets very large positive or very large negative)?
7. turning points

We will not discuss turning points at this stage of the course, as that is more appropriate after we have formally introduced the process of differentiation. However, in the following examples we will show how the other ideas may be introduced in practice.

Example 2.5. Sketch the graph of the following function:

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(x) = \begin{cases} -10 & \text{if } x \leqslant -2 \\ -x & \text{if } x \in (-2, 2) \\ x & \text{if } x \geqslant 2 \end{cases}.$$

Sketching this graph is a perfect instance of the 'first principle' approach. A graph is simply the line formed by all the points with coordinates $(x, f(x))$. In certain simple is examples, we can just think directly what this would look like on the page. When $x \leqslant -2$ this is points of the form $(x, -10)$, which is a straight horizontal line. When $x \in (-2, 2)$ this is points of the form $(x, -x)$, which we can plot as a diagonal line from top-left to bottom-right going through

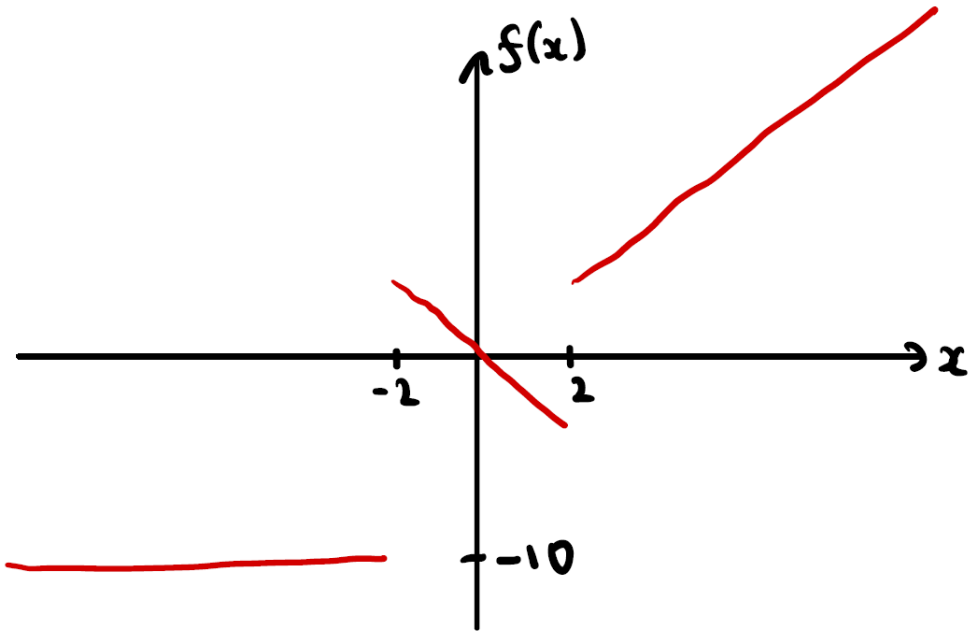


Figure 2.2: Graph of $f(x)$ from Example 2.5

the origin. When $x \geqslant 2$ this is just points of the form (x, x), i.e. with equal coordinates, which we can plot as a diagonal line from bottom-left to top-right.

Regarding first principles, it is important to remember that for each x in the domain of f there is at exactly one value of $f(x)$ for which the point $(x, f(x))$ is in the graph of the function: functions are not 'one-to-many' maps. If your sketch has multiple lines above a single x, as in the sketch below, it cannot be the graph of a function. See Figure 2.3.

Example 2.6. Sketch the graph of the following function:

$$f : \mathbb{R} \setminus \{2\} \to \mathbb{R}, \qquad f(x) = \frac{1}{x-2}.$$

Sketching this graph is an instance of the 'link to known example' approach. We would expect that you know that the graph of $f(x) = 1/x$ (which is defined for all $x \neq 0$) looks like a hyperbola, with a vertical asymptote at $x = 0$ and a horizontal asymptote at $y = 0$. See Figure 2.4 for a sketch. Replacing x by $x - 2$ shifts the graph to the right by 2, so the sketch is as in Figure 2.5. Note we have also observed the axis intercepts. When $x = 0$ we see $f(x) = -\frac{1}{2}$. There is no x for which $f(x) = 0$, so the graph does not intercept the x-axis.

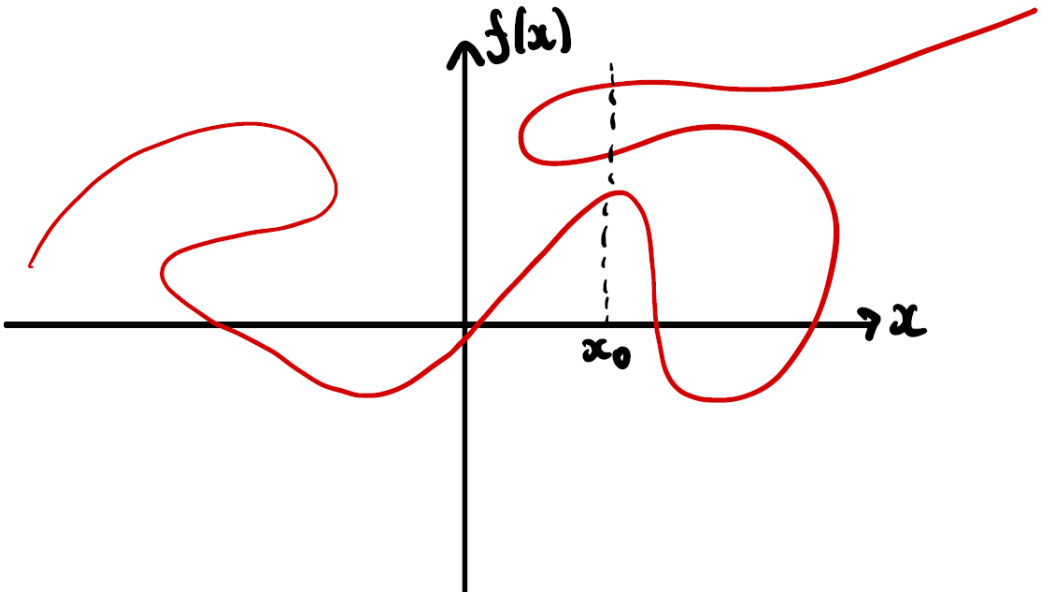


Figure 2.3: This is not the graph of a function, as (for instance) the value of x_0 is associated to 3 different values on the y-axis.

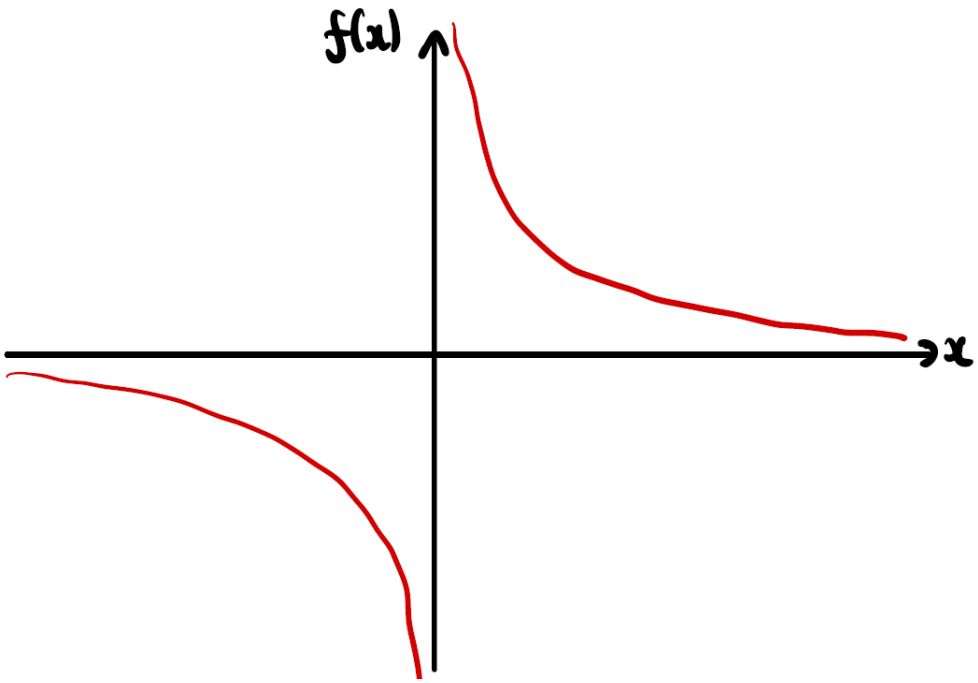


Figure 2.4: Sketch of the graph of $f(x) = \frac{1}{x}$

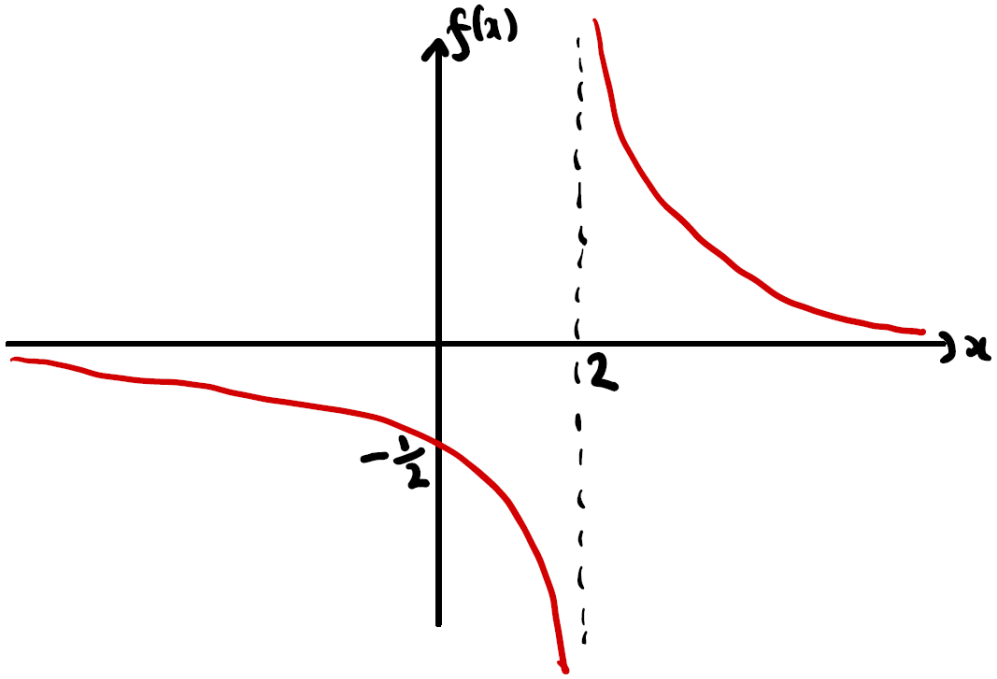


Figure 2.5: Sketch of the graph of $f(x) = \frac{1}{x-2}$

.

Example 2.7. Sketch the graph of the following function:

$$f : \mathbb{R} \setminus \{-2, 3\} \to \mathbb{R}, \qquad f(x) = \frac{x^3 + 1}{(x-3)(x+2)}.$$

This example will demonstrate the other basic principles. This function is pretty complicated, so it's going to be hard to do something directly from first principles or from a simple example that we already know. Regarding the axis intercepts, when $x = 0$ we have $f(x) = -\frac{1}{6}$, and when $f(x) = 0$ we know that $x^3 + 1 = 0$, meaning that $x^3 = -1$ and hence $x = -1$. So $(0, -\frac{1}{6})$ and $(-1, 0)$ are the only axis intercepts.

The function is not defined at $x = -2$ or $x = 3$ (as the denominator is 0). When x is very close to -2 or 3 we see that $f(x)$ involves dividing by a very small quantity, making $x = -2$ and $x = 3$ vertical asymptotes for the function. When $x < -2$, we have $x^3 + 1 < 0$, $(x - 3) < 0$ and $x + 2 < 0$. Therefore $f(x) < 0$. By performing a similar calculation, when $x \in (-2, -1)$ we have $f(x) > 0$; when $x \in (-1, 3)$ we have $f(x) < 0$; and when $x \in (3, \infty)$ we have $f(x) > 0$.

What about the behaviour as $x \to \infty$ and $x \to -\infty$? Well, using polynomial division we

have

$$\begin{aligned} f(x) &= \frac{x^3+1}{x^2-x+6} \\ &= \frac{(x+1)(x^2-x-6)+7x+7}{x^2-x+6} \\ &= x+1+\frac{7x+7}{x^2-x+6} \\ &= x+1+\frac{7+\frac{7}{x}}{x-1+\frac{6}{x}} \\ &\approx x+1 \end{aligned}$$

when x is very large positive or very large negative. We will determine more rigorous ways of determining this approximation later in the course. For now, the rule of thumb is to perform polynomial division to divide the numerator by the denominator, until you are left with

$$f(x) = \text{polynomial} + \text{term that approaches zero.}$$

So $f(x) \approx x+1$ when x is very large positive or negative, so the line $y = x+1$ is another asymptote. It is a little more complicated to work out whether the graph of f ever cuts this asymptote. Fortunately, we note that

$$f(x) - (x+1) = \frac{7x+7}{(x-3)(x+2)}.$$

When $x < -2$ this is a negative divided by a positive, so is negative; when $x > 3$, this is a positive divided by a positive, so is positive. This means that the graph doesn't cut the asymptote as x gets large positive or large negative.

Armed with this analysis, we may complete the sketch in Figure 2.6.

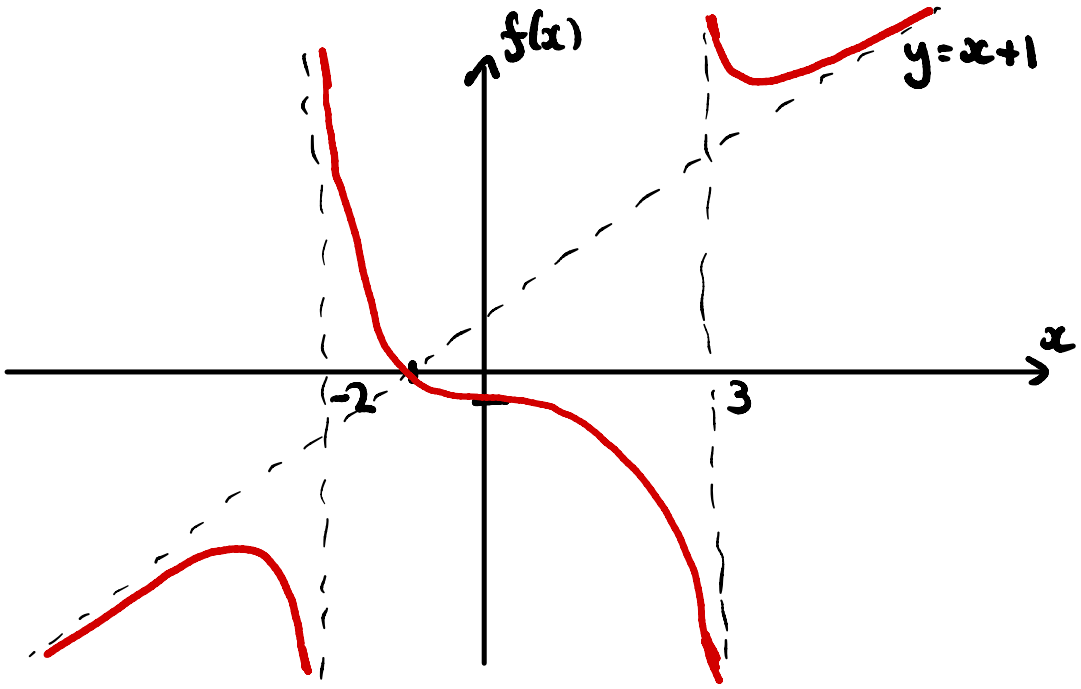


Figure 2.6: Sketch of the graph of $f(x) = \frac{x^3+1}{(x-3)(x+2)}$. The y-axis intercept should be marked at $y = -\frac{1}{6}$.

For examples involving exponential functions like $f(x) = 2^x$ or trigonometric functions like $f(x) = \sin x$, consult the sketches in the next section to understand the behaviour of these functions.

Indeed, we will now turn our attention to trying to write down more formally a large set of 'standard' functions, which will serve as our main examples and building blocks for the remainder of the course.

2.6 Some standard functions

2.6.1 Exponents, Logarithms and Polynomial Functions

Our aim in this section is to define polynomial functions, as well as the exponential function and the logarithm. Unfortunately there is an immediate pedagogical problem. In this course, I hope to introduce you to the idea of being mathematically 'careful'. This means:

- defining your terms clearly and rigorously;

- not taking mathematical statements at face value;
- always seeking an explanation for why a mathematical assertion is true.

However, this is not a course in 'analysis from first principles': that is the Real Analysis course in 2nd year, which combines ideas from this course with the fundamental techniques you are learning in the Sequences and Series module. Therefore, I must occasionally ask you to take some mathematical statements 'on trust', in so much as these statements have not yet been proved rigorously in this module but they can be proved rigorously, using techniques from your Sequences and Series module or from 2nd year modules.

We will first take some time to discuss the operation of raising a number to a power. This will illustrate a common technique in mathematics, that of developing a simple definition in a straightforward setting so that it can be applied in a more general setting. Our aim will be to start with the definition of a real number raised to the power of a positive integer, and then use this to arrive at a definition when the power is also a real number.

Definition 2.6.1 (Exponents, positive whole numbers)**.** Let $x \in \mathbb{R}$ and $n \in \mathbb{N}$. In the expression x^n, n is called the *power (or exponent)* and x is called the *base*. The number x^n denotes the n-fold multiplicative product of x, i.e.

$$x^n := \underbrace{x \times x \times x \ldots \times x}_{n\text{-fold product of } x\text{'s}}$$

Comment(s). The notation x^n to denote the n'th power of x was, like so many things, first introduced by Descartes in his book La Géométrie, published in 1637.

It is clear from the definition of x^n above that n must be a positive integer. So what does it mean to multiply a number by itself a negative number of times? Or a fraction number of times? Or even zero times?

The definition allows us to develop some rules for multiplying powers of numbers, informed by our understanding of multiplication of the real numbers – but written in the short exponent notation. For example,

$$x^n \times x^m = \underbrace{x \times x \times x \ldots \times x}_{n\text{-fold product of } x\text{'s}} \times \underbrace{x \times x \times x \ldots \times x}_{m\text{-fold product of } x\text{'s}} = \underbrace{x \times x \times x \ldots \times x}_{(n+m)\text{-fold product of } x\text{'s}} = x^{n+m}$$

where $n, m \in \mathbb{N}$.

We may also use the exponent notation to divide powers of x. If $n, m \in \mathbb{N}$ with $n \geqslant m$ then

$$\frac{x^n}{x^m} = \frac{\overbrace{x \times x \times x \ldots \times x}^{n\text{-fold product of } x\text{'s}}}{\underbrace{x \times x \times x \ldots \times x}_{m\text{-fold product of } x\text{'s}}} = \underbrace{x \times x \times x \ldots \times x}_{(n-m)\text{-fold product of } x\text{'s}} = x^{n-m} \quad \text{if } n, m \in \mathbb{N} \text{ with } n \geqslant m.$$

This last condition is a little restrictive, and comes from our definition of x^n which is only defined when n is a positive integer. Using the rule for multiplication of powers we realise that we can give meaning to x^{-m} (when $m \in \mathbb{N}$) via the formula

$$x^{n-m} = x^n \times x^{-m} = \frac{x^n}{x^m}.$$

Hence we have constructed a logical definition for the meaning of x^{-n} when $n \in \mathbb{N}$, namely:

$$x^{-n} := \frac{1}{\underbrace{x \times x \times x \ldots \times x}_{n\text{-fold product of } x\text{'s}}}.$$

To recap, we have extended the definition of x^n from n being a positive integer, to n being a positive or a negative integer. From this the definition of x^0 also follows quickly as

$$x^0 := x^{n-n} = x^n \times x^{-n} = \frac{\overbrace{x \times x \ldots \times x}^{n\text{-fold product of } x\text{'s}}}{\underbrace{x \times x \ldots \times x}_{n\text{-fold product of } x\text{'s}}} = 1.$$

Consequently we have extended our definition of x^n to any integer power, i.e. when $n \in \mathbb{Z}$.

Turning again to our knowledge and experience of multiplication of the real numbers we also understand that

$$(x^m)^n = \underbrace{x^m \times x^m \ldots \times x^m}_{n} = \overbrace{(\underbrace{x \times x \times x \ldots \times x}_{m}) \times (\underbrace{x \times x \times x \ldots \times x}_{m}) \times \ldots (\underbrace{x \times x \times x \ldots \times x}_{m})}^{\text{n}} = x^{m \times n}.$$

Comment(s). It is customary to drop the multiplication symbol algebraically and write $m \times n = mn$, where, as here, when the multiplication rule is obvious. Hence one writes $(x^m)^n = x^{mn} = x^{nm}$.

We can now extend the notion of the exponent to any rational power $\frac{p}{q} \in \mathbb{Q}$. If $x^{p/q}$ is to follow the same rules as above, we would have

$$(x^{p/q})^q = x^{qp/q} = x^p.$$

Therefore $x^{p/q}$ must denote[5] a q'th root of x^p. By convention, we always define $x^{p/q}$ to be the positive real q^{th} root of x^p, e.g. if $p = 1$ and $q = 2$ then $x^{1/2} := \sqrt[2]{x} = \sqrt{x}$ is the square root, or if $p = 1$ and $q = 3$ then $x^{1/3} := \sqrt[3]{x}$ is the cube-root, or if $p = 2$ and $q = 5$ then $x^{2/5} := (x^2)^{1/5} = \sqrt[5]{x^2}$. With this interpretation we have extended the range of our definition of x^n to $n \in \mathbb{Q}$, although note that in general $x^{p/q}$ is only well-defined when $x \geqslant 0$ (as otherwise x^p might not have a real q^{th}-root).

[5]Again, we leave the issue of whether the q^{th} root always exists in $\mathbb{R}$ as a question for other modules.

Exercise 2.2. Check carefully that, under these definitions, the laws for the manipulation of powers (i.e. $(x^n)(x^m) = x^{n+m}$ and $(x^m)^n = x^{nm}$) are now valid for all $n, m \in \mathbb{Q}$.

It is a little more challenging to go further and generalise the notation x^n to real powers of x, i.e. when $n \in \mathbb{R}$. A rigorous handling of this definition lies beyond the scope of this course, but we can sketch one option roughly as follows:

Suppose we want to give meaning to the expression 2^π. Here $\pi \approx 3.1415926535...$ is the ratio of a circumference of a circle and its diameter: it is known that π is irrational, so currently we don't have a definition for 2^π. Now, we can express π as a fraction to two decimal places, i.e. $3.14 = \frac{314}{100} = \frac{157}{50}$, and we already know that we can give meaning to $2^{157/50}$; this of course is not the same as 2^π, but given that 3.14 is quite close to π, we might reason that $2^{\frac{157}{50}}$ is a reasonably good approximation to whatever 2^π might be defined to be; we might write this (informally) as $2^{\pi(\text{up to 2.d.p.})}$. If we wanted to find a better approximation of 2^π, we can always find a more accurate rational approximation of π, e.g. to eight decimal places $3.14159265 = \frac{314159265}{10^8}$, and consider the value of $2^{\frac{314159265}{10^8}}$. This will surely be an even better approximation to 2^π (whatever 2^π actually is). As the rational approximations to π improve we will get a better and better approximation of 2^π: we can in fact *define* 2^π to be the 'limit' of these approximations. Of course we have not defined rigorously what a limit is, even though this technique was hopefully reasonably intuitive: the subject of limits will be a major topic later in this course. There was clearly nothing special about the values of 2 or π in the above discussion; we could just as easily have used any positive real number a in place of 2, and any real number x in place of π. We would have thus defined the value of a^x by using rational approximations to x.

2.6.2 Exponential functions

Let $a > 0$ be a positive real number. A function $f : \mathbb{R} \to (0, \infty)$ of the form $f(x) = a^x$ is called an exponential function. (Be careful to distinguish this from the definition of *the* exponential function below). For example, here is the graph of $f(x) = 2^x$ again.

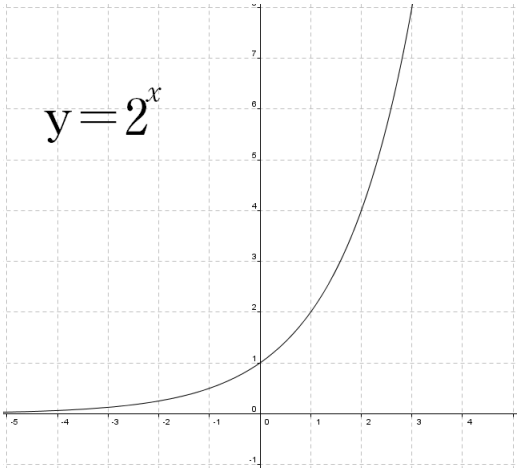


Figure 2.7: A sketch of $y = 2^x$. Observe the axis intercept at $x = 0, y = 1$ (as $2^0 = 1$).

The graphs of other exponential functions a^x look very similar (see Figure 2.8 below).

We will make use of the following facts about these functions, some of which you will formally prove in the Sequences and Series course:

Some Facts: Let $a \in (0, \infty)$. Then the functions $f_a : \mathbb{R} \to (0, \infty)$ given by $f_a(x) = a^x$:

- are well-defined functions;
- satisfy $a^1 = a$, and more generally $a^{\frac{p}{q}}$ agrees with the exponent definition given above for rational $\frac{p}{q} \in \mathbb{Q}$;
- satisfy $a^{x+y} = a^x a^y$ and $(a^x)^y = a^{xy}$ for all $x, y \in \mathbb{R}$;
- if $a \neq 1$, f_a is bijective;
- if $a > 1$ then f_a is monotonically increasing, and if $a < 1$ then f_a is monotonically decreasing;
- $f_a(x)$ is *continuous* at every point x.

The final bullet point requires some further discussion. We have not yet formally defined the notion of what it means for a function to be continuous, and this will be an important concept later in the course. For now, we will satisfy ourselves with the intuitive notion that a function is continuous at every point if you can draw its curve without 'taking the pen of the paper'. Observe the continuous curve in the graph in Figure 2.7. However, it is ***extremely important***

that you supplement this intuitive understanding with the more detailed rigorous definition of the concept of continuity, which will be covered in upcoming lectures.

The Exponential Function.

Definition 2.6.2. The *exponential function* $f : \mathbb{R} \to (0, \infty)$ is denoted

$$f(x) = e^x$$

where

$$e = 1 + \frac{1}{1} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots \approx 2.718281828459045235...$$

is Euler's number. Here, $n! := 1 \times 2 \times \cdots \times n$, so $2! = 2$, $3! = 6$, $4! = 24$ and so on.

Comment(s). (On the exponential function)

1. Euler's number is named after the prolific Swiss mathematician Leonhard Euler, and is usually simply called 'e'. The value of the exponential function at $x = 1$ gives a value for e.

2. Sometime the exponential function is denoted by $\exp(x)$. In this notation, the value of e can be recovered as the value $\exp(1)$.

3. There are alternative and equivalent definitions of the exponential function, such as

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

or as the unique solution $f(x)$ to the differential equation $\frac{df}{dx} = f$ satisfying $f(0) = 1$, or as the value of y which satisfies

$$x = \int_1^y \frac{dt}{t}.$$

These definitions make use the limit, the derivative, and the integral respectively, concepts which we will define later in this course.

4. $e^x > 0$ for all $x \in \mathbb{R}$.

5. e^x is a monotonically increasing function.

6. For $x > 0$, e^x grows very quickly ('exponential growth') and is unbounded. Consequently, for $x < 0$, e^x decreases rapidly as x decreases and approaches zero as x approaches $-\infty$. See the sketch of e^x in figure 2.11.

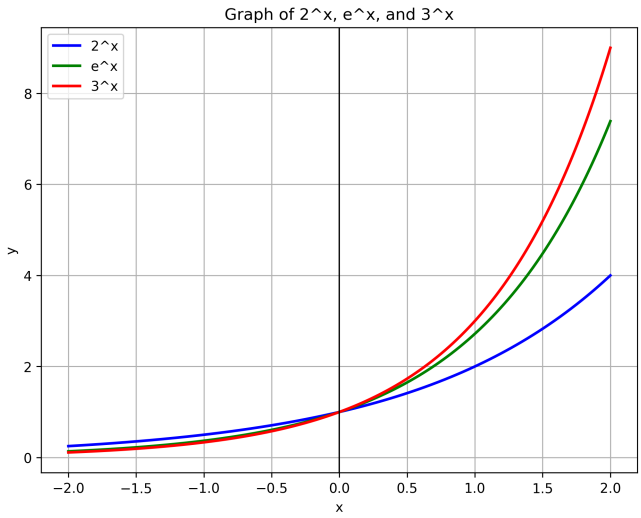


Figure 2.8: The exponential functions 2^x, e^x and 3^x.

2.6.3 The Logarithm

Definition 2.6.3 (Logarithm)**.** Let $a \in (0, \infty) \setminus \{1\}$. The *logarithmic function*

$$\log_a : (0, \infty) \to \mathbb{R},$$

or logarithm, to the base a, is the inverse function of the function $f : \mathbb{R} \to (0, \infty)$ given by $f(x) = a^x$.

Comment(s). (On the logarithm...)

1. In other words, '$\log_a(y)$ gives the power to which I must raise a to get y' i.e. $\log_a(a^x) = x$.
2. Note that the inverse function to $x \mapsto a^x$ exists provided $a \neq 1$, as the exponentiation function is a bijection in these cases.
3. The graph of $y = \log_2 x$ is shown in Figure 2.9. This is the inverse function of $f : x \mapsto 2^x$ viewed as a function $f : \mathbb{R} \to (0, \infty)$. As $f : A \to B$ and $f^{-1} : B \to A$, then you would expect that the inverse function is the mirror image of the original function in the line $y = x$ as indicated in figure 2.9.
4. The natural logarithm is defined as the logarithmic function when the base is e, and is denoted ln, i.e. $\ln(x) := \log_e(x)$. The graph of $y = \ln(x)$ is shown in Figure 2.10. Notice that broadly there is little difference in the sketch of the two logarithm graphs $y = \ln x$ and $y = \log_2 x$.

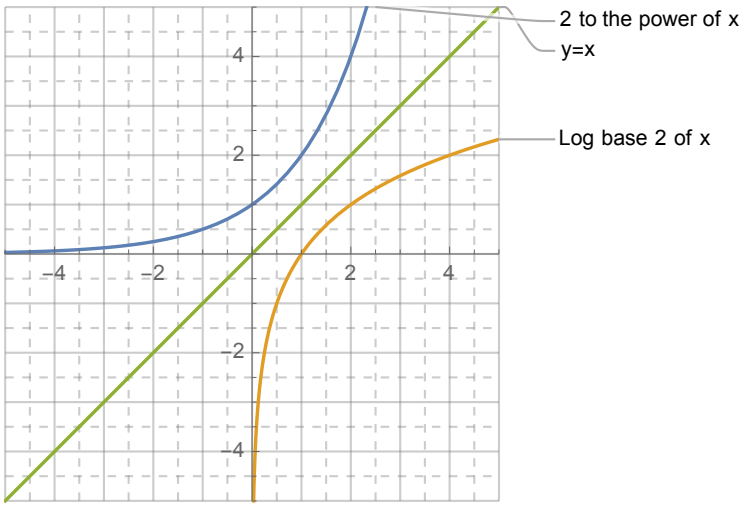


Figure 2.9: A sketch of $y = 2^x$, $y = x$ and $y = \log_2 x$. The curve $y = \log_2 x$ is the reflection in the line $y = x$ of $y = 2^x$.

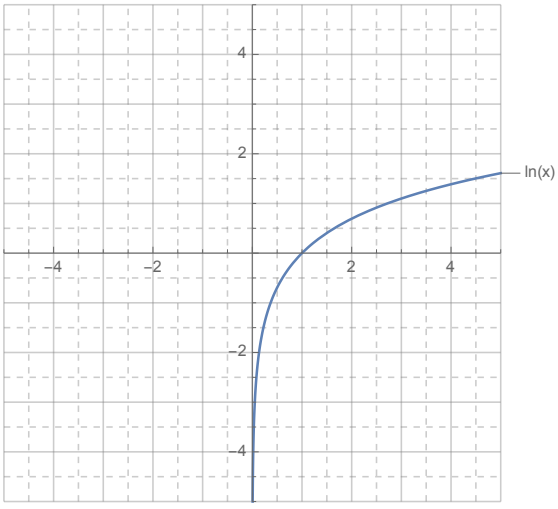


Figure 2.10: A sketch of $y = \ln(x)$. Notice that it is qualitatively similar to the graph of $y = \log_2(x)$ shown in figure 2.9.

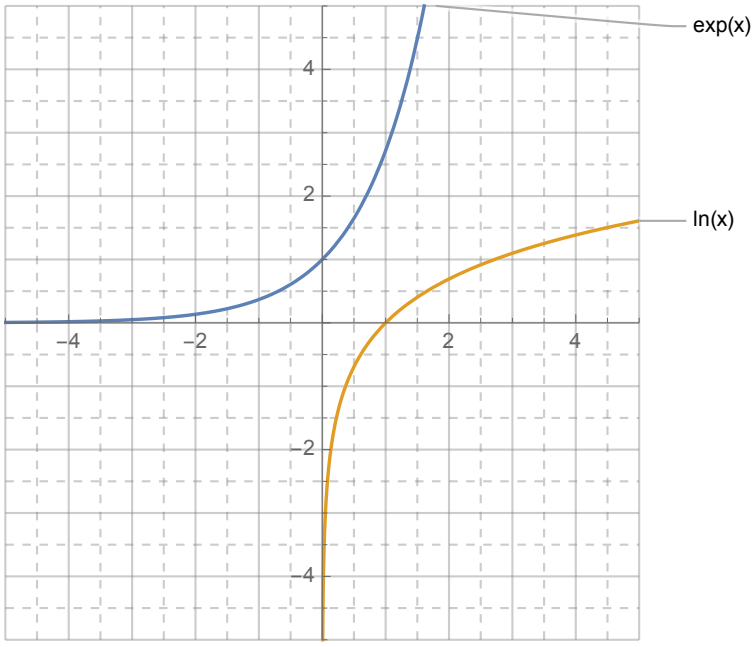


Figure 2.11: The exponential function e^x (or $\exp(x)$) and its inverse the natural logarithm $\ln(x)$.

The logarithm obeys several (very useful) identities, which can be derived from properties of exponential functions. For all $x, y, a, b > 0$ where $a, b \neq 1$, we have:

(i) $\log_a x = \frac{\log_b x}{\log_b a}$

(ii) $\log_a xy = \log_a x + \log_a y$.

(iii) $\log_a x^y = y \log_a x$.

Proof (of the logarithm identities):[6]

(i) We will prove that $\log_a x \log_b a = \log_b x$, by raising both sides of this equation as a power of base b. For the right-hand-side, by definition, we have

$$b^{\log_b x} = x$$

while from the left-hand-side we obtain

$$b^{\log_a x \log_b a} = (b^{\log_b a})^{\log_a x} = a^{\log_a x} = x.$$

[6]Note that proving identities is always a matter of showing that the two sides of the given equation are equal, using pre-existing properties you are assuming to be true – in this case, properties of exponential functions.

Note that exponentiation with the base b gives a unique output for each input (i.e. it is a one-to-one function). Therefore if $b^y = b^z$ we may conclude that $y = z$. Applying this principle here, since $b^{\log_a(x)\log_b(a)} = b^{\log_b(x)}$ we conclude that $\log_a(x)\log_b(a) = \log_b(x)$, and hence $\log_a(x) = \frac{\log_b(x)}{\log_b(a)}$ as claimed.

(ii) We will raise both sides of the identity $\log_a xy = \log_a x + \log_a y$ as a power of a and show the result is identical. From the left-hand-side we have

$$a^{(\log_a xy)} = xy$$

while from the right-hand-side we find

$$a^{(\log_a x + \log_a y)} = a^{(\log_a x)} a^{(\log_a y)} = xy.$$

Since exponentiation with the base a gives a unique output for each input (it is a one-to-one function), we have shown the identity.

(iii) We will use the same device again and raise both sides of the identity $\log_a x^y = y \log_a x$ as a power of a. From the left-hand-side we have

$$a^{(\log_a x^y)} = x^y$$

while from the right-hand-side we find

$$a^{(y \log_a x)} = (a^{(\log_a x)})^y = x^y.$$

As before, this shows the identity.

2.6.4 Polynomial Functions

Definition 2.6.4 (Polynomial function)**.** Let $n \in \{0, 1, 2, 3, \ldots\}$. A *polynomial function in x of degree n* is a function that can be written in the form

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0$$

where $a_i \in \mathbb{R}$ for $i \in \{0, 1, 2, \ldots n\}$ and $a_n \neq 0$.

Comment(s). (On polynomial functions...)

1. A constant function is trivially a polynomial of degree zero.

2. A linear equation $f(x) = ax + b$ is a polynomial of degree one.

3. A quadratic equation $f(x) = ax^2 + bx + c$ is a polynomial of degree two, and so on....

4. We have not mentioned the domain and range of the polynomial functions. This is because these functions might be real polynomial functions (mapping from $\mathbb{R}$ to $\mathbb{R}$) or complex polynomial functions (mapping from $\mathbb{C}$ to $\mathbb{C}$). More generally, there is an algebraic object called a *ring* (which you will meet in the Introduction to Abstract Algebra course), and given a ring A you can define a polynomial map from A to A. The reals $\mathbb{R}$ and complex numbers $\mathbb{C}$ are special examples of rings called *fields*, but there are many more examples (in particular $\mathbb{Z}$). The context in which the polynomial function is used will normally define clearly the range and domain of any particular polynomial function. In this course we will consider real and complex polynomial functions.

2.6.5 A look ahead at derivatives...

We have not formally introduced the derivative yet in these lecture notes. Of course many students will have met the derivative before, but it is unfair to those who haven't to assume too much knowledge in advance. That said, when discussing polynomials and the exponential function (and trigonometric and hyperbolic trigonometric functions in the forthcoming lectures), differentiation can be a very useful perspective. Therefore, we will introduce a rudimentary definition of the differentiation operator, before considering it in much more detail later on in the course.

Definition 2.6.5 (Informal and rudimentary definition of $\frac{d}{dx}$)**.** We define $\frac{d}{dx}$ to be an operator on functions, i.e. if $f : \mathbb{R} \to \mathbb{R}$ is a function then $\frac{d}{dx}(f) : \mathbb{R} \to \mathbb{R}$ is another function. For any $n \in \mathbb{N}$ we define

$$\frac{d}{dx}(x^n) := nx^{n-1}.$$

We extend linearly, meaning that for any polynomial

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

we have

$$\frac{d}{dx}(f) = na_n x^{n-1} + (n-1)a_{n-1}x^{n-2} + \cdots + 2a_2 x + a_1,$$

i.e. we apply differentiation term by term. To make the notation cleaner, we sometimes write $\frac{df}{dx}$ for the function $\frac{d}{dx}(f)$. The operation $\frac{d}{dx}$ can be applied to a function more than once, of course, and we write $\frac{d^m f}{dx^m}$ for $\frac{d}{dx}$ applied to f a total of m times.

For example,

$$\frac{d}{dx}(x^5 + x^2 + 1) := 5x^4 + 2x$$

and

$$\frac{d^2}{dx^2}(x^5 + x^2 + 1) := 20x^3 + 2.$$

Similarly

$$\frac{d}{dx}(3x^4) = 3\frac{d}{dx}(x^4) = 3(4x^3) = 12x^3.$$

I cannot stress highly enough how this is an ***informal*** definition. In particular we have not made any comments on whether all functions $f : \mathbb{R} \to \mathbb{R}$ are valid in the expression $\frac{d}{dx}(f)$ (Hint: they aren't!) nor talked about differentiating functions that can't be written in terms of monomials x^n. All this is to come.

For now, we make the observation that the derivatives of a polynomial function, evaluated at the point $x = 0$, are very closely related to the coefficients a_i themselves. Indeed, note first that $f(0) = a_0$. Next,

$$\frac{df}{dx} = na_n x^{n-1} + (n-1)a_{n-1}x^{n-2} + \cdots + 2a_2 x + a_1$$

so

$$\frac{df}{dx}\Big|_{x=0} = a_1.$$

(Remark: the notation $\frac{df}{dx}|_{x=0}$ means "the function $\frac{df}{dx}$ evaluated at the point $x = 0$".) Furthermore,

$$\frac{d}{dx}\Big(\frac{df}{dx}\Big) = \frac{d^2 f}{dx^2} = n(n-1)a_n x^{n-2} + (n-1)(n-2)a_{n-1}x^{n-3} + \cdots + 6a_3 x + 2a_2.$$

So

$$\frac{d^2 f}{dx^2}\Big|_{x=0} = 2a_2.$$

Continuing this process, we get the following table:

m	$\frac{d^m f}{dx^m}\big\vert_{x=0}$
0	a_0
1	a_1
2	$2a_2$
3	$6a_3$
$\vdots$	$\vdots$
$n-1$	$(n-1)!\,a_{n-1}$
n	$n!a_n$
$n+1$	0
$n+2$	0
$\vdots$	$\vdots$

We see that from the values of the multiple derivatives at a point we have extracted all the data about the coefficients, from which we could reconstruct the polynomial function itself. This illustrates an alternative way to define a polynomial function.

The polynomial functions are ubiquitous in mathematics, and we will meet many examples in this course. However, their definition immediately suggests a way to generalise them: could we consider polynomial functions of infinite degree? This would mean the function would be an infinite sum of terms, with each term given in the form of $a_m x^m$ for various $m \in \{0, 1, 2, \dots\}$ and coefficients $a_m \in \mathbb{R}$. We will meet such functions, known as power series, frequently in the course – particularly in the final section. However we raise the idea here, because it is a useful lens through which to study the exponential function.

2.6.6 The Exponential Function: power series

In our comments following the definition of the exponential function $f(x) = e^x$, we commented that an equivalent definition would be to say that $f(x) = e^x$ is the unique solution to the differential equation

$$\frac{df}{dx} = f$$

with $f(0) = 1$. Let us explore a little further what this might mean – allowing that the rigorous justification of much of the manipulations will not be achieved until the very end of the course.

We will start by testing an assumption that e^x can be written as a polynomial in x of finite degree n (and we will see how this fails). Let us suppose that there are coefficients $a_k \in \mathbb{R}$

(with $a_n \neq 0$) such that for all x we can write

$$e^x = \sum_{k=0}^{n} a_k x^k.$$

Now we will take the derivative with respect to x. This means that, as a function,

$$\frac{d}{dx}(e^x) = \frac{d}{dx}(\sum_{k=0}^{n} a_k x^n) = \sum_{k=0}^{n} \frac{d}{dx}(a_k x^k) = \sum_{k=0}^{n} a_k k x^{k-1} = \sum_{k=1}^{n} a_k k x^{k-1} = \sum_{j=0}^{n-1} a_{j+1}(j+1)x^j$$

Now by the definition of the exponential function this derivative must equal our proposed polynomial expression for e^x i.e.

$$\sum_{k=0}^{n} a_k x^k = \sum_{k=0}^{n-1} a_{k+1}(k+1)x^k \tag{2.1}$$

for all x. Notice that the left-hand-side is a polynomial of degree n while the right-hand-side is a polynomial of degree $n-1$. In particular, (2.1) contradicts the fact that if two polynomials agree for all $x \in \mathbb{R}$ they must have the same coefficients[7] and the same degree. Therefore e^x cannot be written as a polynomial.

We have hinted already how we can resolve this: we will allow the maximum power of x to tend to infinity, and the sum becomes infinite. In other words, we seek to find coefficients $a_k \in \mathbb{R}$ for which, for all x, the exponential function can be written as

$$e^x = \sum_{k=0}^{\infty} a_k x^k.$$

We are neglecting to mention any rigorous definition of what it means to sum infinitely many numbers together, nor the conditions under which such an expression is valid – this is a matter for later on, and in particular for the Sequences and Series course.

The condition that

$$\frac{d}{dx} e^x = e^x$$

for all x becomes

$$\sum_{k=0}^{\infty} a_k x^k = \sum_{k=0}^{\infty} a_{k+1}(k+1)x^k$$

(where we have assumed that we can differentiate term by term, like for polynomials). Can we pick coefficients a_k so that the left-hand-side equals the right-hand-side? Well, if we try to

[7] When you meet polynomials over finite fields $\mathbb{F}$ this fact no longer holds, but this subtlety will not feature in this course.

make the coefficient of each power of x to be equal on both sides, we require $a_k = (k+1)a_{k+1}$ for all k. In other words, $a_{k+1} = \frac{a_k}{k+1}$ for all k. We can repeatedly use this to find:

$$a_{k+1} = \frac{a_k}{k+1} = \frac{a_{k-1}}{(k+1)k} = \ldots = \frac{a_0}{(k+1)!}$$

where we again remind you that $(k+1)! = (k+1) \times k \times (k-1) \times (k-2) \times \ldots 2 \times 1$ is called the factorial of $(k+1)$ and is defined as indicated for all $k+1 \in \mathbb{Z}^+$. We supplement the definition by defining $0! := 1$. Putting together our findings we now have

$$e^x = \sum_{k=0}^{\infty} \frac{a_0}{k!} x^k$$

for all x. The final condition is $e^0 = 1$, which is satisfied when $a_0 = 1$. Therefore, via this somewhat ad-hoc method, we have derived the series form of the exponential function

$$e^x = \sum_{k=0}^{\infty} \frac{1}{k!} x^k = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \ldots. \tag{2.2}$$

More precisely, we have deduced that *if* e^x has a representation as a power series, and *if* this power series can be differentiated term by term, then the power series must have the above form.

We can use this expression to evaluate Euler's number as

$$e = e^1 = 1 + 1 + \frac{1}{2} + \frac{1}{6} + \frac{1}{24} + \frac{1}{120} + \frac{1}{720} \ldots = 2.718\ldots$$

as we did earlier.

To find e with greater accuracy one can sum more of the first terms (convince yourself that the first terms are the largest terms in the sum) e.g. to 50 decimal places we have

$$e = 2.71828182845904523536028747135266249775724709369995\ldots.$$

As an aside, you may be interest to know that e is an irrational, transcendental[8] number.

Let us pause, and again sound a note of alarm: what does an infinite sum mean? After all we can never write down the sum in completion: it does not end. How do we know that the sum does give a finite number? Shouldn't we be concerned that an infinite sum does not exist nor make sense? Shouldn't be concerned about whether, even if the infinite sum does exist and make sense, whether it is equal to the claimed function? Well, yes we should and yes we are.

[8] A transcendental number is any number which is not the solution of any non-constant polynomial equation with rational coefficients.

It will be the subject of the final chapter of this course, where we will address this concern and lose some of our worries.

For now, we introduce an important definition, to which we'll return at the very end of the course.

Definition 2.6.6 (Analytic functions, informal definition)**.** A function which can be locally written as a convergent power series is called an *analytic function*.

The word 'locally' in the above definition really means that the function is well-defined in the vicinity of a certain point. It might be that there is not a power series p such that $p(x) = f(x)$ for all x (as was the case with the exponential function). Rather, we may need to take a collection of different power series $p_1, \ldots, p_k$, where for each x we can find one of the power series p_i such that $p_i(x) = f(x)$. We will understand the reason for including this in the definition when we study Taylor series.

Here is the formal version of the definition, which is just a rephrasing of the one you saw at the very start of the course (when discussing mathematical definitions in general):

Definition 2.6.7 (Analytic functions, formal definition)**.** Let $A \subset \mathbb{R}$. Then a function $f : A \to \mathbb{R}$ is *analytic* if for all $x_0 \in A$ there is a positive $\delta > 0$, an interval $I_\delta := (x_0 - \delta, x_0 + \delta)$, and coefficients $a_0, a_1, a_2, \ldots, \in \mathbb{R}$ such that for all $x \in I_\delta$

$$f(x) = \sum_{k=0}^{\infty} a_k (x - x_0)^k.$$

You won't need to understand this definition fully now! All that matters is that you know that, although not all functions are analytic, all polynomial functions, the exponential function, the logarithm function, and the functions introduced in the next section (the trigonometric functions, and the hyperbolic trigonometric function) are examples of analytic functions.

2.6.7 The Trigonometric Functions.

There are many equivalent definitions for the trigonometric functions, and we will introduce both a geometric and an algebraic definition here. The natural introduction to the trigonometric functions is via geometry and the circle.

Definition 2.6.8. The *unit circle* is the set of points (x, y) satisfying $x^2 + y^2 = 1$.

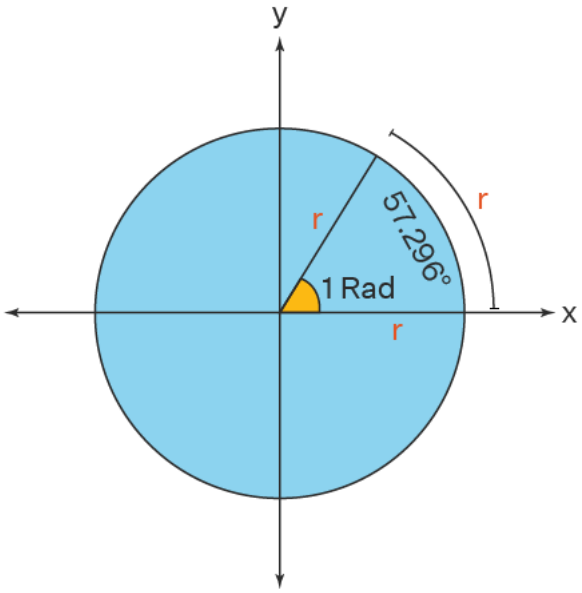
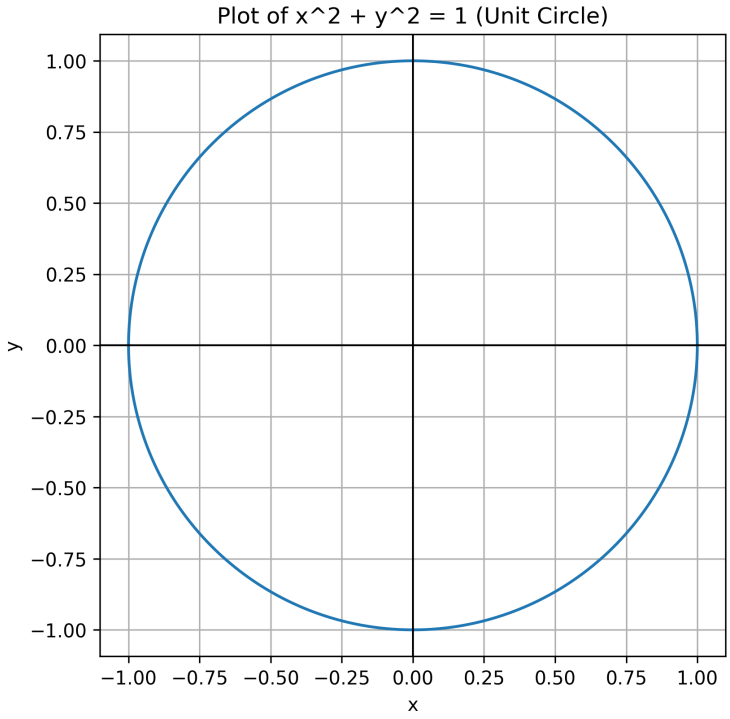


Figure 2.12: A radian



Definition 2.6.9 (Radian)**.** One *radian* is the angle subtended at the centre of a circle by an arc that is equal in length to the radius of the circle.

In other words, the arc length on a circle of radius R is equal to $R\theta$, where θ is the angle

(in radians) swept through by the radial vector for points along the arc. The radian is the standard unit of angular measurement. Since the circumference of the the circle $2\pi R$ for a circle of radius R, 2π is the length of the circumference of the unit circle. If one wishes to convert degrees to radians then using the conversion for the unit circle $360° = 2\pi$ radians, or 1 radian $= \frac{360°}{2\pi} = 57.295779513\ldots°$.

Now we can define the trigonometric functions $\cos\theta$, $\sin\theta$ and $\tan\theta$. It can be useful to see all these functions on the same diagram, which we give in Figure 2.13 below.

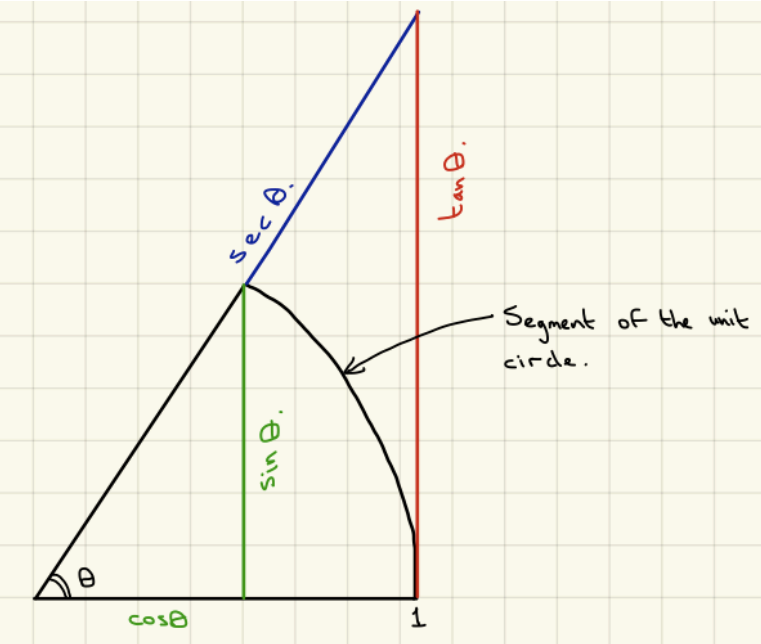


Figure 2.13: A segment of the unit circle, subtending an angle of θ. The length $\sec\theta$ refers to the entire hypoteneuse of the large right-angled triangle, and indeed the scale factor to map the small right-angled triangle up to the large one is $\sec\theta$.

Where do all these unusual names come from? Well, Latin is the short answer... the sine function is derived from 'sinus' meaning 'a bend', while cosine is short for 'complementary sine', the tangent originates from 'tangere' meaning 'touching' (see the sketch in figure 2.13) and secant from 'secare' meaning 'to cut' (again see the sketch in figure 2.13 to get the sense behind these names).

Definition 2.6.10 (Cos and Sin)**.** The points $(x, y) = (\cos\theta, \sin\theta)$ are the Cartesian coordinates of points on the unit circle whose radial vector (i.e. the straight line from the point $(0, 0)$ to (x, y)) is at an angle θ to the x-axis.

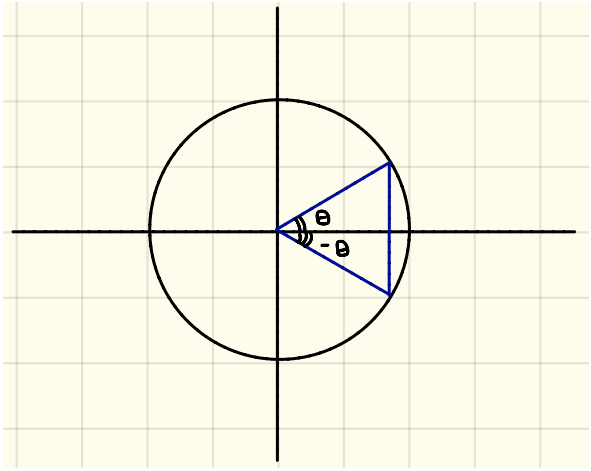
Definition 2.6.11 (Tan)**.** The tangent function is given by $\tan\theta := \frac{\sin\theta}{\cos\theta}$.

Comment(s). (On the trigonometric functions...)
Due to the definition above we immediately see that

1. $\cos^2\theta + \sin^2\theta = 1$.
2. $\cos : \mathbb{R} \to [-1, 1]$ and $\sin : \mathbb{R} \to [-1, 1]$.
3. $\tan\theta$ is not defined for $\theta = \frac{\pi}{2} + n\pi$ where $n \in \mathbb{Z}$: for these values of the angle θ, $\cos\theta = 0$.
4. $\cos\theta$ and $\sin\theta$ are both periodic, with period 2π, in other words $\cos(\theta + 2\pi) = \cos\theta = \cos(\theta + n2\pi)$ where $n \in \mathbb{Z}$, and similarly for $\sin\theta$.
5. $\tan\theta$ is periodic, but its period is π as $\sin(\theta + \pi) = -\sin\theta$ and $\cos(\theta + \pi) = -cos(\theta)$ (see below for proof of these statements).
6. $\cos\theta$ and $\sin\theta$ have simple values at $\theta = 0, \frac{\pi}{2}, \pi, \ldots$ and also at $\theta = \frac{\pi}{4}, \frac{3\pi}{4}, \frac{5\pi}{4}, \ldots$.

Sketches of the trigonometric functions can be derived quickly from the definitions and are shown in figure 2.14. In the comments above we made use of some fundamental properties of sine and cosine that originate from their definition as defining the coordinate $(x, y) = (\cos\theta, \sin\theta)$ on the unit circle. Since any point on the unit circle can be written in terms of sine and cosine, transformations which generate symmetries of the circle maps one point on the circle (x, y) to another point (x', y') also on the circle and so also expressible in terms of sine and cosine. That is to say, the symmetry transformations of the circle give rise to interesting transformations of sine and cosine. We consider the effect of some simple symmetry transformations:

(i) Reflection in the x-axis, i.e. $(x, y) \to (x', y') = (x, -y)$



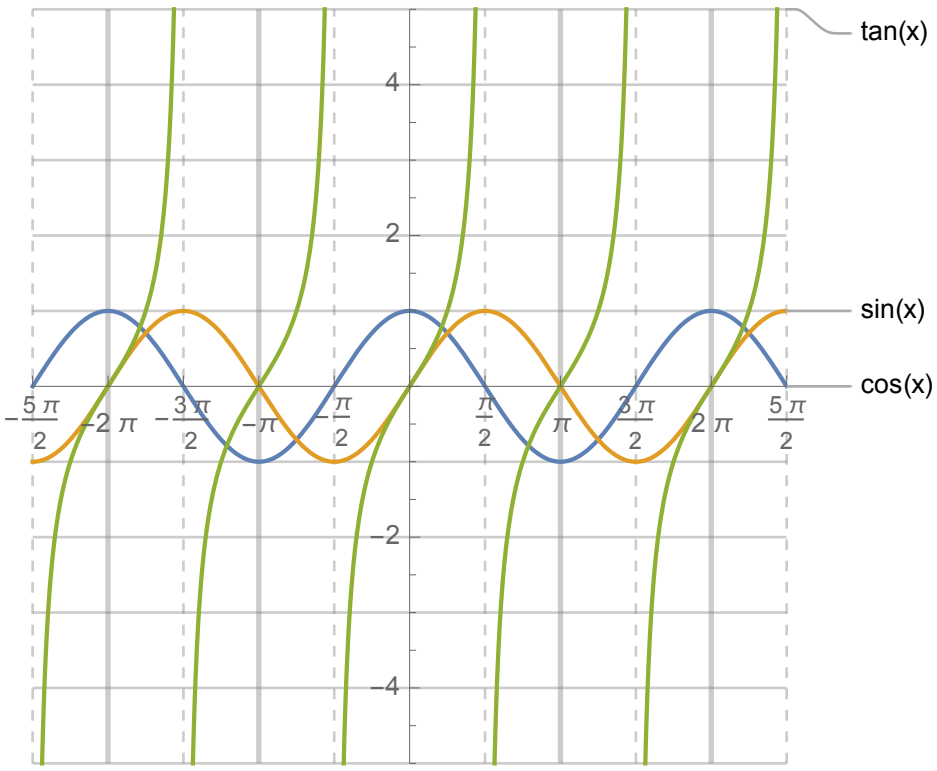


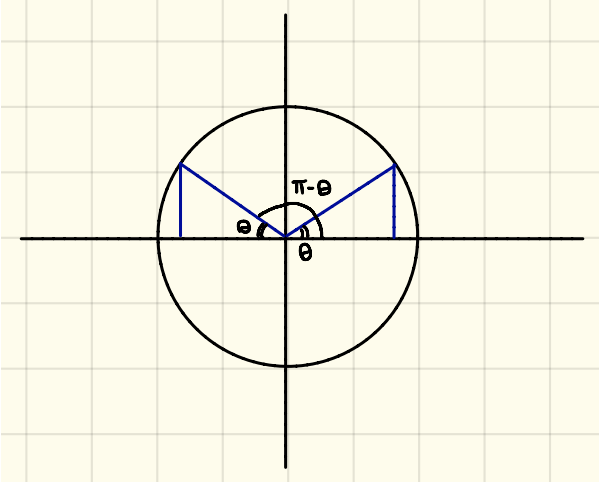
Figure 2.14: The trigonometric functions $\cos(x)$, $\sin(x)$ and $\tan(x)$.

$$\cos\theta = x = x' = \cos(-\theta)$$
$$\sin\theta = y = -y' = -\sin(-\theta).$$

In summary:

$$\cos(-\theta) = \cos\theta$$
$$\sin(-\theta) = -\sin\theta.$$

(ii) Reflection in the y-axis, i.e. $(x, y) \to (x', y') = (-x, y)$

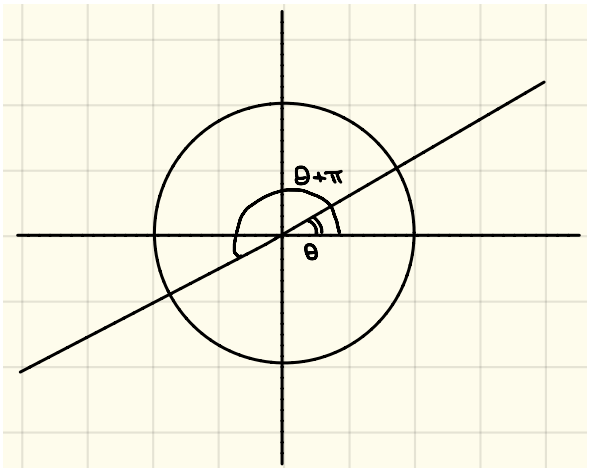


$$\cos\theta = x = -x' = -\cos(\pi - \theta)$$
$$\sin\theta = y = y' = \sin(\pi - \theta).$$

In summary:

$$\cos(\pi - \theta) = -\cos\theta$$
$$\sin(\pi - \theta) = \sin\theta.$$

(iii) Rotation by π i.e. $(x, y) \to (x', y') = (-x, -y)$

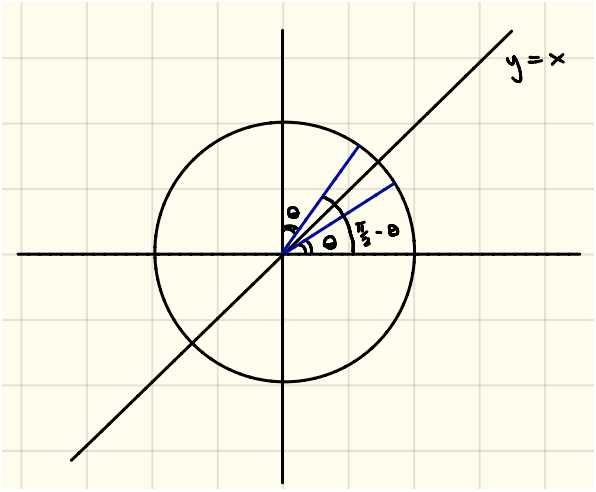


$$\cos\theta = x = -x' = -\cos(\theta+\pi)$$
$$\sin\theta = y = -y' = -\sin(\theta+\pi).$$

In summary:

$$\cos(\theta+\pi) = -\cos\theta$$
$$\sin(\theta+\pi) = -\sin\theta.$$

(iv) Reflection in the line $y = x$ i.e. $(x, y) \to (x', y') = (y, x)$



$$\cos\theta = x = y' = \sin(\frac{\pi}{2} - \theta)$$
$$\sin\theta = y = x' = \cos(\frac{\pi}{2} - \theta).$$

In summary:

$$\cos(\frac{\pi}{2} - \theta) = \sin\theta$$
$$\sin(\frac{\pi}{2} - \theta) = \cos\theta.$$

The identities above agree with the double-angle formulae for sine and cosine, which are often taught without proof in pre-university calculus. We will return to the trigonometric functions and derive the double-angle formulae in a later chapter.

Much of the modern analytic theory of trigonometric functions comes from Euler's work in the 18^{th} century. However, he was building on 2000 years of knowledge across most of the major mathematical traditions of the world. In particular, mathematicians of Europe were rediscovering much that had already been worked out in the flourishing Indian mathematical tradition of the Kerala school.

We will now state a second (equivalent) algebraic definition of the basic trigonometric functions. Again this will use the derivative operation $\frac{d}{dx}$, which has not yet been formally introduced! However, we have seen how $\frac{d}{dx}$ acts on power series, and this will be enough for the present discussion.

Definition 2.6.12 (Sin and Cos, algebraic)**.** *The solutions $f : \mathbb{R} \longrightarrow \mathbb{R}$ to the second order differential equation*

$$\frac{d^2f}{dx^2} = -f$$

are

$$f = \begin{cases} \cos x & \text{if } f(0) = 1 \text{ and } \frac{df}{dx}(0) = 0 \\ \sin x & \text{if } f(0) = 0 \text{ and } \frac{df}{dx}(0) = 1. \end{cases}$$

The conditions on the right ($f(0) = 1$ etc.) are sometimes called the 'boundary conditions' of the differential equation. You will think much more about boundary conditions in the dynamical systems course and in Calculus II.

We have asserted, in passing, that $\cos x$ and $\sin x$ are analytic functions. While we are not in a position to check convergence yet, we can certainly use the algebraic definition to identify the correct power series.

Example 2.8. Use the definition of $\cos x$ and $\sin x$ as solutions to $\frac{d^2f}{dx^2} = -f$ with the appropriate boundary conditions to find their power series.

Suppose there are some coefficients $a_0, a_1, \ldots$, for which $f(x) = \sum_{n=0}^{\infty} a_n x^n$ for all x. Then, using the differential equation for f,

$$\frac{d^2f}{dx^2} = \sum_{n=0}^{\infty} a_n n(n-1)x^{n-2} = \sum_{n=0}^{\infty} a_{n+2}(n+2)(n+1)x^n = -\sum_{n=0}^{\infty} a_n x^n$$

for all x. Equating coefficients of each power of x gives $a_{n+2}(n+2)(n+1) = -a_n$ or

$$a_{n+2} = -\frac{a_n}{(n+2)(n+1)}.$$

Rather like when establishing the power-series form of the exponential function, we can use this relationship recursively[9]. In this case, the outcome will depend on whether $n+2$ is even or odd. After a little thought and taking care with the minus signs, we find

$$a_{n+2} = \begin{cases} (-1)^{(n+2)/2}\frac{a_0}{(n+2)!} & n \text{ even} \\ (-1)^{(n+1)/2}\frac{a_1}{(n+2)!} & n \text{ odd} \end{cases}$$

hence we have

$$f(x) = a_0 \sum_{n\,even} (-1)^{n/2}\frac{x^n}{n!} + a_1 \sum_{n\,odd} (-1)^{(n-1)/2}\frac{x^n}{n!}.$$

Applying the boundary conditions for $\sin(x)$, where we have $f(0) = 0$ and $\frac{df}{dx}|_{x=0} = 1$ gives $a_0 = 0$ and $a_1 = 1$, therefore

$$\sin(x) = \sum_{n\,odd} (-1)^{(n-1)/2}\frac{x^n}{n!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots. \tag{2.3}$$

While the boundary conditions for $\cos(x)$ give $a_0 = 1$ and $a_1 = 0$ hence

$$\cos(x) = \sum_{n\,even} (-1)^{n/2}\frac{x^n}{n!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \ldots \tag{2.4}$$

where we have used $0! := 1$.

We can plot graphs of some of the partial sums of the first few terms of $\sin(x)$ and compare the result against the graph of $\sin(x)$. Bearing in mind that we expect the full sum to reproduce the sine function exactly, we may anticipate that the sum of the terms up to and including order n in x, given by $x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots + (-1)^{(n-1)/2}\frac{x^n}{n!}$ (for odd n) gives a better and better approximation to the sine function as n increases. Sketches of the partial sums are shown in figure 2.15.

In addition to the basic trigonometric functions there are a set of related and commonly used functions. They appear so frequently in mathematics that we will define them now. The first (secant) already appears on the diagram Figure 2.13.

[9] A full analysis of such recurrence relations is performed in the 2nd year Discrete Mathematics course.

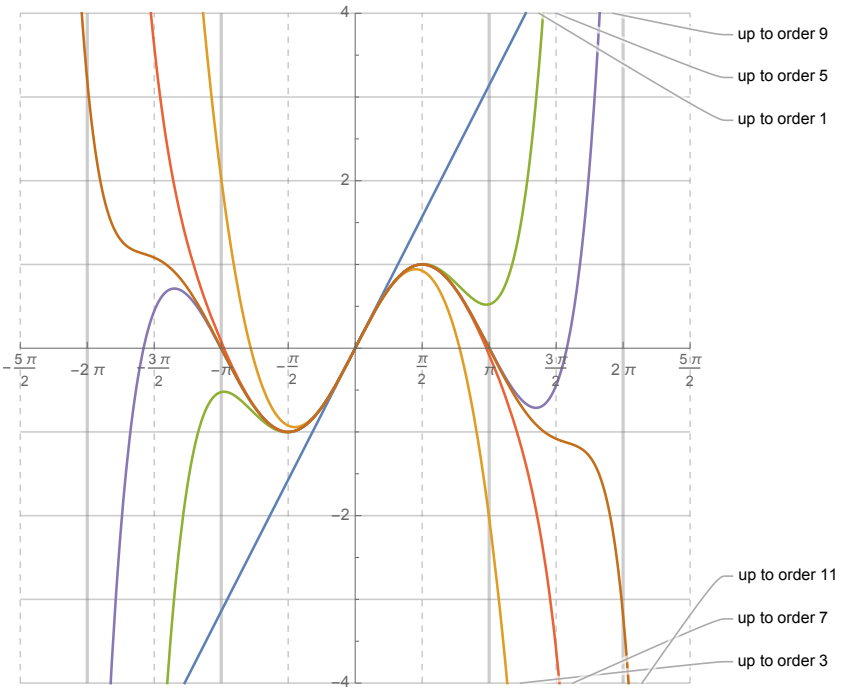


Figure 2.15: Building $\sin(x)$ as a power series, by taking more and more terms in the summation. Solid lines: $f(x) = \sum_{k\,odd}^{n} (-1)^{\frac{k-1}{2}} x^k/k!$ for different choices of n (values of n are indicated at the edges of the sketch by the order).

Definition 2.6.13 (Secant)**.** The *secant* function is denoted sec and defined as

$$\sec\theta = \frac{1}{\cos\theta}.$$

Comment(s). (On $\sec\theta$...)

1. It is not defined for $\theta = \frac{\pi}{2} + n\pi$ where $n \in \mathbb{Z}$. These are the same points where $\tan\theta$ is not defined, so you might suspect there is a connection between $\tan\theta$ and $\sec\theta$ and you would be correct. Indeed, dividing both sides of the identity $\cos^2\theta + \sin^2\theta = 1$ by $\cos^2\theta$ gives the related identity $1 + \tan^2\theta = \sec^2\theta$.

2. $\sec(\theta + 2\pi) = \sec\theta$ and $\sec(\theta + \pi) = -\sec\theta$.

Definition 2.6.14 (Cosecant)**.** The *cosecant* function is denoted cosec and defined as

$$\operatorname{cosec}\theta = \frac{1}{\sin\theta}.$$

Comment(s). (On $\operatorname{cosec}\theta$...)

1. It is not defined for $\theta = n\pi$ where $n \in \mathbb{Z}$.

2. $\operatorname{cosec}(\theta + 2\pi) = \operatorname{cosec}\theta$ and $\operatorname{cosec}(\theta + \pi) = -\operatorname{cosec}\theta$.

Definition 2.6.15 (Cotangent)**.** The *cotangent* function is denoted cot and defined as

$$\cot\theta = \frac{1}{\tan\theta}.$$

Comment(s). (On $\cot\theta$...)

1. It is not defined for $\theta = n\pi$ where $n \in \mathbb{Z}$.

2. $\cot(\theta + \pi) = \cot\theta$.

You may be wondering again where all these weird and wonderful names come from, and whether $\operatorname{cosec}\theta$ and $\cot\theta$ have any geometric significance. Figure 2.16 below sheds some light on the matter, showing how the tangent and cotangent form part of the same line, and the secant and cosecant are at right-angles. I must admit to not thinking about such diagrams very frequently myself, but they are nonetheless beautiful illustrations (to my taste) of the geometric underpinnings of the ideas.

The Inverse Trigonometric Functions.

The trigonometric functions are all periodic: sine and cosine have period 2π while the tangent has period π, see the sketch in figure 2.17. Consequently for a given θ there are infinitely many values θ', such that $\theta \neq \theta'$ but $\sin(\theta') = \sin(\theta)$, $\cos(\theta') = \cos(\theta)$ and $\tan(\theta') = \tan(\theta)$. This means that the trigonometric functions are not bijective functions, so they are not invertible functions on their full domain. However by restricting the domain, one can still construct well-defined inverse functions. In this section we will give the details of these definitions.

Definition 2.6.16 (Names of inverse trigonometric functions)**.** The inverse sine function is denoted arcsin or $\sin^{-1}$, the inverse cosine function is denoted arccos or $\cos^{-1}$ and the inverse tangent function is denoted arctan or $\tan^{-1}$.

From the graph in Figure 2.17, and our knowledge of the continued periodicity of the trigonometric functions, we see that, for example, $\sin\theta = \frac{1}{2}$ has an infinite number of solutions for θ, i.e. $\theta = \frac{\pi}{6} + n2\pi$ and $\theta = \frac{5\pi}{6} + m2\pi$ for all $n, m \in \mathbb{Z}$. That is, there exist $\theta_1 \neq \theta_2$ such that $\sin\theta_1 = \sin\theta_2 = \ldots$, so that on the domain of the real numbers $\mathbb{R}$ the sine function is not an injective function and so the inverse sine function is not defined.

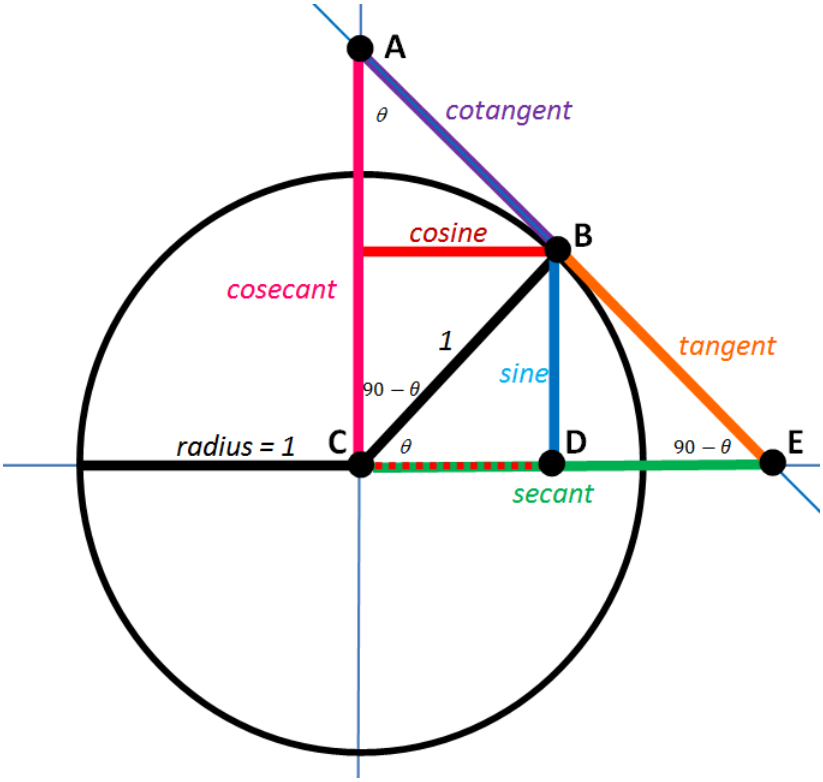


Figure 2.16: The coloured lengths represent $\sin\theta, \cos\theta$ etc. Apologies for the angles being given in degrees, this was the best diagram I could find!

Our aim is to reduce the domain and range of the trigonometric function such that each input gives a different unique output. Graphically this corresponds to restricting the sine and cosine graphs to a domain between consecutive turning points. There are many ways one could conceivably do this, but there are standard choices.

Definition 2.6.17 (Arcsin)**.** The function arcsin (also denoted $\sin^{-1}$) is defined by

$$\arcsin : [-1,1] \to [-\pi/2, \pi/2] \qquad \arcsin(\sin x) = x.$$

The function $\arcsin x$ when sketched as a graph is a reflection in the line $y = x$ of $\sin x$ over the domain $[-\pi/2, \pi/2]$. This is shown in figure 2.18.

Definition 2.6.18 (Arccos)**.** The function arccos (also denoted $\cos^{-1}$) is defined by

$$\arccos : [-1,1] \to [0, \pi] \qquad \arccos(\cos x) = x.$$

The sketch of $\cos(x)$ and $\arccos(x)$ are shown on the same graph in figure 2.19.

Definition 2.6.19 (Arctan)**.** The function arctan (also denoted $\tan^{-1}$) is defined by

$$\arctan : \mathbb{R} \to (-\pi/2, \pi/2) \qquad \arctan(\tan x) = x.$$

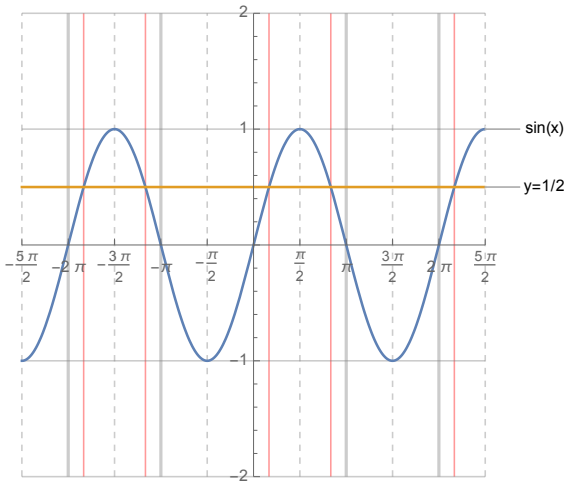


Figure 2.17: A sketch of $\sin x$ highlighting its many-to-one nature. The horizontal line indicates the line $y = \frac{1}{2}$ and its multiple intersections with $\sin x$ are shown over the domain $[-5\pi/2, 5\pi/2]$.

A plot of $y = \arctan x$ and $\tan(x)$ is shown in figure 2.20.

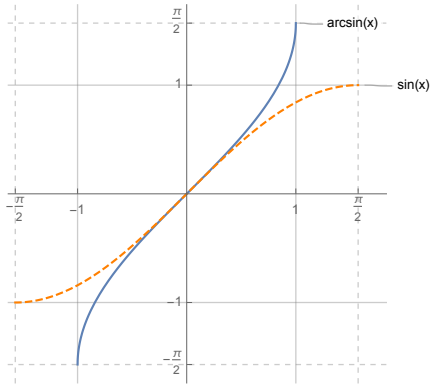


Figure 2.18: A sketch of $\arcsin x$ and $\sin x$ over the domains $[-1, 1]$ and $[-\pi/2, \pi/2]$ respectively. Notice how they are mirrored in the line $y = x$.

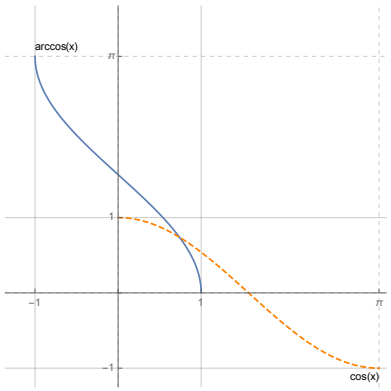


Figure 2.19: A sketch of $\arccos x$ and $\cos x$ over the domains $[-1, 1]$ and $[0, \pi]$ respectively.

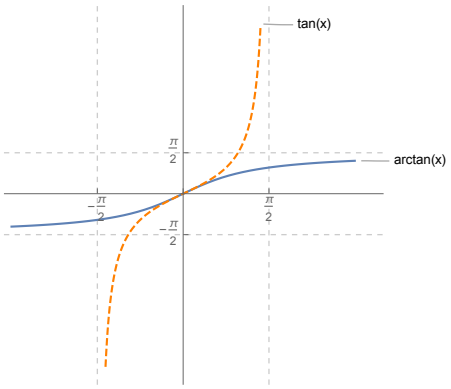


Figure 2.20: A sketch of $\arctan x$ and $\tan x$ over the domains $[-\pi, \pi]$ and $(-\pi/2, \pi/2)$ respectively.

2.6.8 The Hyperbolic Functions.

A major theme in classical geometry (since the work of Apollonius in the 3^{rd} century BCE) involved curves called 'conic sections'. These were curves that you could get by placing two cones of the same slope, balancing one on top of the other at their points, and then slicing through the whole ensemble with a plane. This is shown in Figure 2.21.

There are four types of conic sections, and you are probably familiar with all of them – even if you have never called them conic sections before. They are: circles, ellipses, parabolas, and hyperbolas. After Descartes, we are more used to defining these curves by algebraic relations such as $x^2 + y^2 = 1$ (a circle) or $x^2 - y^2 = 1$ (a hyperbola).

We have seen how the properties of trigonometric functions are naturally defined in terms of properties of a circle. Now, it turns out that similar functions may be defined with respect to

a different conic section – the hyperbola. These functions are called, unimaginatively enough, *hyperbolic functions* or *hyperbolic trig functions*. We introduce them formally below.

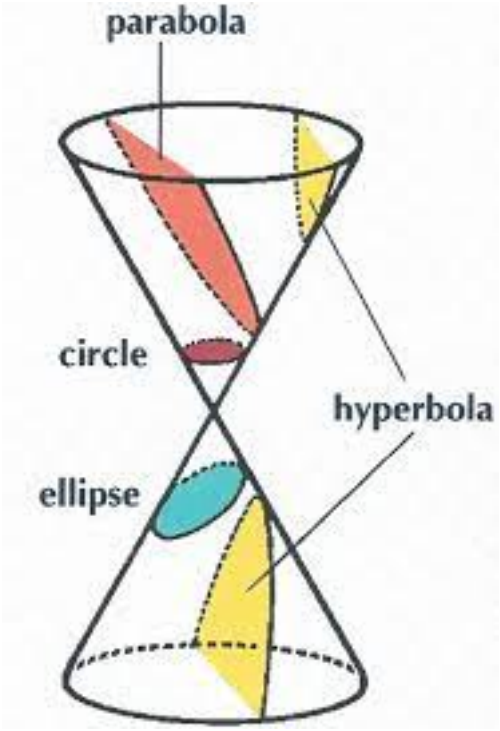


Figure 2.21: The four types of conic section marked on a double-cone. Observe how the hyperbola has two piece, corresponding to where the plane hits the top cone and the bottom cone.

Recall that we geometrically defined the trigonometric functions $x = \cos\theta$ and $y = \sin\theta$ as the coordinates of points on the unit circle $x^2 + y^2 = 1$, we will now define the hyperbolic (trigonometric) functions in a similar manner.

Definition 2.6.20 (Hyperbolic functions)**.** The hyperbolic functions $x = \cosh(a)$ and $y = \sinh(a)$ are the coordinates of points on the right-hand branch of the hyperbola $x^2 - y^2 = 1$, as depicted in Figure 2.22.

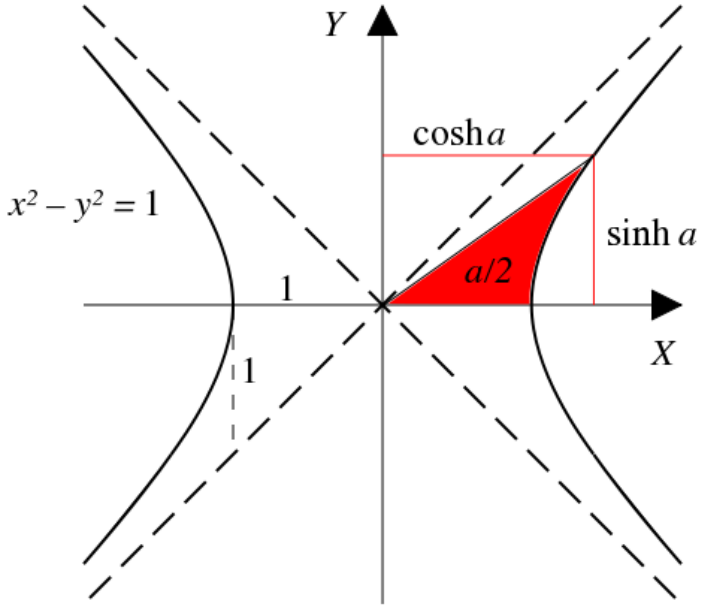


Figure 2.22: The hyperbolic functions $x = \cosh(a)$ and $y = \sinh(a)$ shown as points on the right-hand branch of the hyperbola $x^2 - y^2 = 1$.

Comment(s). (On the hyperbolic functions...)

1. The notation cosh and sinh are shorthand for 'hyperbolic cosine' and 'hyperbolic sine'.
2. There are multiple ways of pronouncing sinh (pronounced as either "shine" or "sinch") while cosh is pronounced as written ("kosh").
3. The definition of cosh and sinh in terms of an area is rather old-fashioned (more usual is to use the algebraic definition below). But it shows that, like trigonometric functions, there are both geometric and algebraic ways to define them.
4. By definition we have the identity:
$$\cosh^2(a) - \sinh^2(a) = 1.$$
5. The parameter a used to define the point $(\cosh(a), \sinh(a))$ is not an angle but has the magnitude of an area and maybe positive or negative (if the point (x, y) is below the x-axis). Its magnitude is twice the area shaded in Figure 2.22 i.e. up to a sign it is the area[10] enclosed between the x-axis, the right-hand branch of the hyperbola and the

[10]The simplest way to prove this is by integration - try and show this once you have met the integral later in the course.

straight line from the origin to the point $x = \cosh(a)$ and $y = \sinh(a)$. It is worth comparing this definition of a with that of the argument of the standard trigonometric functions. Indeed, looking at Figure 2.12 again and using the formula πr^2 for the area of a circle with radius r, the area of a segment of the unit circle subtending an angle θ at the origin is $\frac{1}{2}\theta r^2$, which for the unit circle is equal to $\frac{1}{2}\theta$. So you see that these definitions are directly analogous.

6. The functions are defined on the domain and ranges:

$$\sinh : \mathbb{R} \to \mathbb{R}$$

and

$$\cosh : \mathbb{R} \to [1, \infty).$$

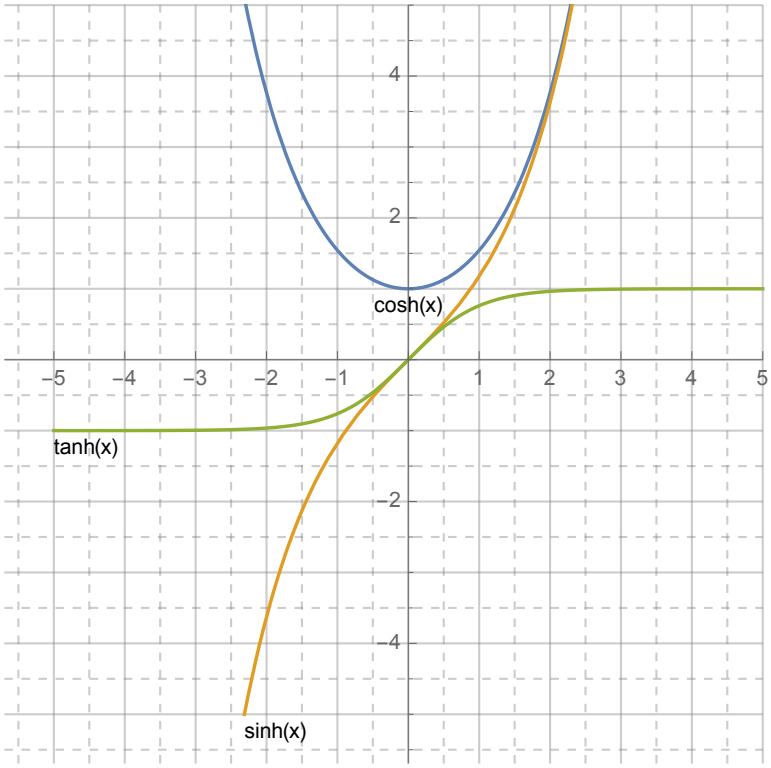


Figure 2.23: Sketches of the hyperbolic functions as the curves $y = \cosh(x)$, $y = \sinh(x)$ and $y = \tanh(x)$.

Definition 2.6.21 (Hyperbolic tangent)**.** The *hyperbolic tangent* function is denoted tanh and defined by

$$\tanh(a) = \frac{\sinh(a)}{\cosh(a)}.$$

Comment(s). (On tanh)

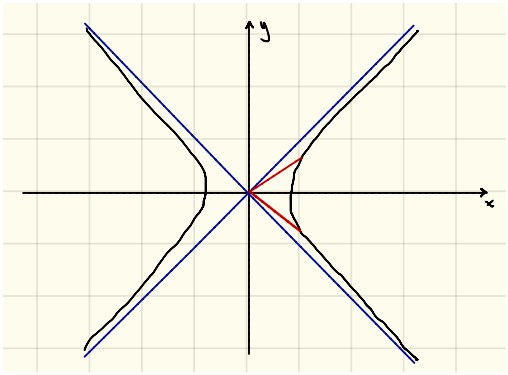
1. tanh is pronounced "tanch" or "than".
2. $\tanh(a)$ is the gradient of the straight line through the origin and the point with coordinates $(\cosh(a), \sinh(a))$ on the hyperbola.
3. As the hyperbola is tangent[11] to the lines $y = x$ and $y = -x$ then as y approaches $\pm\infty$, $\tanh(a)$ approaches ± 1.
4. $\tanh : \mathbb{R} \to (-1, 1)$.

Sketches of the hyperbolic trigonometric functions are shown in figure 2.23. Related hyperbolic functions are given by

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

One can derive some properties of the hyperbolic functions from the symmetries of the hyperbola:

(i) Reflection in the x-axis, i.e. $(x, y) \to (x', y') = (x, -y)$.



$$\cosh(a) = x = x' = \cosh(-a)$$
$$\sinh(a) = y = -y' = -\sinh(-a).$$

[11] The equation defining the hyperbola can be rewritten as the pair of curves $y = \sqrt{x^2 - 1}$ and $y = -\sqrt{x^2 - 1}$ and for large x these equations approximate closely the lines $y = x$ and $y = -x$.

In summary:

$$\cosh(-a) = \cosh(a)$$
$$\sinh(-a) = -\sinh(a).$$

The circle has an infinite number of symmetries, but the hyperbola only obviously has two: reflection in the x-axis and reflection in the y-axis. Why does reflection of the hyperbola in the y-axis not generate an identity of the hyperbolic functions?

One can also define the hyperbolic functions algebraically as solutions to differential equations. Again, we remark that we have not yet formally defined the operation $\frac{d}{dx}$ nor applied it to functions that are not power series. However we can make use of it (again assuming only that $\frac{d}{dx}(ax^n) = anx^{(n-1)}$) to find power series.

Definition 2.6.22 (Cosh and Sinh, algebraic definition)**.** The solutions to the second order differential equation

$$\frac{d^2f}{dx^2} = f$$

are

$$f = \begin{cases} \cosh x & \text{if } f(0) = 1 \text{ and } \frac{df}{dx}(0) = 0 \\ \sinh x & \text{if } f(0) = 0 \text{ and } \frac{df}{dx}(0) = 1. \end{cases}$$

Note how similar this definition is to Definition 2.6.12 above.

Example 2.9. Use the definition of $\cosh(x)$ and $\sinh(x)$ as solutions to $\frac{d^2f}{dx^2} = f$ with the appropriate boundary conditions to find their power series.

Suppose there are some coefficients $a_0, a_1, \ldots$, for which $f(x) = \sum_{n=0}^{\infty} a_n x^n$ for all x. Then, using the differential equation for f,

$$\frac{d^2f}{dx^2} = \sum_{n=0}^{\infty} a_n n(n-1)x^{n-2} = \sum_{n=0}^{\infty} a_{n+2}(n+2)(n+1)x^n = \sum_{n=0}^{\infty} a_n x^n$$

for all x. Equating coefficients of each power of x gives $a_{n+2}(n+2)(n+1) = a_n$ or

$$a_{n+2} = \frac{a_n}{(n+2)(n+1)}.$$

Up to a sign this is the same as the relation we found when deriving the power series for $\cos(x)$ and $\sin(x)$, and via a similar analysis we may show that

$$f(x) = a_0 \sum_{n\,even} \frac{x^n}{n!} + a_1 \sum_{n\,odd} \frac{x^n}{n!}.$$

Applying the boundary conditions for $\sinh(x)$, where we have $f(0) = 0$ and $\frac{df}{dx}|_{x=0} = 1$ gives $a_0 = 0$ and $a_1 = 1$. Therefore

$$\sinh(x) = \sum_{n\,odd} \frac{x^n}{n!} = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \frac{x^7}{7!} + \dots. \tag{2.5}$$

Similarly the boundary conditions for $\cosh(x)$ give $a_0 = 1$ and $a_1 = 0$ hence

$$\cosh(x) = \sum_{n\,even} \frac{x^n}{n!} = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \frac{x^6}{6!} + \dots \tag{2.6}$$

where we have used $0! := 1$.

The Inverse Hyperbolic Functions

Definition 2.6.23 (Inverse hyperbolic functions)**.** The *inverse hyperbolic functions* are called arcsinh or $\sinh^{-1}$, arccosh or $\cosh^{-1}$ and arctanh or $\tanh^{-1}$ and are the inverse sinh, cosh and tanh functions respectively.

Recall the graphs of the hyperbolic functions which are shown in figure 2.23. These functions are not periodic, and only cosh "doubles back on itself" and is the only one of the three which is not injective and whose inverse function will have a restricted domain. There are some further pleasant qualities of the hyperbolic functions which will allow us to find simple analytic expressions for their inverse functions. Our construction will rest upon the power series for the exponential function[12] which we recall is:

$$e^x = \sum_{n=0}^{\infty} \frac{1}{n!} x^n = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \dots$$

and hence we also have

$$e^{-x} = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} x^n = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} - \frac{x^5}{5!} + \dots.$$

Compare these series with ones found earlier, in a similar analysis in equations (2.5) and (2.6), where we found that

$$\sinh(x) = \sum_{n\,odd} \frac{x^n}{n!} = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \frac{x^7}{7!} + \dots \qquad \text{and}$$

$$\cosh(x) = \sum_{n\,even} \frac{x^n}{n!} = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \frac{x^6}{6!} + \dots.$$

[12] We should note that we are building up a lot of debt here: we have assumed that the derivative of ax^n is anx^{n-1} and that the infinite sum for e^x converges - we will pay off our debts in due course.

Hence we notice that our power series for the two hyperbolic functions can be rewritten in terms of the exponential function, that is,

$$\sinh(x) = \frac{1}{2}(e^x - e^{-x}) \qquad \text{and} \tag{2.7}$$

$$\cosh(x) = \frac{1}{2}(e^x + e^{-x}). \tag{2.8}$$

Consequently we also have

$$\tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{e^{2x} - 1}{e^{2x} + 1}. \tag{2.9}$$

Using these expressions we can find simple analytic expressions for arcsinh , arccosh and arctanh .

Example 2.10. Find an analytic expression for the inverse sinh function.

We follow the algebraic recipe from previously in the notes. Let us write $y = \sinh x = \frac{1}{2}(e^x - e^{-x})$ and rearrange this expression to find x as a function of y. Putting all terms on the same side of the equation we find

$$e^x - e^{-x} - 2y = 0$$

and multiplying this by e^x gives

$$e^{2x} - 1 - 2ye^x = 0.$$

We may complete the square to find

$$(e^x - y)^2 - y^2 - 1 = 0.$$

Observe that as $y = \frac{1}{2}(e^x - e^{-x})$ and $e^{-x} > 0$ we have $y < \frac{1}{2}e^x$. In particular $e^x - y > 0$, and so rearranging the 'completed square' expression and taking the square root we conclude

$$e^x - y = \sqrt{1 + y^2}.$$

Therefore

$$x = \ln(y + \sqrt{1 + y^2}),$$

That is

$$\operatorname{arcsinh} y = \ln(y + \sqrt{1 + y^2}) \qquad \forall\, y \in \mathbb{R}.$$

A sketch of the graph of arcsinh is given in in figure 2.24, you should take some time to confirm the plot agrees with your expectations of the analytic function we have derived in the previous example.

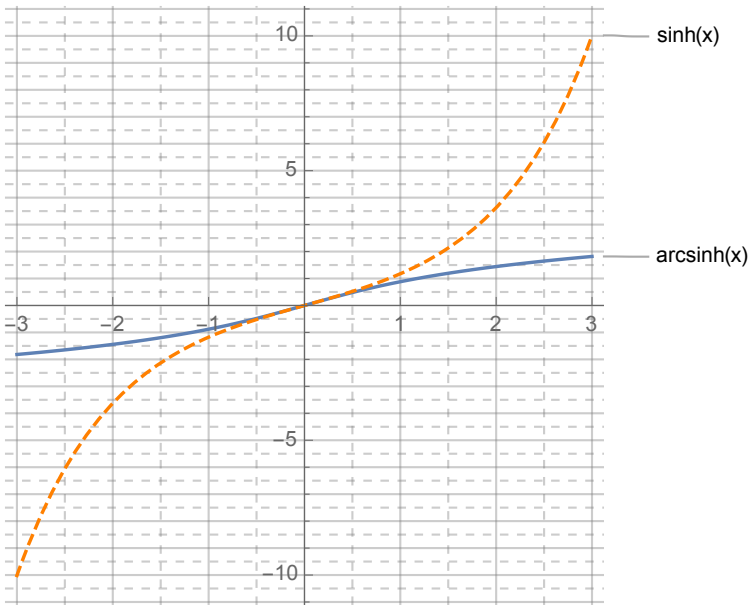


Figure 2.24: A sketch of $\operatorname{arcsinh}(x)$ and $\sinh(x)$.

Exercise 2.3. Repeat the procedure above to show that

$$\operatorname{arccosh} x = \ln(\sqrt{x^2 - 1} + x) \qquad \forall\, x \in [1, \infty).$$

Exercise 2.4. Repeat the procedure again to show that

$$\operatorname{arctanh} x = \frac{1}{2} \ln\left(\frac{1+x}{1-x}\right) \qquad \forall\, x \in (-1, 1).$$

2.6.9 Trigonometric and Hyperbolic Functions with Complex Variables

What more do we have to say about the trigonometric and hyperbolic functions? There are a number of identities which will prove convenient for integration later on, and to develop them we will consider the useful role played by the complex numbers in the description of the trigonometric functions. I will assume that you have already met the complex numbers $\mathbb{C}$ to some extent, not least because they are covered at the start of the Linear Algebra and Geometry I course this semester.

Euler's Formula

The complex numbers $\mathbb{C}$ introduces the imaginary number $i = \sqrt{-1}$ to the real numbers and consists of all numbers of the form $x + iy$ where $x, y \in \mathbb{R}$. The complex numbers $\mathbb{C}$ is an

extension of $\mathbb{R}$, and we may wonder if we can we extend the range of validity of the exponent a^n to include the case where $n \in \mathbb{C}$. For example just what could an expression such as a^i mean? What is the sense in saying it means we multiply a by itself i times? Not much, but this is just the kind of loose end we might have expected when we investigating the complex numbers, and we must hope that we can give such expressions some logical meaning.

Recall that we can write e^x, $\cosh x$ and $\sinh x$ as infinite sums, and even relate the formulae to find exponential expressions for the hyperbolic functions. Now that we have introduced the complex number i we can also re-express our power series for sine in equation (2.3) and for cosine in equation (2.4) in terms of the exponential in a similar way. Recalling our power series for e^x given in equation (2.2) we can now add in the imaginary number to find/define:

$$e^{ix} := \sum_{k=0}^{\infty} \frac{1}{k!}(ix)^k = 1 + ix - \frac{x^2}{2!} - i\frac{x^3}{3!} + \frac{x^4}{4!} + i\frac{x^5}{5!} - \frac{x^6}{6!} - i\frac{x^7}{7!} + \dots$$

and

$$e^{-ix} := \sum_{k=0}^{\infty} \frac{1}{k!}(-ix)^k = 1 - ix - \frac{x^2}{2!} + i\frac{x^3}{3!} + \frac{x^4}{4!} - i\frac{x^5}{5!} - \frac{x^6}{6!} + i\frac{x^7}{7!} + \dots$$

where we have used our definition of i, that $i^2 = -1$ to expand out and simplify the powers of i. By comparing with the power series proposed for $\sin(x)$ and $\cos(x)$ in equations (2.3) and (2.4), we deduce that

$$\sin x = \frac{1}{2i}(e^{ix} - e^{-ix}) \tag{2.10}$$

$$\cos x = \frac{1}{2}(e^{ix} + e^{-ix}). \tag{2.11}$$

We may now invert these equations to find an expression for e^{ix} in terms of $\cos x$ and $\sin x$. This gives us Euler's formula.

Definition 2.6.24. *Euler's formula* is

$$e^{i\theta} = \cos\theta + i\sin\theta$$

where $\theta \in \mathbb{R}$.

[13]This portrait is by Jakob Handmann. Euler is winking at you. No he isn't, in later life he had problems with his eye, although it is understood he developed a cataract in his left eye, so his closed right eye in this portrait is puzzling.

Leonhard Euler (1707-1783) was a Swiss mathematician who played a pivotal role in the development of analysis. Euler, pronounced "Oiler" rather than "U-ler"[14], was a wildly productive mathematician. He wrote over 800 papers(!), not just in maths but across most branches of existing science, and fathered 13 children. It was said of Euler that "he calculated just as men breathe, as eagles sustain themselves in the air". I do not believe that referring to mathematicians as 'great' is a helpful perspective on the subject. However, it is certainly the case that Euler was one of the most influential mathematicians to have lived. His work acted as a bridge between the inventive but confusing mathematical advances of the 17^{th} century and the formal rigor of the 19^{th} century.

Figure 2.25
Leonhard Euler[13]

Our development of Euler's formula has been based on a number of assumptions which we have yet to prove, in particular regarding power series and convergence. So, let us at this stage use some of our prior knowledge of calculus to check that the formula is, at least, sensible.

Forgoing again the idea that we have yet to discover the derivative, we note from the power series expansion that for any constant a,

$$\frac{d}{dx}e^{ax} = \sum_{n=1}^{\infty} \frac{na^n x^{n-1}}{n!} = \sum_{n=1}^{\infty} \frac{a^n x^{n-1}}{(n-1)!} = a\sum_{n=0}^{\infty} \frac{a^n x^n}{n!} = ae^{ax}.$$

Therefore, differentiating both sides of Euler's formula, we get

$$\frac{d}{d\theta}(e^{i\theta}) = ie^{i\theta} \qquad \text{while} \qquad \frac{d}{d\theta}(\cos\theta + i\sin\theta) = -\sin\theta + i\cos\theta = i(i\sin\theta + \cos\theta) = ie^{i\theta}.$$

Thus this remarkable formula of Euler's passes a first test for internal consistency with our current knowledge of derivatives.

Comment(s). (On Euler's Formula)

1. When $\theta = 0$ we have $e^0 = 1 = \cos(0) + i\sin(0)$ as required.
2. When $\theta = \pi/2$ we have $e^{i\pi/2} = \cos(\pi/2) + i\sin(\pi/2) = i$.
3. When $\theta = \pi$ we have $e^{i\pi} = -1$. This remarkable statement is known as Euler's identity and is frequently celebrated as it relates e, i, π and -1. See Figure 2.26 for many rapturous comments on this identity on Wikipedia (not all of which I agree with!)

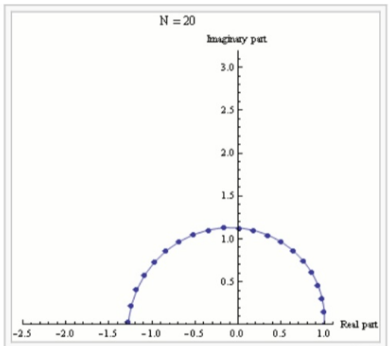
[14]See the 2014 film "The Imitation Game" for a good example of how not to pronounce Euler.

4. $e^{i\theta}$ has a period of 2π inherited from the trigonometric functions, and indeed $e^{i2\pi} = 1$. This will be of concern in defining its inverse function, the natural logarithm, over the complex numbers.

On "Euler's identity" [edit]

In mathematical analysis, **Euler's identity** is the equation "$e^{i\pi} + 1 = 0$."

- **One of the most frequently mentioned equations was Euler's equation, $e^{i\pi} + 1 = 0$. Respondents called it "the most profound mathematical statement ever written"; "uncanny and sublime"; "filled with cosmic beauty"; and "mind-blowing".** Another asked: **"What could be more mystical than an imaginary number interacting with real numbers to produce nothing?" The equation contains nine basic concepts of mathematics — once and only once — in a single expression. These are: e (the base of natural logarithms); the exponent operation; π; plus (or minus, depending on how you write it); multiplication; imaginary numbers; equals; one; and zero.**
  - Robert P. Crease, in "The greatest equations ever" at PhysicsWeb (October 2004)
- Like a Shakespearean sonnet that captures the very essence of love, or a painting that brings out the beauty of the human form that is far more than just skin deep, **Euler's equation reaches down into the very depths of existence.**
  - Keith Devlin, as quoted in *Dr. Euler's Fabulous Formula : Cures Many Mathematical Ills* (2006) ISBN 978-0691118222
- **Our jewel ... one of the most remarkable, almost astounding, formulas in all of mathematics.**
  - Richard Feynman on Euler's formula, of which Euler's identity is a special case, in *The Feynman Lectures on Physics*, Vol. I, p. 22
- There is a famous formula, perhaps the most compact and famous of all formulas — developed by Euler from a discovery of de Moivre: $e^{i\pi} + 1 = 0$. **It appeals equally to the mystic, the scientist, the philosopher, the mathematician.**
  - Edward Kasner and James R. Newman in *Mathematics and the Imagination* (1940)
- **Gentlemen, that is surely true, it is absolutely paradoxical; we cannot understand it, and we don't know what it means. But we have proved it, and therefore we know it must be the truth.**
  - Benjamin Peirce, as quoted in notes by W. E. Byerly, published in *Benjamin Peirce, 1809-1880 : Biographical Sketch and Bibliography* (1925) by R. C. Archibald; also in *Mathematics and the Imagination* (1940) by Edward Kasner and James Newman



Gentlemen, that is surely true, it is absolutely paradoxical; we cannot understand it, and we don't know what it means. But we have proved it, and therefore we know it must be the truth. ~ Benjamin Peirce

Figure 2.26: Quotes on the beauty of Euler's identity and formula.

Double-Angle Formulae

When trying to integrate complicated functions, it will be useful to us to develop the double-angle formulae for sine and cosine (i.e. to rewrite the expressions $\cos(\theta + \phi)$ and $\sin(\theta + \phi)$ in terms of $\cos\theta$, $\cos\phi$, $\sin\theta$ and $\sin\phi$). Due to Euler's formula we now have a very simple way to quickly write these formulae down.

Let us multiply two complex numbers, both of modulus 1:

$$e^{i\theta}e^{i\phi} = e^{i(\theta+\phi)}.$$

There appears to be nothing to this, but notice that if we use Euler's formula on the left we will have trigonometric functions of single angles, while on the right it will give us trigonometric functions of the sum of the angles:

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

If we equate the real part on each side and the imaginary part on each side of the above we arrive at the double-angle formulae for cosine and sine:

$$\cos(\theta+\phi)=\cos\theta\cos\phi-\sin\theta\sin\phi \tag{2.12}$$

$$\sin(\theta+\phi)=\cos\theta\sin\phi+\sin\theta\cos\phi. \tag{2.13}$$

In some sense, these formulae are the real mathematical truth lying behind Euler's formula (which packages this truth in a convenient way using the imaginary unit i).

Exercise 2.5. Use the double angle formulae for cosine and sine to show that

$$\tan(\theta+\phi)=\frac{\tan\theta+\tan\phi}{1-\tan\theta\tan\phi}.$$

Comment(s). (On the double angle formulae...)

1. When $\phi=\theta$ we have

$$\begin{aligned}\cos(2\theta) &= \cos^2\theta-\sin^2\theta\\ \sin(2\theta) &= 2\cos\theta\sin\theta\\ \tan(2\theta) &= \frac{2\tan\theta}{1-\tan^2\theta}\end{aligned}$$

2. Adding $\cos^2\theta+\sin^2\theta=1$ and $\cos^2\theta-\sin^2\theta=\cos(2\theta)$ gives

$$\cos^2\theta=\frac{1}{2}(1+\cos(2\theta))$$

while the difference gives:

$$\sin^2\theta=\frac{1}{2}(1-\cos(2\theta)).$$

These formula reduce the power of the trigonometric function at the expense of doubling the argument. This can be very useful when solving integrals.

3. The difference formula are quick to derive (using $\sin(-x) = -sin(x)$ and $\cos(-x) = \cos(x)$):

$$\begin{aligned}\cos(\theta-\phi) &= \cos\theta\cos\phi + \sin\theta\sin\phi\\ \sin(\theta-\phi) &= -\cos\theta\sin\phi + \sin\theta\cos\phi\\ \tan(\theta-\phi) &= \frac{\tan\theta - \tan\phi}{1+\tan\theta\tan\phi}\end{aligned}$$

4. One can convert sums of trigonometric functions into products and vice-versa:

$$\begin{aligned}\frac{1}{2}(\cos(\theta+\phi)+\cos(\theta-\phi)) &= \cos\theta\cos\phi\\ \frac{1}{2}(\cos(\theta-\phi)-\cos(\theta+\phi)) &= \sin\theta\sin\phi\\ \frac{1}{2}(\sin(\theta+\phi)+\sin(\theta-\phi)) &= \sin\theta\cos\phi\\ \frac{1}{2}(\sin(\theta+\phi)-\sin(\theta-\phi)) &= \sin\phi\cos\theta\end{aligned}$$

To give the details in one case

$$\begin{aligned}\frac{1}{2}(\cos(\theta+\phi)+\cos(\theta-\phi)) &= \frac{1}{2}((\cos\theta\cos\phi-\sin\theta\sin\phi)+(\cos\theta\cos(-\phi)-\sin\theta\sin(-\phi))\\ &= \frac{1}{2}\cos\theta\cos\phi-\sin\theta\sin\phi+\cos\theta\cos\phi+\sin\theta\sin\phi)\\ &= \cos\theta\cos\phi.\end{aligned}$$

5. By writing $\alpha = \theta+\phi$ and $\beta = \theta-\phi$, so that $\alpha+\beta = 2\theta$ and $\alpha-\beta = 2\phi$ one can use the product formulae above to find formulae relating half-angles to angles:

$$\begin{aligned}\cos(\alpha)+\cos(\beta) &= 2\cos(\frac{\alpha+\beta}{2})\cos(\frac{\alpha-\beta}{2})\\ \sin(\alpha)+\sin(\beta) &= 2\sin(\frac{\alpha+\beta}{2})\cos(\frac{\alpha-\beta}{2})\\ \cos(\alpha)-\cos(\beta) &= -2\sin(\frac{\alpha+\beta}{2})\sin(\frac{\alpha-\beta}{2})\\ \sin(\alpha)-\sin(\beta) &= 2\sin(\frac{\alpha-\beta}{2})\cos(\frac{\alpha+\beta}{2})\end{aligned}$$

Exercise 2.6. Write $\cos\theta$, $\sin\theta$ and $\tan\theta$ in terms of $t = \tan\frac{\theta}{2}$. Motivation: these formulae can come in extremely useful for solving otherwise difficult integrals, and are known as the '$t = \tan\frac{\theta}{2}$ trick'

Trigonometric to Hyperbolic Functions and Back Again

There are some simple relations between the trigonometric functions and the hyperbolic functions, which are evident in the similarity of the analytic expressions. By allowing the arguments of cos, sin, cosh and sinh to be extended to the complex numbers we can use the analytic expressions to find direct relations between the trigonometric functions with argument θ and the hyperbolic functions with argument $i\theta$:

$$\cos\theta = \frac{1}{2}(e^{i\theta} + e^{-i\theta}) = \cosh(i\theta)$$
$$\sin\theta = \frac{1}{2i}(e^{i\theta} - e^{-i\theta}) = -i\sinh(i\theta)$$
$$\tan\theta = \frac{-i\sinh(i\theta)}{\cosh(i\theta)} = -i\tanh(i\theta).$$

Notice that we can use our relations above to convert the fundamental trigonometric identity

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

into

$$\cosh^2(i\theta) - \sinh^2(i\theta) = 1.$$

Similarly we can also rewrite hyperbolic functions with argument θ as trigonometric functions of $i\theta$ with complex coefficients:

$$\cosh\theta = \cos(-i\theta) = \cos(i\theta)$$
$$\sinh\theta = i\sin(-i\theta) = -i\sin(i\theta)$$
$$\tanh\theta = \frac{-i\sin(i\theta)}{\cos(i\theta)} = -i\tan(i\theta).$$

2.6.10 Functions of Functions

How do we construct more functions? Given two functions $f(x)$ and $g(x)$ there are some obvious ways to combine them to form a new function. We may add the functions

$$f(x) + g(x)$$

or we might multiply them

$$f(x)g(x).$$

Furthermore since the output of both $f(x)$ and $g(x)$ are real numbers for any $x \in \mathbb{R}$ then we may even raise one function as a power of the other:

$$f(x)^{g(x)}.$$

This gives us a clue to consider a more abstract way to combine functions: functions of functions.

Definition 2.6.25. *Function composition* is the application of one function f to the output of another function g. Suppose $g : A \to B$ and $f : C \to D$. Then the function $f \circ g$ is defined to be the composition $f(g(x))$, for all x where this expression is valid: namely, $x \in A$ such that $g(x) \in C$.

By composing two functions we link them together, in the sense that the output of one function ($g(x)$ above) becomes the input of another function ($f(x)$ above).

Important warning: Note that composition is 'right-to-left' in the notation, in the sense that to compute the function $f \circ g$ one must apply the function g first, *then* the function f.

Function composition is very common. Consider classical mechanics, which was a major motivation for the development of the calculus. One might expect the speed v of a particle to be defined as a function of a position x of the particle. But x in turn may be defined as a function of time, t. In other words, we might view speed instead as a function composition $v(x(t))$, and analyse the speed as a function of time t. (Some people say that this means that the speed is an explicit function of position x and an implicit function of t.) Of course if we know $x(t)$ then we can write the composition $v(x(t))$ as an explicit function of t. For example if $v = ax + b$ and $x = t^2$ then we can also define $v = at^2 + b$.

2.7 A Zoo of Functions

How many functions $f : \mathbb{R} \to \mathbb{R}$ are there? Obviously there are an infinite number of linear functions (i.e. $f(x) = ax + b$ where $a, b \in \mathbb{R}$). So rather than try to count functions, can we classify them and write down the classes? This is a difficult question, which we will make some inroads to answering in this course.

The first problem in classifying functions is: how do we know whether two functions $f(x)$ and $g(x)$ are identical? The answer is that we check all the values of each function and see if they are the same for each identical input! For certain kinds of functions, there are some tricks; in particular, we have already seen that if we know that f and g are polynomials of degree at most n we can establish whether they are equal by computing the values of $f(0), \frac{df}{dx}|_{x=0}, \frac{d^2f}{dx^n}|_{x=0}, ..., \frac{d^nf}{dx^2}$, and the same for g, and checking if these two sequences of numbers are the same. More generally, if f and g were power series then we could check whether all derivatives were the same.

For two linear functions $f(x) = ax + b$ and $g(x) = cx + d$, say, we can evaluate them both

at $x = 0$ and find all their derivatives:

$$f(0) = b \qquad\qquad g(0) = d$$
$$\left.\frac{df}{dx}\right|_{x=0} = a \qquad\qquad \left.\frac{dg}{dx}\right|_{x=0} = c.$$

So if $b = d$ and $a = c$ then the two linear functions are the same. Since we have not introduced the derivative yet we will leave this idea on the back-burner. However, note for now that if we cannot find the derivative of a function (and there are indeed functions for which the derivative does not exist) then we will not be able to invoke this check of whether two functions are the same.

Rather than compare all values of two functions (or even all derivatives), it can be tempting instead to manipulate one algebraic expression for a function $f(x)$ into another for $g(x)$. However, there are some traps with this, of which we should be aware.

Example 2.11. Is the function $f(x) = \frac{x^2}{x}$ the same function as $g(x) = x$?

We are tempted to simplify $f(x)$ as follows

$$f(x) = \frac{x^2}{x} = x = g(x) \tag{2.14}$$

and argue that the two functions are identical. But we should be more careful. According to the definition, what is the value of $f(0)$? Well, we might write $f(0) = \frac{0}{0}$, but such expressions are *undefined*. On the otherhand, according to the definition we have $g(0) = 0$. So the two functions are *not identical*. The message is that one must be careful when simplifying algebraic expressions where one might be implicitly dividing by zero.

Aside: Later on in your mathematical education, you may learn so-called 'Complex Analysis'. This is the theory of differentiation/integration/limits etc. but for functions $f : \mathbb{C} \to \mathbb{C}$ rather than for functions $f : \mathbb{R} \to \mathbb{R}$. In this theory you may learn that values of x such as $x = 0$ in the above example form something called a 'removable singularity'. There is a systematic way of dealing with such things, but we will not cover this idea in this course.

3. The Limit

In which we encounter the most important idea in the course: the limit. We develop a definition, gain experience in computing limits, and use limits to define continuous functions.

This will be covered during weeks 4 and 5 of the lectures.

Our first main goal is to present a formal definition of a limit. However, when first presented with such a definition it can seem technical, off-putting, and unnecessary. To soften the impact, we will build up to the statement with some general motivation.

3.1 Motivation

Here is the sort of question that mathematicians loved to ask (before about 1900 and the time of Hilbert): what *is* $\frac{0}{0}$? This is in the general vein of questions like 'what *is* $\log i$?' or 'what *is* a continuous function?', and all three questions lead to rich mathematical investigations, the like of which their posers never imagined. The style of mathematical discovery is slightly different now. On the whole, mathematical researchers feel less like they are discovering immortal truths of the universe and more like they are exploring the bounds of a certain logical framework. This was Hilbert's doing, for better or worse. Nevertheless, for pedagogical purposes, it can be useful to think like a mathematician of yesteryear.

We intuitively feel that $\frac{0}{0}$ can't possibly be a number, because we can see the evident (and ill-defined) division by zero. Yet, given that the numerator is also zero, we may be tempted to take any of the following stances in opposition to the idea that it is not a number:

(a.) it is zero, as any number multiplied by zero gives zero;

(b.) it is one, as any number of the form $\varepsilon/\varepsilon = 1$;

(c.) it tends to either $\pm\infty$, as any number divided by a small, positive number blows up towards plus infinity, while division by a small, negative number blows up towards negative infinity. Yet, since zero is equally close to a small positive and a small negative number, even with this stance we remain confused over which number $\frac{1}{\epsilon}$ approaches as $\epsilon \to 0$: is it plus or minus infinity?

In light of such confusion, we declare that $\frac{0}{0}$ is undefined.

By way of example, consider the function

$$f(x) = \frac{x^2 - 9}{x - 3}$$

and resist the strong inclination to algebraically simplify the function to $(x + 3)$ using the difference of two squares as follows

$$\frac{x^2 - 9}{x - 3} = \frac{(x + 3)(x - 3)}{x - 3} = x + 3 = g(x).$$

Why should we not do this? Well, it is the same issue as with the final example of the previous chapter. The function $f(x)$ is almost but not quite the same function as $g(x) = x + 3$: the functions are the same apart from at the point $x = 3$. To rewrite the line above properly, we should add a condition so that it now reads

$$f(x) = \frac{(x + 3)(x - 3)}{x - 3} = x + 3 = g(x) \qquad \text{for } x \neq 3,$$

otherwise we would be implicitly dividing by zero. The function $f(x)$ is identical to $g(x)$ except at the point $x = 3$ where it involves a division by zero and so becomes undefined. If one plotted all the points of $f(x)$ for $x \in \mathbb{R}$ it would be identical to the straight line $g(x) = x + 3$ apart from missing a single point at $x = 3$. If we plugged in some numbers to see what happens to $f(x)$ as x gets close to 3 we could build up a table of data to help us understand what happens before and after $f(3)$ and compare it with $g(x)$.

x	$f(x) = \frac{x^2-9}{x-3}$
2.9	5.9
2.99	5.99
2.999	5.999
2.9999	5.9999
3	Not defined.
3.0001	6.0001
3.001	6.001
3.01	6.01
3.1	6.1

x	$g(x) = x + 3$
2.9	5.9
2.99	5.99
2.999	5.999
2.9999	5.9999
3	6
3.0001	6.0001
3.001	6.001
3.01	6.01
3.1	6.1

In the table above we have evaluated $f(x)$ and $g(x)$ at and near $x = 3$. We could have picked a value for x very close to 3, but not quite 3, e.g. $x = 3 + \delta$ such that $\delta \neq 0$. Here the Greek letter *delta*, denoted δ, stands for a notion of 'distance'. We would have found that $f(3 + \delta) = 6 + \delta = g(3 + \delta)$. Furthermore, as δ approaches 0, we intuitively see that $f(3 + \delta)$ approaches 6. In words, as x approaches 3, $f(x)$ approaches 6. In symbols, as $x \to 3$, $f(x) \to 6$.

As mathematicians we are interested in fixing problems. A resolution to the problem for $f(x)$ at $x = 3$ that we are tempted towards is to discard the point $x = 3$ and to treat $f(x)$ as we did algebraically. We would be able to work with the function $f(x)$ to an arbitrary degree of accuracy as close to $x = 3$ as we would care to choose. So long as we do not have to do any computations at exactly $x = 3$ then we could replace $f(x)$ in all our calculations with the better behaved function $g(x) = x + 3$.

This reasoning is potentially very disturbing. Of course it would be nice to replace a function which is an ill-defined function on $\mathbb{R}$ with one which is well-defined everywhere by adding in a point, but how do we know when we can do this? And what about other functions which have ill-defined, singular points[1]? We would like to know how to go about finding the well-behaved function g that we could work with instead.

Let us think about this last problem. If we had been presented with $f(x)$ how would we have known that we could add in a point and it would become $g(x)$? In the case we considered the algebra was simple to do and we could immediately find a way to get rid of the singularity in $f(x)$ and find $g(x)$, but what about more complicated examples? Suppose instead we considered

[1]Points which become ill-defined are called singularities or singular points.

the function

$$f(x) = \frac{\sin x}{x} \qquad (x \neq 0),$$

whose graph looks like an oscillating $\sin x$ function but whose amplitude varies as $\frac{1}{x}$ (see Figure 3.1). With the aid of the computer-generated graph, we can guess that approaching the singular

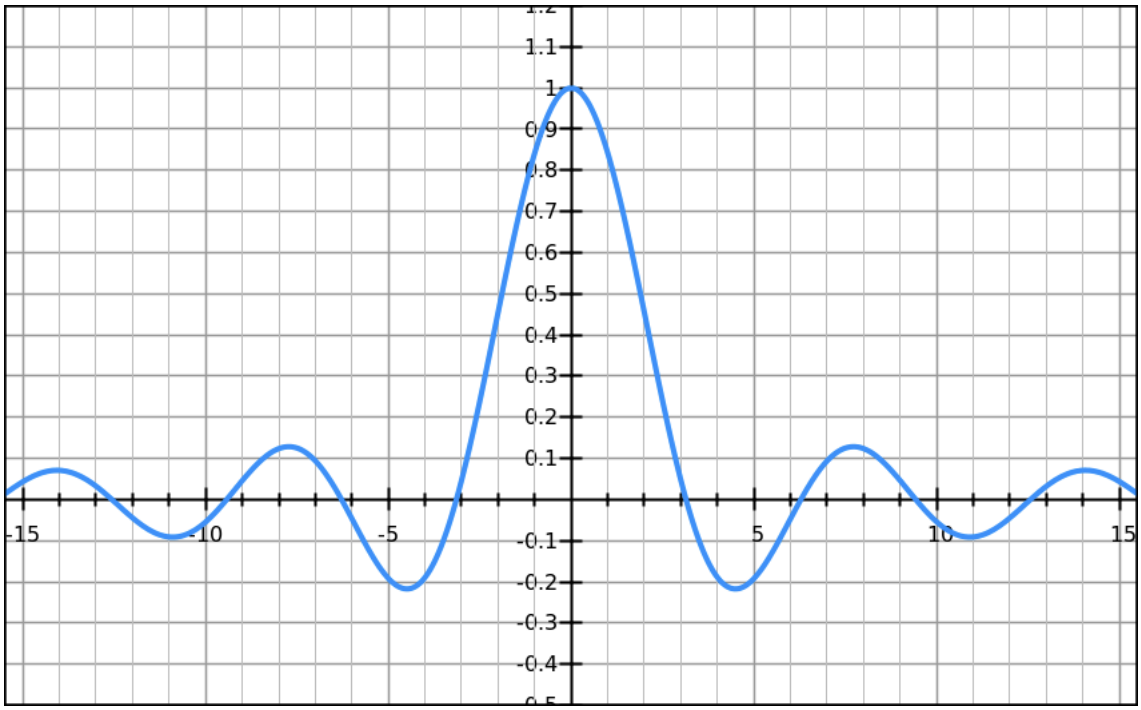


Figure 3.1: The graph of $f(x) = \frac{\sin x}{x}$.

point (as $x \to 0$) then $\frac{\sin x}{x} \to 1$. With this example we guess that the well-behaved function is found by adding the point 1 to the function to obtain:

$$g(x) \equiv \begin{cases} \frac{\sin x}{x} & x \in \mathbb{R} \setminus 0 \\ 1 & x = 0 \end{cases}$$

This is all very well, but we have been entirely loose with our method and our mathematical language. Let's try and rectify this[2].

3.2 The Existence of the Limit.

In terms of the language we have been speaking of "well-behaved" functions but we have not spelled out exactly what we mean by this. Suppose a function $f(x)$ is singular at a single point

[2]We will return later to $\frac{\sin x}{x}$ to see if our intuition from the graph was correct, but first we need to develop our mathematical technology.

$x = x_0$. We have seen an example where we added a single point (to the definition of the function $f(x)$); we could determine the value of the function at x_0 because it was the value that the function was approaching as its argument approached arbitrarily close to the singular point x_0. The value that the function approaches is called the *limit* of $f(x)$ as x approaches x_0 and we denote this as

$$\lim_{x \to x_0} f(x).$$

We will develop this terminology to give its full definition shortly. However before doing so let's address any doubts we have about whether what we are proposing has any mathematical value. We may think it is sufficient to be able to sketch a graph and read off its tendencies as it approaches any singular points, but can we trust our intuition about the output data from our computers and calculators to work out the value of a limit? Let us look at a slightly different example of an ill-defined function which will trouble us on this point. Consider now

$$f(x) = \sin\left(\frac{\pi}{x}\right) \qquad (x \neq 0),$$

whose argument π/x is singular at $x = 0$. Near the singularity at $x = 0$ we may compute the value of the function to try to deduce the limit of the function as x approaches 0. We show some data in the table below:

x	$f(x) = \sin(\frac{\pi}{x})$
0.1	0.0000000000020665
0.01	0.0000000000206930
0.001	0.0000000002065886
0.0001	0.0000000020658863
0	Not defined.
-0.0001	-0.0000000020658863
-0.001	-0.0000000002065886
-0.01	-0.0000000000206930
-0.1	-0.0000000000020665

For our computations shown in the table we have approximated $\pi \approx 3.14159265359$ and we realise that if our calculator had only shown eight decimal places we would have thought this function to be zero in the domain above. This would have led us to conclude that as x approaches zero $f(x)$ also approaches zero. But we'd have been wrong. With a little thought (and without a calculator) we realise that the values of x we chose to evaluate $f(x)$ are special points for the sine function, as for $x = \pm\left(\frac{1}{10}\right)^n$ for $n = 1, 2, 3, 4$ we see that we were evaluating $\sin(\frac{\pi}{x}) = \sin(\frac{\pm\pi}{10^{-n}}) = \sin(\pm 10^n \pi) = 0$. The only reason we didn't get exactly zero was because we were using a numerical approximation to the value of π.

If we had chosen different values of x we would have found wildly different answers as we approached $x = 0$:

x	$f(x) = \sin(\frac{\pi}{x})$
0.7	−0.9749279121818890
0.07	0.7818314824698670
0.007	0.4338837390909680
0.0007	0.9749279121161520
0	Not defined.
−0.0007	−0.9749279121161520
−0.007	−0.4338837390909680
−0.07	−0.7818314824698670
−0.7	0.9749279121818890

Again we have used the approximation $\pi \approx 3.14159265359$ to evaluate the function in the table. Notice how the output of the function now varies wildly between -1 and 1. In fact, had we chosen values of x of the form $x = \frac{1}{2n+\frac{1}{2}}$ (for $n \in \mathbb{Z}$), for all such values of x we would have found that $f(x) = \sin(\pi(2n + \frac{1}{2})) = 1$. Similarly, if we had chosen values of x of the form $x = \frac{1}{2n+\frac{3}{2}}$ (for $n \in \mathbb{Z}$), for all such values of x we would have found that $f(x) = \sin(\pi(2n + \frac{3}{2})) = -1$. So there is no fixed number L for which $f(x)$ approaches L as x approaches 0.

This behaviour of the function is (slightly) better appreciated by considering the graph of the function which is shown in Figure 3.2. Around $x = 0$ the graph becomes a solid blob. We might think this is a consequence of the thick line of the graph, but if we zoom in, the dense lines of the graph only persist (see Figure 3.3). Again, we are left to conclude that neither our calculator nor our graphing software is helping us to find the limit of $\sin\left(\frac{\pi}{x}\right)$ as x approaches

zero.

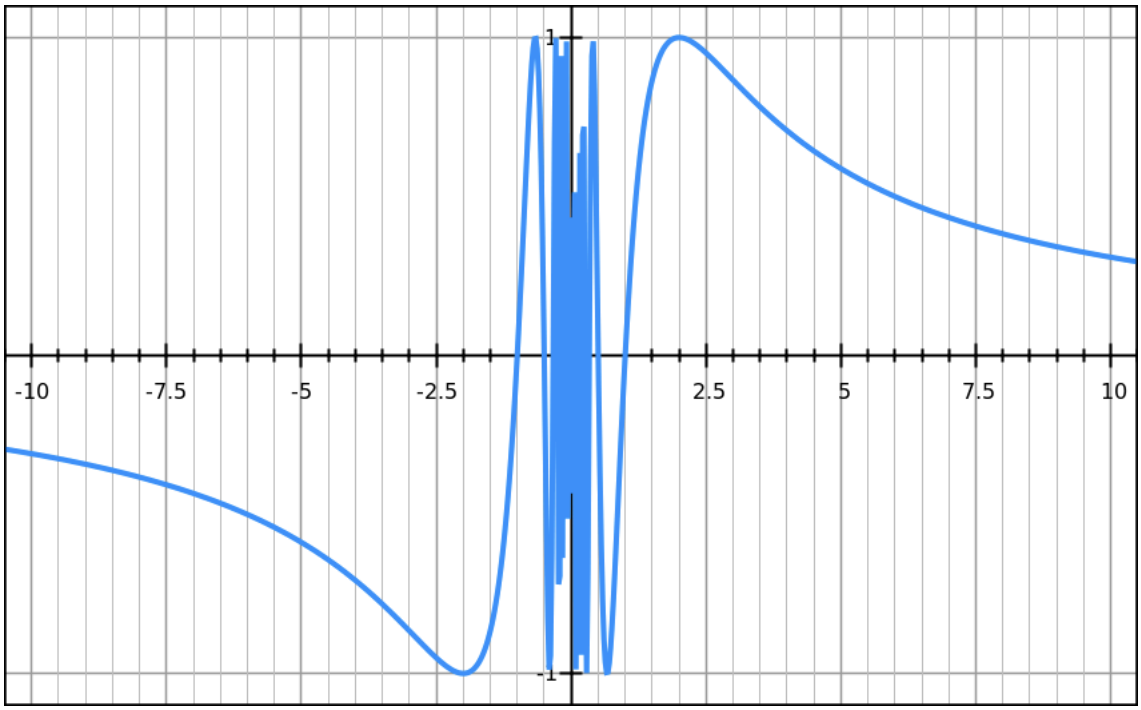


Figure 3.2: The graph of $f(x) = \sin\left(\frac{\pi}{x}\right)$.

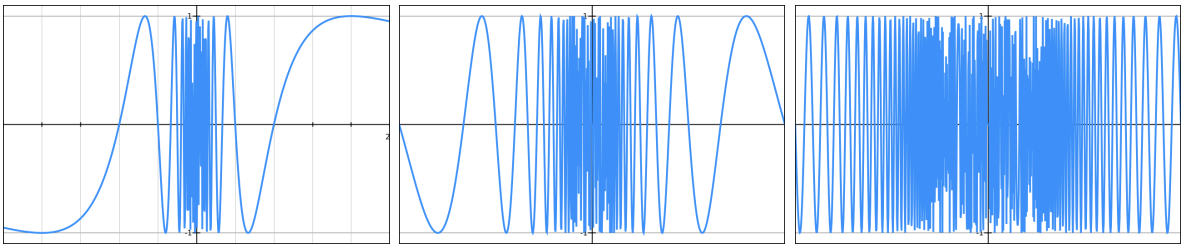


Figure 3.3: Zooming in on the graph of $f(x) = \sin\left(\frac{\pi}{x}\right)$ around $x = 0$: the blob persists!

Our rough definition of $\lim_{x \to x_0}[f(x)]$ was 'the value $f(x)$ approaches as its argument approaches arbitrarily close to x_0'. In making this definition we imagined that functions do always have such limit points: we were picturing a graph given by a continuous curve with one point missing. But the graph in Figure 3.2 shows that functions can be much more complicated than this, and maybe the limit doesn't always exist. But how can we tell if no such limit exists? We must bear in mind that we will not be able to learn much from the graph alone (as the picture can get even more complicated than Figure 3.2 in general): we need to formulate an algebraic definition that nonetheless respects the intuitive sense of the limit that we received

from dealing with simple graphs of continuous curves.

Having built up our motivation, we may finally give a formal definition.

Definition 3.2.1 (Limit of a function)**.** Let $L \in \mathbb{R}$ and $A \subset \mathbb{R}$. Let $f : A \to \mathbb{R}$ be a function, and suppose that $x_0 \in \mathbb{R}$. We say that *$f(x)$ tends to L as x tends to x_0* (or equivalently we say that *$f(x)$ approaches L as x approaches x_0*, or equivalently we say that *L is the limit of $f(x)$ as x tends to x_0*) if the following two properties both hold:

1. there is some $\delta > 0$ such that the set $(x_0 - \delta, x_0 + \delta) \setminus \{x_0\}$ is a subset of A (i.e. there is some interval around the point x_0 entirely contained within the domain of f, except perhaps for the point x_0 itself);

2. for all $\varepsilon > 0$, there exists some $\delta > 0$ for which $|x - x_0| \in (0, \delta)$ implies $|f(x) - L| < \varepsilon$.

Under these circumstances, in symbols we may write "$\lim_{x \to x_0} f(x) = L$", or "$f(x) \to L$ as $x \to x_0$".

Definition 3.2.2 (Existence of a limit)**.** If there is some L for which $\lim_{x \to x_0} f(x) = L$, we say that *the limit of $f(x)$ as x tends to x_0 exists*. We may write *$\lim_{x \to x_0} f(x)$ exists*.

It was Cauchy's analysis textbooks of the 1820s and 1830s that first introduced these definitions to a wide audience. Observe how we carefully avoided making any assumptions about the value of $f(x_0)$ itself, as it could be that x_0 is a singular point for f.

You may get a bit of 'snow-blindness' from all the Greek letters (here the Greek letter epsilon ε is meant to stand for 'error'). However, these definitions are nothing more or less than a formal statement of the intuitive notion we have discussed for several pages now: namely, $f(x)$ tends to a limit L as x approaches x_0 if $f(x)$ gets closer and closer to L as x approaches x_0. In Figure 3.4 we illustrate the definition with a simple function, whose curve is given in blue.

Lets try to apply our definition to the function

$$f(x) = \frac{x^2 - 9}{x - 3} \qquad \text{if } x \neq 3$$

from earlier. We convinced ourselves by some computations that $\lim_{x \to 3} f(x)$ should be equal to 6. Now let us argue more directly from the definition.

The domain of f equals $\mathbb{R} \setminus \{3\}$. In particular the open interval $(2, 4)$ is a subset of the domain, except for the central point 3: hence the first property of the limit definition is satisfied at $x_0 = 3$. Secondly, we have to show that for every $\varepsilon > 0$ there exists a $\delta > 0$ such that $|x - 3| \in (0, \delta)$ implies $|f(x) - 6| < \varepsilon$. Well, for any $\delta > 0$, when $|x - 3| \in (0, \delta)$ we have

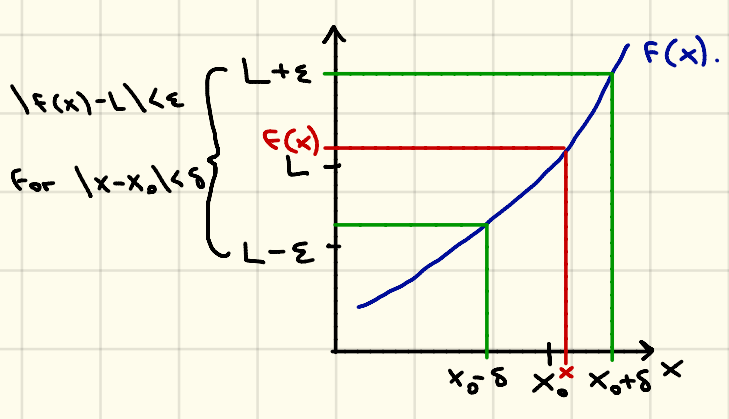


Figure 3.4: For any given $\varepsilon > 0$, there exists some interval $(x_0 - \delta, x_0 + \delta)$ for which $f(x)$ lies in the range $(L - \epsilon, L + \epsilon)$ when x lies in the range $(x_0 - \delta, x_0 + \delta)$.

$f(x) = x + 3$ and so $|f(x) - 6| = |x - 3| < \delta$. Therefore, fixing $\varepsilon > 0$ and picking $\delta = \varepsilon$, we conclude that

$$|x - 3| \in (0, \delta) \text{ implies } |f(x) - 6| < \varepsilon$$

as required.

This example was particularly simple, as we could in fact pick the value of δ by $\delta = \varepsilon$. Sometimes you may need to pick a more complicated value of δ, and you will see many formal arguments such as this in the Sequences and Series course. They are less of a feature of this module, but I felt it was important to show you at least one such example!

Returning our thoughts to the function $f(x) = \sin\left(\frac{\pi}{x}\right)$, we may realise that we can always find any output value in the range $[-1, 1]$, no matter how close x is to zero. We argued previously that, intuitively, this means that f does not have a limit as $x \to 0$. However, let's try to convince ourselves of this by using our mathematical definition of the limit.

We use a proof by contradiction. Suppose that there were some value $L \in \mathbb{R}$ for which $\lim_{x\to 0} f(x) = L$. Our plan is to show that there exists some $\varepsilon > 0$ for which there does not exists $\delta > 0$ for which

$$|\sin\left(\frac{\pi}{x}\right) - L| < \varepsilon$$

when $0 < |x| < \delta$, i.e the second property of the definition of a limit does not hold. This would contradict the assumption that $\lim_{x\to 0} f(x) = L$. In fact, we will show this contradiction with the very explicit value $\varepsilon = \frac{1}{2}$.

Let us enact the plan. Since we are assuming that $\lim_{x\to 0} f(x) = L$, there exists some $\delta > 0$ for which

$$|\sin\left(\frac{\pi}{x}\right) - L| < \frac{1}{2}$$

when $0 < |x| < \delta$. There are two cases.

- If $L \leqslant 0$, observe that if we define x via the equation
$$\frac{\pi}{x} = \frac{\pi}{2} + 2n\pi$$
(with $n \in \mathbb{N}$ sufficiently large) then $|x| < \delta$ but $\sin\left(\frac{\pi}{x}\right) = 1$, implying that $|\sin\left(\frac{\pi}{x}\right) - L| \geqslant 1 > \frac{1}{2} = \varepsilon$. This contradicts our previous assumptions.
- If $L \geqslant 0$, observe that if we define x via the equation
$$\frac{\pi}{x} = \frac{3\pi}{2} + 2n\pi$$
(with $n \in \mathbb{N}$ sufficiently large) then $|x| < \delta$ but $\sin\left(\frac{\pi}{x}\right) = -1$, implying that $|\sin\left(\frac{\pi}{x}\right) - L| \geqslant 1 > \frac{1}{2} = \varepsilon$. This contradicts our previous assumptions.

So a contradiction has been obtained in all cases, proving that the original assumption (that $\lim_{x\to 0} f(x) = L$) must be false. Since L was arbitrary, we conclude that $\lim_{x\to 0} f(x)$ does not exist.

Let it be stressed that we do not have to repeat the above proof for any other values of ε, as the key inequality has already been violated for a single value $\varepsilon = \frac{1}{2}$; if the limit exists, the key inequality must be true for all $\varepsilon > 0$.

Thus far we have obscured an important point. In the notation $\lim_{x\to x_0} f(x)$ we have required that x be allowed to approach x_0 from above or from below (as the key inequality has to hold for all $x \in (x_0 - \delta, x_0 + \delta)$, save for the value $x = x_0$ itself). However, there are certain settings when it is convenient to use milder assumptions, and only consider whether the limit exists as x approaches x_0 from one direction: either from above or from below. Because of how this looks on the number line, we talk about x approaching x_0 'from the left' (i.e. from below) or x approaching x_0 'from the right' (i.e. from above). This process might generate two limits (the 'left limit' and the 'right limit'), and these limits may be different. Here are the formal definitions:

Definition 3.2.3 (Left limit)**.** The *left limit* (or *limit from below*) of $f(x)$ is denoted $\lim_{x\to x_0^-}[f(x)]$ or $\lim_{x\uparrow x_0}[f(x)]$, if it exists. Formally, the limit $\lim_{x\to x_0^-}[f(x)] = L$ if for all $\epsilon > 0$ there exists $\delta > 0$ such that $|f(x) - L| < \epsilon$ for $x \in (x_0 - \delta, x_0)$.

Definition 3.2.4 (Right limit)**.** The *right limit* (or *limit from above*) of $f(x)$ is denoted $\lim_{x\to x_0^+}[f(x)]$ or $\lim_{x\downarrow x_0}[f(x)]$, if it exists. Formally, the limit $\lim_{x\to x_0^+}[f(x)] = L$ if for all $\epsilon > 0$ there exists $\delta > 0$ such that $|f(x) - L| < \epsilon$ for $x \in (x_0, x_0 + \delta)$.

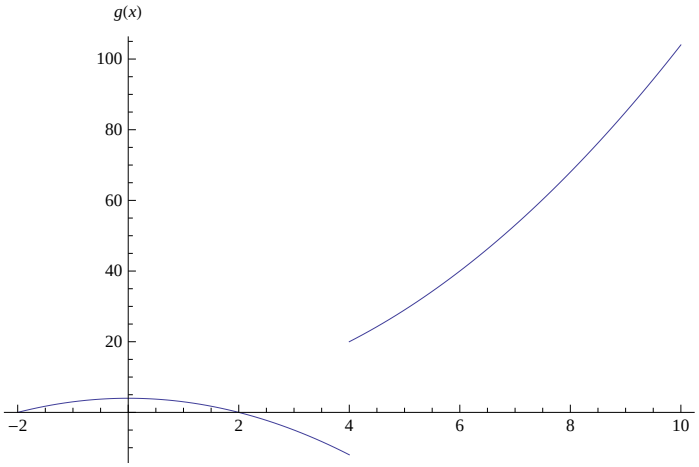
We'll consider two examples.

Example 3.1. Show that the limits from above and below as x approaches 4 for the function

$$g(x) = \begin{cases} x^2 + 4 & x > 4 \\ -x^2 + 4 & x \leq 4 \end{cases}$$

are different.

We can sketch the graph around $x = 4$.



Notice the jump that occurs at $x = 4$. Consequently the limit will differ if we approach $x = 4$ from above or below. Specifically we have

$$\lim_{x \to 4^-} (g(x)) = -12 \qquad \text{and} \qquad \lim_{x \to 4^+} (g(x)) = 20.$$

This is an example of a function which is not continuous, and we will soon use the limit to give a rigorous definition to the continuity of a function.

The second example involves a very common function, $f(x) = 1/x$. To talk about it, we need a slight generalisation of limits.

Definition 3.2.5 (Tending to infinity)**.** Let $A \subset \mathbb{R}$. Let $f : A \to \mathbb{R}$ be a function, and suppose that $x_0 \in \mathbb{R}$. We say that *$f(x)$ tends to infinity as x tends to x_0* (or equivalently we say that *$f(x)$ approaches infinity as x approaches x_0*) if the following two properties both hold:

1. there is some $\delta > 0$ such that the set $(x_0 - \delta, x_0 + \delta) \setminus \{x_0\}$ is a subset of A (i.e. there is some interval around the point x_0 entirely contained within the domain of f, except perhaps for the point x_0 itself);

2. for all $K > 0$, there exists some $\delta > 0$ for which $|x - x_0| \in (0, \delta)$ implies $f(x) \geqslant K$.

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

In other words, $f(x)$ gets larger and larger the closer x gets to x_0.

The property $\lim_{x \to x_0} f(x) = -\infty$ is defined similarly (with everything the same except the final part being $f(x) \leqslant -K$). The definition of left and right limits follows the same patterns as above.

Example 3.2. Show that the limits from above and below as x approaches zero for the function $f(x) = \frac{1}{x}$ are different.

Of course, 1 divided by any small number is large in absolute value. However, if the small number is negative then its reciprocal will be a large negative number, but if the small number is positive its reciprocal will be a large positive number. This reasoning implies that

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

whereas

$$\lim_{x \to 0^+} (\frac{1}{x}) = \infty.$$

Comment(s). (On the existence of a limit)

1. The definition of the limit as stated above implicitly means that the limit from above is equal to the limit from below, i.e. if $f(x)$ tends to a limit as $x \to x_0$ then

$$\lim_{x \to x_0^-} [f(x)] = \lim_{x \to x_0^+} [f(x)] = \lim_{x \to x_0} [f(x)].$$

The converse is also true, namely that if the left and right limits of f exist at the point x_0 and moreover are equal, i.e.

$$\lim_{x \to x_0^-} [f(x)] = \lim_{x \to x_0^+} [f(x)],$$

then $\lim_{x \to x_0} f(x)$ exists and

$$\lim_{x \to x_0^-} [f(x)] = \lim_{x \to x_0^+} [f(x)] = \lim_{x \to x_0} [f(x)].$$

2. Our definitions are the so called $\varepsilon - \delta$ definitions of the limit. This definition was in fact first given by Bolzano in 1817, but is most closely associated with Augustin-Louis Cauchy[3] (1789-1857).

[3]Even though it was only expressed in this modern form by Karl Weierstrass.

We will now take a short digression from our general discussion of limits to focus on one important use of the limit: to define continuous functions. We introduced this idea in an intuitive way when discussing the exponential function a^x earlier in the course, but now we have the language to talk about the concept rigorously.

3.2.1 Continuous Functions.

Definition 3.2.6 (Continuous at a point)**.** Let $A \subset \mathbb{R}$ and $x_0 \in A$. A function $f : A \to$ is *continuous at the point* $x = x_0$ if $\lim_{x \to x_0}[f(x)]$ exists and is equal to $f(x_0)$.

Definition 3.2.7 (Continuous function)**.** A function $f(x)$ is *continuous* if it is continuous at all points x_0 in its domain.

Comment(s). (On continuous functions...)

1. Informally speaking, continuity implies that a small change in the argument of a function always gives a small change in its output. For example, the function $g(x)$ used in Example 3.1 is not continuous, as it is not continuous at the point $x_0 = 4$. A small change in the input (for $x = 4$ to $x = 4 + \delta$) leads to a large change in the output. Try and prove this formally for yourself, using the definitions above.

2. Colloquially it is often said that continuous functions are those for which the graph can be drawn without lifting the pen from the paper. However, we have already seen how sometimes graphs are very hard to draw, e.g. the graph of $\sin\left(\frac{\pi}{x}\right)$. The ***key point*** about this definition of continuity is that it is algebraic: it does not rely on us having to draw a picture of the function.

3. Some functions which we might not think of as being so well-behaved are continuous, for example $f(x) = |x|$. See the following example.

Example 3.3. The function $f(x) = |x|$ is a continuous function.

The graph of the function is shown in figure 3.5.

It should be obvious that it is continuous at all points $x_0 \neq 0$ (though you are welcome to write the argument down formally for practice). The point $x_0 = 0$ is our main concern. If we take the limit from above we have

$$\lim_{x \to 0^+}[|x|] = \lim_{x \to 0^+}[x] = 0$$

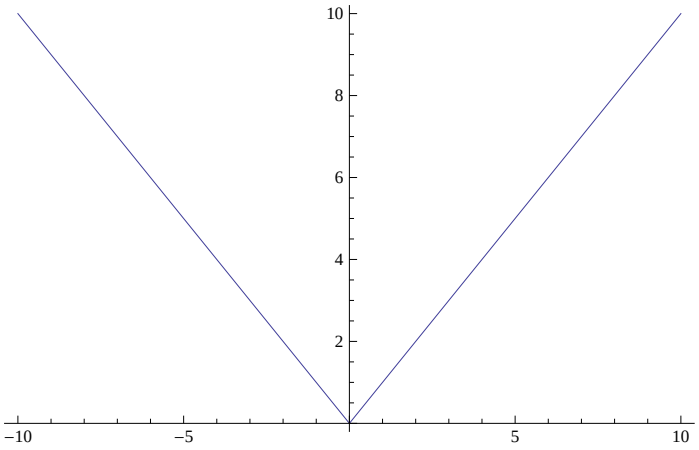


Figure 3.5: The graph of $f(x) = |x|$.

and from below we have

$$\lim_{x\to 0^-}[|x|] = \lim_{x\to 0^-}[-x] = 0.$$

Hence the limits from above and from below are both equal to zero and hence the limit $\lim_{x\to 0}[|x|]$ exists and is also equal to zero. Since this is also the value of the function at zero (i.e. $|0| = 0$) then we have

$$\lim_{x\to 0}[|x|] = |0|.$$

Hence the modulus function is a continuous function, despite its nasty-looking right-angle at $x = 0$.

Exercise 3.1. Show that the function $f : \mathbb{R} \to \mathbb{R}$

$$f(x) = \begin{cases} e^x & x \leq 2 \\ e^{4-x} & x > 2 \end{cases}$$

is continuous. The graph of the function is shown in Figure 3.6.

The Intermediate Value Theorem

I have argued that the key property of the formal definition of continuity is that it is an algebraic definition. However, it would be concerning if this algebraic definition did not recover some intuitive geometric properties of continuous curves, in particular this ephemeral 'draw without taking your pen off the paper' idea. Fortunately this algebraic definition does recover many such properties. Of course these must be proved. The first such property is called the *intermediate value theorem*, and it is the algebraic version of 'not taking your pen off the paper'.

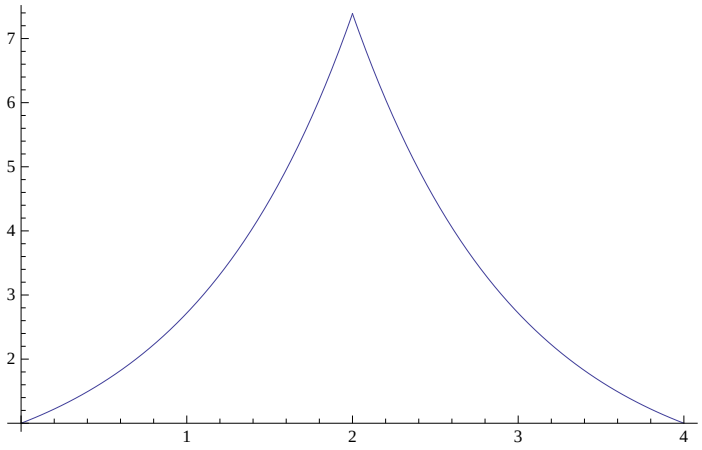


Figure 3.6: The function $f(x)$ defined in Exercise 3.1 about $x = 2$.

In order to state it, we need to make a slightly generalisation of our notion of what it means for a function to be continuous. Suppose $f : [a, b] \to \mathbb{R}$ is a function on a closed interval. Accord to Definition 3.2.6, f cannot be continuous at a because $\lim_{x\to a} f(x)$ doesn't exist. But $\lim_{x\to a} f(x)$ does not exist for a rather silly reason, namely that there is no $\delta > 0$ for which $(a - \delta, a + \delta) \setminus \{a\}$ lies in the domain of the function. It could well be that $\lim_{x\to a^+} f(x)$ exists, and perhaps even $\lim_{x\to a^+} f(x) = f(a)$. It would be a shame if our theory had no way of considering such 'one-sided continuity'. Therefore, in this special case when f is defined on a closed interval, we give a supplement to Definition 3.2.6 and Definition 3.2.7

Definition 3.2.8 (Continuity on a closed interval)**.** Suppose $[a, b]$ is a closed interval and $f : [a, b] \to \mathbb{R}$. We say that *f is continuous (on $[a, b]$)* if the following three properties hold:

- f is continuous at all points $x_0 \in (a, b)$ (according to Definition 3.2.6);
- $f(a) = \lim_{x\to a^+} f(x)$;
- $f(b) = \lim_{x\to b^-} f(x)$.

For example, the function $\arcsin : [-1, 1] \longrightarrow \mathbb{R}$ is continuous (on $[-1, 1]$).

We can now state:

Theorem 3.1 (The Intermediate Value Theorem)**.** *Let $f(x)$ be a continuous function on $[a, b]$. If $f(a) < f(b)$ suppose that y is in the range $f(a) < y < f(b)$, and if $f(a) > f(b)$ suppose that y is in the range $f(b) < y < f(a)$. Then there exists a value c with $c \in (a, b)$ such that $f(c) = y$.*

As with the definition of a limit, the theorem is probably best understood via an explanatory picture, given as a graph in Figure 3.7.

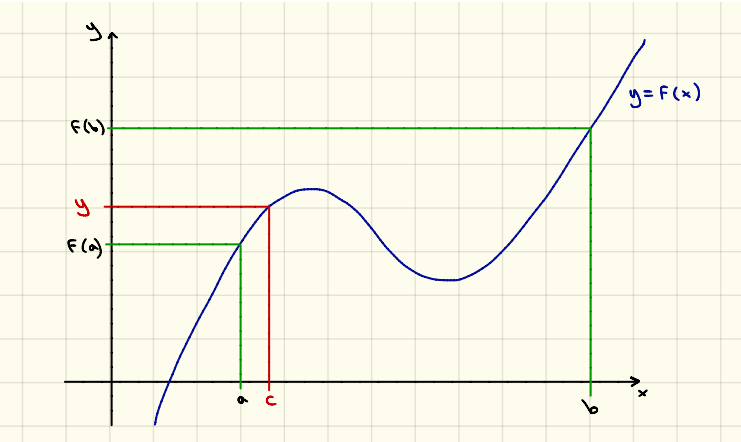


Figure 3.7: An illustration of the intermediate value theorem.

This theorem appears almost unworthy of the dignified title of a theorem. It does after all seem 'obvious' (once the mathematical jargon has been decoded). It is saying that if a continuous function $f(x)$ is defined at a and at b then as it is continuous it 'has no gaps' and so must pass through all output values between $f(a)$ and $f(b)$ as x increases from a to b. The function $f(x)$ may even pass through certain values many times (as in Figure 3.7 the only constraint is that it moves continuously from $f(a)$ to $f(b)$). However, the point of the theorem is (again) that we may derive it as a result of a rigorous algebraic definition, without relying on a picture that neither we nor a computer may be able to draw.

To get to grips with the ideas, let's explore some consequences of the intermediate value theorem. For a start, let us return to an issue we discussed right at the start of the course: how do we even known that square roots $\sqrt{n}$ always exist? For instance, why is $\sqrt{2} \in \mathbb{R}$ (despite the fact that we know that $\sqrt{2} \notin \mathbb{Q}$)?

Consider the function $f(x) = x^2 - 2$ defined on the interval $[1, 2]$. We know that when $f(x) = 0$ then $x = \sqrt{2}$ (if x is positive), and we can use the intermediate value theorem to show that there does indeed exist such an x.

Firstly, we must check that $f : [1, 2] \to \mathbb{R}$ is continuous. This can be done by hand (although we will soon develop useful quick methods for deducing continuity of such functions). Then, the intermediate value theorem tells us (with $a = 1$ and $b = 2$) that for every $y \in (f(a), f(b))$ there is a value $c \in (a, b)$ with $f(c) = y$. Since $f(a) = -1$ and $f(b) = 2$ the theorem tells us that for every $y \in (-1, 2)$ there is a value $c \in (1, 2)$ with $f(c) = y$. In particular, we can apply the theorem with the value $y = 0$, concluding that there really does exist some $c \in (1, 2)$ with $f(c) = 0$. From our previous discussion, this implies that $\sqrt{2}$ exists in $\mathbb{R}$. Moreover, the argument implies that $\sqrt{2}$ lies in the interval $(1, 2)$.

Of course, the trick that made this argument work was picking the values a and b such that $f(a) < 0 < f(b)$. If we wanted to find a better approximation to $\sqrt{2}$, we would need to reduce the range (a, b). In principle this method works for other polynomial functions (not just $x^2 - 2$), though takes some effort to write down in full generality.

Another neat observation that follows from a particular version of the intermediate value theorem. Assume we have a continuous function $f(x)$ on the interval $[0, 1]$ such that $0 \leq f(x) \leq 1$. Then we claim there exists $c \in [0, 1]$ such that $f(c) = c$. In other words, at (at least) one point c, the function behaves like the identity function! Now $f : [0, 1] \to [0, 1]$ so you can picture f as a set of reassignments of all the numbers in the interval $[0, 1]$, and we are arguing that one number c is mapped to itself: hence such a c is also sometimes called a *fixed point* of f.

Let's use the intermediate value theorem to prove our claim. To do so, consider the function $g(x) = f(x) - x$. It turns out that g is continuous, as $f(x)$ is continuous by assumption and the identity function $h(x) = x$ is continuous too. We will soon see the general result that if f and h are continuous then $f - h$ is also continuous: this is a special case. Next, note that $g(0) = f(0) - 0 = f(0) \geq 0$. In fact, with $g(0) = 0$ then we conclude that $f(0) = 0$, hence the claim is satisfied with $c = 0$. So we can assume that $g(0) > 0$. Similarly, $g(1) = f(1) - 1 \leq 0$, and if $g(1) = 0$ we conclude that $f(1) = 1$ and so the claim is satisfied with $c = 1$. Therefore

$$g(1) < 0 < g(0).$$

Applying the intermediate value theorem, we know that there exist some value $c \in (0, 1)$ such that $g(c) = 0$. Therefore we have that $g(c) = f(c) - c = 0$, hence $f(c) = c$ as claimed.

For example consider the sine function or the cosine function on the interval $[0, 1]$. This theorem gives us the insight that there are values c and $\hat{c}$ for which $\sin(c) = c$ and $\cos(\hat{c}) = \hat{c}$. Can you find a way to estimate these fixed points of sine and cosine?

The intermediate value theorem is seemingly simple but its applications are manifold. In particular it is the intermediate value theorem that lies behind a result[4] called *Brouwer's fixed point theorem*, which has applications in the investigation of differential equations, differential geometry, economics, and even game theory. From small acorns, mighty oak trees grow.

[4]Brouwer's fixed point theorem is a result about functions f from a closed disc to itself. One can define a notion of continuity for such maps, and the theorem states that all such continuous maps have a fixed point. The intermediate value theorem is the '1-dimesional' version of Brouwer's fixed point theorem.

3.2.2 Limits Involving Infinity.

The interplay between the notation used for limits and the notion of infinity is interesting and can be confusing. Earlier we remarked how it was sensible to write that $\lim_{x\to x_0}[f(x)] = \infty$ even though infinity is not a number (and so in particular ∞ is not in the range of f). This phenomenon occurs even with limits that don't mention infinity. For example, $\frac{\sin x}{x}$ is undefined at $x = 0$ and it is possible to show that $|\frac{\sin x}{x}| < 1$ for all x. However, we claimed earlier (though still have to prove) that $\lim_{x\to 0}[\frac{\sin x}{x}] = 1$. It does not trouble us that 1 is not in the range of $\frac{\sin x}{x}$ and yet its limit is 1; the limit tells us only that the function is approaching a value, not that it attains that value. In the same way although, a function maps to $\mathbb{R}$ (and not to $\mathbb{R} \cup \pm\infty$) it does not present a problem to say that its value tends towards infinity.

It is also possible to consider the value of a function as its argument tends towards infinity, i.e. as $x \to \infty$ (or similarly as $x \to -\infty$), as the 'tending to' is not the same as trying to evaluate the function at the 'value' $x = \infty$ (which would be nonsensical). We may define the following so-called *asymptotic limits*:

Definition 3.2.9 (Limits at infinity, asymptotic limits)**.** Let $L \in \mathbb{R}$. Let $A \subset \mathbb{R}$ and suppose that there is some value $c \in \mathbb{R}$ for which $(c, \infty) \subset A$. Let $f : A \to \mathbb{R}$. We write

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

and say *the limit of $f(x)$ as x tends to infinity is L* if

$$\text{for all } \varepsilon > 0, \text{ there exists } K > 0 \text{ such that } |f(x) - L| < \varepsilon \text{ for all } x > K.$$

We may also say *$f(x)$ is asymptotic to L as x tends to ∞*

Comment(s). (On limits to infinity...)

1. Compare this definition with the definition for approaching a limit from below, i.e. contrast $x \in (x_0 - \delta, x_0)$ with $x \in (K, \infty)$ used here. Of course ∞ can only be approached from below which explains the change of form for the definition above.

2. Similarly one can take the limit $x \to -\infty$, which takes the form of a limit from above:

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

means

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

Example 3.4. Consider the function

$$f(\alpha) = \tanh(\alpha)$$

and show that

$$\lim_{\alpha\to\infty}[f(\alpha)] = 1.$$

We have to show that for all $\varepsilon > 0$ there exists an $K > 0$ such that $|f(\alpha) - 1| < \varepsilon$ when $\alpha > K$. We will first give a sketch proof using the graph of $\tanh\alpha$ shown in Figure 2.23. Our aim is to show that $\tanh\alpha$ approaches 1 to a specified degree of accuracy encoded in ε. For example if we started with $\varepsilon = \frac{1}{100}$ then we aim to show that we can identify a range of $\alpha > X$ such that $|\tanh\alpha - 1| < \frac{1}{100}$. From the graph we understand this is possible immediately, as visually we see that $\tanh\alpha$ asymptotes to $y = 1$ as α approaches infinity. Hence it is only a small matter of computation to identify when $|\tanh K - 1| = \frac{1}{100}$, i.e. we aim to solve $\tanh K = \frac{99}{100}$ which gives, roughly $K = 2.65$. This is not a proof, but by this stage we should be relatively convinced that we could repeat the computation of K for any positive ε.

However, we should consider a more challenging version of the same question. Let us try to show that $\lim_{\alpha\to\infty} f(\alpha) = 1$ directly. Recalling the definition of the hyperbolic functions as points on the right hand branch of the hyperbola $x^2 - y^2 = 1$, where $x = \cosh\alpha$ and $y = \sinh\alpha$ then we have

$$\tanh\alpha = \frac{\sinh\alpha}{\cosh\alpha} = \frac{y}{x} = \frac{y}{\sqrt{1+y^2}}$$

As $y = \sinh\alpha$, we know that as $\alpha \to \infty$ then $y \to \infty$. (Consult the graph of sinh if this observation is not immediate for you.) We have therefore recast the limit as

$$\lim_{\alpha\to\infty}[\tanh\alpha] = \lim_{y\to\infty}[\frac{y}{\sqrt{1+y^2}}] = \lim_{y\to\infty}[\frac{y}{\sqrt{1+y^2}} \times \frac{\left(\frac{1}{y}\right)}{\left(\frac{1}{y}\right)}] = \lim_{y\to\infty}[\frac{1}{\sqrt{\frac{1}{y^2}+1}}] = 1$$

as $\lim_{y\to\infty}[\frac{1}{y^2}] = 0$. In the above line we have made a number of standard manipulations of limits and while it should certainly be possible to read and understand the line of working above, it should also make us a little worried as we have not yet discussed the fundamentals of working with limits. The working above is correct, but what we will discuss next is why these lines of working are valid.

3.3 Working with Limits

In this section we will develop the standard set of tools for being able to compute limits.

3.3.1 Rules for Limits of Composite Expressions.

Our first rules concern the limits of sums, products and quotients of functions. Before giving the rules, let us think about the sum of two functions $h(x) := f(x) + g(x)$. Now, as a definition of a new function $h(x)$,this makes sense so long as both $f(x) \in \mathbb{R}$ and $g(x) \in \mathbb{R}$; we can use the addition of numbers in $\mathbb{R}$ to build up the function $h(x)$ for all points $x \in \mathbb{R}$. Now if it happens that the functions, rather than being well-defined in $\mathbb{R}$, are slightly weaker but do still have well-defined limits which lie in $\mathbb{R}$, then we can quickly say that the limit of $h(x)$ is equal to the sum of the limits $f(x)$ and $g(x)$: in other words"the limit of the sum is the sum of the limits", so long as the limits both lie in $\mathbb{R}$. Bad things can happen if one or more of the limits are $\pm\infty$, which we shall go into later.

With the same constraint that the limit must be a well-defined number, we can convince ourselves that also the limit of a product is the product of the limits. But what about division? Well then we must take care, as although zero is a well-defined number the operation of division by zero is ill-defined. Let's list the rules. You will prove some of these rules carefully in the Sequences and Series course.

Sums, Product and Quotients of Limits

Let $a, b \in \mathbb{R}$, and $x_0 \in \mathbb{R} \cup \{\infty, -\infty\}$. If $\lim_{x\to x_0}[f(x)] = a$ and $\lim_{x\to x_0}[g(x)] = b$ then

1. $$\lim_{x\to x_0}[f(x) \pm g(x)] = \lim_{x\to x_0}[f(x)] \pm \lim_{x\to x_0}[g(x)] = a \pm b$$

2. $$\lim_{x\to x_0}[f(x)g(x)] = \lim_{x\to x_0}[f(x)] \lim_{x\to x_0}[g(x)] = ab$$

3. $$\lim_{x\to x_0}[\frac{f(x)}{g(x)}] = \frac{\lim_{x\to x_0}[f(x)]}{\lim_{x\to x_0}[g(x)]} = \frac{a}{b} \text{ only if } b \neq 0.$$

We have laboured the point that these rules for limits are not generally valid if $\lim_{x\to x_0}[f(x)] = \infty$. Why? As infinity is not a real number we do not, for example, have a way to define its addition and subtraction in $\mathbb{R}$. Let's convince ourselves that this is a problem. Suppose $\lim_{x\to x_0}[f(x)] = \infty$ and $\lim_{x\to x_0}[g(x)] = 1$. Then, if we could apply the rules above for the sum of limits in a carefree fashion, we would find

$$\lim_{x\to x_0}[f(x) + g(x)] = \text{“}\infty + 1\text{”} = \infty \tag{3.1}$$

$$\lim_{x\to x_0}[f(x) - f(x)] = \text{“}\infty - \infty\text{”} = 0. \tag{3.2}$$

As it happens the first of the above equations presents no problems, and you can indeed carefully prove from first principles that $f(x)+g(x) \to \infty$ as $x \to x_0$. The second equation also seems to be fine, as after all the function $f(x) - f(x)$ is identically 0 everywhere, so of course $\lim_{x\to x_0}(f(x) - f(x)) = 0$. However, subtracting two 'infinities' from each other can cause all manner of illogical deductions. For example, if this second rule were correct we could write

$$1 = \lim_{x\to x_0}(g(x)) = \lim_{x\to x_0}[f(x) + g(x) - f(x)] = \text{“}\infty + 1 - \infty\text{”} = \text{“}\infty - \infty\text{”} = 0,$$

i.e. we could show that $1 = 0$. This is clearly nonsense. Hence we must take more care when manipulating sums of infinite limits, and so these rules above are not generally valid when the limit is infinite.

Let's use these rules to evaluate limits of some rational functions (i.e. one polynomial divided by another). Let's start with an example we have seen twice already.

Example 3.5. Evaluate

$$\lim_{x\to 3}\left[\frac{x^2-9}{x-3}\right].$$

$$\lim_{x\to 3}\left[\frac{x^2-9}{x-3}\right] = \lim_{x\to 3}\left[\frac{(x-3)(x+3)}{x-3}\right] = \lim_{x\to 3}\big[(x+3)\big] = \lim_{x\to 3}\big[x\big] + \lim_{x\to 3}\big[3\big] = 3+3 = 6.$$

Note that the division $\frac{(x-3)(x+3)}{x-3} = x+3$ is valid, because we are only applying this when $x \neq 3$.

Example 3.6. Evaluate

$$\lim_{x\to -1}\left[\frac{x^2-4x-5}{x(x+1)}\right].$$

$$\lim_{x\to -1}\left[\frac{x^2-4x-5}{x(x+1)}\right] = \lim_{x\to -1}\left[\frac{(x+1)(x-5)}{x(x+1)}\right] = \lim_{x\to -1}\left[\frac{(x-5)}{x}\right] = \frac{\lim_{x\to -1}\big[(x-5)\big]}{\lim_{x\to -1}\big[x\big]} = \frac{-6}{-1} = 6.$$

Again note how the division of numerator and denominator by $x+1$ was valid because we were not evaluating these expressions at $x = -1$, so $x + 1$ was not equal to 0.

Example 3.7. Evaluate

$$\lim_{x\to \infty}\left[\frac{x^2+7x-3}{3x^2-18x+24}\right].$$

$$\begin{aligned}
\lim_{x\to\infty}\left[\frac{x^2+7x-3}{3x^2-18x+24}\right] &= \lim_{x\to\infty}\left[\frac{x^2+7x-3}{3x^2-18x+24}\times\frac{1/x^2}{1/x^2}\right] \\
&= \lim_{x\to\infty}\left[\frac{1+\frac{7}{x}-\frac{3}{x^2}}{3-\frac{18}{x}+\frac{24}{x^2}}\right] \\
&= \frac{\lim_{x\to\infty}\left[1+\frac{7}{x}-\frac{3}{x^2}\right]}{\lim_{x\to\infty}\left[3-\frac{18}{x}+\frac{24}{x^2}\right]} \\
&= \frac{\lim_{x\to\infty}\left[1\right]+\lim_{x\to\infty}\left[\frac{7}{x}\right]-\lim_{x\to\infty}\left[\frac{3}{x^2}\right]}{\lim_{x\to\infty}\left[3\right]-\lim_{x\to\infty}\left[\frac{18}{x}\right]+\lim_{x\to\infty}\left[\frac{24}{x^2}\right]} \\
&= \frac{1+0-0}{3-0+0} \\
&= \frac{1}{3}.
\end{aligned}$$

Notice in the first line that we have multiplied the function inside the limit by 1 written in the form $\frac{1/x^2}{1/x^2}$. Doing this to a function would normally change it (so it is no longer defined at $x=0$), so we might be concerned about the validity of the equalities written above. However, taking the limit of a function as $x\to\infty$ means that we are not considering the functions near $x=0$, but rather as x becomes extremely large. So changing $f(x)$ at $x=0$ does not affect the value of the limit of $f(x)$ as $x\to\infty$.

We will introduce a final rule that can be very useful in evaluating limits, this is a rule for the limit of a composite function.

Let $x_0, a, b \in \mathbb{R}$. If $\lim_{x\to x_0}[g(x)] = b$, $\lim_{x\to b}[f(x)] = a$, and f is continuous at b we have:

4.
$$\lim_{x\to x_0}[f(g(x))] = f(\lim_{x\to x_0} g(x)) = f(b) = a.$$

You should convince yourself that this rule is numerically plausible for some examples of well-behaved functions; confident students may try and formally prove the rule using the definitions of limits, though this type of exercise is mostly the preserve of the Sequences and Series Course.

Comment(s). (On limits of composite functions)

1. The conditions surrounding this limit are particularly constrained: in particular the asymptotic limit $x_0\to\infty$ is not necessarily covered by the rule. However when facing such a limit one may always make a change of variables $x=\frac{1}{y}$ such that the limit operation becomes $y\to 0$. Then the situation may be covered by the rule. We will discuss changing the limit variable in more detail in the upcoming section on the evaluation of limits.

2. Notice that functions $f(x)$ which are not continuous are not covered by this rule. Why is this? Consider the following functions:

$$f(x) = \begin{cases} x^2 + 1 & x \neq 0 \\ 7 & x = 0 \end{cases} \qquad \text{and} \qquad g(x) = 0 \quad \text{for all } x$$

As $\lim_{x\to 0}[f(x)] = 1 \neq f(0) = 7$, we see that $f(x)$ is not a continuous function. However, let us note that the composite function $f(g(x))$ is equal (for all values of x) to $f(0)$, which is identically 7. Constant functions are continuous, so in particular we have

$$\lim_{x\to 0}[f(g(x))] = \lim_{x\to 0}[7] = 7.$$

However if we had attempted (recklessly) to use the rule above for this discontinuous function f we would have used the fact that $\lim_{x\to 0}[g(x)] = 0$ and $\lim_{x\to 0}[f(x)] = 1$ to conclude (incorrectly) that

$$\lim_{x\to 0}[f(g(x))] = f(\lim_{x\to 0}[g(x)]) = 1.$$

We draw attention to that the constraint on using the rule for taking limits of composite functions is only on the function $f(x)$ being continuous, there is no constraint that $g(x)$ be a continuous function.

We will illustrate its utility with two examples below.

Example 3.8. Evaluate

$$\lim_{x\to\frac{\pi}{2}}\left[e^{\left(\cos^2 x\right)}\right].$$

$$\lim_{x\to\frac{\pi}{2}}\left[e^{\left(\cos^2 x\right)}\right] = e^{\lim_{x\to\frac{\pi}{2}}\left[\cos^2 x\right]} = e^0 = 1.$$

If we wished to compare this with the abstract formulation of the limit rule we would define

$$f(x) = e^x \qquad \text{and} \qquad g(x) = \cos^2 x$$

and then note that $\lim_{x\to\frac{\pi}{2}}[g(x)] = 0$ and $\lim_{x\to 0}[f(x)] = 1$.

Example 3.9. Evaluate

$$\lim_{x\to 1}\left[\sin\left(\frac{\pi(x^2-1)}{4(x-1)}\right)\right]$$

$$\lim_{x\to 1}\left[\sin\left(\frac{\pi(x^2-1)}{4(x-1)}\right)\right] = \sin\left(\lim_{x\to 1}\left[\left(\frac{\pi(x^2-1)}{4(x-1)}\right)\right]\right) = \sin\left(\lim_{x\to 1}\left[\frac{\pi}{4}(x+1)\right]\right) = 1.$$

To further convince ourselves that this is correct we show the graph in figure 3.8

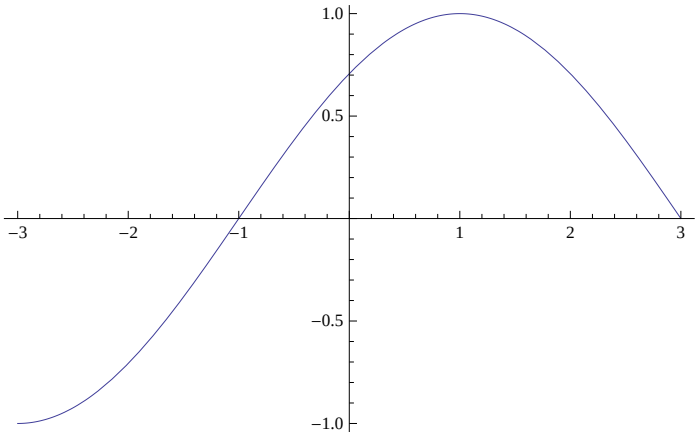


Figure 3.8: The graph of $f(x) = \sin\left(\frac{\pi(x^2-1)}{4(x-1)}\right)$ in the vicinity of $x = 1$.

A very common trick that can be used to manipulate expressions is to replace a function with the exponential function acting on the natural logarithm, e.g.

$$\lim_{x\to x_0}[f(x)^{g(x)}] = \lim_{x\to x_0}[e^{\ln(f(x)^{g(x)})}] = \lim_{x\to x_0}[e^{g(x)\ln(f(x))}] = e^{\lim_{x\to x_0}[g(x)\ln(f(x))]}.$$

It's not clear at this stage that we have much call for such a rearrangement, but as the functions we consider become more complicated it will be useful for us to try such operations in order to split complicated limits into simpler ones. The logarithm allows powers to be re-cast as coefficients, which can prove very useful; for example, a limit that you may evaluate in the tutorial exercises can be rearranged as

$$\lim_{x\to 0}[(1+x)^{\frac{1}{x}}] = e^{\lim_{x\to 0}[\frac{1}{x}\ln(1+x)]}.$$

Of course, since we do not know how to evaluate $\lim_{x\to 0}[\frac{1}{x}\ln(1+x)]$ it seems that we haven't gained that much! However, $\lim_{x\to 0}[\frac{1}{x}\ln(1+x)]$ will be one of the standard limits that, a little later in the course, we will learn how to evaluate – tt turns out that $\lim_{x\to 0}[\frac{1}{x}\ln(1+x)] = 1$.

3.3.2 Multiple Limits.

Functions of multiple variables may have their limit taken for each variable. Of course, when one takes a limit the global structure of the function is lost, and only some of the local information remains about the limit point. Consequently, for functions of multiple variables it will in general make a difference in which order limits are taken. There is an implicit ordering in which the notation indicates the limits will be taken:

$$\lim_{x\to x_0}\lim_{y\to y_0} f(x,y) := \lim_{x\to x_0}\left[\lim_{y\to y_0}\left[f(x,y)\right]\right].$$

Example 3.10. Show that changing the order the limits are taken in the following double-limit changes its value:

$$\lim_{x\to\infty}\lim_{y\to-\infty}(1+\tanh(x+y)).$$

$$\lim_{x\to\infty}\lim_{y\to-\infty}(1+\tanh(x+y)) = \lim_{x\to\infty}(1-1) = 0$$

while

$$\lim_{y\to-\infty}\lim_{x\to\infty}(1+\tanh(x+y)) = \lim_{y\to-\infty}(1+1) = 2.$$

We include a sketch of $(1+tanh(x+y))$ in Figure 3.9 to help us visualise the limits.

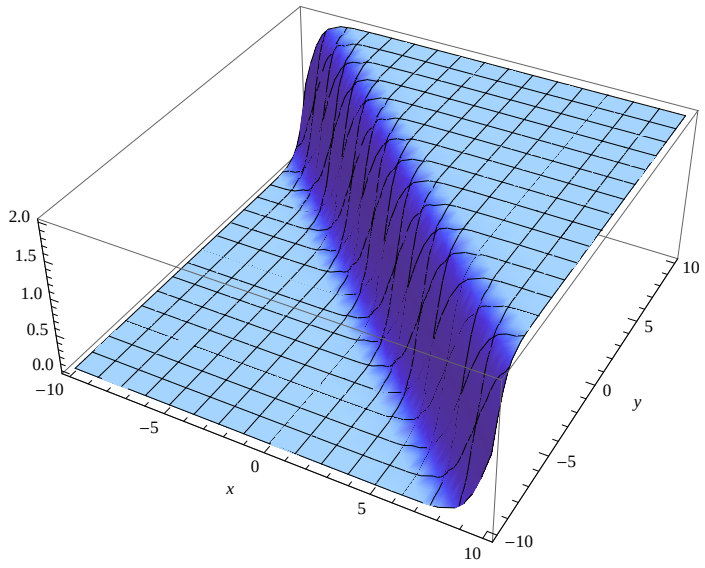


Figure 3.9: A sketch of the function $f(x,y) = 1+\tanh(x+y)$.

However there are many functions for which the order of the limits does not change the evaluation.

Example 3.11. Show that changing the order the limits are taken in the following double-limit does not change its value:

$$\lim_{x\to0}\lim_{y\to1}(x^2y - e^{-x-y}).$$

$$\lim_{x\to0}\lim_{y\to1}(x^2y - e^{-x-y}) = \lim_{x\to0}(x^2 - e^{-x-1}) = -\frac{1}{e}$$

and

$$\lim_{y\to1}\lim_{x\to0}(x^2y - e^{-x-y}) = \lim_{y\to1}(-e^{-y}) = -\frac{1}{e}.$$

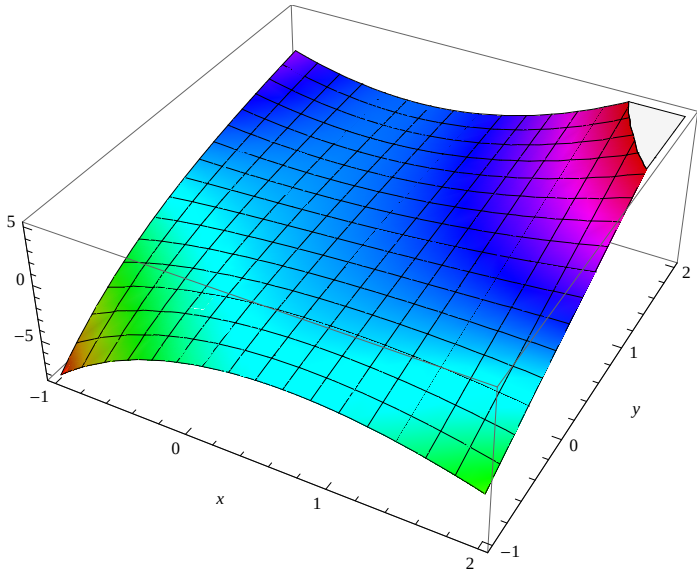


Figure 3.10: A sketch of the function $f(x, y) = x^2y - e^{-x-y}$.

We include a sketch of $x^2y - e^{-x-y}$ in figure 3.10 to help us visualise the limits.

We do not intend to draw any conclusions about the class of functions and double-limits for which it is possible to change the order of the limits. However you can begin to see that the order of the limits specifies the path one takes in approaching a limit on a surface, or on a more general multi-variable function. Consequently the condition for a multi-variable function to be continuous (i.e. the limit is independent of the path) is constraining. This will be a subject of study in some of the following courses in your degree.

3.3.3 Evaluating standard limits

So far we have spent some time establishing how one might tell if a limit exists, but we have only found limits for a small class of functions. Many limits can be evaluated by using the rules for manipulating sums and products of limits. The interesting cases are those which cannot be investigated this way. We have met only a few of these interesting cases so far, including those of the form

$$\lim_{x \to x_0} \left[\frac{f(x)}{g(x)} \right]$$

where $\lim_{x \to x_0}[f(x)] \in \mathbb{R}$, $\lim_{x \to x_0}[g(x)] = 0$, and $g(x)$ is a polynomial factor of $f(x)$. In such cases, if we can identify the function $h(x)$ such that $h(x) = \frac{f(x)}{g(x)}$ inside the limit (meaning for x close to x_0 but not equal to x_0), and if we can show that $\lim_{x \to x_0}[h(x)]$ exists, we can evaluate the limit $\lim_{x \to x_0} \frac{f(x)}{g(x)}$. We saw this method used in Examples 3.5 and 3.6. When the limit is

taken as $x \to \pm\infty$ on ratios of polynomial functions of the same degree we can evaluate the limit by looking at the terms of highest order, i.e.

$$\lim_{x\to\infty}\left[\frac{\sum_{j=0}^{n} a_j x^j}{\sum_{k=0}^{n} b_k x^k}\right] = \lim_{x\to\infty}\left[\frac{a_n x^n}{b_n x^n}\right] = \lim_{x\to\infty}\left[\frac{a_n}{b_n}\right] = \frac{a_n}{b_n}.$$

We saw this method used in example 3.7.

Inspired by the graph, shown in Figure 3.1, we claimed earlier that $\lim_{x\to 0}\left[\frac{\sin x}{x}\right] = 1$. Now we could at this stage invoke the infinite power series for x and convince ourselves that our guess was correct. To show this, we need to introduce a small piece of notation.

Definition 3.3.1 ("Big-Oh notation")**.** *Let $A \subset \mathbb{R}$. If $f, g : A \to \mathbb{R}$ are two functions, we write $f(x) = \mathcal{O}(g(x))$ if there is a positive constant $C > 0$ for which*

$$|f(x)| \leqslant C|g(x)|$$

for all $x \in A$.

If g is a much simpler function than f, and if all we care to know about f is its rough size, then it can be very convenient when manipulating limits to replace $f(x)$ by the expression $\mathcal{O}(g(x))$.

Example 3.12. Show that on the range $|x| \leqslant \frac{1}{2}$ we have

$$\sum_{n=1}^{\infty} x^n = x + \mathcal{O}(x^2).$$

Using the summation formula for a geometric series (which is valid since $|x| < 1$), we have

$$\sum_{n=1}^{\infty} x^n = x + \sum_{n=2}^{\infty} x^n = x + x^2 \sum_{n=0}^{\infty} x^n = x + \frac{x^2}{1-x}.$$

Since $|x| \leqslant \frac{1}{2}$ we have $1 - x \geqslant \frac{1}{2}$, and hence $|\frac{x^2}{1-x}| \leqslant 2x^2$. Therefore, by the definition of the Big-Oh notation, we have

$$\sum_{n=1}^{\infty} x^n = x + \mathcal{O}(x^2)$$

as claimed.

Attempting to use this notation on the power series for $\sin x$, one can prove that on the range $|x| \leqslant \frac{1}{2}$

$$\sin x = \sum_{n,\ \text{odd}} (-1)^{\frac{n-1}{2}} \frac{x^n}{n!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots = x + \mathcal{O}(x^3).$$

In other words, the sum of all the remaining terms in the power series is, in absolute value, at most Cx^3 for some constant C. Therefore

$$\begin{aligned}\lim_{x\to 0}\big[\frac{\sin x}{x}\big] &= \lim_{x\to 0}\left[\frac{\sum_{n,\,\text{odd}}(-1)^n\frac{x^n}{n!}}{x}\right]\\ &= \lim_{x\to 0}\left[\frac{x+\mathcal{O}(x^3)}{x}\right]\\ &= \lim_{x\to 0}\big[1+\mathcal{O}(x^2)\big]\\ &= \lim_{x\to 0}\big[1\big]+\lim_{x\to 0}\big[\mathcal{O}(x^2)\big]\\ &= 1\end{aligned}$$

However our formula for the sine function as an infinite power series rests on some conjectures we have yet to prove about the exponential function, and we have not justified why the tail of the power series can be bounded by $\mathcal{O}(x^3)$. This might trouble us. Fortunately there is another common method used to evaluate limits, which we can use to rigorously confirm such limits.

The Sandwich Theorem

This theorem is sometimes called the Squeeze Theorem or the Pinching Theorem.

Theorem 3.2 (The Sandwich Theorem.)**.** *Let $x_0, L \in \mathbb{R}$ and $\delta > 0$. If f, g, h are functions such that*

$$f(x) \le g(x) \le h(x)$$

for all x such that $0 < |x - x_0| < \delta$, and if $\lim_{x\to x_0}[f(x)] = \lim_{x\to x_0}[h(x)] = L$, then $\lim_{x\to x_0}[g(x)] = L$ as well.

That is: if we can identify two functions $f(x)$ and $h(x)$ that sandwich the function whose limit we are interested in (namely $g(x)$) in the vicinity of x_0, and such that f and h are have equal limits at x_0, then the limit of g at x_0 must exist and be equal to the same limit. This theorem is very helpful if we can find two sandwiching functions such that their limit is simpler to evaluate than the function in the middle of the sandwich. Let us look at some examples.

Example 3.13. Use the sandwich theorem to prove that

$$\lim_{x\to 0}\big[\frac{\sin x}{x}\big] = 1.$$

The trickiest part in applying the sandwich theorem is identifying the bounding functions which lie above and below the function we are interested in, in the vicinity of the limit. To find some useful inequalities we will look again at the two right-angled triangles (with sides of length $\{\cos\theta, \sin\theta, 1\}$ and the second with edges of length $\{1, \tan\theta, \sec\theta\}$) which bound the segment of the unit circle subtending an angle θ at the centre of the circle. This is hard to comprehend without a diagram, fortunately we have drawn the central idea in an earlier image – see figure 2.13.

The length of the arc shown in the figure in radians is θ. So from the diagram we immediately have that $\sin\theta \leqslant \theta$ for $\theta \in [0, \pi/2]$. Furthermore, the area of the segment of the circle is $\frac{\theta}{2}$, which is less than the area of the large enveloping right-angled triangle. The area of the triangle, as half the base times the height, is $\frac{\tan\theta}{2}$. Therefore

$$\sin\theta \leq \theta \qquad \text{and} \qquad \theta \leq \tan\theta \qquad \text{for } 0 \leq \theta \leq \frac{\pi}{2}.$$

From the first inequality we may divide through (recall $\theta > 0$ here) to obtain

$$\frac{\sin\theta}{\theta} \leq 1 \qquad \text{for } 0 \leq \theta \leq \frac{\pi}{2}$$

which will give the upper bound we will use when we invoke the sandwich theorem. While from $\theta \leq \tan\theta = \frac{\sin\theta}{\cos\theta}$ we have

$$\cos\theta \leq \frac{\sin\theta}{\theta} \qquad \text{for } 0 \leq \theta \leq \frac{\pi}{2}.$$

Altogether we have

$$\cos\theta \leq \frac{\sin\theta}{\theta} \leq 1 \qquad \text{for } 0 \leq \theta \leq \frac{\pi}{2}.$$

To apply the sandwich theorem we need to have such an inequality for all θ in a neighbourhood of 0 (not just for θ that lie to the right-hand side of 0). Fortunately, we may use the symmetry properties of the trigonometric functions to extend the inequality to negative θ, $\cos(-\theta) = \cos(\theta)$ and $\frac{\sin(-\theta)}{-\theta} = \frac{\sin(\theta)}{\theta}$. So we have

$$\cos\theta \leq \frac{\sin\theta}{\theta} \leq 1 \qquad \text{for } -\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2}.$$

Using the sandwich theorem and taking the limit as $x \to 0$, we have

$$1 = \lim_{\theta\to 0}(\cos\theta) \leq \lim_{\theta\to 0}(\frac{\sin\theta}{\theta}) \leq \lim_{\theta\to 0}(1) = 1$$

hence as the upper bound and the lower bound both limit to one as x approaches zero: by the sandwich theorem we conclude that

$$\lim_{\theta\to 0}(\frac{\sin\theta}{\theta}) = 1$$

as required.

Example 3.14. Use the sandwich theorem to prove that

$$\lim_{x\to 0}\left[x\sin\left(\frac{1}{x}\right)\right] = 0.$$

In this example we are at an advantage as it features a trigonometric function (about which we know a lot). Our first thought might be to try to use

$$-1 \leq \sin\left(\frac{1}{x}\right) \leq 1.$$

To turn this into something close to our expression, without losing the order of the inequality, we must multiply by a positive number. Hence let's multiply by $|x|$ rather than x to obtain:

$$-|x| \leq |x|\sin\left(\frac{1}{x}\right) \leq |x|.$$

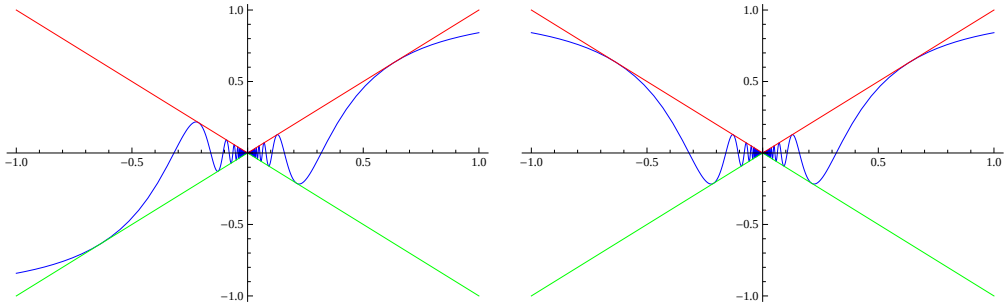


Figure 3.11: In the left-hand graph we show $-|x| \leq |x|\sin(1/x) \leq |x|$, while on the right we show $-|x| \leq x\sin(1/x) \leq |x|$

In the graph on the left of figure 3.11 we sketch the graphs of these functions to see the inequality in terms of the curves. How do we arrive at an inequality useful for the problem at hand? Now we notice that $-|x| \leq x \leq |x|$, hence we now can deduce

$$-|x| \leq x\sin\left(\frac{1}{x}\right) \leq |x|$$

and this inequality is shown graphically on the right in figure 3.11. Now the function sandwiched in the middle of the inequality is the one we are interested in, and furthermore the inequalities are of the form of the sandwich theorem as in the limit we have $\lim_{x\to 0}(|x|) = \lim_{x\to 0}(-|x|) = 0$. Hence if take the limit on the inequality we find:

$$0 = \lim_{x\to 0}(-|x|) \leq \lim_{x\to 0}((x\sin\left(\frac{1}{x}\right)) \leq \lim_{x\to 0}(|x|) = 0.$$

Therefore we surmise that

$$\lim_{x\to 0}((x\sin\big(\frac{1}{x}\big)) = 0$$

as required.

We did a lot of work to find these limits: can we learn anything else from these results? The two limits in the example do look rather similar, and we can modify them to show that they tell us about two different limits of the same function.

Let us commence with the result of the second limit

$$\lim_{x\to 0}((x\sin\big(\frac{1}{x}\big)) = 0.$$

If we change the variable used in the limit to $y = \frac{1}{x}$ then the function in the limit becomes $\frac{1}{y}\sin y$, the same function, albeit written in terms of a different variable, as in the first example. We also have to change the limit $x \to 0$ into a limit in terms of y, using $y = \frac{1}{x}$ as x approaches a very small number so y approaches a very large number, but there is a small concern for as $x \to 0^+$ then $y \to \infty$, while as $x \to 0^-$, $y \to -\infty$. So in translating the limit from one in terms of x we in fact generate two limits in terms of y:

$$\lim_{y\to-\infty}(\frac{\sin y}{y}) = 0 \quad \text{and} \quad \lim_{y\to\infty}(\frac{\sin y}{y}) = 0.$$

Given that these are limits to infinity it is in fact necessary that the limits are one-sided (i.e. from above and from below) so what seemed like a complexity is rather satisfying and we can be pleased with this two-for-one result[5]. We may also use the same change of variables to show that the limit $\lim_{x\to 0}(\frac{\sin x}{x}) = 1$ can be rewritten as the two asymptotic limits:

$$\lim_{y\to-\infty}(y\sin\big(\frac{1}{y}\big)) = 1 \quad \text{and} \quad \lim_{y\to\infty}(y\sin\big(\frac{1}{y}\big)) = 1.$$

Furthermore we may even consider changing to imaginary variables if we are comfortable that the limit remains well-defined, for example,

$$1 = \lim_{x\to 0}(\frac{\sin x}{x}) = \lim_{x\to 0}\left(\frac{\frac{1}{2i}(e^{ix} - e^{-ix})}{x}\right) = \lim_{y\to 0}\left(\frac{(e^{y} - e^{-y})}{2y}\right) = \lim_{y\to 0}(\frac{\sinh y}{y})$$

where we have substituted $y = ix$, in order to find another useful limit.

Exercise 3.2. Prove that $\lim_{x\to 0}[\frac{\tan x}{x}] = 1$. Hint: you may assume, without further proof, that $\lim_{x\to 0}[\frac{\sin x}{x}] = 1$ in your answer.

[5]We make the side comment that, given the symmetries of the sine function in the numerator and the linear function in the denominator, even if we had only one of these results we could have determined the second by substituting $z = -y$.

Exercise 3.3. Find the following limit:

$$\lim_{x\to\frac{\pi}{2}}[\frac{2\cos x}{2x-\pi}].$$

Hint: Try a change of the limit variable.

The result in example 3.14 above may seem rather exciting to us, as previously we argued that $\sin(\frac{\pi}{x})$ had no limit as $x \to 0$. Now if we rescale the variable by defining $y = \frac{x}{\pi}$, then as $x \to 0$ we have $y \to 0$. We therefore derive that $\lim_{y\to 0}(\sin(\frac{1}{y}))$ does not exist. However, from the results above we see that just by multiplying the function by y we discover a well-defined limit i.e. $\lim_{y\to 0}(y\sin\left(\frac{1}{y}\right)) = 0$. Let us emphasise that the fact that 'the limit of a product is the product of the limits' i.e.

$$\lim_{y\to 0}\left(y\sin\left(\frac{1}{y}\right)\right) \neq \lim_{y\to 0}(y)\lim_{y\to 0}(\sin\left(\frac{1}{y}\right))$$

does **not** apply here, as the limit $\lim_{y\to 0}(\sin\left(\frac{1}{y}\right))$ is undefined and so the rule for limits of products is not valid. What has happened is that the linear function y has gone to zero "faster" than $(\sin\left(\frac{1}{y}\right))$ has oscillated about zero as $y \to 0$, thus killing the oscillatory expression $y\sin(\frac{1}{y})$ and forcing it to tend to zero.

It can be useful to this mode of thinking. Ask yourself how parts of functions grow or shrink as the limit is taken, and which types of function dominate. This thinking does not replace a detailed analysis of a function, but it can prove a helpful guide. We will now compare some other common types of function in the limit, to get a sense of which dominate in specified limits.

Logarithmic vs. Polynomial vs. Exponential Functions: Which grows faster?

The answer to the question in the title is given away by these terms use in modern language where it is common to speak of exponential growth (or decay) to refer to something which grows incredibly fast (or shrinks very quickly). As $\ln(x) = \ln(\ln(e^x))$, and e^x grows rapidly, it seems that ln greatly reduces the rate of growth: we may guess that $\ln(x)$ grows more slowly as x increases than the identity function x itself (which is an example of a polynomial function). In turn, as $x = \ln(e^x)$, we may guess that polynomial functions grow more slowly than exponential functions as x grows.

In each case it is common to speak of *logarithmic growth*, *polynomial growth* and *exponential growth*. Of course the linguistic meaning of the terms is confirmed by plotting the graphs of $f(x) = \ln(x)$, $g(x) = x$ and $h(x) = e^x$. The result shown in figure 3.12 is what we should expect: $f(x)$ is the mirror image of $h(x)$ (where it is defined) in $g(x)$ as $h(x)$ is the inverse function of $f(x)$.

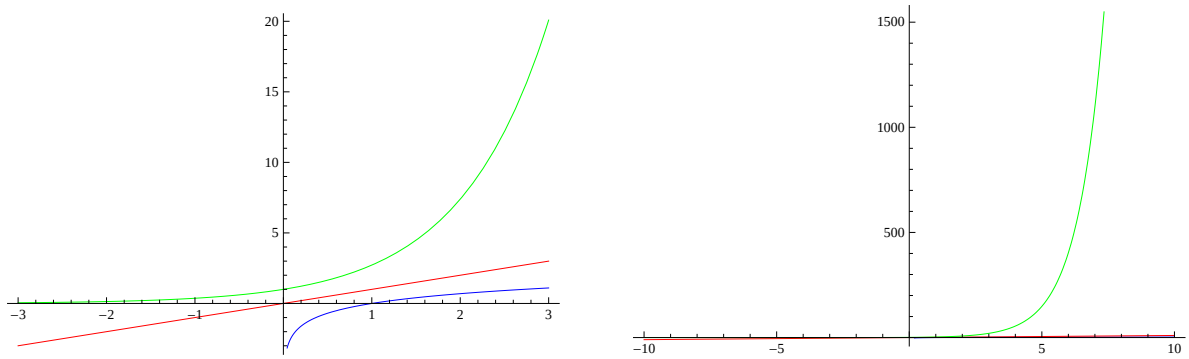


Figure 3.12: The graphs of $f(x) = \ln(x)$ (blue), $g(x) = x$ (red) and $h(x) = e^x$ (green) above, near zero on the left and for a slightly larger domain on the right. Notice that for positive x $f(x) < x < e^x$ and that exponential growth is really very fast indeed!

Let us now investigate how these types of functions behave in the limit by way of the following examples.

Example 3.15. Show that

$$\lim_{x\to 0}\left[\frac{\ln(1+x)}{x}\right] = 1.$$

Let us reflect and notice that we are comparing the speed with which $\ln(1+x)$ approaches zero (this is a shifted version of the logarithm function that is sketched in figure 3.12). Hence, as $x \to 0$ then $\ln(1+x) \to \ln(1) = 0$. All in all, the limit in question is approaching the dreaded $0/0$, and we cannot obviously split the function as a product. Instead let us try making a substitution to change the variable. We will substitute $1 + x = e^y$ so that the logarithm is simplified, and we end up with

$$\lim_{y\to 0}\left[\frac{y}{e^y-1}\right] = \lim_{y\to 0}\left[\frac{y}{e^{\frac{y}{2}}(e^{\frac{y}{2}}-e^{-\frac{y}{2}})}\right] = \lim_{y\to 0}\left[\frac{ye^{-\frac{y}{2}}}{2\sinh\left(\frac{y}{2}\right)}\right] = \lim_{y\to 0}\left[\frac{y/2}{\sinh\left(\frac{y}{2}\right)}\right]\lim_{y\to 0}\left[e^{-\frac{y}{2}}\right] = 1$$

where we have used the fact that $\lim_{x\to 0}[\frac{\sinh x}{x}] = 1$, which was shown earlier.

If you did not find this proof instinctive, do not worry: there are often many ways to analyse a limit. For example we could have used the power series for e^y (and the Big-Oh notation introduced earlier) to simplify the limit in a different way:

$$\lim_{y\to 0}\left[\frac{y}{e^y-1}\right] = \lim_{y\to 0}\left[\frac{y}{y+\frac{y^2}{2!}+\mathcal{O}(y^3)}\right] = \lim_{y\to 0}\left[\frac{1}{1+\frac{y}{2!}+\mathcal{O}(y^2)}\right] = \frac{\lim_{y\to 0}\left[1\right]}{\lim_{y\to 0}\left[1+\frac{y}{2!}+\mathcal{O}(y^2)\right]} = 1.$$

This is a perfectly rigorous kind of argument, and frequently used in physics calculations. Which route you prefer to take will depend upon your tastes: there is an artistry to analysis.

Exercise 3.4. Prove that

$$\lim_{x\to\infty}\left[x\ln(1+\frac{1}{x})\right]=1.$$

Hint 1: You may assume all the limits proven in the lecture notes this far and try a substitution.
Hint 2: This should be a very quick exercise!

Example 3.16. Show that for all $n\in\mathbb{N}$ and for all $x\geqslant 0$,

$$x^n\leqslant n^n e^x.$$

We first claim that $e^x\geqslant x+1$ for all $x\geqslant 0$. There are (at least) two possible proofs, though both rely on properties that we haven't completely rigorously established yet.

1. if we allow ourselves the power series expression for e^x, then
$$e^x=1+x+\frac{x^2}{2}+\frac{x^3}{6}+\cdots\geqslant 1+x$$
since all the rest of the terms are non-negative.

2. if we allow ourselves to use differentiation, we may argue as follows. Let f be the function $f(x)=e^x-x-1$. Note
$$\frac{df}{dx}=\frac{d}{dx}(e^x)-\frac{d}{dx}(x)-\frac{d}{dx}(1)=e^x-1-0=e^x-1$$
which is non-negative for all $x\geqslant 0$. Therefore the function $f(x)$ always has 'positive gradient' (though of course we haven't yet formally identified the derivative $\frac{df}{dx}$ with the gradient of f), meaning that f is monotonically increasing on $[0,\infty)$. Since $f(0)=0$, we conclude that $f(x)\geqslant 0$ for all $x\in[0,\infty)$, and hence $e^x\geqslant x+1$ for all such x.

With this claim in place, we certainly have $e^x\geqslant x$ for all $x\geqslant 0$, and so $e^{\frac{x}{n}}\geqslant\frac{x}{n}$. Raising both sides to the n^{th} power, we obtain $e^x\geqslant\frac{x^n}{n^n}$. Rearranging, $x^n\leqslant n^n e^x$ as claimed.

In fact, a direct argument via the power series immediately shows that $e^x\geqslant\frac{x^n}{n!}$ for all $x\geqslant 0$, leading to the improved inequality $x^n\leqslant n!e^x$.

This example showed that the exponential function grows faster than all polynomial functions. In particular

$$0\leqslant\frac{x^n}{e^x}\leqslant\frac{(n+1)^{n+1}}{x}$$

for all $x\geqslant 0$, so by sandwiching we get

$$\lim_{x\to\infty}\frac{x^n}{e^x}=0.$$

We can use this to prove an analogous statement about the natural logarithm function.

Exercise 3.5. Show that for all $p > 0$,

$$\lim_{y\to\infty} \Big[\frac{\ln(y)}{y^p}\Big] = 0.$$

In other words, show that any positive power y^p grows faster than $\ln(y)$ as $y \to \infty$.
Hint: set $x = p \ln y$ and use the previous limit.

Example 3.17. Show that

$$\lim_{x\to 0^+} \big[x \ln(x)\big] = 0.$$

This limit is telling us that the polynomial function x (which tends to zero as $x \to 0$) dominates the logarithmic function $\ln(x)$ (which tends to $-\infty$ as $x \to 0$). We give the sketch of the graph near zero in figure 3.13. Notice that the limit is taken from above as $ln(x)$ is undefined for $x \leq 0$.

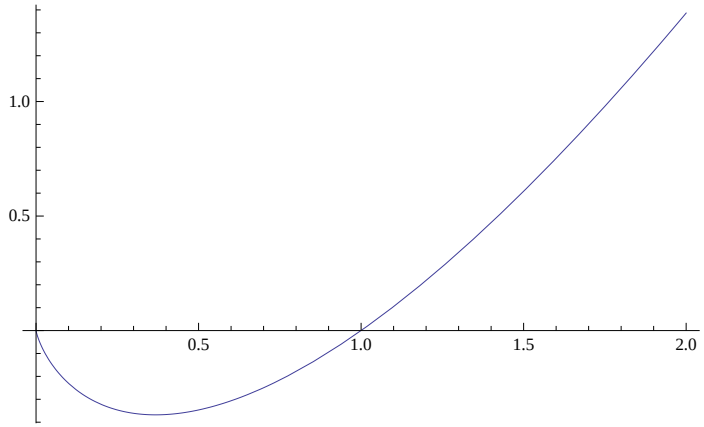


Figure 3.13: The graph of $f(x) = x \ln x$ in the vicinity of $x = 0$.

Let $y = \frac{1}{x}$. Then $y \to \infty$ as $x \to 0^+$. It follows from the previous exercise that

$$\lim_{x\to 0^+} x \ln x = \lim_{y\to\infty} \frac{\ln(1/y)}{y} = \lim_{y\to\infty} \Big(-\frac{\ln y}{y}\Big) = -\lim_{y\to\infty} \frac{\ln y}{y} = 0.$$

Finding intelligent ways to discover limits of functions is an art-form and requires practice: do try out all the limit questions you can find. We now turn our attention to the a fundamental object in calculus: the derivative. The limit will play a central role in this concept.

4. The Derivative

In which we will meet the 'derivative function': the function which gives the slope of the tangent to a curve. We will see that the limit will play a central role in the definition of the derivative, and we will derive the derivatives of many of the most common functions from first principles.

The material in this chapter will be covered in weeks 7 and 8 of the course.

Is it ever important to know the value a function is approaching rather than the actual value of the function? One of the motivating factors in developing the calculus was the desire to understand physical properties in the natural world, and it turns out that many of our intuitive notions of physical quantities are in fact a form of limit. In particular, we might very naturally enquire as to the speed of an object, but speed itself can be surprisingly difficult to define, and will in fact be defined as a certain limit.

The average speed of an object is defined as

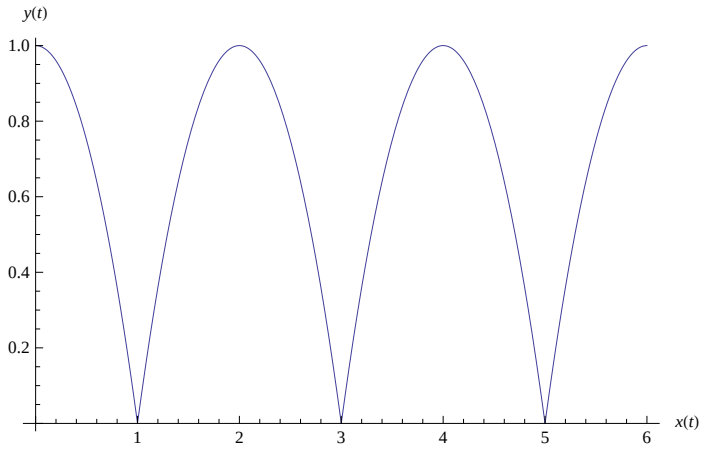
$$\text{Average speed} := \frac{\text{Change in the object's position.}}{\text{Change in time.}}$$

If an object moves from position $\vec{r}(t_1)$ to $\vec{r}(t_2)$ over the time interval $[t_1, t_2]$ (where we are writing the position $\vec{r}$ as a function of time), then we can use this definition to find that its average speed is

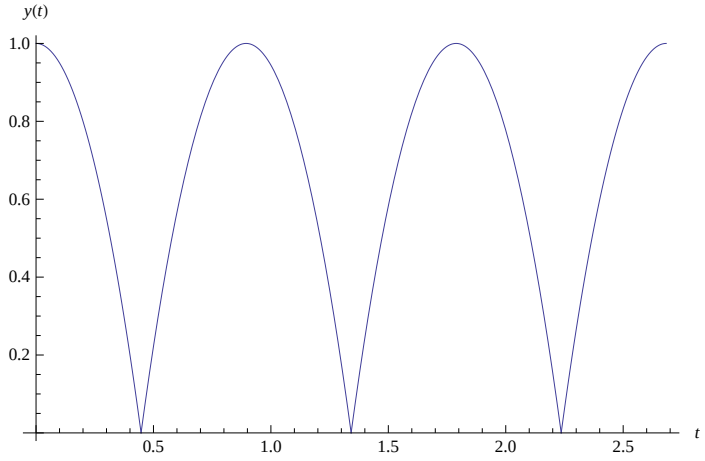
$$\frac{|\vec{r}(t_2) - \vec{r}(t_1)|}{t_2 - t_1}.$$

Obviously this presents some problems for objects whose speed changes rapidly: the average speed loses a lot of information. For an extreme example of this phenomenon, consider an 'idealised bouncing ball' moving in the x-direction with constant speed[1]:

[1] In case you were wondering, for a simple illustration we have neglected resistance in the ball's motion, and presumed gravity is $10ms^{-2}$ acting vertically downwards, started the ball at a height of $1m$ with zero initial vertical speed and a horizontal speed of $\sqrt{5}ms^{-1}$. The equations of motion give us $y = -x^2 + 1$, before any



The horizontal speed is constant by design but the vertical speed is varying from zero at the top of the bounce and its maximum speed when it hits the floor. Above we have a graph of the vertical position y against the x coordinate of the ball, but let's also look at the plot of its vertical position y against time t:



As $x(t) = \sqrt{5}t$ this graph looks like a squashed version of the last sketch.

Let us make a computation of the ball's average speed for the time interval until the first bounce, namely $[0, \frac{1}{\sqrt{5}}]$. From the graph (or better from the equation $y(t) = -5t^2 + 1$) we can read off $y(0) = 1$ and $y(\frac{1}{\sqrt{5}}) = 0$. In Figure 4.1 we have annotated the graph of the first bounce (in a plot of $y(t)$ against t), to help compute the average speed. Reading off from the graph for $t_1 = 0$ and $t_2 = \frac{1}{\sqrt{5}}$ we have $y(t_2) - y(t_1) = -1$. The average vertical speed (in ms^{-1}) is therefore

$$\frac{|y(t_2) - y(t_1)|}{t_2 - t_1} = \frac{|-1|}{\frac{1}{\sqrt{5}}} = \sqrt{5}\, ms^{-1}$$

bounce occurs. In addition we have assumed a perfect bounce takes place, i.e. an instantaneous reflection of the speed of the ball in the floor-line at the moment of impact. The situation is not very realistic but is good enough for our purpose.

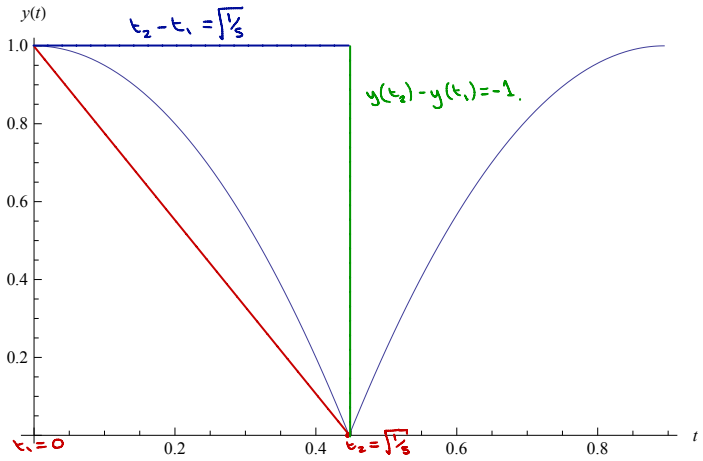


Figure 4.1: The computation of the average speed from $t = 0$ to the first bounce at $t = \sqrt{\frac{1}{5}}$.

This is the (absolute value) of the gradient of the red line in Figure 4.1. More usually we would speak of *velocity* rather than speed, where we keep track of the direction of movement. The average velocity of the ball over the interval $[t_1, t_2]$ is

$$\frac{y(t_2) - y(t_1)}{t_2 - t_1} = \frac{-1}{\frac{1}{\sqrt{5}}} = -\sqrt{5} ms^{-1}.$$

From the graph we can see this $-\sqrt{5} ms^{-1}$ is the velocity of a 'phantom ball' which moved with constant speed from $y = -1$ to $y = 0$ over the $\frac{1}{\sqrt{5}}$ seconds. Our ball actually has this velocity for only one moment during its first descent. We can surmise this because we are actually experts dropping objects on the floor, and we know that the vertical speed starts at $0\, ms^{-1}$ and continually increases due to the gravitational acceleration, reaching maximum speed at the moment of impact to the floor. So the vertical velocity passes through the value $-\sqrt{5}\, ms^{-1}$ exactly once. We will return to this idea right at the end of this section, when discussing the Mean Value Theorem.

Over long intervals, the average velocity calculation no longer even vaguely approximates the behaviour of the function. The average velocity over the time interval $[0, \frac{2}{\sqrt{5}}]$ is $0\, ms^{-1}$, for example, but the ball only has this velocity at the very beginning and very end of this period, and the speed is always at least $0\, ms^{-1}$ for the entire time!

Of course if we were to consider a collection of smaller and smaller time intervals, we would be able to find an average velocity for each such time interval, i.e. we could build up a set of average speeds $\hat{v}_1, \hat{v}_2, \hat{v}_3, \ldots, \hat{v}_n$ for the time intervals $[t_1, t_2], [t_2, t_3], [t_3, t_4], \ldots [t_n, t_{n+1}]$ respectively, where $t_{n+1} := \frac{1}{\sqrt{5}}$. We attempt to illustrate the improved accuracy in Figure 4.2, where one can see that for sufficiently short time intervals the sequence of red straight

lines, whose slopes are the average velocity for each interval, approaches and becomes almost indistinguishable from the curve $y(t)$.

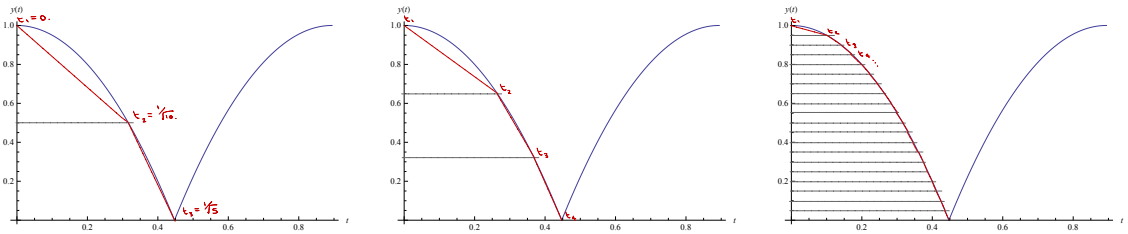


Figure 4.2: Approaching a speed function by finding the average speed for smaller and smaller time intervals.

Let's consider the short lines near a fixed time $t^* \in (0, \sqrt{5})$. So, take a very short interval $[t, t^*]$ (where t is close to t^*), and calculate the average speed over this interval. Then as $t \to t^*$ the average speed for that interval approaches what we might think of as the 'instantaneous velocity' at the time t^*. Geometrically it is the gradient/slope of the tangent to the curve at the time t^*.

Since the curve is relatively'well-behaved' (apart from at the bounces, of which more later) this gives a good practical definition of the speed of the ball at any time t^*: namely,

$$\dot{y}(t^*) := \lim_{t \to t^*} \left[\frac{y(t) - y(t^*)}{t - t^*} \right].$$

The dot on the y is Newton's notation for a time-derivative, and we would still write $\dot{y}$ as the standard notation for the velocity in the y-direction[2]. Of course we can't actually put $t = t^*$ as then the denominator becomes zero and the function is not well-defined (the usual $\frac{0}{0}$ problem). So the instantaneous vertical velocity at time t^* really is best described as a limit.

4.1 Differentiation from first principles

Historically, Newton's early ideas in calculus came directly from 'Fermat's way of drawing tangents'. Fermat had proposed to find the instantaneous slope by simply evaluating the change in a function $f(x)$ as x is varied and divide it by the corresponding change in x and then to let this change be put equal to zero. Precisely, Fermat proposed to find the slope of a function for the interval $[x, x + h]$ by evaluating the fraction

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

[2]Leibniz would have written $\dot{y} = \frac{dy}{dt}$, as we have been doing up to this point

and then setting $h = 0$ in the result. While this is evidently undefined as a function, the development of Newton to consider not the slope function but the limit of the slope function gives the derivative.

Definition 4.1.1. *The derivative of a function $f(x)$ with respect to x is*

$$\frac{df}{dx} \equiv \lim_{h\to 0}\left[\frac{f(x+h)-f(x)}{h}\right]$$

where the limit exists.

Comment(s). *(On the derivative...)*

1. *There are different notations in common use for the derivative of a function $f(x)$, the most common notation is due to Leibniz and is $\frac{df}{dx}$, but one may also commonly write $f'(x)$ for the derivative of $f(x)$ with respect to x. This second 'primed' notation is due to Lagrange and means that the derivative is taken with respect to the argument of the function, hence $f'(y) = \frac{df}{dy}$, when the argument is a complicated function this notation is very useful, e.g. the derivative of $f(x-ct)$, a travelling wave function, with respect to $(x-ct)$ is written $f'(x-ct)$ and frequently the argument is dropped in the notation so one may simply write f' for this same derivative - use of this notation obviously requires the function and its argument to be clearly defined. As mentioned earlier time derivatives have a special "dotted" notation reserved for them, which was first used by Newton who was studying dynamics, so that $\dot{x} = \frac{dx}{dt}$.*

2. *The notation for derivatives of derivatives is written*

$$\frac{d^n f}{dx^n}$$

where $n \in \mathbb{Z}^+$. When $n = 1$ we do not write the number 1, this denotes the first derivative of a function. When $n \geq 2$ we write the value of n explicitly, hence the second derivative of $f(x)$ with respect to x is written

$$\frac{d^2 f}{dx^2}.$$

Other notations for derivatives employ repeated use of the prime or the dot to indicate further derivatives. For example one might write

$$f''(x) \equiv \frac{d^2 f}{dx^2}$$

or

$$\ddot{x} \equiv \frac{d^2 x}{dt^2}.$$

3. *Geometrically the definition of the derivative is delightful and instructive. Suppose you wish to evaluate the derivative of a function $f(x)$ at $x = x_0$, the definition above is equivalent to drawing the straight line that connects $f(x_0)$ to any other point on the curve $f(x_0+h)$. This straight line (so long as it exists) is called a secant and has a well-defined slope, which we may compute. Now in taking the limit one allows the two separate points to become close, so that $f(x_0 + h)$ approaches $f(x_0)$. If the slope is tending to a well-defined limit then the derivative exists: the secant becomes a tangent. As $h \to 0$ we are able to construct the straight line through two points arbitrarily close to each other (but not identical) in this way we may have a well-defined definition of the tangent line to a function: a line apparently given meaning by only a single point on $f(x)$.*

4. *In this definition we have defined a derivative function - the result is a function of x and in doing so we have been efficient but have side-stepped an important notion. The derivative at a point $x = x_0$ is given by*

$$\left[\frac{df}{dx}\right]_{x=x_0} \equiv \lim_{h\to 0}\left[\frac{f(x_0+h)-f(x_0)}{h}\right].$$

This gives a value for the slope of the tangent to $f(x)$ at $x = x_0$, it is not a function of x. To construct the derivative function, which we have defined above we should evaluate the derivative at all points x_0 in the domain of the function, the resulting set of values can be used to then define the derivative as given above. One understands why we have taken the short cut we have in our definition, but we must wonder what happens if the derivative at a point $x = x_0$ does not exist. Such a function is said to not be differentiable at x_0 and we will return to this idea in the following section.

5. *Note that the limit in the derivative will only exist on an open interval $(a, b) \in \mathbb{R}$ when the limit from above is equal to the limit from below.*

Finding the derivative by evaluating the limiting value of the slope is often referred to as 'differentiation from first principles' as it does not rely upon knowing the derivative of any other function. We will now find from first principles some standard derivatives.

Example 4.1. *Find the derivative of the linear function $f(x) = ax + b$ with respect to x where $a, b \in \mathbb{R}$ at the point $x = x_0$.*

Commencing with the definition of the derivative, we compute:

$$\frac{df}{dx} = \lim_{h\to 0}\left[\frac{f(x_0+h)-f(x_0)}{h}\right] = \lim_{h\to 0}\left[\frac{a(x_0+h)+b-a(x_0)-b}{h}\right] = \lim_{h\to 0}\left[\frac{ah}{h}\right] = a.$$

As expected the derivative of the straight line is constant.

Example 4.2. *Find the derivative of the quadratic function*

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

with respect to x where $a, b, c \in \mathbb{R}$ as a function of x.

Commencing with the definition of the derivative, we compute:

$$\begin{aligned}
\frac{df}{dx} &= \lim_{h\to 0}\left[\frac{f(x+h)-f(x)}{h}\right] \\
&= \lim_{h\to 0}\left[\frac{a(x+h)^2+b(x+h)+c-a(x)^2-bx-c}{h}\right] \\
&= \lim_{h\to 0}\left[\frac{ax^2+2axh+ah^2+bx+bh+c-ax^2-bx-c}{h}\right] \\
&= \lim_{h\to 0}\left[\frac{2axh+ah^2+bh}{h}\right] \\
&= \lim_{h\to 0}\left[2ax+ah+b\right] \\
&= 2ax+b
\end{aligned}$$

Example 4.3. *Find the derivative of the function*

$$f(x) = ax^n$$

with respect to x where $a, \in \mathbb{R}$ and $n \in \mathbb{Z}^+$ as a function of x.

Note that this derivative was assumed in earlier chapters when defining the exponential function, the trigonometric functions and the hyperbolic functions as infinite series. This was one key pillar that we built much of the course so far upon, the other was that the infinite sum

for e^x converges. From first principles we have:

$$\begin{aligned}
\frac{df}{dx} &= \lim_{h\to 0}\left[\frac{f(x+h)-f(x)}{h}\right] \\
&= \lim_{h\to 0}\left[\frac{a(x+h)^n - ax^n}{h}\right] \\
&= \lim_{h\to 0}\left[\frac{a}{h}\left(\sum_{k=0}^{n}\left(\binom{n}{k}x^{n-k}h^k\right) - x^n\right)\right] \\
&= \lim_{h\to 0}\left[\frac{a}{h}\sum_{k=1}^{n}\binom{n}{k}x^{n-k}h^k\right] \\
&= \lim_{h\to 0}\left[a\sum_{k=1}^{n}\binom{n}{k}x^{n-k}h^{k-1}\right] \\
&= \lim_{h\to 0}\left[a\binom{n}{1}x^{n-1} + a\binom{n}{2}x^{n-2}h + \mathcal{O}(h^2)\right] \\
&= anx^{n-1}
\end{aligned}$$

Or, alternatively, without using the binomial expansion (so that the following becomes a proof for all $n \in \mathbb{R}$):

$$\begin{aligned}
\frac{df}{dx} &= \lim_{h\to 0}\left[\frac{f(x+h)-f(x)}{h}\right] \\
&= \lim_{h\to 0}\left[\frac{a(x+h)^n - ax^n}{h}\right] \\
&= ax^n \lim_{h\to 0}\left[\frac{(1+\frac{h}{x})^n - 1}{h}\right] \\
&= ax^n \lim_{h\to 0}\left[\frac{e^{\ln[(1+\frac{h}{x})^n]} - 1}{h}\right] \\
&= ax^n \lim_{h\to 0}\left[\frac{e^{n\ln(1+\frac{h}{x})} - 1}{h}\right].
\end{aligned}$$

Let us now change the limit variable using $h \equiv xg$ so we have:

$$\begin{aligned}
\frac{df}{dx} &= ax^n \lim_{g\to 0}\left[\frac{e^{n\ln(1+g)} - 1}{xg}\right] \\
&= ax^{n-1} \lim_{g\to 0}\left[\frac{e^{n\ln(1+g)} - 1}{g}\right].
\end{aligned}$$

Here we now choose to insert $\ln(1+g)/\ln(1+g) = 1$, so that we may make use of the standard limits (evaluated in the previous chapter): $\lim_{x\to 0}[\frac{\ln(1+x)}{x}] = 1$ and $\lim_{x\to 0}[\frac{e^x-1}{x}] = 1$ - see

example 3.14 for the proof of both limits. So, upon inserting $\ln(1+g)/\ln(1+g)$ *we have*

$$\begin{aligned}\frac{df}{dx} &= ax^{n-1}\lim_{g\to 0}\left[\left(\frac{e^{n\ln(1+g)}-1}{\ln(1+g)}\right)\left(\frac{\ln(1+g)}{g}\right)\right]\\ &= ax^{n-1}\lim_{g\to 0}\left[\frac{e^{n\ln(1+g)}-1}{\ln(1+g)}\right]\lim_{g\to 0}\left[\frac{\ln(1+g)}{g}\right]\\ &= ax^{n-1}\lim_{k\to 0}\left[\frac{e^{k}-1}{k/n}\right]\\ &= anx^{n-1}\end{aligned}$$

where we substituted $k \equiv n\ln(1+g)$*. This second method is much longer than using the binomial expansion, but it used two nice and common "tricks" to manipulate the limit into the form of some standard limits.*

Example 4.4. *Find the derivative of the function*

$$f(x) = a^x$$

with respect to x *where* $a, \in \mathbb{R}$ *as a function of* x*.*

$$\begin{aligned}\frac{df}{dx} &= \lim_{h\to 0}\left[\frac{f(x+h)-f(x)}{h}\right]\\ &= \lim_{h\to 0}\left[\frac{a^{x+h}-a^x}{h}\right]\\ &= \lim_{h\to 0}\left[a^x\left(\frac{a^h-1}{h}\right)\right]\\ &= \lim_{h\to 0}[a^x]\lim_{h\to 0}\left[\frac{a^h-1}{h}\right]\\ &= a^x\lim_{h\to 0}\left[\frac{e^{\ln(a^h)}-1}{h}\right]\\ &= a^x\lim_{k\to 0}\left[\frac{e^{k}-1}{k/\ln(a)}\right]\\ &= a^x\ln(a)\lim_{k\to 0}\left[\frac{(1+k+\frac{k^2}{2!}+\mathcal{O}(k^3))-1}{k}\right]\\ &= a^x\ln(a)\lim_{k\to 0}\left[1+\mathcal{O}(k)\right]\\ &= a^x\ln(a)\end{aligned}$$

Where we changed the limit variable using $k = \ln a^h = h\ln a$*. Note that had we chosen* $a = e$*, Euler's number, then we would have found* $\frac{df}{dx} = e^x\ln(e) = e^x$ *which was part of our defining relation for the exponential in earlier chapters.*

Example 4.5. *Find the derivative of the function*

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

with respect to x as a function of x.

$$\begin{aligned}
\frac{df}{dx} &= \lim_{h\to 0}\left[\frac{f(x+h)-f(x)}{h}\right] \\
&= \lim_{h\to 0}\left[\frac{\ln(x+h)-\ln(x)}{h}\right] \\
&= \lim_{h\to 0}\left[\frac{\ln(x(1+\frac{h}{x}))-\ln(x)}{h}\right] \\
&= \lim_{h\to 0}\left[\frac{\ln(1+\frac{h}{x})+\ln(x)-\ln(x)}{h}\right] \\
&= \lim_{h\to 0}\left[\frac{\ln(1+\frac{h}{x})}{h}\right] \\
&= \lim_{g\to 0}\left[\frac{\ln(1+g)}{gx}\right] \\
&= \frac{1}{x}\lim_{g\to 0}\left[\frac{\ln(1+g)}{g}\right] \\
&= \frac{1}{x}
\end{aligned}$$

where we have used $h \equiv gx$ and the standard limit $\lim_{x\to 0}[\frac{\ln(1+x)}{x}] = 1$ (see example 3.14).

Example 4.6. *Find the derivative of the function*

$$f(x) = \sin(x)$$

with respect to x as a function of x.

$$\begin{aligned}\frac{df}{dx} &= \lim_{h\to 0}\left[\frac{f(x+h)-f(x)}{h}\right] \\ &= \lim_{h\to 0}\left[\frac{\sin(x+h)-\sin(x)}{h}\right] \\ &= \lim_{h\to 0}\left[\frac{\sin(x)\cos(h)+\cos(x)\sin(h)-\sin(x)}{h}\right] \\ &= \lim_{h\to 0}\left[\frac{\sin(x)(\cos(h)-1)+\cos(x)\sin(h)}{h}\right] \\ &= \sin(x)\lim_{h\to 0}\left[\frac{\cos(h)-1}{h}\right]+\cos(x)\lim_{h\to 0}\left[\frac{\sin(h)}{h}\right]\end{aligned}$$

Let us comment at this stage that we do have recourse to using the series expansion for $\cos(h) = 1-\mathcal{O}(h^2)$ *to rapidly show that*

$$\lim_{h\to 0}\left[\frac{\cos(h)-1}{h}\right] = \lim_{h\to 0}\left[\frac{\mathcal{O}(h^2)}{h}\right] = \lim_{h\to 0}\left[\mathcal{O}(h)\right] = 0$$

and thence after substituting the limit $\lim_{h\to 0}\left[\frac{\sin(h)}{h}\right] = 1$ *we would find* $\frac{df}{dx} = \cos(x)$*. However since this is an analysis course we are going to think of another proof just to practise using the sandwich theorem again, and of course it will take slightly longer to get to the same answer but will be interesting. As* $-1 \leq \cos(h) \leq 1$ *then we may deduce that* $(\cos(h)-1) \leq 0$*, this gives the upper bound for use in the sandwich theorem. For the lower bound we turn to geometry and the sector shown in figure 2.5. The base of the large right-angled-triangle has length* 1*, and we have embedded a similar triangle within it whose base is* $\cos(\theta)$*. Hence the remaining part of the base line (whose length isn't indicated on the diagram) has length* $1-\cos(\theta)$*. Now consider the right-angled-triangle with base* $1-\cos(\theta)$ *and height* $\sin\theta$*, its hypotenuse has length-squared:*

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

As this triangle would be embedded within the sector of the unit circle shown in figure 2.5, its hypotenuse must have length less than or equal to the arc length shown, which is θ*, i.e. we have*

$$2(1-\cos\theta) \leq \theta^2$$

or

$$(\cos\theta-1) \geq -\frac{\theta^2}{2}$$

which gives our lower bound. Returning to our limit, we are now able to sandwich the limit as follows

$$0 = \lim_{h\to 0}\left[-\frac{h}{2}\right] = \lim_{h\to 0}\left[-\frac{h^2}{2h}\right] \leq \lim_{h\to 0}\left[\frac{\cos(h)-1}{h}\right] \leq 0.$$

Hence by the sandwich theorem we have

$$\lim_{h\to 0}\left[\frac{\cos(h)-1}{h}\right] = 0$$

which we may use to show[3] *that* $\frac{d}{dx}(\sin(x)) = \cos(x)$.

Exercise 4.1. *Find the derivative of* $f(x) \equiv \cos(x)$ *with respect to* x.

4.2 Differentiable Functions

But do all functions have a derivative at all points? The answer is emphatically no. There were some caveats involved in defining the derivative and we did not draw attention to them in the previous section, so we will look again now at the definition of the derivative in definition 4.1.1. It is a good idea now to turn back and re-read the definition. Where does the definition of the derivative encounter problems? There are two situations which spring to mind:

(i) if the function $f(x)$ is not continuous at $x = x_0$ and

(ii) if the limit

$$\lim_{h\to 0}\left[\frac{f(x+h)-f(x)}{h}\right]$$

does not exist for some particular value of $x = x_0$.

If either of these cases occur, the derivative at $x = x_0$ does not exist and the function is said to be 'not differentiable' at $x = x_0$.

Definition 4.2.1. *A function* $f(x)$ *is differentiable at the point* $x = x_0$ *if the derivative* $\frac{df}{dx}$ *exists at* $x = x_0$. *A function for which the derivative exists for all points in its domain is called a differentiable function.*

Theorem 4.1. *All differentiable functions are continuous functions.*

[3] Once again!

Proof: Suppose that $f(x)$ is a differentiable function. Now recall the definition for $f(x)$ to be continuous at the point $x = x_0$, namely that

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

This is what we are aiming to show, we will nevertheless start with this statement and see how it relates to the limit which defines the derivative of $f(x)$. We begin by rearranging the statement of the continuity of $f(x)$ at $x = x_0$:

$$\begin{aligned}
\lim_{x\to x_0}(f(x)) - f(x_0) &= \lim_{x\to x_0}[f(x) - f(x_0)] \\
&= \lim_{x\to x_0}\left[\frac{f(x) - f(x_0)}{(x - x_0)}(x - x_0)\right] \\
&= \lim_{h\to 0}\left[\frac{f(x_0 + h) - f(x_0)}{h}(h)\right] \\
&= \left[\frac{df}{dx}\right]_{x=x_0} \times \lim_{h\to 0}[h] \\
&= 0
\end{aligned}$$

where we redefined the limit variable $h \equiv x - x_0$ and in the penultimate line we were able to split the limit of the product into the product of the limits as we know both limits exist. We only know that both limits exist because we assumed that the function is differentiable, and hence it is differentiable at the point $x = x_0$. Hence by assuming that $f(x)$ is differentiable we have shown that it is also continuous.

At this stage you might be wondering if there are any continuous functions which are not differentiable. Indeed there are, and this point tackles our second problem with the derivative: when the limit of the slope does not exist. The most famous example used to show that a continuous function is not necessarily differentiable is the modulus function (which we defined in definition 2.5.3) and is sketched in figure 4.3.

One can immediately guess where the problem point will be from the graph: the graph of the function although continuous has a right-angle in it at $x = 0$. Our instincts will prove to be correct but let us check it carefully. First we ought to confirm that the modulus function is continuous, namely we check that

$$\lim_{x\to x_0}[|x|] = |x_0|$$

this function is simple enough that we may be confident enough to assert the above from the sketch of the graph: it is a continuous function. But is it differentiable? The derivative can be

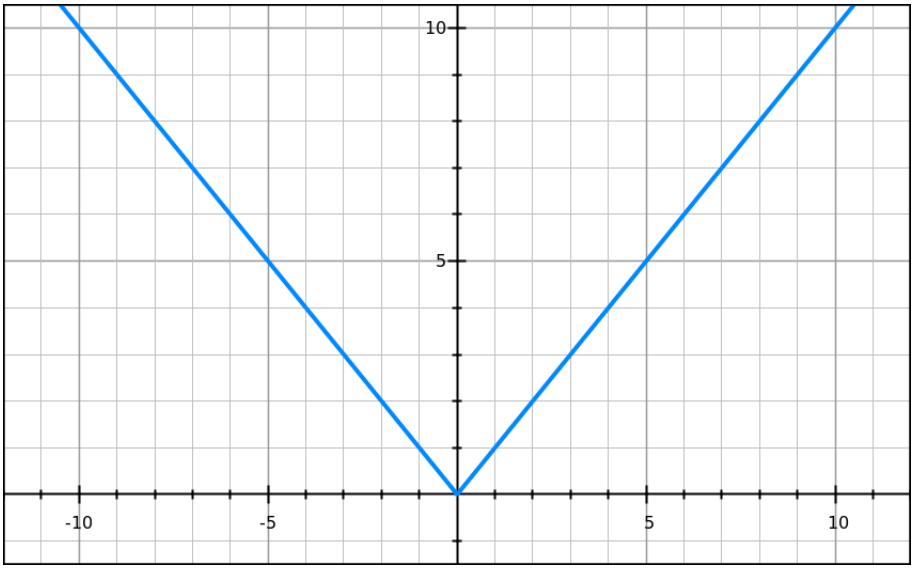


Figure 4.3: The modulus function is a continuous function, but is not a differentiable function.

readily checked to give (for $f(x) = |x|$)

$$\left[\frac{df}{dx}\right]_{x=x_0} = \begin{cases} +1 & x_0 > 0 \\ -1 & x_0 < 0. \end{cases}$$

Hence now we see that at $x_0 = 0$ the limit from above in the definition of the derivative does not equal the limit from below. Above zero the limit in the derivative is $+1$, while from below it is -1, hence

$$\lim_{h\to 0}\left[\frac{f(0+h)-f(0)}{h}\right]$$

does not exist and so the modulus function is an example of a continuous function which is not differentiable at $x = 0$, and hence is not a differentiable function.

These thoughts above about the conditions necessary for a function to have a well-defined derivative, highlight a very simple idea behind the derivative, namely that if one zooms in sufficiently close to the graph of a differentiable function, its graph approaches a straight line. The straight line it approaches is part of the tangent to the function at a point, and the slope of the tangent is the derivative of the function at the point. No matter how much one zoomed into the microscopic detail of the modulus function at $x = 0$, the V shape of the curve would not smooth out into a straight line - so it is not differentiable. Of course when we face questions of differentiability we will rarely be in the position of being able to quickly sketch the function and understand what happens to the curve as we zoom in on a point.

4.3 Properties of the derivative

We have thought about and computed derivatives for many common functions. As we have seen it is possible to create new functions from a basic set by adding the functions, multiplying the function, dividing the functions or composing the functions (functions of functions). Do we have to compute the derivative from first principles for all functions formed in this way? No, fortunately we can rely on properties of the derivative to determine a set of rules which can be used to break down the derivatives of complicated functions into derivatives of simpler functions which we can quickly compute.

4.3.1 The Chain Rule

Given a composite function $f(g(x))$ where both f and g are differentiable functions we can find the derivative $\frac{d}{dx}(f(g(x))$ using our knowledge of the derivatives of f and g. Let us elaborate a little on this idea using some dynamical quantities. Suppose we were interested in computing the acceleration function from the velocity function $v(x)$ written as a function of the displacement functions $x(t)$, by taking its time derivative. In other words we wish to compute:

$$\frac{d}{dt}\bigg(v(x(t))\bigg).$$

At this point it is useful to emphasise the difference between the function $v(x(t))$ which is the velocity written as an explicit function of $x(t)$ (i.e. when one writes $v(x(t))$ it is written as a function of x, it is the same as writing $v(x)$ where the notation indicating that x is a function of t has been suppressed) and the function $v(t)$ which is the velocity function written as an explicit function of t. We can obtain $v(t)$ from $v(x(t))$ by substituting the expression for $x(t)$ into $v(x(t))$ so we obtain a function of t rather than x. No matter which variable we use to express v the velocity is still the same for any particular value of $t = t_0$ or $x = x_0 = x(t_0)$. Hence while it is a simple matter to compute from first principles the derivative $\frac{d(v(t))}{dt}$ we know that it must also be possible to compute $\frac{d(v(x(t)))}{dt}$ - the only change is that we use the variable x to express v before taking the derivative. This is the aim of the chain rule: to give a method to compute $\frac{dv}{dt}$ starting from $v(x)$, without having to substitute $x(t)$ immediately. It is worth commenting that if we begin with $v(x)$ and $x(t)$ and we wish to compute derivatives there are two simple derivatives we can compute without much thought, namely $\frac{dv}{dx}$ and $\frac{dx}{dt}$. The Chain rule is a simple rule for combining these two derivatives to obtain $\frac{dv}{dt}$.

Let us now state and prove the chain rule formula in terms of a pair of differentiable functions $f(x)$ and $g(x)$. The chain rule is:

$$\frac{d(f(g))}{dx} = \left(\frac{df}{dg}\right)\left(\frac{dg}{dx}\right)$$

Proof:

$$\begin{aligned}
\frac{d(f(g))}{dx} &= \lim_{h\to 0}\left[\frac{f(g(x+h)) - f(g(x))}{h}\right] \\
&= \lim_{h\to 0}\left[\left(\frac{f(g(x+h)) - f(g(x))}{g(x+h) - g(x)}\right)\left(\frac{g(x+h) - g(x)}{h}\right)\right] \\
&= \lim_{h\to 0}\left[\left(\frac{f(g(x+h)) - f(g(x))}{g(x+h) - g(x)}\right)\right] \lim_{h\to 0}\left[\left(\frac{g(x+h) - g(x)}{h}\right)\right] \\
&= \lim_{\epsilon\to 0}\left[\left(\frac{f(g+\epsilon) - f(g)}{\epsilon}\right)\right] \lim_{h\to 0}\left[\left(\frac{g(x+h) - g(x)}{h}\right)\right] \\
&= \left(\frac{df}{dg}\right)\left(\frac{dg}{dx}\right)
\end{aligned}$$

where we have made a change of the limit variable so that $\epsilon \equiv g(x+h) - g(x)$, hence $g(x+h) = g(x) + \epsilon$ and

$$\lim_{h\to 0}(\epsilon) = \lim_{h\to 0}(g(x+h) - g(x)) = 0.$$

Exercise 4.2. *Show that*

$$\frac{d}{dx}\Big(\sin(x^2)\Big) = 2x\cos(x^2)$$

both from first principles and by using the chain rule.

Example 4.7. *Use the chain rule to find the derivative of $f(ax)$ with respect to x where a is a constant.*

Let us write $g(x) = ax$ then we have

$$\frac{d(f(g))}{dx} = \left(\frac{df}{dg}\right)\left(\frac{d(ax)}{dx}\right) = af'(ax)$$

e.g. if $f(x) = x^3$ then $f(ax) = a^3x^3$ and $\frac{df(ax)}{dx} = \frac{d(a^3x^3)}{dx} = 3a^3x^2$ while $f'(ax)$ can be found by writing $y \equiv ax$ and computing $f'(ax) = f'(y) = \frac{df(y)}{dy} = \frac{d(y^3)}{dy} = 3y^2 = 3a^2x^2$, hence $af'(ax) = 3a^3x^2$, in agreement with the earlier computation.

Example 4.8. *Find the derivative with respect to x of the function e^{x^n} as a function of x.*

Using the chain rule (with $f(x) = e^x$ and $g(x) = x^n$, so that $f(g(x)) = e^{x^n}$) we have

$$\frac{d}{dx}\left(e^{x^n}\right) = \frac{d(f(g))}{dx} = \left(\frac{df}{dg}\right)\left(\frac{dg}{dx}\right) = nx^{n-1}e^{x^n}.$$

Example 4.9. *Find the derivative with respect to x of the function $\ln(\cosh(x))$ as a function of x.*

Using the chain rule (with $f(x) = \ln(x)$ and $g(x) = \cosh(x)$, so that $\frac{df}{dx} = \frac{1}{x}$ and $\frac{dg}{dx} = \sinh(x)$) we have

$$\frac{d}{dx}\Big(\ln(\cosh(x))\Big) = \frac{d(f(g))}{dx} = \left(\frac{df}{dg}\right)\left(\frac{dg}{dx}\right) = \left(\frac{1}{g}\right)\Big(\sinh(x)\Big) = \frac{\sinh(x)}{\cosh(x)} = \tanh(x).$$

Example 4.10. *Find the derivative with respect to x of the function $f(g(h(x)))$ as a function of x.*

Using the chain rule repeatedly we find:

$$\frac{d}{dx}\Big(f(g(h(x)))\Big) = \left(\frac{df}{dg}\right)\left(\frac{dg}{dx}\right) = \left(\frac{df}{dg}\right)\left(\frac{dg}{dh}\right)\left(\frac{dh}{dx}\right).$$

4.3.2 The Sum Rule

The derivative acts linearly on a sum of differentiable functions $f(x)$ and $g(x)$:

$$\frac{d}{dx}\Big(f(x) + g(x)\Big) = \frac{df}{dx} + \frac{dg}{dx}.$$

Proof: (relies on the properties of limits)

$$\begin{aligned}\frac{d}{dx}\Big(f(x) + g(x)\Big) &= \lim_{h\to 0}\left[\frac{(f(x+h) + g(x+h)) - (f(x) + g(x))}{h}\right] \\ &= \lim_{h\to 0}\left[\frac{f(x+h) - f(x)}{h}\right] + \lim_{h\to 0}\left[\frac{g(x+h) - g(x)}{h}\right] \\ &= \frac{df}{dx} + \frac{dg}{dx}.\end{aligned}$$

Exercise 4.3. *Employ the sum rule and, subsequently, differentiation from first principles to show that*

$$\frac{d}{dx}\Big(\cos^2 x + \sin^2 x\Big) = 0$$

i.e. prove this without using the identity $\cos^2 x + \sin^2 x = 1$.

4.3.3 The Product Rule

This is also known as the Leibniz product rule, and was mentioned (albeit in an alternative notation) in the introduction to our course as one of the puzzles we would aim to demystify. The product rule is defined for the derivative of a product of differentiable functions $f(x)$ and $g(x)$:

$$\frac{d}{dx}\bigg(f(x)g(x)\bigg) = \frac{df}{dx}g(x) + f(x)\frac{dg}{dx}.$$

Proof:

$$\begin{aligned}
\frac{d}{dx}\bigg(f(x)g(x)\bigg) &= \lim_{h\to 0}\left[\frac{f(x+h)g(x+h) - f(x)g(x)}{h}\right] \\
&= \lim_{h\to 0}\left[\frac{f(x+h)g(x+h) - f(x)g(x+h) + f(x)g(x+h) - f(x)g(x)}{h}\right] \\
&= \lim_{h\to 0}\left[\frac{(f(x+h) - f(x))g(x+h) + f(x)(g(x+h) - g(x))}{h}\right] \\
&= \lim_{h\to 0}\left[\frac{(f(x+h) - f(x))g(x+h)}{h}\right] + \lim_{h\to 0}\left[\frac{f(x)(g(x+h) - g(x))}{h}\right] \\
&= \lim_{h\to 0}\left[\frac{(f(x+h) - f(x))}{h}\right]\lim_{h\to 0}\Big[g(x+h)\Big] + f(x)\lim_{h\to 0}\left[\frac{(g(x+h) - g(x))}{h}\right] \\
&= \frac{df}{dx}g(x) + f(x)\frac{dg}{dx}.
\end{aligned}$$

Exercise 4.4. *Use the product rule to find the derivatives of $f(x) = \sin^2 x$ and $g(x) = \cos^2 x$. Make sure your results agree with the calculation of the derivatives from first principles found in answering the previous exercise.*

Example 4.11. *Find the derivative with respect to x of the function $x^2 \sin x$ as a function of x.*

Using the product rule (on the product of functions x^2 and $\sin x$) we have

$$\frac{d}{dx}\bigg(x^2 \sin x\bigg) = \frac{d(x^2)}{dx}\sin x + x^2\frac{d(\sin x)}{dx} = 2x\sin x + x^2\cos x.$$

Example 4.12. *Find the derivative with respect to x of the function x^x as a function of x.*

We will insert a natural logarithm and an exponential to manipulate x^x into a form upon which we may employ the chain rule:

$$\frac{d}{dx}\bigg(x^x\bigg) = \frac{d}{dx}\bigg(e^{\ln(x^x)}\bigg) = \frac{d}{dx}\bigg(e^{x\ln(x)}\bigg).$$

Now we may use the chain rule (and the product rule) with $f(x) = e^x$ *and* $g(x) = x\ln(x)$*, so that* $f(g(x)) = e^{x\ln x}$*. Hence we have:*

$$\frac{d}{dx}\left(e^{x\ln(x)}\right) = e^{x\ln x}\frac{d}{dx}\left(x\ln x\right) = e^{x\ln x}(\ln x + 1) = x^x(1+\ln x).$$

4.3.4 The Quotient Rule

The quotient rule is a method for finding the derivative of a pair of differentiable functions $f(x)$ and $g(x)$ written as a quotient, or fraction, i.e. it is a method for computing

$$\frac{d}{dx}\left(\frac{f(x)}{g(x)}\right).$$

The quotient rule can be derived from the product rule for all points where $\frac{1}{g(x)}$ is well-defined. Instead of treating $f(x)/g(x)$ as a quotient we instead consider it as a product $f(x)\times(1/(g(x)))$, and employ the product rule and the chain rule. Explicitly we have:

$$\begin{aligned}\frac{d}{dx}\left(\frac{f(x)}{g(x)}\right) &= \frac{df}{dx}\frac{1}{g(x)} + f(x)\frac{d}{dx}\left(\frac{1}{g(x)}\right)\\ &= \frac{df}{dx}\frac{1}{g(x)} + f(x)\left(-\frac{1}{(g(x))^2}\right)\frac{dg}{dx}\\ &= \frac{1}{(g(x))^2}\left(\frac{df}{dx}g(x) - f(x)\frac{dg}{dx}\right).\end{aligned}$$

This is not so simple a formula to remember, so it is good news that it can be derived from the product rule and the chain rule together.

Exercise 4.5. *Check the validity of the quotient rule by first computing the derivative of* $\tan(x)$ *from first principles and comparing your result with the derivative* $\frac{d}{dx}\left(\frac{\sin x}{\cos x}\right)$ *computed using the quotient rule.*

4.4 Derivatives of Implicit Functions

The methods we have used so far to find derivatives are effective for functions which are defined, or can be defined, explicitly in terms of a variable, i.e. functions which may be written as $f(x)$. However it is often interesting to be able to find derivatives of functions whose definition are given in terms of their properties without an explicit function being available, or even possible in principle. The prime example of such a function that we have met earlier in the course is

the abstract definition of the inverse function, defined by $f(f^{-1}(x)) = x$, but there are other common implicit functions defined as curves or by parametric equations. The common quality that all implicit functions share is that they are defined by a relation rather than given an explicit functional definition in terms of a parameter. Let us consider these examples in turn and evaluate their derivatives.

4.4.1 Inverse Functions

Earlier in the course we saw that the inverse function $f^{-1}(x)$ of a function $f(x)$ is defined via $f(f^{-1}(x)) = x$. We will invoke the inverse function theorem without proof here (it is not part of this course, but it will prove invaluable in this section):

Theorem 4.2. *(Inverse Function Theorem) If f is a differentiable function with a non-zero derivative at the point x_0 then f is invertible in any neighbourhood of x_0 and f^{-1} is a differentiable function such that*

$$(f^{-1})'(x_0) = \frac{1}{f'(f^{-1}(x_0))}.$$

We will only make use of part of the theorem, but that part is the most interesting part, namely that the inverse function f^{-1} is a differentiable function if f is a differentiable function. By assuming these properties we will be able to find the derivative of the inverse function using the chain rule.

Assume that $f(x)$ is a differentiable function (hence by the inverse function theorem $f^{-1}(x)$ is also a differentiable function) then starting from the definition of the inverse function, and acting with the derivative on both sides of the equation we obtain:

$$\frac{d}{dx}\Big(f(f^{-1}(x))\Big) = \frac{d}{dx}\Big(x\Big) = 1.$$

Now, since both f and f^{-1} are differentiable functions we may use the chain rule to obtain:

$$\frac{d}{dx}\Big(f(f^{-1}(x))\Big) = \left(\frac{df}{df^{-1}}\right)\left(\frac{d(f^{-1}(x))}{dx}\right) = 1.$$

Hence,

$$\left(\frac{d(f^{-1})}{dx}\right) = \frac{1}{\left(\frac{df}{df^{-1}}\right)}$$

or using the primed notation to denote the derivative with respect to the argument we have

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

which is the statement in the last part of the inverse function theorem. The meaning of the above is best seen through some examples.

Example 4.13. *Find the derivative of* $\ln(x)$.

The natural logarithm is the inverse function of the exponential function $f(x) = e^x$ *and is defined for* $x > 0$. *The defining relation is:*

$$e^{\ln(x)} = x.$$

By taking the derivative of both sides of the equation above we have:

$$\frac{d}{dx}\Big(e^{\ln(x)}\Big) = e^{\ln(x)}\frac{d}{dx}\Big(\ln(x)\Big) = 1.$$

Hence,

$$\frac{d}{dx}\Big(\ln(x)\Big) = \frac{1}{e^{\ln(x)}} = \frac{1}{x}.$$

Example 4.14. *Find the derivative of* $\arcsin(x)$.

$\arcsin(x)$ *is the inverse function for* $f(x) = \sin(x)$ *and is defined for* $x \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. *The defining relation is:*

$$\sin\left(\arcsin(x)\right) = x.$$

By taking the derivative of both sides of the equation above we have:

$$\frac{d}{dx}\Big(\sin\left(\arcsin(x)\right)\Big) = \cos\left(\arcsin(x)\right)\frac{d}{dx}\Big(\arcsin(x)\Big) = 1.$$

Hence,

$$\frac{d}{dx}\Big(\arcsin(x)\Big) = \frac{1}{\cos\left(\arcsin(x)\right)} = \frac{1}{\sqrt{1-\sin^2(\arcsin(x))}} = \frac{1}{\sqrt{1-x^2}}.$$

Where we have made use of the identity $\cos^2(x) + \sin^2(x) = 1$.

Example 4.15. *Find the derivative of* $\arctan(x)$.

$\arctan(x)$ *is the inverse function for* $f(x) = \tan(x)$ *and is defined for* $x \in (-\frac{\pi}{2}, \frac{\pi}{2})$. *The defining relation is:*

$$\tan\left(\arctan(x)\right) = x.$$

By taking the derivative of both sides of the equation above we have:

$$\frac{d}{dx}\Big(\tan\left(\arctan(x)\right)\Big) = (1+\tan^2\left(\arctan(x)\right))\frac{d}{dx}\Big(\arctan(x)\Big) = (1+x^2)\frac{d}{dx}\Big(\arctan(x)\Big) = 1.$$

Hence,

$$\frac{d}{dx}\Big(\arctan(x)\Big) = \frac{1}{1+x^2}.$$

4.4.2 Curves

A curve, for the purpose of this course, means the set of solutions to equations defined in terms of x and y which are related by

$$R(x,y) = 0.$$

Typically the sets of points which satisfy the above relation form lines which may be curved (as the name suggests) or straight lines (as the name does not suggest). There is a difference between a relation being used to implicitly define a function as above and the idea that one can find an explicit function $f(x)$ to define the curve as $y = f(x)$. Usually one can only locally find a function $y = f(x)$ which will relate the y-coordinate of a point on a curve to its x-coordinate, but one can cover the curve with these local functions. Consider the example of the unit circle which is defined by the relation:

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

One can rearrange this equation to find two functions which cover the curve for different values of the coordinates, namely

$$y = \begin{cases} +\sqrt{1-x^2} & y \geq 0 \\ -\sqrt{1-x^2} & y < 0 \end{cases}$$

For more interesting curves the principle remains the same: one can locally find functions of the form $y = f(x)$ that are the same shape as the curve (locally). These local functions are the functions which are defined implicitly by $R(x,y) = 0$.

To find the gradient of a tangent to a curve, one can employ the chain rule and the defining relation, together with the notion that $y = f(x)$:

$$\frac{d}{dx}\Big(R(x,y)\Big) = \frac{d}{dx}\Big(R(x,f(x))\Big) = 0.$$

Let us consider some examples.

Example 4.16. *Find the derivative $\frac{dy}{dx}$ for the function $y(x)$ defined implicitly by the relation:*

$$R(x,y) \equiv y + \sin y - x = 0.$$

We treat y as a function of x, i.e. $y = y(x)$ and use the chain rule when taking the derivative of the relation with respect to x:

$$\frac{d}{dx}\Big(R(x,y)\Big) = \frac{d}{dx}\Big(y + \sin y - x\Big) = \frac{dy}{dx} + \cos y\frac{dy}{dx} - 1 = 0$$

which we may rearrange to find

$$\frac{dy}{dx} = \frac{1}{1+\cos y}.$$

This is an unusual result in our experience of derivative functions so far, the derivative function is given as a function of y itself - whereas we are used to obtaining the derivative function as a function of x. However it presents no serious problems for us: to find the derivative of y at $x = x_0$, we first must identify the corresponding point $y = y_0$ from the relation

$$y_0 + \sin y_0 - x_0 = 0$$

armed with the explicit point (x_0, y_0) which lies on $R(x,y) = 0$ we can then compute the gradient of the the tangent line to the curve at that point, i.e.

$$\frac{dy}{dx}(x_0) = \frac{1}{1+\cos y_0}.$$

For example at $x = 0$ we have $y = -sin(y)$ which is satisfied when $y = 0$, so the point $(0,0)$ lies on the curve and at that point the gradient of the tangent line is $\frac{dy}{dx}(0) = \frac{1}{1+\cos 0} = \frac{1}{2}$. As y increases we see that the gradient oscillates between being undefined (i.e. the curve becomes vertical) and $\frac{1}{2}$. So we expect this curve to wiggle its way through the point $(0,0)$ and a plot of the graph can confirm this. The sketch of the graph shown here in figure 4.4 was found by solving the differential equation for the gradient (and setting the constant to zero) to obtain $y(x)$.

Example 4.17. *Find the derivative $\frac{dy}{dx}$ for the function $y(x)$ defined implicitly by the relation:*

$$R(x,y) \equiv y - \tanh(xy) = 0.$$

We have,

$$\frac{d}{dx}\Big(R(x,y)\Big) = \frac{d}{dx}\Big(y - \tanh(xy)\Big) = \frac{dy}{dx} - \frac{1}{\cosh^2(xy)}\frac{d(xy)}{dx} = \frac{dy}{dx} - \frac{1}{\cosh^2(xy)}(y + x\frac{dy}{dx}) = 0$$

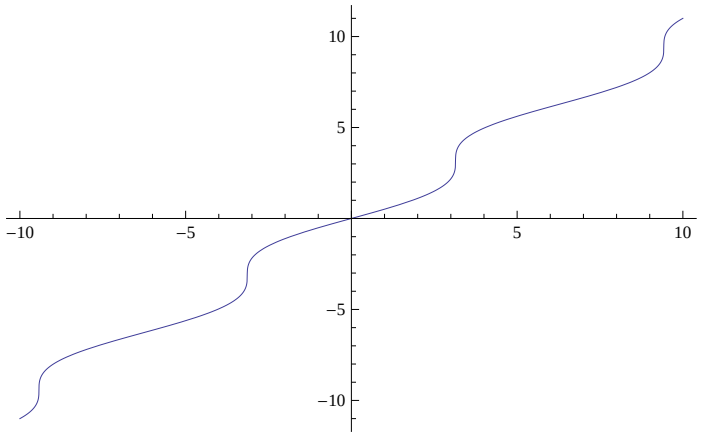


Figure 4.4: The graph of the curve $y(x)$ defined as the solution set of $y + \sin y - x = 0$.

which we may rearrange to find

$$\frac{dy}{dx} = \frac{y}{\cosh^2(xy) - x}.$$

This is a challenging graph to sketch! First we notice that when $y = 0$ *we have* $\frac{dy}{dx}(y = 0) = \frac{0}{1-x}$ *which equals zero apart from at the point* $x = 1$ *where it becomes undefined. For which values of* x *does* $y(x) = 0$*? We must solve* $0 = \tanh 0 \times x$ *which is true for all* x*. Hence we find one set of solutions is given by the* $x-$*axis whose gradient is zero. Let us turn our thoughts to the point* $x = 1$ *where the derivative was undefined: evidently* $y = \tanh(y)$ *is solved when* $y = 0$*, but this we already knew. So we may (correctly) wonder why the derivative is undefined there. Now as* $y = \tanh(xy)$ *then* $y \in (-1, +1)$*. Let us see if there are any points on the curve where* $y(x) = 1$*, this corresponds to* $\tanh(x) = 1$*, which is only true in the limit, i.e.* $\lim_{x\to\infty}(\tanh(x)) = 1$ *where we find*

$$\lim_{x\to\infty, y\to 1}\left(\frac{dy}{dx}\right) = \lim_{x\to\infty, y\to 1}\left(\frac{y}{\cosh^2(xy) - x}\right) = \lim_{x\to\infty}\left(\frac{1}{\cosh^2(x) - x}\right) = 0.$$

So the curve is tangential to $y = 1$ *as* $x \to +\infty$*. We can make a similar argument for* $y = -1$ *when* $x \to \infty$*, so* $y = -1$ *is also a tangent to the curve as* $x \to \infty$*. We have now three tangent lines to the curve as* $x \to \infty$*, given by* $y = \{-1, 0, 1\}$*, we can conclude that the curve has at least three parts to it as* x *grows. We may wonder if there are more branches to this curve? Consider* $y = \delta$ *where* $0 < \delta < 1$*, then we are aiming to solve the equation* $\delta = \tanh(x\delta)$ *which has a unique solution* $x = \frac{1}{\delta}\text{arctanh}(\delta)$*. Hence we now have the picture that there are three parts to the curve: a curve when* $y > 0$ *and* $x > 1$*, the line* $y = 0$ *and a curve when* $y < 0$ *and* $x > 0$ *(for the last curve consider* $y = \delta'$ *where* $-1 < \delta' < 0$*). Finally consider the gradient for the part of the curve where* $y > 0$ *and* $x > 1$*, at* $x \to \infty$ *the gradient approaches zero. We may*

consider the points near $x = 1$ *by substituting* $x = 1 + \epsilon$ *where* $\epsilon \geq 0$ *to find:*

$$\begin{aligned}
\lim_{y\to 0^+, x\to 1^+} \left(\frac{dy}{dx}\right) &= \lim_{y\to 0^+, \epsilon\to 0^+} \left(\frac{y}{\cosh^2((1+\epsilon)y) - (1+\epsilon)}\right) \\
&= \lim_{y\to 0^+, \epsilon\to 0^+} \left(\frac{y}{(1 + \frac{(1+\epsilon)^2 y^2}{2!} + \mathcal{O}(y^4))^2) - (1+\epsilon)}\right) \\
&= \lim_{y\to 0^+} \left(\frac{y}{y^2 + \mathcal{O}(y^4)}\right) \\
&= \infty.
\end{aligned}$$

Hence the upper curve becomes vertical as it approaches the point $(x = 1, y = 0)$*. A similar argument can be made for the curve when* $y < 0$*. Eventually these lengthy observations can allow us to sketch the curve and we show the graph in figure 4.5.*

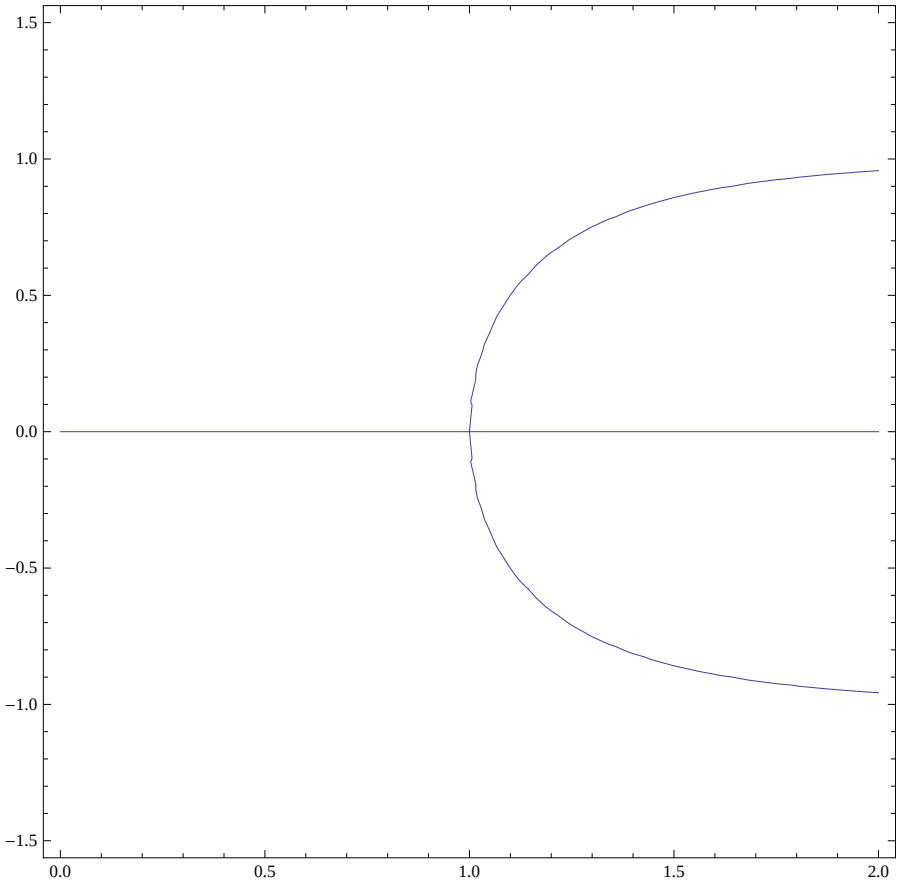


Figure 4.5: The graph of the curve $y(x)$ defined as the solution set of $y - \tanh(yx) = 0$.

4.4.3 Parametric Functions

Given two functions $x(t)$ and $y(t)$ for $t \in \mathbb{R}$ we can define a curve as the set of points $(x(t), y(t))$ which are sketched out as t varies. In other words, t is a single coordinate that defines a point on a curve: one might like to think of t as a distance along a curve as measured from the point on the curve given by $t = 0$. As one moves along the curve the value of t changes, and t is said to parameterise the curve.

One can easily rewrite common curves with a parametric definition, e.g. the parabola $y = x^2$ consists of the points (t, t^2) parameterised by t, or the straight line $y = mx+c$ is the set of points $(t, mt + c)$. The choice of parameterisation is not unique: one can always shift t by a constant t_0 so as to move the point corresponding to $t = 0$ along the curve: if $t' \equiv t + t_0$ then $t = t' - t_0$ and the parameterisation of the straight line becomes $(t, mt+c) = (t'-t_0, mt'-mt_0+c)$, and if we picked the value of $t_0 = \frac{c}{m}$ then the line becomes the points $(t' - \frac{c}{m}, mt')$ now parameterised by t': one can move the origin in t ($t = 0$) to wherever one wishes on a parametric curve in this way, which can be very useful.

Now, given $(x(t), y(t))$ we may wonder if $y(x)$ exists. Suppose that $x = f(t)$, if f^{-1} exists and is well-defined we have $t = f^{-1}(x)$, hence we would then have $y(t) = y(f^{-1}(x)) \equiv y(x(t))$. So if $y(x(t))$ exists then by the chain rule we have:

$$\frac{dy}{dt} = \left(\frac{dy}{dx}\right)\left(\frac{dx}{dt}\right)$$

and hence

$$\frac{dy}{dx} = \frac{\left(\frac{dy}{dt}\right)}{\left(\frac{dx}{dt}\right)}.$$

Now as $x(t)$ and $y(t)$ are differentiable functions we may compute directly $\frac{dx}{dt}$ and $\frac{dy}{dt}$, and combine to obtain $\frac{dy}{dx}$, so long as $\frac{dx}{dt} \neq 0$.

Example 4.18. *Find the derivative function $\frac{dy}{dx}$ for the curve defined parametrically by*

$$x(t) = t + \cos(t)$$
$$y(t) = \ln(\cosh(\sin(t)))$$

where $t \in \mathbb{R}$.

We may compute the following derivatives immediately:

$$\frac{dx}{dt} = 1 - \sin(t)$$
$$\frac{dy}{dt} = \frac{1}{\cosh(\sin(t))}\frac{d}{dt}\Big(\cosh(\sin(t))\Big) = \frac{\sinh(\sin(t))}{\cosh(\sin(t))}\frac{d}{dt}\Big(\sin(t)\Big) = \tanh(\sin(t))\cos(t)$$

Hence we see that the derivative is not defined when $t = \frac{\pi}{2} + n2\pi$ for $n \in \mathbb{Z}$, but is defined elsewhere. We have:

$$\frac{dy}{dx} = \frac{\left(\frac{dy}{dt}\right)}{\left(\frac{dx}{dt}\right)} = \frac{\tanh(\sin(t))\cos(t)}{1 - \sin(t)}.$$

A sketch of the parametric curve is shown in figure 4.6 where it can be observed that the derivative is not well-defined at $x = \frac{\pi}{2} + n2\pi$.

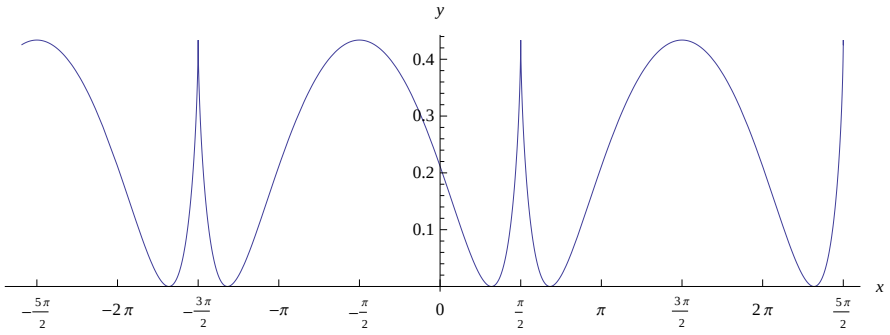


Figure 4.6: The graph of the curve $y(x)$ defined parametrically as the points $(t + \cos(t), \ln(\cosh(\sin(t)))\,)$ for $t \in \mathbb{R}$.

Example 4.19. *Find the derivative function $\frac{dy}{dx}$ for the curve defined parametrically by*

$$x(t) = e^t$$
$$y(t) = \tan(t)$$

where $t \in \mathbb{R}$.

Our preliminary observation is that $x(t) > 0$ so the curve is defined only for $x \in \mathbb{R}^+$. We may compute the following derivatives immediately:

$$\frac{dx}{dt} = e^t = x$$
$$\frac{dy}{dt} = 1 + \tan^2(t) = 1 + y^2$$

Hence we see that the derivative is not defined when $x = 0$ but is defined elsewhere. We have:

$$\frac{dy}{dx} = \frac{\left(\frac{dy}{dt}\right)}{\left(\frac{dx}{dt}\right)} = \frac{1+y^2}{x}.$$

A sketch of this parametric curve is shown in figure 4.7.

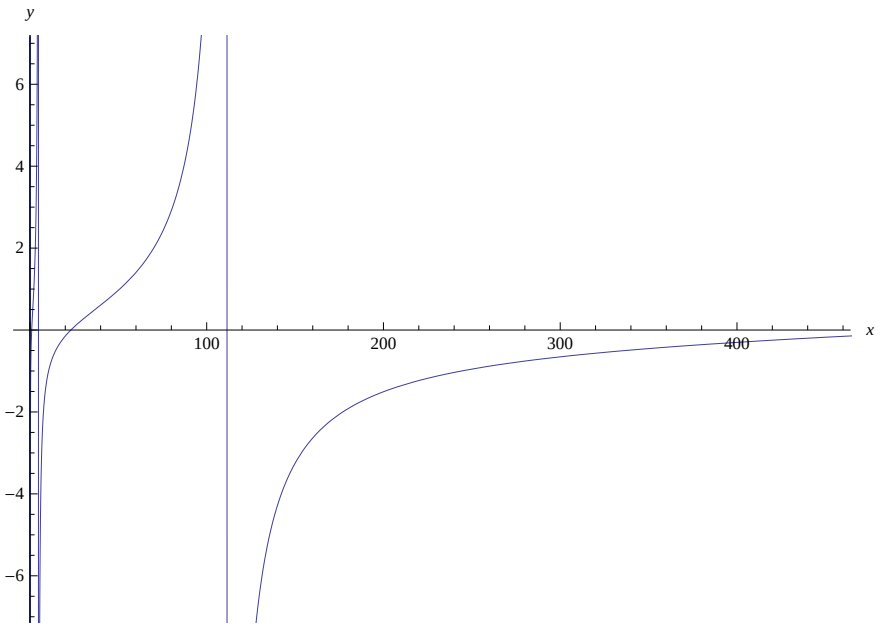


Figure 4.7: The graph of the curve $y(x)$ defined parametrically as the points $(e^t, \tan(t))$) for $t \in [-8, 8]$.

4.5 The Mean Value Theorem

Suppose that we know a lot of information about the derivatives of a function, but we don't know much about the function at all. In the last few sections we saw some examples of how this may come about by defining curves via an equation or defining them parametrically, but in these examples we did in principle have a definition of the function whose derivative we were trying to compute and it just happened to be simpler to work with the derivative due to the way the function was defined. Here we are imagining that we have a table of data defining the derivative function in detail and we are wondering if we can use this data to reconstruct the function. A real world example of this occurs in speed traps for cars, where a car's position is measured at two times and from this information it can be deduced whether the car broke the

speed limit at any point between the two measurement points. All this rests upon the mean value theorem.

Theorem 4.3. *(The Mean Value Theorem) Let f be a continuous function on $[a, b]$ and a differentiable function on (a, b), then there exists a point $c \in (a, b)$ where*

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

The mean value theorem says that at some point c the slope $\frac{f(b)-f(a)}{b-a}$ of the straight line connecting $f(a)$ to $f(b)$ is actually equal to the derivative of f at c. It is helpful to think what this means when the function is the position function and the derivative is taken with respect to time. In this case the mean value theorem tells us that for a journey from $f(a)$ to $f(b)$ then your average speed $\frac{f(b)-f(a)}{b-a}$ is equal to your actual instantaneous speed at least one moment, time c. Let us try some numbers and stick with the setting of the distance, time and speed. Suppose that it takes you 1 hour to travel 60 miles, then your average speed for the journey is 60 miles per hour (mph). The mean value theorem tells you that at least for one moments in the journey you were travelling at 60 mph. Of course you might have travelled the route in a number of different ways, e.g. you may have travelled at 60 mph for the entire journey, or you may have travelled part of the journey at a speed greater than 60 mph and another at a speed less than 60 mph (so that the average speed is still 60 mph), but in that case your speed would have to pass through 60 mph at some point. This is obviously an unpleasant realisation for an apprentice analyst wishing to break the speed limit. Such a mathematician might have thought to argue in court against the evidence of a speed trap by imagining that there existed a position function whose derivative nowhere exceeded the speed limit, but the mean value theorem tells her or him that no such function exists.

Let us return to abstract functions and illustrate the mean value theorem on a graph. In figure 4.8 we see a curve $y = f(x)$ and construct the straight line which passes through the points $f(a)$ and $f(b)$. The slope of this straight line is $\frac{f(b)-f(a)}{b-a}$ and the mean value theorem tells us that this straight line is tangent to the curve $y = f(x)$ for at least one point $x = c$ where $c \in (a, b)$, in figure 4.8 we illustrate this by translating the line passing through $f(a)$ and $f(b)$ until it is tangent to $y = f(x)$. For the curve we have sketched there is at least one other point on $y = f(x)$ whose tangent has the same slope.

A very useful example of the mean value theorem occurs when $f(a) = f(b)$ as in this case we learn that there exists some c such that

$$f'(c) = \frac{f(b) - f(a)}{b - a} = \frac{f(a) - f(a)}{b - a} = 0$$

i.e. there exists a stationary point at $x = c$. This example of the mean value theorem is

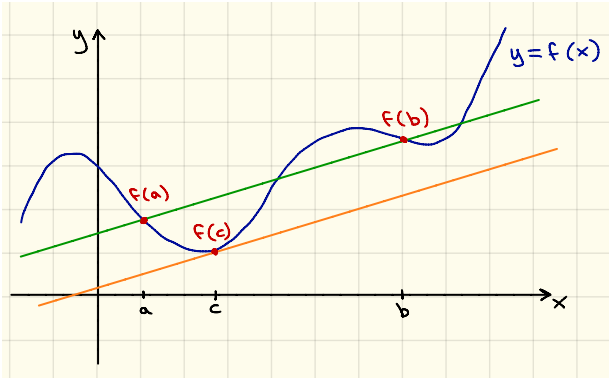


Figure 4.8: The mean value theorem illustrated on the curve $y = f(x)$.

important enough to have its own name, it is called Rolle's theorem after Michel Rolle who first proved this in 1691.

We commenced this section by wondering whether we can construct a function from knowledge of its derivatives. The mean value theorem tells us about at least one derivative on the open interval (a, b), so how might we use this to construct the function on the open interval? There is a very useful theorem that allows us to reconstruct certain functions: the constant functions.

Theorem 4.4. *If $f(x)$ is a differentiable function on an open interval such that*

$$f'(x) = 0$$

for all x then $f(x)$ is constant on the interval.

Proof: We may consider two points a and b defined on the open interval and then we have, by the mean value theorem, that

$$\frac{f(b) - f(a)}{b - a} = f'(c)$$

where $c \in (a, b)$. However we know (from the statement of the theorem) that the function's derivatives are all zero on the open interval, which includes (a, b) hence $f'(c) = 0$. Therefore we have

$$f(b) - f(a) = 0$$

for all a and b, hence $f(a) = f(b)$ for any choice of a and b in the open interval where the derivative vanishes, hence it is a constant function.

In this chapter we developed a definition of the derivative function and studied its properties. The derivative is the instantaneous slope of a function and is defined using the limit. We

developed some properties of the derivative, most importantly the chain rule and the product rule. It was proved that every differentiable function is also continuous and we saw how we can use the derivative on functions defined implicitly, via an inverse or curves defined parametrically and we had some practise sketching functions with the aid of the derivative function. The final section of this chapter presented the mean value theorem and we saw how the vanishing derivative on an open interval implied that the function must be constant on the open interval. We began our final comments by speculating whether we could commence with knowledge of a derivative function and use it to reconstruct the function. Such an operation would be the inverse of the derivative, it would be the process of finding the antiderivative. In the following chapter we will see how the antiderivative is related to the integral which gives a method for finding the area under a curve.

5. Integration

In which we meet the integral: a way of turning a sum of infinitely many infinitesimally thin parallelograms into finite number which gives the area under a curve. Astonishingly this will be related by the fundamental theorem of calculus to the antiderivative of a function!

The material in this chapter will be covered in weeks 9-11.

Figure 5.1: Isaac Newton (top) in 1689 and Gottfried Leibniz (below).

Integration was developed almost independently by both Isaac Newton (1643-1726) and Gottfried Leibniz (1646-1716) in the late 17$^{\text{th}}$ century. The argument over who had priority in the invention of calculus is one of the most famous disputes over the ownership of an idea in all of science: the passion shown by these two intellectual giants over priority rather than joy in their shared legacy of a great discovery leaves a bitter aftertaste. While Newton made his discoveries first he had kept them secret, on the other hand it seems indisputable that Leibniz had found partial inspiration from Newton's wider work. It is put succinctly in James Gleick's biography "Isaac Newton":

> "Newton had made his discoveries first, and he had discovered more, but Leibniz had done what Newton had not: published his work for the world to use and to judge."

It is widely accepted that both men discovered the calculus independently but Newton did his reputation no favours during the dispute. When a committee of the Royal Society gave its report on the argument it said, in the words of Gleick, that:

> "It judged Newton's method to be not only the first - "by many years" - but also more elegant, more natural, more geometrical, more useful and more certain."

The President of the Royal Society at the time the report was made and also the secret author of the report was... Isaac Newton.

5.0.1 Motion at constant speed: the area under a straight line

To intuit that the area under a curve and the operation giving the gradient may be related it is worth calling to mind the dynamical examples that inspired Newton. If we sketch the graph of the position of a particle $x(t)$ which moves at constant speed $\frac{dx}{dt}$ we have a simple straight line (its tangent line at every point has the same, constant slope): The velocity function for this

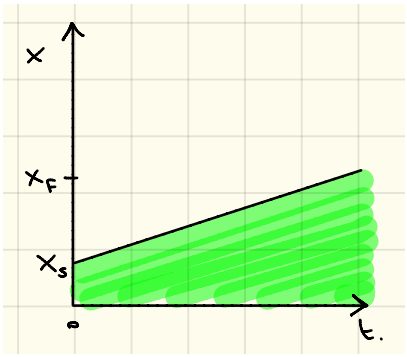


Figure 5.2: The function $x(t)$ where $x'(t)$ is constant. The area under the graph as the position changes from x_S to x_F is shaded, badly, but not particularly dynamically interesting.

motion is a constant function and we sketch it below. The area under the velocity function is

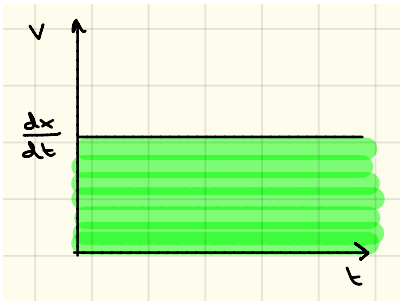


Figure 5.3: The function $v(t)$ is constant. The area under the graph as the position changes from x_S to x_F is shaded.

dynamically interesting, as the speed for this example is constant, we can use the formula for the average speed as the particle moves from x_S at time t_S to x_F at time t_F is

$$\frac{dx}{dt} \times (t_F - t_S) = x_F - x_S.$$

For this simple motion we see that from the graph of the velocity, $\frac{dx}{dt}$, we can obtain information related to the graph of the position function $x(t)$ rather than just $v(t)$.

We have considered a very simple straight line function, but for infinitesimal domains any continuous function is approximately a straight line, so we may expect this feature to continue, that integration (the process of computing the area under a graph) is roughly the inverse procedure to taking the derivative, that is,

$$x(t) \xrightarrow{\frac{d}{dt}} v(t) \qquad \text{and} \qquad x(t) \overset{\text{Integration}}{\longleftarrow} v(t).$$

That integration is the inverse of differentiation is remarkable and is the fundamental theorem of calculus - we will return to this later in this chapter. First we must formalise and develop integration.

5.1 The Riemann Integral

The integral of a function is up to a sign the area "under" a graph of the function. Let us look at the fine-print in the definition of the integral.

Definition 5.1.1. *The integral, denoted $\int_a^b f(x)dx$ is the total area between the graph of $y = f(x)$ and the x-axis in the x, y plane between $x = a$ and $x = b$, counted positively for $f(x) > 0$ (area above the x-axis) and negatively for $f(x) < 0$ (area below the x-axis).*

Figure 5.4: Bernhard Riemann in 1863.

But how does one practically go about computing $\int_a^b f(x)dx$? Perhaps surprisingly, a rigorous definition of the integral was not given until much later than the fundamental theorem of calculus and the surrounding mathematics had been established by Newton and Leibniz. It was not until 1854 that the first rigorous definition of the integral was written down by Bernhard Riemann (1826-1866)[1]. The idea behind the integral is not challenging, namely that one can split the area underneath a curve into a sequence of rectangles and that one can approximate the height of the rectangles to the curve, although the widths of the rectangles will typically become infinitesimal... The integral of a function calculated in this way is called the Riemann integral.

First let us consider a graph which is precisely a sequence of rectangles. Such a graph is called (for obvious reasons) a 'staircase graph' or 'staircase function', let us consider one such arbitrary staircase function as shown in figure 5.5. The staircase function is defined by

[1]Note how young Riemann was at the time of his death. Mathematics surely lost one of its finest minds too young. After his death his housekeeper threw out many of his papers, which Riemann was not content to publish, but which probably contained many correct insights and results.

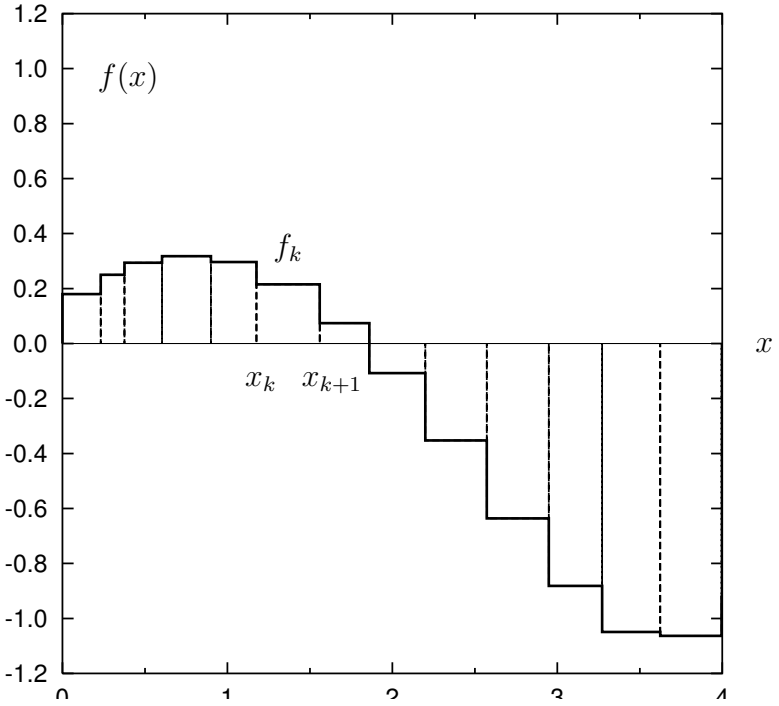


Figure 5.5: A staircase function: let the vertical sides of the rectangles sit at $x_1 < x_2 < x_3 < \ldots < x_n < x_{n+1}$ and let the height of the k'th rectangle by f_k.

$$f(x) = f_k \quad \text{if } x \in [x_k, x_{k+1})$$

where f_k is a sequence of real numbers (not specified here, although it is possible to read off some values of f_k from the graph). Now the area of the k'th rectangle is its width times its height or

$$(x_{k+1} - x_k) f_i$$

and notice that the sign convention agrees with that of the integral: if $f_k < 0$ then the "area" is counted negatively. This unusual function is chosen solely because we can compute its integral exactly:

$$\int_{x_1}^{x_{n+1}} f(x)dx = \sum_{k=1}^{n} (x_{k+1} - x_k) f_k.$$

On the left-hand-side we have the notation denoting the integral, while on the right-hand-side we have been able to write an expression for the sum of the areas of the n rectangles under the staircase function between $x = x_1$ and $x = x_{n+1}$. This exact integral is the foundation upon which all of our Riemann integrals will be constructed.

In the previous example we considered a staircase function and showed that we could compute its integral exactly. This was because the staircase function's special shape allowed it to be partitioned into rectangles whose areas could be rapidly evaluated. We would like to be able to find the integral of an arbitrary curve, not only staircase functions, and what we will do is to try to come up with a method of approximating curves to staircase functions. What we will achieve ultimately will be a way of sandwiching the integral of any function between two staircase functions, such that in a limit we can compute the integral. We will worry what this means soon enough, first let us try to see how we might construct staircase functions which closely approximate an arbitrary continuous function.

So how would we approximate a continuous function to a staircase function? What we would like is a staircase whose steps are the same height as the function. But as the function is continuous it cannot have any jumps in values as occurs for the staircase function when one moves between steps. But what if we made the width of the steps infinitesimally thin? If we could do this then we could approximate a continuous function by a staircase function whose steps have infinitesimal width and whose heights are equal to the function. This will be our goal, but first let us think a little about some staircase functions which begin to approximate an arbitrary continuous function.

Imagine writing an algorithm to associate a staircase function to a curve: it would be difficult to know where to begin, indeed there are many possible beginnings. There are a number of decisions to make: how wide should the steps be? Should the step width vary or be constant? How do decide the height of each step so that it is related to the function? For example should the step height be the value of the function at the mid-point of the step width (as depicted in figure 5.6), or it could be given by the top left-hand-corner of each step, or the right-hand corner or indeed any point on the step could be put equal to the function at the same point.

How will we find a unique way to associate a staircase function to a continuous curve? Well, the answer is that for steps of finite width we will not find a unique staircase function, but once we take the limit so that the steps go to infinitesimal width then all staircase functions will approach the continuous function. The algorithm we will follow is:

(i) consider all possible staircase functions where each step touches the curve we are interested in (typically this will be an infinite set of staircase functions),

(ii) find the area under each staircase graph in the limit where the width of all steps goes to zero and

(iii) check whether all these areas for the limits of the variety of staircase functions give the same area.

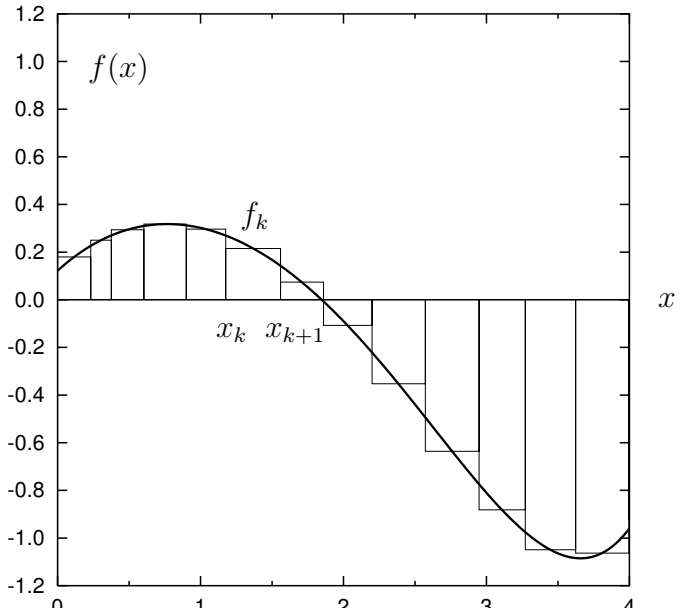


Figure 5.6: An example of a staircase function (split into rectangular strips) for which the mid-point of the step is equal to the function, note the varying width of the steps.

The method proposed is tedious and, potentially, will take an infinite time to actually carry out but since such a method may exist it gives us courage that, for a continuous function $f(x)$ defined on the interval $[a, b]$ the integral $\int_a^b f(x)\,dx$ exists. This encapsulates the formal definition of the Riemann integral:

Definition 5.1.2. *The (Riemann) integral of a function $f(x)$ on the interval $[a, b]$, if it exists, is equal to the sum of the areas of the steps of any staircase function $S(x)$ in the limit that the step width, w, approaches zero so long as $\lim_{w\to 0}(S(x)) = f(x)$.*

Comment(s). *(On the Riemann integral...) Notice that we have generalised the discussion and the Riemann integral is defined for any function, not just continuous functions in this definition and we have introduced the idea that the integral might not exist for general functions. As we have seen earlier in the course the limit does not always exist and it is therefore possible for the integral of a function to not be defined. For example, think of integrating $1/x$ on the interval $[0, b]$ where $b > 0$, from the graph we know that the integral will be infinite and so not exist. The short story (but not the full story) is that if a function is continuous on the interval then the Riemann integral exists and the function is said to be integrable on the interval.*

To construct the Riemann integral we will first notice that out of the infinite set of staircase

functions which we may associate with a function we are trying to integrate, there are two classes of staircase function which stand out:

- $S^-(x)$ those which have the minimum area (for a given, not necessarily constant, set of step-widths) and
- $S^+(x)$ those which have the maximum area (for a given, not necessarily constant, set of step-widths).

Before defining these functions let us consider the graphs in figure 5.7 showing examples of $S^-(x)$ and $S^+(x)$ for an arbitrary continuous function $f(x)$. Notice that, when $f(x) > 0$, the

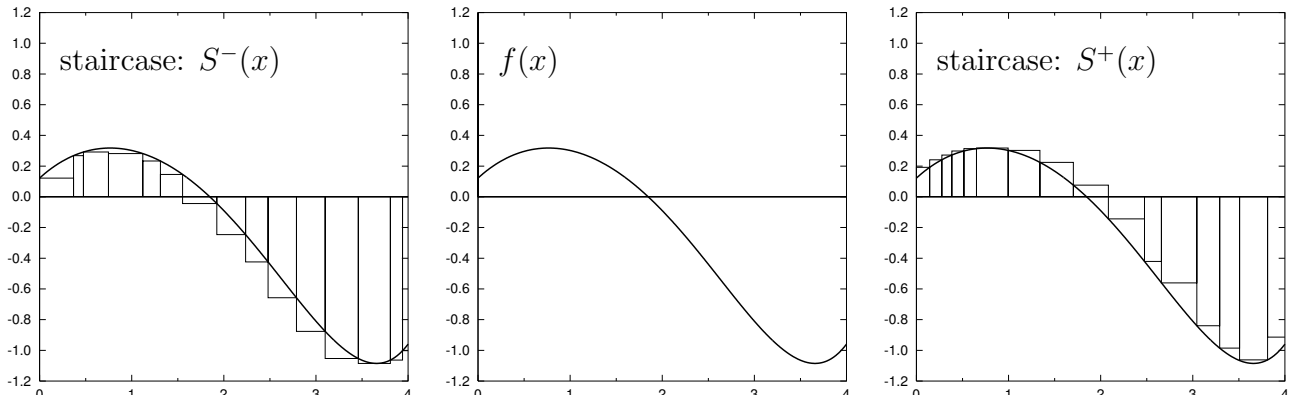


Figure 5.7: Staircase functions constructed for the function $f(x)$ (shown in the centre and repeated in the diagrams on the left and on the right). The area of the staircase function $S^-(x)$ underestimates the integral of $f(x)$, while the area of $S^+(x)$ overestimates the same integral.

area of each rectangular step in the staircase function $S^-(x)$ is less than the area under the graph evaluated on the interval covered by the step's width. When $f(x) < 0$ the integral of each step in $S^-(x)$ is more negative (less than) that of $f(x)$ on the same interval. Now consider $S^+(x)$ and notice that the integral of each step in the staircase function is always greater than the integral of $f(x)$. In other words the staircase functions satisfy the relation

$$S^-(x) \leq f(x) \leq S^+(x) \qquad \forall\, x \text{ in the domain of } f(x).$$

Practically how do we define these staircase functions? It is a simple matter of finding a definition that agrees with the inequality given above. Recall from our first use of the staircase function that a staircase function $S(x)$ is defined by specifying the height of its steps on the intervals $[x_k, x_{k+1})$, such that $x_1 < x_2 < x_3 < \ldots < x_n < x_{n+1}$ are the x coordinates of each

side of the n steps. We will consider the integral over the interval $x \in [a, b]$ so $x_1 = a$ and $x_{n+1} = b$. Hence we can define $S^-(x)$ and $S^+(x)$ for a given function $f(x)$ as follows:

$$S^-(x) = S_k^- \equiv \min[f(x)] \qquad \forall\, x \in [x_k, x_{k+1})$$
$$S^+(x) = S_k^+ \equiv \max[f(x)] \qquad \forall\, x \in [x_k, x_{k+1}).$$

In words the $k'th$ step of the staircase function has constant height over the range $x \in [x_k, x_{k+1})$. We pick the height of each step of $S^-(x)$ to just touch $f(x)$ at its minimum point, so it will always have an integral less than that of $f(x)$, while $S^+(x)$ has steps which just touch $f(x)$ at its maximum point in the width of the step, so the integral of $S^+(x)$ is always greater than the integral $f(x)$. Notice that the width of each step has not been fixed in any way, the steps have arbitrary width, so there still remain an infinite set of choices for $S^-(x)$ and $S^+(x)$. However we are now sure that:

$$A^- \equiv \sum_{k=1}^{n} S_k^-(x_{k+1} - x_k) = \int_a^b S^-(x)dx \leq \int_a^b f(x)dx \leq \int_a^b S^+(x)dx = \sum_{k=1}^{n} S_k^+(x_{k+1} - x_k) \equiv A^+.$$

We have "sandwiched" the integral of $f(x)$ between the integrals for the two staircase functions which we are able to evaluate as summations. The idea now is to use the sandwich theorem. If we can take a limit so that $S^+(x)$ and $S^-(x)$ both approach $f(x)$ then if the limits of their integrals exist and are equal then that limiting value is the value of the integral of $f(x)$. The real question is what is the limit that we should take? By now we should expect to take a limit such that the width of the steps in $S^{\pm}(x)$ become infinitesimal. But of course if we simply make the step widths thinner without increasing the number of steps then the staircase functions will no longer cover the domain of the function. So we must do two things at once when we take the limit:

- the width of each step must become infinitesimal i.e. $(x_{k+1} - x_k) \to 0$ and
- the number of steps must approach infinity i.e. $n \to \infty$.

How do we know that the number of steps must grow infinitely large? Well we are interested in the integral over $x \in [a, b]$, hence the width of the staircase functions in total must remain $b - a$ as we take the limit, i.e. we need $b - a = \sum_{k=1}^{n+1}(x_{k+1} - x_k)$ and so as $(x_{k+1} - x_k)$ shrinks, so the number of steps n must increase as $b - a$ is constant.

You might think of this limit as moving through the space of staircase functions $S^{\pm}(x)$ to arrive at the pair of staircase functions whose steps are infinitesimally thin. For example pictorially for $S^-(x)$ the process of taking this limit and moving through the staircase functions is shown in figure 5.8. If we now take this (abstract) limit on the inequality which sandwiches

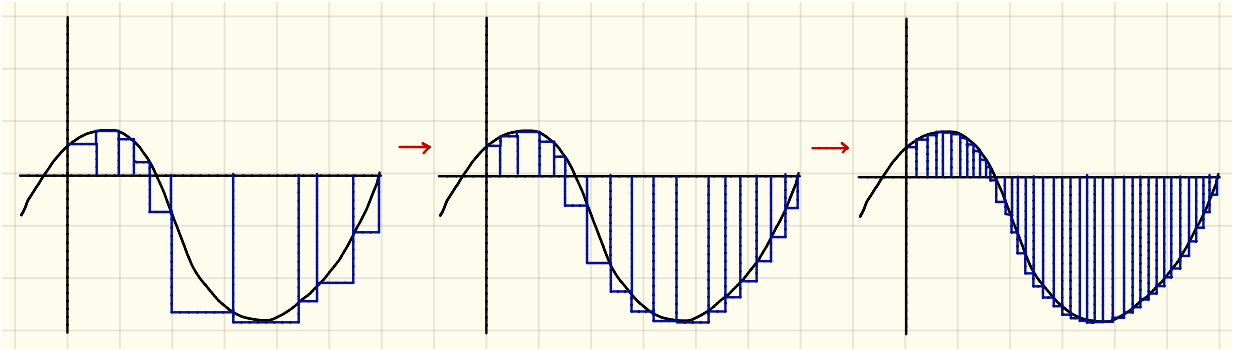


Figure 5.8: The width of each step in the staircase function $S^-(x)$ is reduced and the number of steps is increased, so that the integral of $S^-(x)$ approaches from below that of $f(x)$.

$\int_a^b f(x)dx$ we have:

$$\lim_{(x_{k+1}-x_k)\to 0, n\to\infty}[A^-] \leq \int_a^b f(x)dx \leq \lim_{(x_{k+1}-x_k)\to 0, n\to\infty}[A^+].$$

Now if the two limits tend to the same value, i.e.

$$\lim_{(x_{k+1}-x_k)\to 0, n\to\infty}[A^-] = A \qquad \text{and} \qquad \lim_{(x_{k+1}-x_k)\to 0, n\to\infty}[A^+] = A$$

then we conclude that $\int_a^b f(x)dx = A$. This is the process by which we can construct and evaluate the Riemann integral of a function, if it exists, but how hard is it to apply this construction in practise?

Example 5.1. *Let*

$$\int_0^b \cos(x)dx = A \qquad \text{for } b \leq \pi.$$

Construct the staircase functions and take the appropriate limit to evaluate the Riemann integral and find A.

For $x \in [0, b]$, the cosine function decreases monotonically (i.e. if $x, x' \in [0, b]$ and $x' > x$ then $\cos(x') < \cos(x)$). Now we may construct the staircase functions $S^{\pm}(x)$ via

$$S_k^- = \min[\cos(x)] \qquad \text{for } x \in [x_k, x_{k+1}) \implies S_k^- = \cos(x_{k+1})$$
$$S_k^+ = \max[\cos(x)] \qquad \text{for } x \in [x_k, x_{k+1}) \implies S_k^+ = \cos(x_k).$$

Now, for convenience, we may choose steps of equal width, w, with $x_1 = 0$ and $x_{n+1} = b$ so that as

$$b = x_{n+1} = nw \qquad \text{then} \qquad w = \frac{b}{n}.$$

We can also find expressions for each of the x-coordinates of the sides of the steps, x_k,

$$x_k = x_1 + (k-1)w = (k-1)w = \frac{b(k-1)}{n}$$

i.e. $x_1 = 0$, $x_2 = w$, $x_3 = 2w, \ldots$, $x_n = (n-1)w$ *and* $x_{n+1} = nw = b$. *Hence in this notation we have,*

$$A^- = \sum_{k=1}^{n} S_k^-(x_{k+1} - x_k) = \sum_{k=1}^{n} w\cos(x_{k+1}) = w\sum_{k=1}^{n}\cos(kw).$$

The limit we wish to take involves simultaneously taking $n \to \infty$ and $w \to 0$, which will be problematic unless we can rewrite the summation, in order to move n from the summation and into an expression where we can take the limit.

Exercise 5.1. *Show that*

$$\sum_{k=0}^{n}\cos(k\theta) = \frac{1 - \cos((n+1)\theta) - \cos\theta + \cos(n\theta)}{2 - 2\cos\theta}.$$

[Hint: First prove that $\sum_{k=0}^{n} z^k = \frac{1-z^{n+1}}{1-z}$ then substitute $z = e^{i\theta} = \cos(\theta) + i\sin(\theta)$ into this result.]

Hence we have,

$$\begin{aligned}
A^- &= w\sum_{k=1}^{n}\cos(kw) \\
&= w(\sum_{k=0}^{n}\cos(kw) - 1) \\
&= w\left(\frac{1 - \cos((n+1)w) - \cos(w) + \cos(nw)}{2 - 2\cos(w)} - 1\right) \\
&= w\left(\frac{-1 - \cos((n+1)w) + \cos(w) + \cos(nw)}{2 - 2\cos(w)}\right) \\
&= w\left(\frac{-1 - \cos(b+w) + \cos(w) + \cos(b)}{2 - 2\cos(w)}\right) \\
&= w\left(\frac{-1 - \cos(b)\cos(w) + \sin(b)\sin(w) + \cos(w) + \cos(b)}{2 - 2\cos(w)}\right)
\end{aligned}$$

where we have used $b = nw$ to eliminate n from the expression. Hence we may now attempt to

take the limit $w \to 0$*:*

$$\begin{aligned}
\lim_{w\to 0}[A^-] &= \lim_{w\to 0}\left[w\left(\frac{-1-\cos(b)\cos(w)+\sin(b)\sin(w)+\cos(w)+\cos(b)}{2-2\cos(w)}\right)\right] \\
&= \frac{1}{2}\lim_{w\to 0}\left[w\left(\frac{(-1+\cos(b))(1-\cos(w))+\sin(b)\sin(w)}{1-\cos(w)}\right)\right] \\
&= \frac{1}{2}\lim_{w\to 0}\left[w(-1+\cos(b))\right] + \frac{\sin(b)}{2}\lim_{w\to 0}\left[\frac{w\sin(w)}{1-\cos(w)}\right] \\
&= \frac{\sin(b)}{2}\lim_{w\to 0}\left[\frac{w(w-\mathcal{O}(w^3))}{1-(1-\frac{w^2}{2!}+\mathcal{O}(w^4))}\right] \\
&= \frac{\sin(b)}{2}\lim_{w\to 0}\left[\frac{1-\mathcal{O}(w))}{\frac{1}{2!}-\mathcal{O}(w^2))}\right] \\
&= \sin(b).
\end{aligned}$$

For A^+ *we have*

$$A^+ = \sum_{k=1}^{n} S_k^+(x_{k+1}-x_k) = \sum_{k=1}^{n} w\cos(x_k) = w\sum_{k=1}^{n}\cos((k-1)w).$$

Hence

$$\begin{aligned}
A^+ &= w\sum_{k=1}^{n}\cos((k-1)w) \\
&= w\sum_{\ell=0}^{n-1}\cos(\ell w) \\
&= w(1+\sum_{\ell=1}^{n}\cos(\ell w)-\cos(nw)) \\
&= w(1-\cos(b))+A^-
\end{aligned}$$

where we used $\ell \equiv k-1$ *in writing the second line above. Hence we have,*

$$\lim_{w\to 0}\left[A^+\right] = \lim_{w\to 0}\left[w(1-\cos(b))\right] + \lim_{w\to 0}\left[A^-\right] = \lim_{w\to 0}\left[A^-\right] = \sin(b).$$

Therefore as we have

$$\sin(b) = \lim_{w\to 0}\left[A^-\right] \le \int_0^b \cos(x)dx \le \lim_{w\to 0}\left[A^+\right] = \sin(b)$$

then

$$\int_0^b \cos(x)dx = \sin(b).$$

Exercise 5.2. *The construction of the staircase functions in the example above rested on the observation that* $\cos(x)$ *is monotonically decreasing for* $x \in [0, b]$ *where* $0 \leq b \leq \pi$*. Outline the procedure for showing that*

$$\int_0^b \cos(x)dx = \sin(b)$$

for $b > \pi$*.*

Exercise 5.3. *Construct the staircase functions one would use to evaluate*

$$\int_0^b \sin(x)dx$$

first for $0 \leq b \leq \pi$ *and more generally when* $b > 0$*.*

Example 5.2. *Use staircase functions and the sandwich theorem to evaluate*

$$A = \int_a^b x^m dx$$

where $0 < a < b$ *and* $m \in \mathbb{Z}^+$*.*

We choose the bounding staircase functions as follows:

$$\begin{aligned} S_k^- &= \min[x^m] \qquad \text{for } x \in [x_k, x_{k+1}) \implies S_i^- = x_k^m \\ S_k^+ &= \max[x^m] \qquad \text{for } x \in [x_k, x_{k+1}) \implies S_k^+ = x_{k+1}^m. \end{aligned}$$

In this example we will follow a seemingly more flamboyant path to evaluate the integral, namely we will consider steps of varying size which grow as a geometric progression in w*. Given that* $x_1 = a$ *and* $x_{n+1} = b$ *then we define the step sides to lie at*

$$x_1 = a,\ x_2 = aw\ x_3 = aw^2,\ \ldots x_k = aw^{k-1},\ \ldots x_{n+1} = aw^n = b.$$

Therefore we have

$$w = \left(\frac{b}{a}\right)^{\frac{1}{n}} \qquad \implies \qquad x_k = a\left(\frac{b}{a}\right)^{\frac{k-1}{n}}.$$

Let us make a few mathematical comments about this choice and a few consequences that will prove useful.

(i) *as* $b > a$ *then* $w > 1$*,*

(ii) $\ln(b/a) = \ln(w^n) = n\ln(w)$ *therefore* $n = \frac{\ln(b)-\ln(a)}{\ln(w)}$*,*

(iii) $x_{k+1} - x_k = aw^k - aw^{k-1} = aw^{k-1}(w-1)$ *is the k'th step's width,*

(iv) the largest step size is the final one which has width $x_{n+1} - x_n = aw^{n-1}(w-1) = b(1-\frac{1}{w})$, *and*

(v) the limit in which the largest step's width is reduced to zero is the limit when $w \to 1^+$ *(the limit is taken from above as by point (i)* $w > 1$*).*

An upper bound on the integral A is given by

$$\begin{aligned} A^+ &= \sum_{k=1}^{n} S_k^+ (x_{k+1} - x_k) \\ &= \sum_{k=1}^{n} x_{k+1}^m (x_{k+1} - x_k) \\ &= \sum_{k=1}^{n} (aw^k)^m (aw^{k-1}(w-1)) \\ &= a^{m+1} \frac{(w-1)}{w} \sum_{k=1}^{n} w^{(m+1)k} \\ &= a^{m+1} \frac{(w-1)}{w} \sum_{k=1}^{n} (w^{m+1})^k \end{aligned}$$

We can rewrite the geometric sum to find

$$\begin{aligned} A^+ &= a^{m+1} \frac{(w-1)}{w} \left(\frac{w^{m+1}((w^{m+1})^n - 1)}{w^{m+1} - 1} \right) \\ &= a^{m+1} \frac{(w-1) w^m ((w^n)^{m+1} - 1)}{w^{m+1} - 1}. \end{aligned}$$

We now eliminate n using $w^n = \frac{b}{a}$,

$$\begin{aligned} A^+ &= a^{m+1} w^m (w-1) \frac{((b/a)^{m+1} - 1)}{w^{m+1} - 1} \\ &= a^{m+1} w^m (w-1) \frac{((b/a)^{m+1} - 1)}{w^{m+1} - 1}. \end{aligned}$$

Turning to the lower bound and noting that

$$S_k^- = x_k^m = (aw^{k-1})^m = (aw^k)^m w^{-m} = S_k^+ w^{-m}$$

we see that the lower bound on A is related to the upper bound by

$$A^- = \sum_{k=1}^{n} S_k^- (x_{k+1} - x_k) = \sum_{k=1}^{n} S_k^+ w^{-m} (x_{k+1} - x_k) = w^{-m} A^+.$$

We have,

$$A^- = w^{-m}A^+ \leq \int_a^b x^m dx \leq A^+ = a^{m+1}w^m(w-1)\frac{((b/a)^{m+1}-1)}{w^{m+1}-1}.$$

This looks simultaneously promising and horrendous: the good news is that in the limit $w \to 1$ it is clear that we will have $\lim_{w\to 1^+}[A^+] \leq A \leq \lim_{w\to 1^+}[A^+]$ so that A will be sandwiched by the limit if it exists, the bad news is that we will need to evaluate the $\lim_{w\to 1^+}[A^+]$. So,

$$\begin{aligned}
\lim_{w\to 1^+}[A^+] &= \lim_{w\to 1^+}\left[a^{m+1}w^m(w-1)\frac{((b/a)^{m+1}-1)}{w^{m+1}-1}\right] \\
&= \lim_{w\to 1^+}\left[(b^{m+1}-a^{m+1})\frac{w^m(w-1)}{w^{m+1}-1}\right] \\
&= (b^{m+1}-a^{m+1})\lim_{w\to 1^+}\left[w^m\right]\lim_{w\to 1^+}\left[\frac{(w-1)}{w^{m+1}-1}\right] \\
&= (a^{m+1}-b^{m+1})\lim_{w\to 1^+}\left[\frac{1-w}{w^{m+1}-1}\right] \\
&= -(a^{m+1}-b^{m+1})\lim_{w\to 1^+}\left[\left(\frac{1-w^{m+1}}{1-w}\right)^{-1}\right] \\
&= -(a^{m+1}-b^{m+1})\left(\lim_{w\to 1^+}\left[\frac{1-w^{m+1}}{1-w}\right]\right)^{-1} \\
&= -(a^{m+1}-b^{m+1})\left(\lim_{w\to 1^+}\left[\sum_{j=0}^{m} w^j\right]\right)^{-1} \\
&= -(a^{m+1}-b^{m+1})(m+1)^{-1} \\
&= \frac{b^{m+1}-a^{m+1}}{m+1}.
\end{aligned}$$

We also have (trivially),

$$\lim_{w\to 1^+}\left[A^-\right] = \lim_{w\to 1^+}\left[w^{-m}A^+\right] = \lim_{w\to 1^+}\left[w^{-m}\right]\lim_{w\to 1^+}\left[A^+\right] = \lim_{w\to 1^+}\left[A^+\right].$$

Consequently as $\lim_{w\to 1^+}\left[A^-\right] = \lim_{w\to 1^+}\left[A^+\right]$ then we conclude that

$$\int_a^b x^m dx = \frac{b^{m+1}-a^{m+1}}{m+1}.$$

It is worth noting a small point about generalising this result. If we had considered $x < 0$ then the defining relations of S_k^- and S_k^+ would have depended crucially on whether n is odd or even, e.g. for $x < 0$ and odd n then

$$\begin{aligned}
S_k^- &= \min[x^n] \qquad \text{for } x \in [x_k, x_{k+1}) \implies S_i^- = x_k^n \\
S_k^+ &= \max[x^n] \qquad \text{for } x \in [x_k, x_{k+1}) \implies S_k^+ = x_{k+1}^n.
\end{aligned}$$

while for $x < 0$ and even n we would have defined

$$S_k^- = \min[x^n] \qquad \text{for } x \in [x_k, x_{k+1}) \implies S_i^- = x_{k+1}^n$$
$$S_k^+ = \max[x^n] \qquad \text{for } x \in [x_k, x_{k+1}) \implies S_k^+ = x_k^n.$$

But notice that in the long run this fact would not cause us much concern as the change of n from even to odd amounts to interchanging the definition of S_k^+ and S_k^-. Hence if the integral exists and the pair of staircase functions limit to the same sum, then it will not change the result if the bounding functions are swapped depending upon even or odd n.

Of course it is very satisfying that meaning can be given to an integral and that we can construct the Riemann integral from first principles in this way. However it is a tedious and lengthy computation and it will come as a relief to learn that we do not need to repeat this process for each and every function we have met in the course so far. Instead we can rely on the fundamental theorem of calculus to give us a quick way to integrate functions.

5.2 The Fundamental Theorem of Calculus

The fundamental theorem of calculus was understood in various forms by John Gregory (1638-1675) and Isaac Barrow (1630-1677) before Isaac Newton and Gottfried Leibnitz formalised the calculus. The idea was understood long before Riemann put the integral on a rigorous foundation. The basic idea is one that you will already be familiar with, namely, that the integral is the antiderivative, i.e. that the integral of the derivative of a function gives back the function, the integral "undoes" the derivative. This is quite incredible and we are a little anaesthetised to how remarkable an idea it is because we are very familiar with the idea. Let us comment on why it is a surprising idea.

Differentiation and integration are the major operations within calculus and if we did not have a practical experience of working with these operations we would not think that they would be closely related. After all, differentiation gives the *local* value of the gradient of a function at a point and is determined using infinitesimal data while integration involves finding the area under a function as it runs the finite or even infinite distance between the limits of the integration and it is a *global* measure related to the curve (i.e. it requires more than local data from one point). Infinitesimal vs. finite (or infinite), local vs. global: it seems unlikely that the two operations will be related but yet, incredibly, integration and differentiation are inverse operations - this fact is called the fundamental theorem of calculus. The theorem is stated in two parts called the first and second fundamental theorems of calculus.

Theorem 5.1. *(The first fundamental theorem of calculus.) Let $f : [a, b] \to \mathbb{R}$ be a continuous function and let $F : [a, b] \to \mathbb{R}$ be defined by*

$$F(x) = \int_a^x f(t)dt.$$

Then, F is a continuous function on $[a, b]$, differentiable on the open interval (a, b) and

$$\frac{dF}{dx} = f(x).$$

Theorem 5.2. *(The second fundamental theorem of calculus.) Let $f : [a, b] \to \mathbb{R}$ be a continuous function and let $F(x)$ be defined by*

$$\frac{dF}{dx} = f(x) \qquad \forall\, x \in [a, b].$$

Then, if $f(x)$ is integrable on $[a, b]$,

$$\int_a^b f(x)dx = F(b) - F(a).$$

Comment(s). *(On the fundamental theorem of calculus...)*

1. *The function $F(x)$ whose derivative gives $f(x)$, i.e. $\frac{dF}{dx} = f(x)$, is called an antiderivative or a primitive of $f(x)$. For example if $f(x) = x^2$ then the antiderivative (think: integral) is $F(x) = x^3/3$ as $\frac{d}{dx}(x^3/3) = x^2 = f(x)$.*

2. *The first fundamental theorem states that the integral of a function (from a to x) is the antiderivative $F(x)$. At this stage it should be clear that*

$$\int_a^x f(t)dt$$

is a function of x, but it should not yet be clear that this function will be the antiderivative and the appearance of a as a limit of the integral is not yet explained.

3. *The second fundamental theorem tells us that if we know the antiderivative of a function, then we can use this to rapidly evaluate the integral (rather than resort to the time-consuming construction of staircase functions and the use of the sandwich theorem, as we have seen in the previous section).*

We will now sketch some proofs of these theorems.

Proof: (the first fundamental theorem of calculus) In this proof we will limit ourselves to the case where $f(x)$ is a differentiable function. Now,

$$\begin{aligned}\frac{dF}{dx} &= \lim_{h\to 0}\left[\frac{F(x+h)-F(x)}{h}\right] \\ &= \lim_{h\to 0}\left[\frac{1}{h}\left(\int_a^{x+h} f(t)dt - \int_a^{x} f(t)dt\right)\right] \\ &= \lim_{h\to 0}\left[\frac{1}{h}\int_x^{x+h} f(t)dt\right].\end{aligned}$$

Next we will introduce an upper and lower bound on $f(t)$ where $t \in [x, x+h]$ and make use of the sandwich theorem to in order to take the limit above. We will make use of

$$C \equiv \max_{t\in[x,x+h]}\left(\left|\frac{df}{dt}\right|\right)$$

this is the maximum of the modulus of the derivative in the range $[x, x+h]$. Now we don't need to evaluate this number, it will prove very useful in evaluating the limit, but will vanish when we take the limit. As C is the maximum rate of increase (or decrease) of the function $f(t)$ on the interval $[x, x+h]$ we know that

$$f(x) - Ch \le f(t) \le f(x) + Ch \qquad \forall\, t \in [x, x+h].$$

Where did this inequality come from, the sketch in figure 5.9 will help us to understand it. In figure 5.9 we have marked in the straight lines with gradient $\pm C$ which pass through $f(x)$. As, by the definition of C, $\frac{df}{dt} \le C$ and $\frac{df}{dt} \ge -C$ for all $t \in [x, x+h]$, then the straight lines with the gradients $\pm C$ always are greater/less than or equal to $f(t)$ for $t \in [x, x+h]$. The maximum/minimum point on these lines lie at $t = x+h$ and hence we have the inequality (which we repeat here):

$$f(x) - Ch \le f(t) \le f(x) + Ch \qquad \forall\, t \in [x, x+h].$$

Returning to our proof, we now have

$$\lim_{h\to 0}\left[\frac{1}{h}\int_x^{x+h}(f(x)-Ch)dt\right] \le \lim_{h\to 0}\left[\frac{1}{h}\int_x^{x+h} f(t)dt\right] \le \lim_{h\to 0}\left[\frac{1}{h}\int_x^{x+h}(f(x)+Ch)dt\right].$$

To take the limit is a simple matter as

$$\lim_{h\to 0}\left[\frac{1}{h}\int_x^{x+h}(f(x)+Ch)dt\right] = \lim_{h\to 0}\left[\frac{1}{h}(f(x)+Ch)h\right] = \lim_{h\to 0}\left[f(x)+Ch\right] = f(x)$$

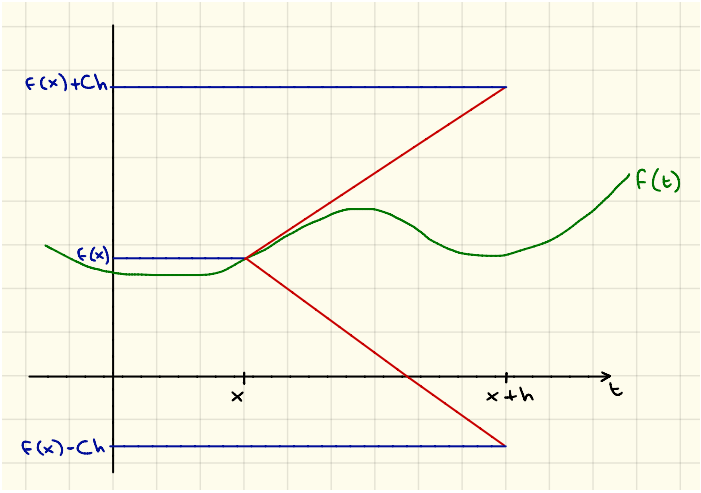


Figure 5.9: An arbitrary differentiable function $f(t)$ and we are interested in its values when $t \in [x, x+h]$. C is the absolute value of the maximum gradient of $f(x)$ and the straight lines marked in are those with gradient $\pm C$ passing through $f(x)$.

as $f(x) + Ch$ are constants in the integral with respect to t and $\int_x^{x+h} dt = h$ (we can evaluate the integral of the constant function: in this case it evaluates the area of a rectangle of width h and height 1). We can also evaluate the lower bound in the limit to find:

$$\lim_{h\to 0}\left[\frac{1}{h}\int_x^{x+h}(f(x)-Ch)dt\right] = \lim_{h\to 0}\left[\frac{1}{h}(f(x)-Ch)h\right] = \lim_{h\to 0}\left[f(x)-Ch\right] = f(x).$$

Hence, by the sandwich theorem, we have:

$$\frac{dF}{dx} = f(x)$$

as required.

Proof: (the second fundamental theorem of calculus) If $f(x) = \frac{dF}{dx}$ is integrable on $[a, b]$ then we can construct the Riemann integral as the limit of a sum of n rectangular areas of width $w_i \equiv x_{i+1} - x_i$ such that $x_0 = a$ and $x_{n+1} = b$. That is we know that,

$$\begin{aligned}\int_a^b f(x)dx &= \lim_{w_i\to 0, n\to\infty}\left[\sum_{i=1}^{n} f(x)\Big|_{x\in[x_i,x_{i+1}]}(x_{i+1}-x_i)\right]\\ &= \lim_{w_i\to 0, n\to\infty}\left[\sum_{i=1}^{n} \frac{dF}{dx}\Big|_{x\in[x_i,x_{i+1}]}(x_{i+1}-x_i)\right]\end{aligned}$$

Now $\frac{dF}{dx}\Big|_{x\in[x_i,x_{i+1}]}$ is some value of the function $\frac{dF}{dx}$ evaluated on the interval $[x_i, x_{i+1}]$ and by

the mean value theorem we know one special value of $\frac{dF}{dx}$ on the interval, namely,

$$\frac{F(x_{i+1}) - F(x_i)}{x_{i+1} - x_i} = \frac{dF}{dx}(c_i) \qquad \text{for some } c_i \in [x_i, x_{i+1}]$$

and we may use this value for the derivative of $F(x)$ in the interval $[x_i, x_{i+1}]$ in the Riemann sum above. We have:

$$\begin{aligned}
\int_a^b f(x)dx &= \lim_{w_i \to 0, n \to \infty} \left[\sum_{i=1}^{n} \frac{F(x_{i+1}) - F(x_i)}{(x_{i+1} - x_i)} (x_{i+1} - x_i) \right] \\
&= \lim_{w_i \to 0, n \to \infty} \left[\sum_{i=1}^{n} (F(x_{i+1}) - F(x_i)) \right] \\
&= \lim_{w_i \to 0, n \to \infty} \Big[(F(x_1) - F(x_0)) + (F(x_2) - F(x_1) + \ldots + (F(x_{n+1}) - F(x_n)) \Big] \\
&= \lim_{w_i \to 0, n \to \infty} \Big[F(x_{n+1}) - F(x_0) \Big] \\
&= \lim_{w_i \to 0, n \to \infty} \Big[F(b) - F(a) \Big] \\
&= F(b) - F(a)
\end{aligned}$$

as required.

Comment(s). *(On another proof of the second fundamental theorem of calculus...) If we knew that $f(x)$ had an antiderivative (and not just that it is integrable[2]) we could have proved the second fundamental theorem by using the first fundamental theorem. As $f(x)$ is a continuous function on $[a, b]$ then the integral $\int_a^x f(t)dt$ exists for $x \in [a, b]$. Let us give this useful function a name*

$$G(x) \equiv \int_a^x f(t)dt$$

and by the first fundamental theorem we have

$$\frac{dG}{dx} = f(x) \qquad \forall\, x \in [a, b].$$

Notice that $G(a) = 0$ and $G(b) = \int_a^b f(t)dt$, i.e. $G(b)$ is the integral of interest in the statement of the second fundamental theorem of calculus. Now consider a second antiderivative of $f(x)$, denoted $F(x)$, such that

$$\frac{dF}{dx} = f(x) \qquad \forall\, x \in [a, b].$$

[2]There exist functions which are integrable, but which do not have an antiderivative e.g. Thomae's function ($f(x) = 0$ if $x \in \mathbb{R} \setminus \mathbb{Q}$ and $f(x) = 1/q$ if $x = p/q \in \mathbb{Q}$ and p and q have no common factors), and there also exist functions who have an antiderivative but which are not Riemann integrable e.g. Volterra's function, see `https://en.wikipedia.org/wiki/Volterra%27s_function`.

In this proof we aim to show that $G(x) = F(x) - F(a)$, so let us introduce a third function, the difference between $F(x)$ and $G(x)$:

$$H(x) \equiv F(x) - G(x)$$

then

$$\frac{dH}{dx} = f(x) - f(x) = 0 \qquad \forall\, x \in [a, b].$$

So we have learnt that $H(x)$ has a gradient of zero for $x \in [a, b]$, therefore $H(x) = H(a)$ which implies that

$$F(x) - G(x) = F(a) - G(a) \qquad \forall\, x \in [a, b].$$

Now, as $G(a) = 0$ then we have $G(x) = F(x) - F(a)$, i.e.

$$\int_a^b f(t)dt = G(b) = F(b) - F(a)$$

as required.

5.2.1 Indefinite and Definite Integrals.

Definition 5.2.1. *The definite integral $A \equiv \int_a^b f(x)dx$ has specific limits $x = \{a, b\}$.*

Comment(s). *(On definite integrals...)*

1. *If $f(x)$ is integrable, then, by the (second) fundamental theorem of calculus:*

$$A = \int_a^b f(x)dx = F(b) - F(a)$$

   where $\frac{dF}{dx} = f(x)$ and A is a number.

2. *So far we have only defined $\int_a^b f(x)dx$ for $a \leq b$ but we can generalise this to include $b \leq a$ via the (second) fundamental theorem of calculus:*

$$\int_a^b f(x)dx = F(b) - F(a) = -(F(a) - F(b)) = -\int_b^a f(x)dx.$$

Definition 5.2.2. *The indefinite integral $\int f(x)dx$ is a function: it has no specified limits.*

Comment(s). *(On indefinite integrals...) The indefinite integral at first sight is an abuse of notation: the integral is defined (up to a sign) as an area via its limits, but the indefinite integral*

has no limits - so what does it mean? It is another notation for the antiderivatives or primitives of $f(x)$. Recall that by the (first) fundamental theorem of calculus that the antiderivative is

$$F(x) = \int_0^x f(t)dt = \int_0^x \frac{dF}{dt}dt = F(x) - F(0)$$

and note that $F(0) = 0$, hence

$$F(x) - F(a) = \int_a^x f(t)dt$$

so that $\frac{d}{dx}(F(x) - F(a)) = \frac{dF}{dx} = f(x)$, in other words there are multiple antiderivatives (antiderivatives are specified only up to a constant). Now it is useful to think of the class of antiderivative functions for $f(x)$ and while one could write $F(x) - F(a)$ to denote the antiderivatives it is useful to have a briefer notation, hence the notation for the indefinite integral, instead of writing $\int_a^x f(t)dt$ we write $\int f(x)dx$ to denote the class of antiderivative functions and emphasise that it is a function of x, hence the indefinite integral is

$$\int f(x)dx \equiv F(x) + K \qquad K \in \mathbb{R}$$

where K is called the constant of integration and $F(x)$ is an antiderivative of $f(x)$.

Using the fundamental theorem(s) of calculus we can now try to solve

$$\int_a^b f(x)dx$$

by identifying the antiderivative $F(x)$, i.e. by solving $\frac{dF}{dx} = f(x)$.

5.3 Properties of the Integral and some Techniques for Integration

Our knowledge of the derivatives of standard functions will prove invaluable when trying to identify a primitive, but it will also prove necessary to understand the basic properties of the integral, so that we may manipulate it into a form where we may identify the antiderivative or primitive.

The integral is a linear operation, i.e. $\int(af(x) + bg(x))dx = a\int f(x)dx + b\int g(x)dx$ where $a, b \in R$ are constants with respect to x (i.e. they do not depend on the integration variable x). We can prove the linearity of the integral by using our knowledge of derivatives and the fundamental theorem of calculus. In the same manner we can also show how we can transform

an integral by a substitution or by using the technique known as integration by parts. Both these methods will open up new ways to find primitives or antiderivatives and evaluate definite integrals. Let $F(x) \equiv \int f(x)dx$ and $G(x) \equiv \int g(x)dx$ then:

(i) Multiplication by a constant:

$$\int (af(x))dx = a \int f(x)dx$$

where a is a constant.

Proof:

$$\frac{d}{dx}(aF(x)) = a\frac{dF}{dx} = af(x).$$

Hence,

$$\int \frac{d}{dx}(aF(x))dx = aF(x) + K = a \int f(x)dx.$$

(ii) Integrals of sums of functions are sums of integrals:

$$\int (f(x) + g(x))dx = F(x) + G(x).$$

Proof:

$$\frac{d}{dx}(F(x) + G(x)) = \frac{dF}{dx} + \frac{dG}{dx} = f(x) + g(x).$$

By points (i) and (ii), integration is a linear operation.

(iii) Transforming an integral with respect to x into an integral with respect to t by substituting $x = x(t)$ to find:

$$\int_{x_1}^{x_2} f(x)dx = \int_{t_1}^{t_2} f(x(t))\frac{dx}{dt}dt$$

where $x(t_1) = x_1$ and $x(t_2) = x_2$.

Proof: We know that

$$\int_{x_1}^{x_2} f(x)dx = \int_{x_1}^{x_2} \frac{dF}{dx}dx = F(x_2) - F(x_1),$$

and if $x_1 = x(t_1)$ and $x_2 = x(t_2)$ then we have

$$\begin{aligned} F(x_2) - F(x_1) &= F(x(t_2)) - F(x(t_1)) \\ &= \Big[F(x(t))\Big]_{t_1}^{t_2} \\ &= \int_{t_1}^{t_2} \frac{d}{dt}(F(x(t)))dt \\ &= \int_{t_1}^{t_2} \frac{dF}{dx}\frac{dx}{dt}dt \end{aligned}$$

where we have used the chain rule $\frac{dF(x(t))}{dt} = \frac{dF}{dx}\frac{dx}{dt}$.

(iv) Integration by parts:

$$\int \frac{df}{dx} g(x)dx = f(x)g(x) - \int f(x)\frac{dg}{dx}dx + K.$$

Proof:

$$\frac{d}{dx}(f(x)g(x)) = \frac{df}{dx}g + f\frac{dg}{dx},$$

hence,

$$\int \frac{d}{dx}(f(x)g(x))dx = f(x)g(x) + K = \int \frac{df}{dx}g(x)dx + \int f(x)\frac{dg}{dx}dx$$

as required.

We will now practise these techniques to evaluate definite integrals or identify the antiderivative for an indefinite integral with multiple examples.

5.3.1 Solving Integrals by Substitution

Example 5.3. *Express the indefinite integral*

$$\int \sin^m(x)\cos(x)dx$$

as a function of x by substituting $sin(x) = t$.

Here $x(t) = \arcsin(t)$, but it is simpler to find $\frac{dx}{dt}$ by acting with $\frac{d}{dt}$ on $sin(x) = t$ to obtain $\frac{d}{dt}(\sin(x)) = \cos(x)\frac{dx}{dt} = 1$, hence,

$$\begin{aligned}
\int \sin^m(x)\cos(x)dx &= \int \sin^m(x(t))\cos(x(t))\frac{dx}{dt}dt \\
&= \int t^m \cos(x(t))\frac{1}{\cos(x(t))}dt \\
&= \int t^m dt \\
&= \frac{t^{m+1}}{m+1} + K \\
&= \frac{\sin^{m+1}(x)}{m+1} + K
\end{aligned}$$

It is rather lengthy to present the substitution as above, it is perfectly acceptable instead to observe that $\sin(x) = t \implies \cos(x)dx = dt$.

Example 5.4. *Express the indefinite integral*

$$\int e^{x^2} x dx$$

as a function of x by using an appropriate substitution.

There are many options, but we are guided to simplify the terms in the integral and so choose the substitution

$$x^2 = t$$

hence $2xdx = dt$. Therefore we have upon substituting into the integral

$$\begin{aligned}\int e^{x^2} x dx &= \int e^t \frac{1}{2} dt \\ &= \frac{1}{2} e^t + K \\ &= \frac{1}{2} e^{x^2} + K.\end{aligned}$$

Example 5.5. *Express the indefinite integral*

$$\int (1 - x^2)^{-\frac{1}{2}} dx$$

as a function of x by using an appropriate substitution.

Let

$$x = \sin\theta$$

hence $dx = \cos\theta d\theta$. Our motivation for choosing this substitution is the trigonometric identity $1 - \sin^2\theta = \cos^2\theta$, i.e. $(1 - x^2)^{-\frac{1}{2}} = 1/\cos\theta$, hence we have

$$\begin{aligned}\int (1 - x^2)^{-\frac{1}{2}} dx &= \int \frac{1}{\cos\theta} (\cos\theta d\theta) \\ &= \int d\theta \\ &= \theta + K \\ &= \arcsin(x) + K.\end{aligned}$$

Of course we may have recognised immediately the antiderivative by recalling that $\frac{d}{dx}(\arcsin(x)) = \frac{1}{\sqrt{1-x^2}}$.

Example 5.6. *Express the indefinite integral*

$$\int (1+x^2)^{-\frac{1}{2}} dx$$

as a function of x by using an appropriate substitution.

Let

$$x = \sinh\theta$$

hence $dx = \cosh\theta d\theta$. Hence we have

$$\begin{aligned}\int (1+x^2)^{-\frac{1}{2}} dx &= \int (1+\sinh^2\theta)^{-\frac{1}{2}} (\cosh\theta) d\theta \\ &= \int d\theta \\ &= \theta + K \\ &= \operatorname{arcsinh}(x) + K.\end{aligned}$$

Where we have used the identity $1+\sinh^2\theta = \cosh^2\theta$.

Example 5.7. *Express the indefinite integral*

$$\int (1+x^2)^{-1} dx$$

as a function of x by using an appropriate substitution.

Let

$$x = \tan\theta$$

hence $dx = (1+\tan^2\theta)d\theta$. Hence we have

$$\begin{aligned}\int (1+x^2)^{-1} dx &= \int (1+\tan^2\theta)^{-1}(1+\tan^2\theta) d\theta \\ &= \int d\theta \\ &= \theta + K \\ &= \arctan(x) + K.\end{aligned}$$

Example 5.8. *Express the indefinite integral*

$$\int (1-x^2)^{-1} dx$$

as a function of x by using an appropriate substitution.

Let

$$x = \tanh\theta$$

hence $dx = (1 - \tanh^2\theta)d\theta$. *Hence we have*

$$\begin{aligned}\int (1-x^2)^{-1}dx &= \int (1-\tanh^2\theta)^{-1}(1-\tanh^2\theta)d\theta \\ &= \int d\theta \\ &= \theta + K \\ &= \operatorname{arctanh}(x) + K.\end{aligned}$$

Example 5.9. *Evaluate the definite integral*

$$\int_2^3 (x^2+2x)^{-\frac{1}{2}}dx$$

by using an appropriate substitution.

First note that

$$\int_2^3 (x^2+2x)^{-\frac{1}{2}}dx = \int_2^3 ((x+1)^2-1)^{-\frac{1}{2}}dx.$$

Now the integral suggests the substitution

$$x + 1 = \cosh\theta$$

so that $dx = \sinh\theta d\theta$. *Hence we have*

$$\begin{aligned}\int_2^3 ((x+1)^2-1)^{-\frac{1}{2}}dx &= \int_{\operatorname{arccosh}(3)}^{\operatorname{arccosh}(4)} (\cosh^2\theta - 1)^{-\frac{1}{2}}\sinh\theta d\theta \\ &= \int_{\operatorname{arccosh}(3)}^{\operatorname{arccosh}(4)} d\theta \\ &= \Big[\theta\Big]_{\operatorname{arccosh}(3)}^{\operatorname{arccosh}(4)} \\ &= \operatorname{arccosh}(4) - \operatorname{arccosh}(3)\end{aligned}$$

Example 5.10. *Evaluate the definite integral*

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

by using an appropriate substitution.

First note that

$$\int_{-\frac{3}{2}}^{-\frac{1}{2}} (-x^2 - 2x)^{-\frac{1}{2}} dx = \int_{-\frac{3}{2}}^{-\frac{1}{2}} (-(x+1)^2 + 1)^{-\frac{1}{2}} dx.$$

Now the integral suggests the substitution

$$x + 1 = \sin\theta$$

so that $dx = \cos\theta d\theta$. *Hence we have*

$$\begin{aligned}
\int_{-\frac{3}{2}}^{-\frac{1}{2}} (-(x+1)^2 + 1)^{-\frac{1}{2}} dx &= \int_{\arcsin(-1/2)}^{\arcsin(1/2)} (-\sin^2\theta + 1)^{-\frac{1}{2}} \cos\theta d\theta \\
&= \int_{\arcsin(-1/2)}^{\arcsin(1/2)} d\theta \\
&= \arcsin(1/2) - \arcsin(-1/2) \\
&= \frac{\pi}{6} - \frac{-\pi}{6} \\
&= \frac{\pi}{3}.
\end{aligned}$$

Example 5.11. *Express the indefinite integral*

$$\int (6 + 4x - 2x^2)^{-\frac{1}{2}} dx$$

as a function of x *by using an appropriate substitution.*

First note that

$$\int (6 + 4x - 2x^2)^{-\frac{1}{2}} dx = \int (-2(x-1)^2 + 8)^{-\frac{1}{2}} dx = \frac{1}{\sqrt{8}} \int (-\frac{1}{4}(x-1)^2 + 1)^{-\frac{1}{2}} dx$$

Now the form of the integral suggests the substitution

$$\frac{1}{2}(x-1) = \sin\theta$$

so that $\frac{1}{2}dx = \cos\theta d\theta$. *Hence we have*

$$\begin{aligned}\frac{1}{\sqrt{8}}\int(-\frac{1}{4}(x-1)^2+1)^{-\frac{1}{2}}dx &= \frac{1}{\sqrt{8}}\int(-\sin^2\theta+1)^{-\frac{1}{2}}(2\cos\theta d\theta)\\ &= \frac{2}{\sqrt{8}}\int d\theta\\ &= \frac{1}{\sqrt{2}}(\arcsin(\frac{x-1}{2}))+K\end{aligned}$$

Example 5.12. *Express the indefinite integral*

$$\int \tan(x)dx$$

as a function of x *by using an appropriate substitution.*

First note that

$$\int \tan(x)dx = \int \frac{\sin(x)}{\cos(x)}dx.$$

With a little thought we realise that the numerator is the derivative of the denominator and so the substitution

$$\cos(x) = t$$

so that $-\sin(x)dx = dt$, *will greatly simplify the integral. We have*

$$\begin{aligned}\int \frac{\sin(x)}{\cos(x)}dx &= \int \frac{-dt}{t}\\ &= \int \frac{d}{dt}(-\ln(|t|))dt\\ &= -\ln(|t|)+K\\ &= -\ln(|\cos(x)|)+K\end{aligned}$$

Exercise 5.4. *Find the integral*

$$\int \frac{3x^2+2x}{x^3+x^2+5}dx$$

as a function of x *by making an appropriate substitution. [Hint: compare the numerator and the denominator.]*

Example 5.13. *Express the indefinite integral*

$$\int \frac{dx}{\sin(x)}$$

as a function of x by using an appropriate substitution.

There are many correct ways to approach this integral. We will first try to simplify the integral with the substitution

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

so that $dt = -\frac{\cos(x)}{\sin^2(x)}dx = -\frac{\sqrt{1-\frac{1}{t^2}}}{\frac{1}{t^2}}dx = -t^2\sqrt{1-\frac{1}{t^2}}dx = -t\sqrt{t^2-1}dx$. *Upon substitution we find*

$$\begin{aligned}\int \frac{dx}{\sin(x)} &= \int t\left(\frac{-dt}{t\sqrt{t^2-1}}\right) \\ &= -\int \frac{dt}{\sqrt{t^2-1}}.\end{aligned}$$

Now we will use another substitution:

$$t = \cosh(u)$$

so that $dt = \sinh(u)du$ *and we have*

$$\begin{aligned}-\int \frac{dt}{\sqrt{t^2-1}} &= -\int \frac{\sinh(u)du}{\sqrt{\cosh^2(u)-1}} \\ &= -\int du \\ &= -u + K \\ &= -\text{arccosh}(t) + K \\ &= -\text{arccosh}(\frac{1}{\sin(x)}) + K.\end{aligned}$$

This answer can be written in many different forms using the various identities we have seen for the trigonometric and hyperbolic functions. Similarly if we had pursued a different path to find the antiderivative, by making different substitutions for example then we would have ended up with the answer expressed in a different form. The solution here is frequently written in

terms of $\tan(x/2)$ *and the logarithm, let us convert our answer into this common format:*

$$\begin{aligned}
-\operatorname{arccosh}(\frac{1}{\sin(x)}) &= -\ln(\left|\sqrt{\frac{1}{\sin^2 x}-1}+\frac{1}{\sin x}\right|)\\
&= -\ln(\left|\sqrt{\frac{1-\sin^2 x}{\sin^2 x}}+\frac{1}{\sin x}\right|)\\
&= -\ln(\left|\frac{\cos x+1}{\sin x}\right|)\\
&= \ln(\left|\frac{\sin x}{\cos x+1}\right|)\\
&= \ln(\left|\frac{2\sin(x/2)\cos(x/2)}{\cos^2(x/2)-\sin^2(x/2)+1}\right|)\\
&= \ln(\left|\frac{2\sin(x/2)\cos(x/2)}{2\cos^2(x/2)}\right|)\\
&= \ln(\left|\frac{\sin(x/2)}{\cos(x/2)}\right|)\\
&= \ln(|\tan(x/2)|).
\end{aligned}$$

5.3.2 Integration by Parts

This method can be very useful when integrating product. Let us rewrite the statement of an integration by parts but this time for a definite integral:

$$\int_a^b f(x)\frac{dg}{dx}dx = -\int_a^b \frac{df}{dx}g(x)dx + \Big[f(x)g(x)\Big]_a^b.$$

The technique allows us to move the derivative from $g(x)$ to its coefficient $f(x)$ (and introduce a minus sign plus a boundary term). This will be very useful if $\frac{df}{dx}$ is a simple function, e.g. a constant. Let us practise using the method by looking at some examples.

Example 5.14. *Find, up to a constant, the function of* x *given by the indefinite integral*

$$\int \arcsin(x)dx.$$

First we will make a substitution aimed at making the integral more palatable, we substitute $x = \sin\theta$*, so that* $dx = \cos\theta d\theta$ *hence*

$$\int \arcsin(x)dx = \int \theta\cos\theta d\theta = \int \theta\frac{d}{d\theta}(\sin\theta)d\theta.$$

In the last line we have rewritten it so that a derivative is applied to one term under the integral in order to emphasise that we are now in a situation where using integration by parts is a good idea. Why is it a good idea, because integration by parts gives us a way to move the derivative from $\sin\theta$ *onto* θ *and* $\frac{d\theta}{d\theta} = 1$ *is simple. Let us be careful and identify the functions* f *and* g *given in our abstract derivative with the functions in our present example*[3]*: we take* $f(\theta) = \theta$, $g(\theta) = \sin\theta$, *so that*

$$\int f(\theta)\frac{dg}{d\theta}d\theta = -\int \frac{df}{d\theta}g(\theta)d\theta + f(\theta)g(\theta)$$

hence

$$\begin{aligned}\int \theta\frac{d}{d\theta}(\sin\theta)d\theta &= -\int \frac{d\theta}{d\theta}\sin\theta d\theta + \theta\sin\theta \\ &= -\int \sin\theta d\theta + \theta\sin\theta \\ &= -\int \frac{d}{d\theta}(-\cos\theta)d\theta + \theta\sin\theta \\ &= \cos\theta + \theta\sin\theta + K \\ &= \cos(\arcsin(x)) + x\arcsin x + K \\ &= \sqrt{1-x^2} + x\arcsin x + K\end{aligned}$$

Example 5.15. *Find, up to a constant, the function of* x *given by the indefinite integral*

$$\int x^2\cos(2x)dx.$$

First we note that

$$\int x^2\cos(2x)dx = \int x^2\frac{d}{dx}(\frac{1}{2}\sin(2x))dx.$$

Now we integrate by parts. It is a good idea to practise integrating by parts without the need to identify $f(x)$ *and* $g(x)$ *before carrying out the procedure. Specifically notice that integration by parts tells you to add a minus sign and shift the derivative from one term onto its coefficients in the integral, and of course do not forget to add the boundary term - this one should practise*

[3]Note that in the abstract definition the variable was x but in our example it is now θ.

doing immediately almost as an algebraic manoeuvre on the integral, so that:

$$\begin{aligned}
\int x^2 \frac{d}{dx}(\frac{1}{2}\sin(2x))dx &= -\int \frac{d}{dx}(x^2)(\frac{1}{2}\sin(2x))dx + \frac{x^2}{2}\sin(2x) \\
&= -\int x\sin(2x)dx + \frac{x^2}{2}\sin(2x) \\
&= \int x\frac{d}{dx}(\frac{1}{2}\cos(2x))dx + \frac{x^2}{2}\sin(2x) \\
&= -\int \frac{dx}{dx}(\frac{1}{2}\cos(2x))dx + \frac{x}{2}\cos(2x) + \frac{x^2}{2}\sin(2x) \\
&= -\int \frac{1}{2}\cos(2x)dx + \frac{x}{2}\cos(2x) + \frac{x^2}{2}\sin(2x) \\
&= -\int \frac{1}{2}\frac{d}{dx}(\frac{1}{2}\sin(2x))dx + \frac{x}{2}\cos(2x) + \frac{x^2}{2}\sin(2x) \\
&= -\frac{1}{4}\sin(2x) + \frac{x}{2}\cos(2x) + \frac{x^2}{2}\sin(2x) + K
\end{aligned}$$

Example 5.16. *Find, up to a constant, the function of x given by the indefinite integral*

$$\int xe^{-x}dx.$$

$$\begin{aligned}
\int xe^{-x}dx &= \int x\frac{d}{dx}(-e^{-x})dx \\
&= -\int x\frac{d}{dx}(e^{-x})dx \\
&= \int e^{-x}dx - xe^{-x} \\
&= -e^{-x} - xe^{-x} + K.
\end{aligned}$$

Example 5.17. *Find, up to a constant, the function of x given by the indefinite integral*

$$\int \ln(x)dx.$$

This example involves a good trick: to use integration by parts we need a derivative under the integral, now since $\frac{dx}{dx} = 1$ we can always insert this trivial derivative without changing the

integral, so that,

$$\begin{aligned}\int \ln(x)dx &= \int \ln(x)\frac{dx}{dx}dx \\ &= -\int \frac{d}{dx}(\ln(x))xdx + x\ln(x) \\ &= -\int \frac{1}{x}xdx + x\ln(x) \\ &= -\int dx + x\ln(x) \\ &= -x + x\ln(x) + K\end{aligned}$$

Exercise 5.5. *Use integration by parts to find*

$$\int \arctan x dx$$

as a function of x, up to a constant. [Hint: insert $\frac{dx}{dx} = 1$ as in the last example.]

5.3.3 Partial Fractions

We saw among our examples of solving integrals by substitution, that if we were integrating a fraction where the numerator was the derivative of the denominator then a substitution could greatly simplify the integral, i.e.

$$\int \frac{f'(x)}{f(x)}dx = \int \frac{du}{u}$$

where we have made the substitution $u = f(x)$, so that $du = f'(x)dx$. But how do we go about integrating fractions which do not have this form? If we are working with fractions of polynomial functions, we can try to split the fraction up into a sum of simpler fractions. This process is known as finding the partial fractions. We will investigate the use of partial fractions through a few simple examples to illustrate the central idea.

Example 5.18. *Find the function of x given by the indefinite integral*

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

As mentioned our aim is to split the fraction into a sum of fractions, now we know that,

$$\frac{A}{f(x)} + \frac{B}{g(x)} = \frac{Ag(x) + Bf(x)}{f(x)g(x)}$$

and our aim is to reverse this process. Therefore it will aid us to identify the factorisation of the denominator in the problem. We have:

$$\int \frac{3x+1}{x^2-3x+2}dx = \int \frac{3x+1}{(x-1)(x-2)}dx \equiv \int \left(\frac{A}{x-1}+\frac{B}{x-2}\right)dx.$$

Our aim now is to find A and B, we know (from summing the fractions) that

$$A(x-2)+B(x-1) = 3x+1$$

hence we may now identify A and B by equating the coefficients of x and the constants on each side of the above and solving the simultaneous equations in A and B that doing so gives us. We find:

$$\begin{aligned} A+B &= 3 \\ -2A-B &= 1 \end{aligned}$$

Hence $A=-4$ and $B=7$, so we have:

$$\int \frac{3x+1}{(x-1)(x-2)}dx = \int \left(\frac{-4}{x-1}+\frac{7}{x-2}\right)dx = -4\ln(|x-1|)+7\ln(|x-2|)+K.$$

Most integrals which can be quickly solved using partial fractions are similar to the above example: one attempts to factorise the denominator, to find the partial fractions and then integrate these simpler fractions. The simplest complication emerges when the denominator has a repeated factor as in the following example.

Example 5.19. *Find the function of x given by the indefinite integral*

$$\int \frac{2x+3}{(x-1)^2}dx.$$

If we commence by trying to split the fraction into partial fractions with the standard procedure we find

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

and so we have the contradiction that $A+B=2$ and $A+B=-3$, something has gone astray here, and the source of the problem is the repeated factor in the denominator. Notice that if it had been possible to find the partial fractions as above it would have meant that a factor of $(x-1)$ could be removed from both the numerator and the denominator. From the form of the fraction that we start with we see that this is not possible (otherwise we would have simplified

the fraction). Hence our first move was amiss, instead of splitting the fraction into partial fractions whose denominators are identical we seek A and B such that:

$$\frac{2x+3}{(x-1)^2} = \frac{A}{x-1} + \frac{B}{(x-1)^2} = \frac{Ax - A + B}{(x-1)^2}.$$

Now we find that $A = 2$ and $B = 5$, hence,

$$\int \frac{2x+3}{(x-1)^2}dx = \int \frac{2}{x-1}dx + \int \frac{5}{(x-1)^2}dx = 2\ln(|x-1|) - \frac{5}{x-1} + K.$$

One further common complication concerns what happens when the factor in the denominator is a polynomial of order greater than one. Consider the following example.

Example 5.20. *Find the function of x given by the indefinite integral*

$$\int \frac{2x^2 - 5x + 7}{x(x^2+3)}dx.$$

If we commence naïvely by trying to split the fraction with constant numerators we find

$$\frac{2x^2 - 5x + 7}{x(x^2+3)} = \frac{A}{x} + \frac{B}{x^2+3} = \frac{Ax^2 + 3A + Bx}{x(x^2+3)}$$

and we see that the simultaneous equations for A and B are contradictory: $A = 2$, $B = -5$ and $3A = 7$, our choice for the partition was too tightly constraining. Notice also that in every other example we have chosen partial fractions such that the numerator was of order one less than the denominator, hence we are motivated to try the following:

$$\frac{2x^2 - 5x + 7}{x(x^2+3)} = \frac{A}{x} + \frac{Bx + C}{x^2+3} = \frac{Ax^2 + 3A + Bx^2 + Cx}{x(x^2+3)}$$

so that the simultaneous equations to solve are

$$\begin{aligned} A + B &= 2 \\ C &= -5 \\ 3A &= 7. \end{aligned}$$

These equations are not contradictory and we can solve them to find $A = \frac{7}{3}$, $B = -\frac{1}{3}$ and $C = -5$. Hence,

$$\begin{aligned} \int \frac{2x^2 - 5x + 7}{x(x^2+3)}dx &= \frac{7}{3}\int \frac{1}{x}dx + \int \frac{-\frac{1}{3}x - 5}{x^2+3}dx \\ &= \frac{7}{3}\int \frac{1}{x}dx - \frac{1}{3}\int \frac{x}{x^2+3}dx - \int \frac{5}{x^2+3}dx \\ &= \frac{7}{3}\ln(|x|) - \frac{1}{6}\ln(|x^2+3|) - \frac{5}{\sqrt{3}}\arctan\left(\frac{x}{\sqrt{3}}\right) + K. \end{aligned}$$

There are some further variations which we will highlight here, namely, what happens when the degree of the polynomial in the numerator is the same as that in the denominator, or even greater? We will consider two simple examples in theses classes and also make some comments on what to do if one cannot factorise the denominator.

Example 5.21. *Find the function of x given by the indefinite integral*

$$\int \frac{x+2}{x-2} dx.$$

Here the denominator has the same power as the numerator, our aim will be to write the numerator as a multiple of the denominator plus some remainder to simplify the problem. If we can do this, we may then cancel out common factors and hope to find a simpler problem. Hence we must first find $f(x)$ and the remainder R such that

$$(x-2)f(x) + R = x + 2$$

we may carry out the long division or alternatively we may write $f(x) = ax + b$ and find a and b, i.e. from $(x-2)(ax+b) + R = ax^2 + bx - 2ax - 2b + R = x + 2$ we have $a = 0$, $b = 1$ and so $R = 4$. Hence,

$$\int \frac{x+2}{x-2} dx = \int \frac{(x-2)+4}{x-2} dx = \int \frac{x-2}{x-2} dx + \int \frac{4}{x-2} dx = x + 4\ln(|x-2|) + K.$$

The factorisation of the numerator allowed us to split the fraction into a constant part and a term in which the order of the numerator was less than the denominator.

Example 5.22. *Find the function of x given by the indefinite integral*

$$\int \frac{(x+2)^3}{(x-2)^2} dx.$$

We will aim to repeat the tactic of the previous example, but now there is some increased complication. Recall that the success of the previous example rested upon removing factors in the denominator from the numerator, so for this example we will aim to solve

$$(ax+b)(x-2)^2 + c(x-2) + d = (x+2)^3$$

or, after multiplying out,

$$ax^3 + (-4a+b)x^2 + (4a-4b+c)x + 4b - 2c + d = x^3 + 6x^2 + 12x + 8.$$

From which we may find that $a = 1$, $b = 10$, $c = 48$ and $d = 64$ and so we have,

$$\begin{aligned}\int \frac{(x+2)^3}{(x-2)^2}dx &= \int \frac{(x+10)(x-2)^2 + 48(x-2) + 64}{(x-2)^2}dx \\ &= \frac{1}{2}x^2 + 10x + 48\ln(|x-2|) - \frac{64}{x-2} + K.\end{aligned}$$

5.3.4 Recursion Relations

Integration by parts is useful when there is a product appearing in an integral and integration by parts can be used to simplify one of the products. Frequently it is necessary to apply the integration by parts many times before an integral becomes simple enough to be computed directly. The repeated procedure can be lengthy and it is often useful to derive a formula which encodes the action of integration by parts which can be used repeatedly to simplify an integral. Such a relation, which describes one integral in terms of another (simpler) integral is called a recursion relation or also a reduction formula.

For example, consider the integral

$$I_n \equiv \int_0^\infty x^n e^{-x} dx$$

where $n \in \mathbb{Z}^+$. Integrating by parts we find a recursion relation,

$$\begin{aligned}I_n &= \int_0^\infty x^n e^{-x} dx \\ &= \int_0^\infty x^n \frac{d}{dx}(-e^{-x})dx \\ &= -\int_0^\infty \frac{d}{dx}(x^n)(-e^{-x})dx + \left[(x^n)(-e^{-x})\right]_0^\infty \\ &= \int_0^\infty n x^{n-1} e^{-x} dx \\ &= nI_{n-1}.\end{aligned}$$

This is a recursion formula for the integral I_n: it expresses I_n as a function of I_{n-1} and more general recursion formulae may involve other integrals of lower order in n, e.g. $\hat{I}_n = f(\hat{I}_{n_1}, \hat{I}_{n_2}, \ldots \hat{I}_0)$ where $n > n_1 > n_2 > \ldots > 0$. A recursion formula allows one to replace more complex integrals (where n is large) with integrals (where n is smaller) which one hopes will be simpler to compute. We can repeatedly use $I_n = nI_{n-1}$ for successively smaller values of n until we find a simple integral that we can solve:

$$I_n = nI_{n-1} = n(n-1)I_{n-2} = n(n-1)(n-2)I_{n-3} = \ldots = n(n-1)(n-2)\ldots 2I_1 = n!I_0.$$

Now we can solve I_0 as

$$I_0 = \int_0^\infty x^0 e^{-x} dx = \int_0^\infty e^{-x} dx = \big[-e^{-x} \big]_0^\infty = 0 - (-1) = 1.$$

Therefore,

$$I_n \equiv \int_0^\infty x^n e^{-x} dx = n!$$

This is a beautiful expression, but notice how simple it is to compute all I_n (i.e. for any $n \in \mathbb{Z}^+$) once a recursion formula has been identified. What is odd about this particular integral is that $\int_0^\infty x^n e^{-x} dx$ is well-defined for any real value of n, not only positive integer values of n, while the computation of the recursion relation is left unchanged. In fact mathematicians use this integral to define what is meant by $n!$ when $n \in \mathbb{R}^+$. For this reason, this integral occurs frequently in mathematics and is called the gamma function and is defined as follows:

$$\Gamma(n+1) \equiv \int_0^\infty x^n e^{-x} dx \equiv n! \qquad \forall\, n \in \mathbb{R}^+.$$

By repeating the integration by parts, one can find the recursion relation for the gamma function is the same as for the original integral, namely,

$$\Gamma(n+1) = n\Gamma(n) \qquad \forall\, n \in \mathbb{R}^+.$$

Of course if one begins with a non-integer value for n then by repeated use of the recursion relation the value of n (appearing in the integral) can be lowered until it lies between 0 and 1, at which point one can only try to compute the remaining integral. A very interesting set of examples occur when n is half-integer, then the final integral to compute (after using the recursion relation) is $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, so, example:

$$\Gamma(\frac{9}{2}) = \frac{7}{2}\Gamma(\frac{7}{2}) = \frac{7\times 5}{2^2}\Gamma(\frac{5}{2}) = \frac{7\times 5\times 3}{2^3}\Gamma(\frac{3}{2}) = \frac{7\times 5\times 3\times 1}{2^4}\Gamma(\frac{1}{2}) = \frac{105}{16}\sqrt{\pi}.$$

The appearance of π through $\Gamma(\frac{1}{2})$ is interesting, and so this is the subject of the following (optional) exercise.

Exercise 5.6. *Compute the integral*

$$\Gamma(\frac{1}{2}) = \int_0^\infty x^{-\frac{1}{2}} e^{-x} dx$$

by using the substitution $x = y^2$ and the integral of (half of) the Gaussian function given by

$$\int_0^\infty e^{-y^2} dy = \frac{\sqrt{\pi}}{2}.$$

Exercise 5.7. *Show that the integral defined by*

$$I_n = \int \cos^n x dx \qquad \forall n \in \mathbb{Z}^+$$

has the following recursion relation:

$$I_n = \frac{n-1}{n} I_{n-2} + \frac{1}{n} \cos^{n-1} x \sin x.$$

Hence find I_5 as a function of x. [Hint: commence by writing $\cos^n x = \cos^{n-1} x \cos x$.]

5.4 Some Applications: Length, Area and Volume.

5.4.1 Areas of Circles and Ellipses

We defined the integral to measure (up to a sign) the area under a curve $f(x)$. It produces the area under a curve if the curve lies above the x-axis, but it produces minus the area between the curve and the x-axis where $f(x) < 0$. This seems to present a problem to anyone trying to use integration to find for example an expression for the area of a circle using integration. If the circle has its centre located on the x-axis then the integral of the circle will be zero, as the positive half-area above the axis cancels out the negative half-area found by integrating the part of the circle beneath the x-axis. Hence we need a trick to use integration to evaluate the area of a circle, or indeed the area of any shape which is left unchanged by a reflection in the x-axis. However the trick is very simple, since we know that half the area is found by integrating under the positive part of the curve which defines the shape, we may simply evaluate this integral and then double the result to find the area of the full shape.

Let us consider the example of the circle of radius R, centred at the origin. It is defined as the set of points (x, y) satisfying

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

Hence writing the curve $y(x)$ we have $y(x) = \pm\sqrt{R^2 - x^2}$ for $x \in [-R, R]$. There are a pair of curves which are mapped into each other by reflection in the x-axis. Our aim will be to find the area under the positive curve, $y(x) = \sqrt{R^2 - x^2}$, this is the area of the semi-circle as shaded in figure 5.10. Now we may compute the area of the semicircle by integration:

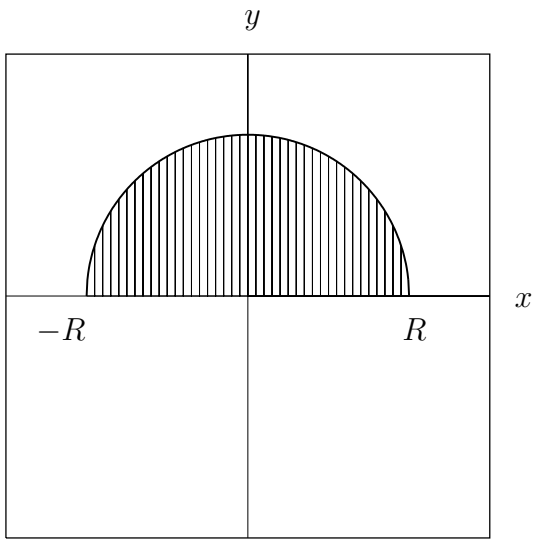


Figure 5.10: The integral of $y(x) = \sqrt{R^2 - x^2}$ from $x = -R$ to $x = R$ gives the area of the semicircle.

$$\begin{aligned}
A_{\text{semicircle}} &= \int_{-R}^{R} y(x)dx \\
&= \int_{-R}^{R} \sqrt{R^2 - x^2}dx \\
&= \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \sqrt{R^2 - R^2\sin^2\theta}(R\cos\theta d\theta) \\
&= \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} R^2\cos^2\theta d\theta \\
&= R^2\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{1}{2}(1+\cos(2\theta))d\theta \\
&= \frac{R^2}{2}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{d}{d\theta}(\theta + \frac{1}{2}\sin(2\theta))d\theta \\
&= \frac{R^2}{2}\left[\theta + \frac{1}{2}\sin(2\theta)\right]_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \\
&= \frac{R^2}{2}((\frac{\pi}{2}+0) - (-\frac{\pi}{2}+0)) \\
&= \frac{1}{2}\pi R^2
\end{aligned}$$

where we made the substitution $x = R\sin\theta$, hence $dx = R\cos\theta d\theta$. This is the area of half the

circle, hence the area of the circle of radius R is

$$A_{\text{circle}} = 2A_{\text{semicircle}} = \pi R^2.$$

One can carry out a similar computation to find the area of an ellipse, this is the subject of the following example.

Example 5.23. *Find the area of the ellipse with foci located at* $(\pm a, 0)$ *where* $a \in \mathbb{R}^+$.

The ellipse is defined by the equation

$$\sqrt{(x-a)^2+y^2} + \sqrt{(x+a)^2+y^2} = 2R$$

where $R \in \mathbb{R}^+$ *and* $R > a$. *We would like to work with an equation of the form* $y = f(x)$, *so we rearrange the equation to obtain*[4]

$$y^2 = R^2 - x^2 - a^2 + \frac{a^2x^2}{R^2}.$$

Hence the upper edge of the ellipse is given by

$$y = \sqrt{R^2 - x^2 - a^2 + \frac{a^2x^2}{R^2}} = \sqrt{(R^2-a^2) - x^2(1-\frac{a^2}{R^2})}$$

. We would like to use an integral to find the area of the shaded region shown in figure 5.11 The area of the ellipse will be twice the area of upper half of the ellipse, i.e. twice the integral of $y(x)$, *hence,*

$$\begin{aligned}
A_{ellipse} &= 2\int_{-R}^{R} y(x)dx \\
&= 2\int_{-R}^{R} \sqrt{(R^2-a^2) - x^2(\frac{R^2-a^2}{R^2})}dx \\
&= 2\int_{-R}^{R} \sqrt{R^2-a^2}\sqrt{1-\frac{x^2}{R^2}}dx \\
&= 2\sqrt{R^2-a^2}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \sqrt{1-\sin^2\theta}(R\cos\theta d\theta) \\
&= 2\sqrt{R^2-a^2}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} R\cos^2\theta d\theta \\
&= 2R\sqrt{R^2-a^2}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{1}{2}(1+\cos(2\theta))d\theta \\
&= \pi R\sqrt{R^2-a^2}
\end{aligned}$$

[4]This is a reasonably lengthy algebraic manipulation, there is no trick used here to simplify the work but we are only presenting the result

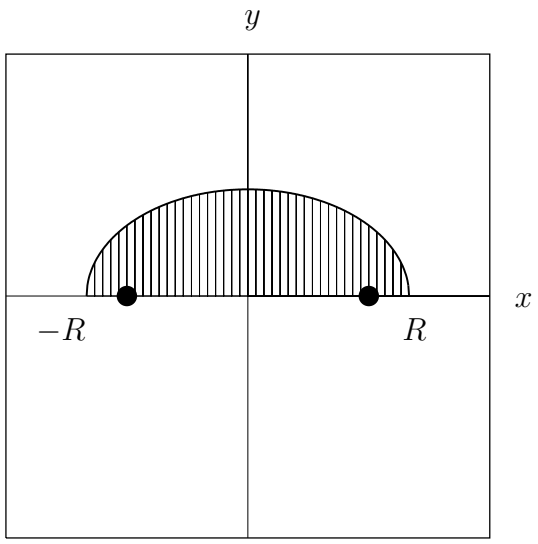


Figure 5.11: The integral of $y(x) = \sqrt{(R^2 - a^2) - x^2(1 - \frac{a^2}{R^2})}$ from $x = -R$ to $x = R$ gives the area of half of the ellipse.

where we have used the substitution $x = R\sin\theta$. Notice that this formula can be used to find the formula for the circle in the limit $a = 0$. Also note that the y-intercept of $y(x)$ is given by $y(0) = \sqrt{R^2 - a^2}$, which is a helpful way to recall the formula for the ellipse.

5.4.2 Volumes of Revolution

By dragging an area along a line we fill out a volume and we may use integration to compute volumes constructed in this way. Some of the most simple volumes (which remain interesting) are constructed by rotating an area about the x-axis: these are highly symmetric and are known as volumes of revolution. We may commence with a continuous function $f(x)$ on $x \in [a, b]$ and then consider slicing it into cross-sectional circles of width w and varying radius as depicted in figure 5.12. Let the cross-sectional area of the solid at $x = x_i$ be denoted $A(x_i)$ then the volume is approximated by

$$V = \sum_{i=1}^{n} wA(x_i) = \sum_{i=1}^{n} w\pi(f(x_i))^2$$

as $A(x_i) = \pi(f(x_i))^2$, the area of a circle of radius $f(x_i)$. If we split the volume into n cylinders of equal width w then we have $x_i = a + (i-1)w$ and as $w = \frac{b-a}{n}$ then $x_i = a + (i-1)\frac{b-a}{n}$, so that $x_{n+1} = b$ and $x_1 = a$. Now in the limit $w \to 0$ we have (by the definition of the Riemann

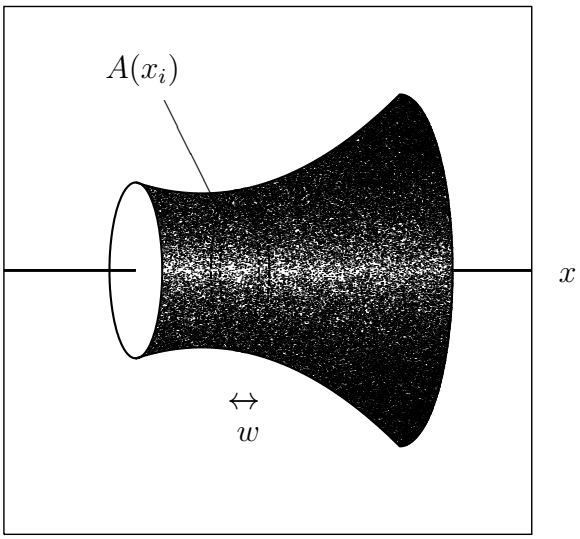


Figure 5.12: A volume whose surface is formed by rotating a continuous curve $y = f(x)$ about the x-axis may be sliced into thin circles of width w and radius $f(x_i)$ at $x = x_i$. In the limit when $w \to 0$ the circle has area $A(x_i) = \pi f(x_i)^2$.

integral):

$$\lim_{w\to 0, n\to\infty}(V) = \lim_{w\to 0, n\to\infty}(\sum_{i=1}^{n} w\pi(f(x_i))^2) = \pi\int_a^b f(x)^2 dx.$$

This integral above is the definition of the volume of revolution of the continuous function $f(x)$ about the x-axis from $x = a$ to $x = b$.

Example 5.24. *Use the formula for the volume of revolution to find the volume of the sphere of radius R.*

By rotating the circle defined by $x^2 + y^2 = R^2$ around the x-axis we find the volume of the

sphere of radius R:

$$\begin{aligned} V &= \pi \int_{-R}^{R} (f(x))^2 dx \\ &= \pi \int_{-R}^{R} (R^2 - x^2) dx \\ &= \pi \int_{-R}^{R} \frac{d}{dx}(R^2 x - \frac{1}{3}x^3) dx \\ &= \pi \left[R^2 x - \frac{1}{3}x^3 \right]_{-R}^{R} \\ &= \pi((R^3 - \frac{1}{3}R^3) - (-R^3 + \frac{1}{3}R^3)) \\ &= \pi(\frac{2}{3}R^3 + \frac{2}{3}R^3) \\ &= \frac{4}{3}\pi R^3. \end{aligned}$$

Exercise 5.8. *The volume bounded by the ellipse when it is rotated about the x-axis is called an ellipsoid or sometimes a cigar. Use the formula for the volume of revolution to show that the volume formed by rotating the ellipse defined by*

$$\sqrt{(x-a)^2 + y^2} + \sqrt{(x+a)^2 + y^2} = 2R$$

about the x-axis is

$$\frac{4}{3}\pi R(R^2 - a^2).$$

5.4.3 The Length of a Curve

A seemingly simple question can now be answered using the integral: how does one measure the length of a curve $f(x)$ from $x = a$ to $x = b$? The answer is that one can use Pythagorus' theorem to compute the straight line distance between points (x_i, y_i) and (x_{i+1}, y_{i+1}) on a curve. If we follow our usual construction and split the curve into n segments of equal width w for the length between $f(a)$ and $f(b)$ we have $w = \frac{b-a}{n}$, and $x_i = a + (i-1)w$. Now the straight line distance from $(x_i, f(x_i))$ to $(x_{i+1}, f(x_{i+1}))$ is given by

$$l_i = \sqrt{(x_{i+1} - x_i)^2 + (f(x_{i+1}) - f(x_i))^2} = w\sqrt{1 + \left(\frac{f(x_{i+1}) - f(x_i)}{w}\right)^2}$$

as $w = x_{i+1} - x_i$. Now the full length of the curve is found by summing the l_i and taking the limit as $w \to 0$ while $x_1 = a$ and $x_{n+1} = b$ are held fixed (i.e. this is the usual prescription for

constructing the Riemann integral), so we find

$$\begin{aligned}\text{Length, L} &\equiv \lim_{w\to 0, n\to\infty}\left(\sum_{i=1}^{n} l_i\right)\\ &= \lim_{w\to 0, n\to\infty}\left(\sum_{i=1}^{n} w\sqrt{1+\left(\frac{f(x_{i+1})-f(x_i)}{w}\right)^2}\right)\\ &= \int_a^b dx\sqrt{1+\left(\frac{df}{dx}\right)^2}.\end{aligned}$$

Example 5.25. *Find the length of the circumference of a circle of radius R.*

The circumference is twice the length of the curve $y=\sqrt{R^2-x^2}$, from $x=-R$ to $x=R$. We compute the ingredients used to find the length of the curve, namely, by taking the derivative with respect to x on both sides of $y^2=R^2-x^2$:

$$2y\frac{dy}{dx} = -2x \qquad \Longrightarrow \qquad \frac{dy}{dx} = -\frac{x}{y}.$$

Therefore,

$$1+\left(\frac{dy}{dx}\right)^2 = 1+\frac{x^2}{y^2} = \frac{x^2+y^2}{y^2} = \frac{R^2}{y^2} = \frac{R^2}{R^2-x^2}.$$

Hence,

$$\begin{aligned}L &= 2\int_{-R}^{R}\sqrt{\frac{R^2}{R^2-x^2}}dx\\ &= 2\int_{-R}^{R}\frac{R}{\sqrt{R^2-x^2}}dx\\ &= 2\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\frac{1}{\sqrt{R^2-R^2\sin^2\theta}}(R\cos\theta d\theta)\\ &= 2R\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}d\theta\\ &= 2\pi R.\end{aligned}$$

Of course one can use this technique to evaluate the lengths of more interesting curves, for example the length of a spiral. However we will leave the examples of greater complexity to the problem sheets. Instead here we will make a small digression to make an interesting comment that is contained in one of the more seemingly paradoxical mathematical results of the twentieth century which was commented on in a famous mathematical paper.

The famous paper was called "How Long is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension." It was written by Benoît Mandelbrot and published in 1967 and involved self-similar curves called fractals. One of his motivations was an observation of Lewis Fry Richardson[5] that the length of a coastline depends on the measurement scale that is used to measure it. Empirical measurement confirms that the smaller the ruler used the longer the coastline is measured to be! Our integral expressions for the length of a curve allow us to understand the idea as $w \to 0$ (our length scale is decreased) the curve length increases to a maximum (but not infinite) length. The curves we are considering are well-behaved, continuous curves so we might conclude that result concerning the length of the coastline of Britain is perfectly sensible, however we might also wonder whether the coastline of Britain is not given by a continuous function. The paradox contained in the measurement leads us to ponder the question: what is the smallest unit of measurement in the natural world and how long is a coastline measured using this length as a basic unit. We might speculate that this basic unit is of sub-nuclear length and then we might wonder further where exactly is the boundary of the nucleus - i.e. what boundary should we measure along at these small scales? We might take heart if we know that most fundamental physics involves descriptions of the world (often) in terms of continuous functions. So there may be hope yet of measuring the coastline of Britain in the 21$^{\text{st}}$ century.

[5]Not the Hollyoaks character but a real English mathematician who lived from 1881 to 1953.

6. Power Series

In which we return to study the infinite sum and give mathematically sound criteria when such a sum exists. We meet the radius of convergence and show that the function e^x and the trigonometric and hyperbolic functions are well-defined. We state Taylor's theorem which gives a way to write a function locally as a power series and finally we show that an application of Taylor's theorem helps in the computation of many limits.

The material in this chapter will be covered in weeks 11-12.

At the start of this course we asked the question: how many functions are there? Our aim was not to give an answer to this question but to guide us to thinking about classes of functions. In the course we have met many large classes of functions, for example polynomial functions, trigonometric functions, inverse functions and so on. Throughout this tour of functions we were haunted by the notion that some functions were more fundamental than others (i.e. they could be used to form the other functions through addition, multiplication or composition) and indeed that some functions were better behaved than other well-defined functions (for example continuous functions and differentiable functions). In this chapter we will combine some of these thoughts to define a purist's notion of a well-behaved function: this will be the class of analytic functions. To define an analytic function we will also face another puzzle that has been with us in the course since we introduced the definition of the exponential function as a series: how do we know when an infinite sum is well-defined and converges? We will use these ideas to give the definition of a Taylor series expansion of a function about a point and then be able to say precisely what the properties of an analytic function are.

6.1 Infinite sums

Definition 6.1.1. *An infinite series is an expression of the form* $\sum_{n=n_0}^{\infty} a_n$ *where* $n, n_0 \in \mathbb{Z}$. *For a real series* $a_n \in \mathbb{R}$.

Comment(s). *(On infinite series...)*

1. *Typically* $n_0 = 0$ *or* 1, *hence a series is simply a sum with an infinite number of terms, for example*
$$\sum_{n=0}^{\infty} \frac{1}{2^n} = 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots$$
where here $a_n = \frac{1}{2^n}$.

2. *A partial sum of an infinite series* $S \equiv \sum_{n=n_0}^{\infty} a_n$ *has the form* $S_N \equiv \sum_{n=n_0}^{N} a_n$. *This will be a useful tool to investigate properties of a series* S.

3. *We can immediately think of infinite series which do not sum to a finite number*[1] *(e.g.* $\sum_{n=0}^{\infty} n$*), so there are infinite series which do not sum to a finite number and are not well-defined. As mathematicians we would like to know when this is sensible.*

6.2 Convergence

As mathematicians we are expected to know whether or not we can find the sum of a series (infinite or not). It is not common in everyday life that sums do not converge but in mathematics this is very common, and such a sum or series is called divergent. Let us formally define these terms.

Definition 6.2.1. *A series* $S \equiv \sum_{n=n_0}^{\infty} a_n$ *is called convergent if the limit* $S \equiv \lim_{N\to\infty}(S_N)$ *exists, where* $S_N \equiv \sum_{n=n_0}^{N} a_n$. *If a series is not convergent it is called divergent.*

What can we say if we are given a series and asked to determine whether it is convergent or divergent? There is something we can deduce immediately about any convergent series, namely

[1] The example used here has made it into popular culture because it is a very important divergent sum that appears, among other places, in string theory, and because the oft-quoted claim that it equals $-\frac{1}{12}$ is disturbing. One can show that this sum can be split in a sensible way into a finite part (equal to $-\frac{1}{12}$) and an infinite part, and there is a method called ζ-regularisation which is used to throw away the infinite part, however sometimes you will see it claimed erroneously that $\sum_{n=0}^{\infty} n = -\frac{1}{12}$, where in reality it is only after the regularisation which removes an infinite part that this is the case. Without any regularisation it is a divergent sum.

that if an infinite series $S \equiv \sum_{n=n_0}^{\infty} a_n$ is convergent then we can say that:

$$\lim_{n\to\infty}(a_n) = 0.$$

This property of a convergent series is known as "the vanishing condition." One can see it is true by considering the difference between two convergent sums and taking a limit:

$$\lim_{n\to\infty}(S_n - S_{n-1}) = \lim_{n\to\infty}(S_n) - \lim_{n\to\infty}(S_{n-1}) = S - S = 0$$

but notice that $S_n - S_{n-1} = a_n$, hence we have the vanishing condition for any convergent series. For example consider the convergent series:

$$\sum_{n=0}^{\infty} \frac{1}{2^n} = 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots$$

where $a_n = \frac{1}{2^n}$. It is clear that the vanishing condition is satisfied on a_n as

$$\lim_{n\to\infty} \frac{1}{2^n} = 0$$

and on the other hand we can compute the partial sum using the geometric identity to find

$$S_N = \sum_{n=0}^{N} \frac{1}{2^n} = \frac{1 - \frac{1}{2^{N+1}}}{1 - \frac{1}{2}} = \frac{2 - \frac{1}{2^N}}{2 - 1} = 2 - \frac{1}{2^N}$$

and hence

$$\lim_{N\to\infty}(S_N) = \lim_{N\to\infty}(2 - \frac{1}{2^N}) = 2.$$

So this infinite sum is convergent and it satisfies the vanishing condition. However the vanishing condition is a *necessary condition for a convergent series but it is not a* sufficient one. That is, the vanishing condition is true for all convergent series, but it is not a criterion to base a test for convergence on, as there are some divergent series which pass the vanishing condition test as well. For example the series

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

does not converge, but does satisfy the vanishing condition as $\lim_{n\to\infty}(\frac{1}{n}) = 0$. There is a cunning argument to show that this sum does not converge. First consider the partial sum for this series:

$$S_N \equiv \sum_{n=1}^{N} \frac{1}{n}$$

and consider the difference between a pair of its partial sums

$$\begin{aligned} S_{2N} - S_N &= (1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{N} + \frac{1}{N+1} + \ldots + \frac{1}{2N}) - (1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{N}) \\ &= \frac{1}{N+1} + \ldots + \frac{1}{2N}. \end{aligned}$$

Now we observe that the smallest term in this difference of the two partial sums is the final one $\frac{1}{2N}$ and as there are N terms in the expression we have:

$$S_{2N} - S_N = \frac{1}{N+1} + \ldots + \frac{1}{2N} \geq \frac{N}{2N} = \frac{1}{2}.$$

But if a series T converges then the limit $\lim_{N\to\infty} T_N = T$ exists for all partial sums and hence if a series converges then we find

$$\lim_{N\to\infty} (T_{2N} - T_N) = T - T = 0$$

but for the series $\sum_{n=1}^{\infty} \frac{1}{n}$ we have found that

$$\lim_{N\to\infty} (S_{2N} - S_N) \geq \frac{1}{2}$$

hence we conclude that the series diverges even though it satisfies the vanishing condition.

The series $\sum_{n=n_0}^{\infty} \frac{1}{n}$ diverges as the terms in the sum do not become small enough at a fast enough rate. Whereas the series $\sum_{n=n_0}^{\infty} \frac{1}{n(n+1)}$ has terms which diminish in size at faster rate than those in the series $\sum_{n=n_0}^{\infty} \frac{1}{n}$ and does converge. Let us convince ourselves of this. The partial sum of this series is

$$\begin{aligned} S_N &\equiv \sum_{n=1}^{N} \frac{1}{n(n+1)} \\ &= \sum_{n=1}^{N} \left(\frac{1}{n} - \frac{1}{(n+1)} \right) \\ &= \left(\frac{1}{1} - \frac{1}{2} \right) + \left(\frac{1}{2} - \frac{1}{3} \right) + \left(\frac{1}{3} - \frac{1}{4} \right) + \ldots \left(\frac{1}{N} - \frac{1}{N+1} \right) \\ &= 1 - \frac{1}{N+1}. \end{aligned}$$

Hence,

$$S \equiv \lim_{N\to\infty} (S_N) = \lim_{N\to\infty} \left(1 - \frac{1}{N+1} \right) = 1.$$

To recap the series $\sum_{n=1}^{\infty} \frac{1}{n}$ is divergent, while the series $\sum_{n=1}^{\infty} \frac{1}{n(n+1)} = 1$ is convergent. It will be an interesting challenge to understand the relation between the exponent of n in the series term a_n and the convergence of the series. However our more immediate challenge is to develop some criteria which we can use to determine whether a series converges or not.

6.2.1 Series Convergence Criteria.

We now state three tests which can be used to determine whether a series $\sum_{n=n_0}^{\infty} a_n$, (we will adopt the shorthand notation $\sum_n a_n \equiv \sum_{n=n_0}^{\infty} a_n$ for the series):

(C1) (*The Limit Comparison Test.*) If both $a_n > 0$ and $b_n > 0$ for all $n \in \mathbb{Z}^+$ and if $\lim_{n\to\infty}\left(\frac{a_n}{b_n}\right) = L$ where $0 < L < \infty$, then either both series $\sum_n a_n$ and $\sum_n b_n$ are convergent or both series are divergent.

(C2) (*The n'th Root Test or Cauchy's Criterion.*)

$$\lim_{n\to\infty} |a_n|^{\frac{1}{n}} \begin{cases} < 1 & \sum_n a_n \quad \text{is convergent} \\ > 1 & \sum_n a_n \quad \text{is divergent} \end{cases}$$

(C3) (*The Ratio Test or D'Alembert's Criterion.*)

$$\lim_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| \begin{cases} < 1 & \sum_n a_n \quad \text{is convergent} \\ > 1 & \sum_n a_n \quad \text{is divergent.} \end{cases}$$

Comment(s). *(On the convergence criteria...)*

1. *The first test uses a comparison against another known convergent series to determine if a series converges, the second and third tests require only knowledge of the series in question.*

2. *We draw attention to the fact that in (C2) and (C3) certain cases are omitted, specifically when* $\lim_{n\to\infty} |a_n|^{\frac{1}{n}} = 1$ *and* $\lim_{n\to\infty} |\frac{a_{n+1}}{a_n}| = 1$*, the root and the ratio tests are both inconclusive.*

3. *The root test (C2) is more powerful than the ratio test (C3), but often (C3) is simpler to work with.*

Proofs of these convergence criteria are not part of this course but they are interesting, so we include a proof of the limit comparison test here in the lecture notes but these are included for fun and will not be examinable!

Proof of (C1): We will make use of the following lemma[2] which is known as the direct comparison test for convergence:

[2] "A subsidiary or intermediate theorem in an argument or proof", from the Greek meaning 'something assumed'. You might like to think of it as meaning a little theorem.

Lemma 6.1. *Suppose that* $0 < a_n \leq b_n$ *for all* $n \in \mathbb{Z}^+$ *then if* $\sum_{n=1}^{\infty} b_n$ *converges then so does* $\sum_{n=1}^{\infty} a_n$.

Proof of lemma: Let the partial sums of the two series be denoted by $S_N \equiv \sum_{n=1}^{N} a_n$ and $T_N \equiv \sum_{n=1}^{N} b_n$ then as $a_n \leq b_n$ for all $n \in \mathbb{Z}$ then $S_N \leq T_N$ for all $N \in \mathbb{Z}$. Consequently (as $0 < a_n \leq b_n$), $\lim_{N\to\infty}(S_N) \leq \lim_{N\to\infty}(T_N) = \sum_n b_n$ as $\sum_n b_n$ is convergent.

Now if $\lim_{n\to\infty}(\frac{a_n}{b_n}) = L$, where $0 < L < \infty$ (as $a_n > 0$ and $b_n > 0$), then (by the definition of the existence of the asymptotic limit) there exists some N such that for all $n > N$

$$-\epsilon < \frac{a_n}{b_n} - L < \epsilon$$

hence,

$$(L-\epsilon)b_n < a_n < (L+\epsilon)b_n.$$

Hence $a_n < (L+\epsilon)b_n$ for all $n > N$ hence if $\sum_n b_n$ is convergent then $(L+\epsilon)\sum_n b_n$ converges and so, by the lemma, $\sum_{n>N} a_n$ is convergent. Now $\sum_{n=1}^{N} a_n$ being a finite sum of real numbers is convergent hence:

$$\sum_n a_n = \sum_{n=1}^{N} a_n + \sum_{n=N}^{\infty} a_n$$

is convergent as it is the sum of two convergent series.

If, alternatively, we know that $\sum_n a_n$ is convergent then by choosing N large enough we can guarantee that $L - \epsilon$ is positive and hence

$$b_n < \frac{a_n}{L-\epsilon}$$

and by the lemma $\sum_n b_n$ is convergent. Hence if either series is convergent, then both are, or, alternatively, if either series is divergent then both are. □

Let us practise using the limit comparison test.

Example 6.1. *Use the limit comparison test (C1) to show that*

$$\sum_{n=1}^{\infty} \frac{1}{n^2}$$

is convergent.

Let $a_n = \frac{1}{n^2}$ *and now we must identify a convergent series* $\sum_n b_n$ *such that* $\lim_{n\to\infty} \frac{a_n}{b_n}$ *exists. We will compare the series with* $\sum_{n=1}^{\infty} \frac{1}{n(n+1)}$ *which we have already shown is convergent, so we have* $b_n = \frac{1}{n(n+1)}$ *and* $a_n \geq 0$, $b_n \geq 0$. *Now,*

$$\lim_{n\to\infty} \frac{a_n}{b_n} = \lim_{n\to\infty} \frac{(\frac{1}{n^2})}{(\frac{1}{n(n+1)})} = \lim_{n\to\infty} \left(\frac{n(n+1)}{n^2}\right) = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right) = 1.$$

Hence as the limit exists and $\sum_n b_n$ is convergent then $\sum_n \frac{1}{n^2}$ is convergent.

Example 6.2. *Show that*

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n}$$

is convergent.

Consider the partial sums:

$$\begin{aligned} S_N &= \sum_{n=1}^{N} \frac{(-1)^{n+1}}{n} \\ &= \sum_{n,\text{odd}}^{N} \frac{1}{n} - \sum_{n,\text{even}}^{N} \frac{1}{n} \\ &= \begin{cases} \sum_{m=1}^{N/2} \frac{1}{2m-1} - \sum_{m=1}^{N/2} \frac{1}{2m} = \sum_{m=1}^{N/2} \frac{1}{2m(2m-1)} & \text{for even } N \\ \sum_{m=1}^{(N+1)/2} \frac{1}{2m-1} - \sum_{m=1}^{(N-1)/2} \frac{1}{2m} = \sum_{m=1}^{(N-1)/2} \frac{1}{2m(2m-1)} + \frac{1}{N} & \text{for odd } N \end{cases} \end{aligned}$$

hence the series will only converge if $\sum_m \frac{1}{2m(2m-1)}$ is convergent. Now we may use the limit comparison test (C1) to compare this series against the convergent series $\sum_n \frac{1}{n^2}$, i.e. taking $a_n = \frac{1}{2n(2n-1)}$ and $b_n = \frac{1}{n^2}$, then

$$\lim_{n\to\infty} \frac{a_n}{b_n} = \lim_{n\to\infty} \frac{\left(\frac{1}{2n(2n-1)}\right)}{\left(\frac{1}{n^2}\right)} = \lim_{n\to\infty} \left(\frac{n^2}{2n(2n-1)}\right) = \lim_{n\to\infty} \left(\frac{n^2}{4n^2-2n}\right) = \frac{1}{4}.$$

Therefore as the limit exists and $\sum_n b_n$ is convergent then $\sum_n \frac{1}{2n(2n-1)}$ is convergent and hence $\sum_n \frac{(-1)^{n+1}}{n}$ is convergent.

6.3 Series as Functions of x.

Definition 6.3.1. *A power series is an expression of the form $S(x) \equiv \sum_{n=n_0}^{\infty} b_n x^n$.*

We may think of a power series as giving a series for any given value of x, that is a power series is a series $\sum_n a_n$ where the terms $a_n = b_n x^n$ each depend on x. We might immediately wonder how $S(x)$ varies as x varies: in particular is $S(x)$ convergent for all or any values of x? Now we may make use of the convergence criteria, in particular Cauchy's $n'th$ root test (C2) and D'Alembert's ratio test (C3) for the series $\sum_n a_n$ by substituting $a_n = b_n x^n$ into the

convergence tests. Doing this we find from (C2) that if

$$\lim_{n\to\infty} |a_n|^{\frac{1}{n}} = \lim_{n\to\infty} |b_n x^n|^{\frac{1}{n}} \begin{cases} < 1 & S(x) \text{ converges} \\ > 1 & S(x) \text{ diverges.} \end{cases}$$

We are interested in finding a condition on x to express the convergence criteria, so it is notationally useful to define $\lim_{n\to\infty} |b_n|^{-\frac{1}{n}} \equiv R_1$ as then the convergence criterion $C2$ is expressed as

$$\lim_{n\to\infty} |b_n x^n|^{\frac{1}{n}} = \lim_{n\to\infty} |b_n|^{\frac{1}{n}} |x^n|^{\frac{1}{n}} = \left(\frac{1}{R_1}\right) \lim_{n\to\infty} |x^n|^{\frac{1}{n}} = \left(\frac{1}{R_1}\right) \lim_{n\to\infty} |x| < 1.$$

Now note that $\lim_{n\to\infty} |x| = |x|$ and therefore the power series $S(x)$ will converge if

$$|x| < R_1$$

and $S(x)$ will diverge if $|x| > R_1$. R_1 is called the radius of convergence as it delimits the points where the power series is well-defined and converges from the points where it diverges. From the convergence criterion (C3) we will find an alternative definition of the radius of convergence. Substituting $a_n = b_n x^n$ into (C3) we find that if

$$\lim_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| = \lim_{n\to\infty} \left|\frac{b_{n+1}x^{n+1}}{b_n x^n}\right| = \lim_{n\to\infty} \left|\frac{b_{n+1}x}{b_n}\right| \begin{cases} < 1 & S(x) \text{ converges} \\ > 1 & S(x) \text{ diverges.} \end{cases}$$

Let us define $\lim_{n\to\infty} \left|\frac{b_{n+1}}{b_n}\right|^{-1} \equiv R_2$ then $S(x)$ will converge for all x such that

$$|x| < R_2$$

and will diverge for all x such that $|x| > R_2$. Naturally we would only expect there to be a single value $R = R_1 = R_2$ which is the radius of convergence for a power series. The relation between R_1 and R_2 is not part of this course, but of course there is an interesting interaction between the pair of radii of convergence. At the end of this chapter we will include an appendix for the interested reader who wishes to understand more about the two formulae for the radius of convergence of a power series (which is not an examinable part of the course, but is interesting).

Definition 6.3.2. *The radius of convergence of a power series $\sum_n b_n x^n$ is a non-negative number R such that the series converges for all x such that $|x| < R$ and diverges for all $|x| > R$, where*

$$R = \lim_{n\to\infty} |b_n|^{-\frac{1}{n}}$$

and, where it exists,

$$R = \lim_{n\to\infty} \left|\frac{b_n}{b_{n+1}}\right|.$$

Comment(s). *(On the radius of convergence...)*

1. *For $|x| = R$ there is no general conclusion to be drawn, one must inspect the series being considered using some alternative method.*
2. *If $R = 0$ then the series diverges for all $x \neq 0$.*
3. *If $R = \infty$ the series converges for all $x \in \mathbb{R}$.*
4. *The limit $R = \lim_{n\to\infty} \left|\frac{b_n}{b_{n+1}}\right|$ is often the simplest limit to compute, but it only exists if the power series has consecutive terms which are non-zero. Many power series have only even or odd terms, in which case this limit is not defined. When both definitions of the radius of convergence exist they agree, and in the appendix you can see that the radius coming from the ratio comparison test implies the existence of the other definition (from the root test) of the radius of convergence, and in that case they coincide.*

Example 6.3. *Show that the power series*

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

is convergent for all $x \in \mathbb{R}$.

Here the series has coefficients $b_n = \frac{1}{n!}$, hence

$$R = \lim_{n\to\infty} \left|\frac{b_n}{b_{n+1}}\right| = \lim_{n\to\infty} \left|\frac{(\frac{1}{n!})}{(\frac{1}{(n+1)!})}\right| = \lim_{n\to\infty} \left(\frac{(n+1)!}{n!}\right) = \lim_{n\to\infty} (n+1) = \infty.$$

Hence e^x converges for all $|x| < \infty$, i.e. for all $x \in \mathbb{R}$. At this stage, we may recall (with some satisfaction) all the results that rested upon this observation, including the exponential notation for the trigonometric and hyperbolic functions, Euler's formula, the polar coordinate notation for complex numbers, the roots of unity, trigonometric identities for double and half angle formulae and hence all the integral and derivatives which required these identities.

Example 6.4. *Find the radius of convergence for the power series*

$$S(x) = \sum_{n=0}^{\infty} x^n.$$

Here the series has coefficients $b_n = 1$, hence

$$R = \lim_{n\to\infty}\left|\frac{b_n}{b_{n+1}}\right| = \lim_{n\to\infty}\left|\frac{1}{1}\right| = 1.$$

Hence $S(x)$ converges for all $|x| < 1$. From the identity for the geometric sum of a finite number of terms we know that

$$\sum_{n=0}^{N} x^n = \frac{1-x^{N+1}}{1-x} \qquad \forall\, x \neq 1.$$

This is a partial sum for $S(x)$ and as $\lim_{N\to\infty}(\sum_{n=0}^{N} x^n) = S(x)$, and if $|x| < 1$ this sum converges. Hence we have

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x} \qquad \forall\, |x| < 1$$

where we have observed that $\lim_{N\to\infty}(x^{n+1}) = 0$ for $|x| < 1$.

To gain confidence in the convergence of power series one might compare $\sum_{n=0}^{N} x^n$ for various N with $\frac{1}{1-x}$ by plotting the graphs for $N = 1, 2, 3, \ldots$. These graphs are plotted together with $y = \frac{1}{1-x}$ in the graph in figure 6.1. Notice how the graphs of $y = \sum_{n=0}^{N} x^n$ coincide with $y = \frac{1}{1-x}$ agreement (improving as N increases) in the region where $|x| < 1$. We might phrase this with a different emphasis and say that the function $\frac{1}{1-x}$ has a power series expansion about $x = 0$ for all $|x| < 1$. This foreshadows the idea of the Taylor power series expansion for a function.

Example 6.5. *Find the radius of convergence for* $\sin x$.

The first problem is to write $\sin x$ *as a power series, we have,*

$$\sin(x) = \sum_{m=0}^{\infty}(-1)^m \frac{x^{2m+1}}{(2m+1)!} = \sum_{n=0}^{\infty} b_n x^n.$$

Now we have

$$b_n = \begin{cases} 0 & \text{if } n \text{ is even} \\ \frac{(-1)^m}{(2m+1)!} = \frac{(-1)^{\frac{n-1}{2}}}{n!} & \text{if } n \text{ is odd} \end{cases}$$

N.B. we have taken $n = 2m+1$ to denote odd n and so $m = \frac{n-1}{2}$ has been substituted to find the expressions above. As $|\frac{b_n}{b_{n+1}}|$ is either zero or undefined for any consecutive pair of coefficients, and so the limit as $n \to \infty$ does not exist. Consequently we will use the alternative definition of the radius of convergence

$$R = \lim_{n\to\infty}\left(|b_n|^{-\frac{1}{n}}\right).$$

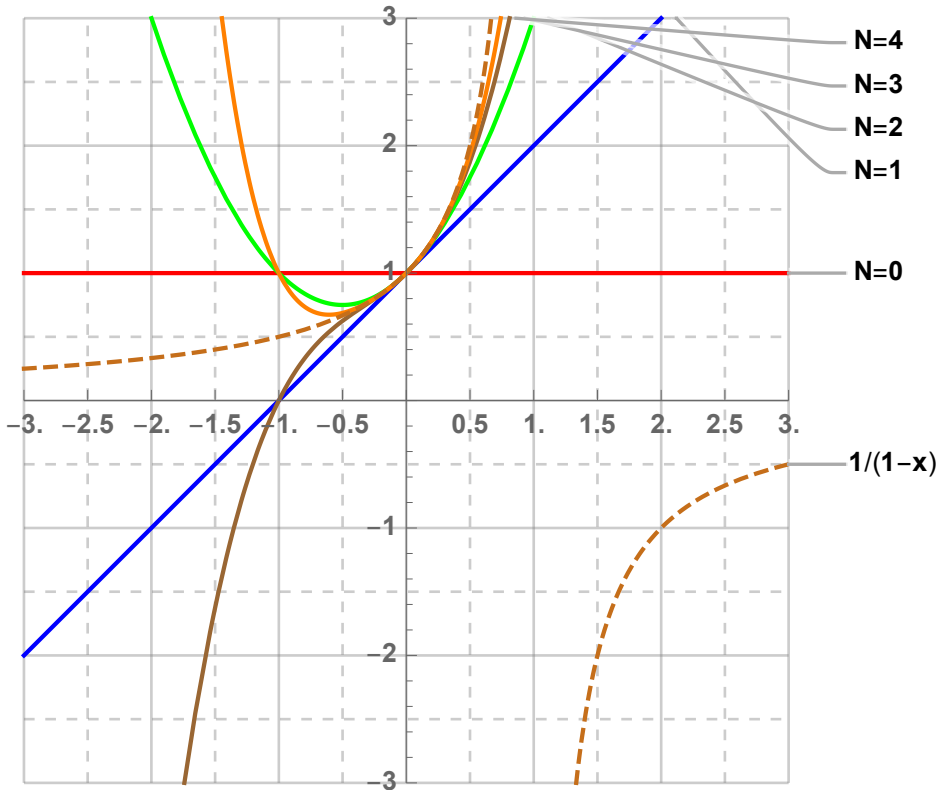


Figure 6.1: The graphs of $y = \sum_{n=0}^{N} x^n$ for $N = 1, 2, 3, 4$ with the various N shown, at the edge of the graph. The dashed line is the graph of $y = \frac{1}{1-x}$.

For odd n this limit becomes

$$R = \lim_{n\to\infty}\left(\left|(-1)^{\frac{n-1}{2}}\frac{1}{n!}\right|^{-\frac{1}{n}}\right) = \lim_{n\to\infty}\left(\left|\frac{1}{n!}\right|^{-\frac{1}{n}}\right) = \lim_{n\to\infty}\left(|n!|^{\frac{1}{n}}\right).$$

To show that this limit exists we can use Stirling's approximation for $n!$ which is given by:

$$n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n} \qquad \textit{for } n >> 1$$

this approximation is valid as $n \to \infty$. Hence we have,

$$R = \lim_{n\to\infty}\left(|n!|^{\frac{1}{n}}\right) \approx \lim_{n\to\infty}\left(|\left(\frac{n}{e}\right)^n\sqrt{2\pi n}|^{\frac{1}{n}}\right) = \lim_{n\to\infty}\left(\frac{n}{e}\right)\lim_{n\to\infty}\left((2\pi n)^{\frac{1}{2n}}\right) = \infty.$$

We are able to split the product of the limits as,

$$\lim_{n\to\infty}(n^{\frac{1}{2n}}) = \lim_{n\to\infty}(e^{\ln(n^{\frac{1}{2n}})}) = \lim_{n\to\infty}(e^{\frac{\ln n}{2n}}) = e^{(\lim_{n\to\infty}(\frac{\ln n}{2n}))} = e^0 = 1.$$

Therefore as the radius of convergence is infinite, the sine function converges for all $x \in \mathbb{R}$. This was a rather difficult conclusion to reach using the root test for convergence, alternatively

we could have argued that $sin(x)$ converges by using the triangle inequality:

$$|\sin(x)| = \left| \sum_{m=0}^{\infty} (-1)^m \frac{x^{2m+1}}{(2m+1)!} \right| \leq \sum_{m=0}^{\infty} \left(\frac{|x|^{2m+1}}{(2m+1)!} \right) \leq \sum_{m=0}^{\infty} \frac{|x|^m}{m!} = e^{|x|}$$

as e^x converges for all $x \in \mathbb{R}$ then $\sin(x)$ converges for all $x \in \mathbb{R}$.

6.4 Taylor's Theorem

Brook Taylor, a contemporary of Isaac Newton, based at Cambridge was one of the finest mathematicians of his generation but he is most frequently remembered as the originator of Taylor series, a way to construct a convergent power series for a function $f(x)$. To express the result efficiently, it is useful to take up some new notation, let

$$f^n(x_0) \equiv \frac{d^n}{dx^n}(f(x))\bigg|_{x=x_0}$$

be a shorthand notation for the n'th derivative of $f(x)$ evaluated at the point $x = x_0$. Using this notation we make the following claim which will form a basis for Taylor's theorem: if a function $f(x)$ can be written as a power series with a radius of convergence $R > 0$ (i.e. for all $x \in \mathbb{R}$ such that $|x| < R$, then $f(x) = \sum_{n=0}^{\infty} b_n x^n$) and if we may exchange the order of the derivative and the summation for an infinite series (as we can for a finite series) then:

$$b_n = \frac{f^n(0)}{n!}$$

if the n'th derivative of $f(x)$ exists at $x = 0$.
Proof of the claim: If

$$S(x) \equiv \sum_{n=0}^{\infty} \frac{f^n(0)}{n!} x^n \qquad \text{for } |x| < R$$

is the same as $f(x)$ then we can confirm this by checking that all of their derivatives at a point $x = 0$ are equal as well as $f(0) = S(0)$. So,

$$\begin{aligned} S^m(x) &= \frac{d^m}{dx^m}\left(\sum_{n=0}^{\infty} \frac{f^n(0)}{n!} x^n \right) \\ &= \sum_{n=0}^{\infty} \frac{f^n(0)}{n!} n(n-1)(n-2)\dots(n-m+1)x^{n-m} \\ &= \sum_{n=m}^{\infty} \frac{f^n(0)}{(n-m)!} x^{n-m}. \end{aligned}$$

Now at $x = 0$ we have

$$S^m(0) = \sum_{n=m}^{\infty} \frac{f^n(0)}{(n-m)!} 0^{n-m} = f^n(0)$$

and $S(0) = f^0(0) = f(0)$. Hence,

$$f(x) = \sum_{n=0}^{\infty} \frac{f^n(0)}{n!} x^n$$

as claimed. □

Now, because we might be (rightly) concerned about moving $\frac{d}{dx}$ (which is a limit) past an infinite sum (which is another limit) we present a proof that sidesteps this issue and uses the fundamental theorem of calculus. Now,

$$\int_0^x f^{n+1}(y)dy = \int_0^x \frac{d}{dy}(f^n(y))dy = f^n(x) - f^n(0)$$

and after rearranging we have,

$$f^n(x) = f^n(0) + \int_0^x f^{n+1}(y)dy.$$

Our intention is to make repeated use of the identity above. By assumption we will presume that f is differentiable (if the function can be written as a convergent series then it will be infinitely differentiable) and we commence with writing the statement above for the case $n = 0$:

$$\begin{aligned} f(x) &= f(0) + \int_0^x f^1(y)dy \\ &= f(0) + x\int_0^1 f^1(xz_1)dz_1 \\ &= f(0) + x\int_0^1 \left(f^1(0) + \int_0^{xz_1} f^2(y_1)dy_1 \right) dz_1 \\ &= f(0) + xf^1(0) + x\int_0^1 \left(\int_0^{xz_1} f^2(y_1)dy_1 \right) dz_1 \end{aligned}$$

where we have made the substitution $y = xz_1$ so that $dy = xdz_1$ and x is a constant. In the final line we have made use of the identity for $f^n(x)$ given above. Our intention is to repeat

this procedure to derive a series expansion for $f(x)$. We have,

$$\begin{aligned}
f(x) &= f(0) + xf^1(0) + x\int_0^1 \left(\int_0^{xz_1} f^2(y_1)dy_1 \right) dz_1 \\
&= f(0) + xf^1(0) + x^2 \int_0^1 z_1 \left(\int_0^1 f^2(xz_1z_2)dz_2 \right) dz_1 \\
&= f(0) + xf^1(0) + x^2 \int_0^1 z_1 \left(\int_0^1 \left(f^2(0) + \int_0^{xz_1z_2} f^3(y_2)dy_2 \right) dz_2 \right) dz_1 \\
&= f(0) + xf^1(0) + x^2 f^2(0) \int_0^1 z_1 dz_1 + x^2 \int_0^1 z_1 \left(\int_0^1 \left(\int_0^{xz_1z_2} f^3(y_2)dy_2 \right) dz_2 \right) dz_1 \\
&= f(0) + xf^1(0) + \frac{f^2(0)}{2}x^2 + x^2 \int_0^1 z_1 \left(\int_0^1 \left(\int_0^{xz_1z_2} f^3(y_2)dy_2 \right) dz_2 \right) dz_1
\end{aligned}$$

where we have substituted $y_1 = xz_1z_2$, where x is a constant and z_1 is treated as constants within the dy_1 integral so that $dy_1 = xz_1dz_2$. Now we may imagine continuing this procedure of substituting $y_n = xz_1z_2 \dots z_n$ at the n'th iteration and using the identity for $f^n(x)$ to expand the integral about $f^n(0)$. Had we carried on we could have confirmed the expansion for the first set of terms, i.e. that

$$f(x) = f(0) + xf^1(0) + \frac{x^2}{2}f^2(0) + \frac{x^3}{3!}f^3(0) + \dots .$$

However if we stopped the procedure after the first N terms of the Taylor series had been generated we would have been left with an integral $R_{N+1}(x)$ which is known as the remainder term:

$$f(x) = \sum_{n=0}^{N} \frac{f^n(0)x^n}{n!} + R_{N+1}(x)$$

where

$$\begin{aligned}
R_1(x) &= x\int_0^1 f^1(xz_1)dz_1 \\
R_2(x) &= x\int_0^1 \left(\int_0^{xz_1} f^2(y_1)dy_1 \right) dz_1 \\
R_3(x) &= x^2 \int_0^1 z_1 \left(\int_0^1 \left(\int_0^{xz_1z_2} f^3(y_2)dy_2 \right) dz_2 \right) dz_1 \\
&\vdots
\end{aligned}$$

The result we have derived above is an example of a Taylor series, a Taylor series $f(x)$ about $x = 0$ is known as the Maclaurin series for $f(x)$. Let us summarise:

Definition 6.4.1. *The Maclaurin series for a function $f(x)$ is a Taylor series expansion about $x = 0$ and to order N is written*

$$f(x) = \sum_{n=0}^{N} \frac{f^n(0)}{n!} x^n + R_{N+1}(x) \qquad \text{where } R_{N+1}(x) = \frac{1}{N!} \int_0^x f^{N+1}(y)(x-y)^N dy.$$

Why can we write the remainder term this way? We can prove this by induction. Let S_N be the statement that

$$R_{N+1}(x) = f(x) - \sum_{n=0}^{N} \frac{f^n(0)}{n!} x^n.$$

We will prove the inductive step first, namely that $S_N \implies S_{N+1}$.

$$\begin{aligned}
R_{N+2}(x) &= \frac{1}{(N+1)!} \int_0^x f^{N+2}(y)(x-y)^{N+1} dy \\
&= \frac{1}{(N+1)!} \int_0^x \frac{d}{dy}(f^{N+1}(y))(x-y)^{N+1} dy \\
&= -\frac{1}{(N+1)!} \int_0^x f^{N+1}(y) \frac{d}{dy}((x-y)^{N+1}) dy + \frac{1}{(N+1)!} \Big[f^{N+1}(y)(x-y)^{N+1} \Big]_{y=0}^{y=x} \\
&= \frac{1}{N!} \int_0^x f^{N+1}(y)(x-y)^N dy - \frac{1}{(N+1)!} f^{N+1}(0)(x)^{N+1} \\
&= R_{N+1}(x) - \frac{1}{(N+1)!} f^{N+1}(0) x^{N+1} \\
&= f(x) - \sum_{n=0}^{N} \frac{f^n(0)}{n!} x^n - \frac{1}{(N+1)!} f^{N+1}(0) x^{N+1} \\
&= f(x) - \sum_{n=0}^{N+1} \frac{f^n(0)}{n!} x^n
\end{aligned}$$

which is the statement S_{N+1}. Now let us check the basis step for $N = 0$, the left-hand-side of the statement S_0 is

$$R_1(x) = \frac{1}{0!} \int_0^x f^1(y)(x-y)^0 dy = \int_0^x f^1(y) dy = f(x) - f(0).$$

While the right-hand-side of S_0 reads

$$f(x) - \sum_{n=0}^{0} \frac{f^n(0)}{n!} \frac{1}{0!} = f(x) - f(0).$$

Therefore S_0 is a true statement and our proof by induction of the structure of the remainder term is complete.

The Maclaurin series give an approximation for $f(x)$ around zero with the discrepancy being given by the remainder term. In general one is interested in expanding about points $x = a$ where a may be non-zero. This more general case is called the Taylor series. The expansion parameter is no longer $x = (x - 0)$ but in the parameter $(x - a)$:

Definition 6.4.2. *The Taylor series to order N around the point $x = a$ of a function $f(x)$ is*

$$f(x) = \sum_{n=0}^{N} \frac{f^n(a)}{n!}(x-a)^n + R_{N+1}(x) \qquad \text{where } R_{N+1}(x) = \frac{1}{N!}\int_a^x f^{N+1}(y)(x-y)^N dy.$$

One can set about proving this theorem in the same manner as we did for the Maclaurin series, but instead of using x one works with the shifted variable $(x - a)$.

Our aim is to get some familiarity with using the Taylor series by computing the Taylor series for some familiar functions.

Example 6.6. *Find the Maclaurin series for the exponential function up to order N and state the remainder term.*

We have $f(x) = e^x$, hence $f^n(x) = \frac{d^n}{dx^n}(e^x) = e^x$. For the Maclaurin series we need to expand about $x = 0$, so we need $f^n(0) = e^0 = 1$ for all n. Hence the Maclaurin series is

$$e^x = \sum_{n=0}^{N} \frac{x^n}{n!} + R_{N+1}(x)$$

where

$$R_{N+1}(x) = \frac{1}{N!}\int_0^x e^y(x-y)^N dy.$$

Although it is not required for the example, it is reassuring to note that the remainder term vanishes as $N \to \infty$ by the following argument. Let $M \in \mathbb{Z}$ such that $M > (x - y) > 0$ (as $y \in [0, x]$ in the integral) then for all $N > M$

$$(x-y)^{N-M} < M^{N-M} < M(M+1)(M+2)\dots(M+(N-M-1)) = \frac{(N-1)!}{(M-1)!}$$

and so,

$$(x-y)^N < (x-y)^M \frac{(N-1)!}{(M-1)!} < M^M \frac{(N-1)!}{(M-1)!}.$$

Therefore

$$0 \le \lim_{N\to\infty}\left(\frac{(x-y)^N}{N!}\right) \le \lim_{N\to\infty}\left(M^M \frac{(N-1)!}{(M-1)!N!}\right) = \frac{M^M}{(M-1)!}\lim_{N\to\infty}\left(\frac{1}{N}\right) = 0.$$

Consequently we see that in the limit $N \to \infty$ that the remainder term vanishes,

$$\lim_{N\to\infty}\left(R_{N+1}(x)\right) = \int_0^x e^y \lim_{N\to\infty}\left(\frac{(x-y)^N}{N!}\right)dy = 0.$$

This is consistent with our observation that the infinite power series for the exponential function is convergent for all $x \in \mathbb{R}$.

Example 6.7. *Find the Maclaurin series for the sine function up to order N, where N is an odd, positive integer, and state the remainder term.*

We have $f(x) = \sin x$, hence

$$f^n(x) = \frac{d^n}{dx^n}(\sin x) = \begin{cases} (-1)^{\frac{n}{2}} \sin x & \text{for even } n \\ (-1)^{\frac{n-1}{2}} \cos x & \text{for odd } n. \end{cases}$$

Hence we have

$$f^n(0) = \begin{cases} 0 & \text{for even } n \\ (-1)^{\frac{n-1}{2}} & \text{for odd } n. \end{cases}$$

Hence to find the terms up to order $n = N$ we may write odd $n = 2m+1$ for $m \in \{0, 1, 2, 3, \ldots, \frac{N-1}{2}\}$, (recall that N is odd by the assumption stated in the question) so that

$$\sin x = \sum_{m=0}^{\frac{N-1}{2}} \frac{(-1)^m}{(2m+1)!} x^{2m+1} + R_{N+2}(x)$$

where

$$R_{N+2} = \frac{1}{(N+1)!}\int_0^x f^{N+2}(y)(x-y)^{N+1}dy = \frac{1}{(N+1)!}\int_0^x (-1)^{\frac{N+1}{2}} \cos(y)(x-y)^{N+1}dy.$$

We observe here that as the terms of even order in x are zero, the remainder term after the x^N term (where N is odd) begins with terms proportional to x^{N+2}, hence the remainder term is R_{N+2} rather than R_{N+1}.

Example 6.8. *Find the first N (non-zero) terms, where N is even, in the Taylor series for the cosine function expanded about the point $x = a$.*

We have $f(x) = \cos x$, hence

$$f^n(x) = \frac{d^n}{dx^n}(\cos x) = \begin{cases} (-1)^{\frac{n+1}{2}} \sin x & \text{for odd } n \\ (-1)^{\frac{n}{2}} \cos x & \text{for even } n. \end{cases}$$

Hence we have

$$f^n(a) = \begin{cases} (-1)^{\frac{n+1}{2}} \sin a & \text{for odd } n \\ (-1)^{\frac{n}{2}} \cos a & \text{for even } n. \end{cases}$$

To find the first N terms we do not need to consider the remainder term, and including the first term we need only construct terms up to and including order $N-1$ hence we have

$$\begin{aligned} \cos x &= \cos(a + (x-a)) \\ &= \sum_{n=0}^{N-2} \frac{f^n(a)}{n!}(x-a)^n \\ &= \sum_{n,\text{even}}^{N-2} \frac{(x-a)^n}{n!}(-1)^{\frac{n}{2}} \cos a + \sum_{n,\text{odd}}^{N-1} \frac{(x-a)^n}{n!}(-1)^{\frac{n+1}{2}} \sin a \\ &= \sum_{m=0}^{\frac{N-2}{2}} \frac{(x-a)^{2m}}{(2m)!}(-1)^m \cos a + \sum_{m=0}^{\frac{N-2}{2}} \frac{(x-a)^{2m+1}}{(2m+1)!}(-1)^{m+1} \sin a \end{aligned}$$

Where we have written odd $n = 2m+1$ and even $n = 2m$.

Example 6.9. *Find the first 4 (non-zero) terms, in the Maclaurin series for the function*

$$f(x) = \ln(1-x).$$

We have $f(x) = \ln(1-x)$, hence we compute and so we find,

n	$f^n(x) = \frac{d^n}{dx^n}(\ln(1-x))$	$f^n(0)$
0	$\ln(1-x)$	0
1	$\frac{-1}{(1-x)}$	-1
2	$\frac{-1}{(1-x)^2}$	-1
3	$\frac{-2}{(1-x)^3}$	-2
4	$\frac{-6}{(1-x)^4}$	-6

$$\begin{aligned} \ln(1-x) &= \sum_{n=0}^{4} \frac{f^n(0)}{n!}x^n + R_5(x) \\ &= -x - \frac{1}{2}x^2 - \frac{1}{3!}(2)x^3 - \frac{1}{4!}(6)x^4 + R_5(x) \\ &= -x - \frac{1}{2}x^2 - \frac{1}{3}x^3 - \frac{1}{4}x^4 + R_5(x). \end{aligned}$$

If we had continued the expansion to higher order terms the apparent pattern would continue and we would find

$$\ln(1-x) = -\sum_{n=1}^{\infty} \frac{x^n}{n}$$

so long as the general remainder term $R_{N+1}(x)$ vanishes as $N \to \infty$. So let us check when this is the case:

$$\begin{aligned}
\lim_{N\to\infty}(R_{N+1}(x)) &= \lim_{N\to\infty}\left(\frac{1}{N!}\int_0^x f^{N+1}(y)(x-y)^N dy\right) \\
&= -\lim_{N\to\infty}\left(\frac{1}{N!}\int_0^x N!\frac{(x-y)^N}{(1-y)^{N+1}} dy\right) \\
&= -\lim_{N\to\infty}\left(\int_0^x \frac{(x-y)^N}{(1-y)^{N+1}} dy\right) \\
&= -\int_0^x \lim_{N\to\infty}\left(\frac{(x-y)^N}{(1-y)^N}\right)\frac{1}{1-y} dy
\end{aligned}$$

Now

$$\lim_{N\to\infty}\left(\frac{(x-y)^N}{(1-y)^N}\right)$$

converges to zero if $|\frac{x-y}{1-y}| < 1$, i.e. if $|x| < 1$, but if $x > 1$ the remainder term diverges. This is a sign that the power series

$$\ln(1-x) = -\sum_{n=1}^{\infty} \frac{x^n}{n}$$

converges only if $|x| < 1$, we can show this is the case by identifying the radius of convergence. We have $b_n = -\frac{1}{n}$ and so,

$$R = \lim_{n\to\infty}\left|\frac{b_{n+1}}{b_n}\right| = \lim_{n\to\infty}\left|\frac{(\frac{1}{n+1})}{(\frac{1}{n})}\right| = \lim_{n\to\infty}\left|\frac{n}{n+1}\right| = \lim_{n\to\infty}\left|\frac{1}{1+\frac{1}{n}}\right| = 1.$$

Therefore, the series expansion,

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

converges for $|x| < 1$.

We can now re-state the definition of the analytic function

Definition 6.4.3. *A function which can be locally written as a convergent power series is called an analytic function.*

That is if the Taylor series expansion for a function about a point $x = a$ is convergent for some non-zero radius R about the point $x = a$, i.e. for all x $|x - a| < R$, and if this is true for all $a \in \mathbb{R}$, then the function is analytic. For example, as the exponential function has a convergent power series for all x it is an analytic function.

Now we can make a few simple observations about such analytic functions, as we can always find a convergent power series expansion, then the function is differentiable (as we have no trouble in taking the derivative of a power series or infinite polynomial function) and so it is also continuous. Hence the analytic functions are the purist's idea of a well-defined function of one variable. This observation presents the end of the long-term quest for this course.

6.4.1 Short-cuts for finding a Taylor Expansion

There are many ways to find Taylor series which build upon well-known Taylor series without needing to compute the series term by term. Here we list a few neat observations and short-cuts to Taylor and Maclaurin series.

Now

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

is an analytic function (it converges for all $x \in \mathbb{R}$) and the map $x \to -x$ is a symmetry of $\mathbb{R}$ so e^{-x} is also convergent, i.e. we have,

$$e^x = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \dots.$$

We may make a similar observation for

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

which is convergent for $x \in (-1, 1)$, hence

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

is convergent for $x \in (-1, 1)$. We may use (well-defined) operations on analytic functions to find other series expansions, e.g.

$$\frac{1}{1+x} = \frac{d}{dx}(\ln(1 + x)) = \frac{d}{dx}(x - \frac{x^2}{2} + \frac{x^3}{3} - \dots) = 1 - x + x^2 - x^3 + \dots$$

which we might have deduced by summing the geometric series on the right when $|x| < 1$. A

final, more involved example, of manipulating convergent series is

$$\begin{aligned}\frac{1}{\cos(x)} &= \frac{1}{1-\frac{x^2}{2!}+\frac{x^4}{4!}-\ldots} \\ &\equiv \frac{1}{1+Y} \\ &= 1-Y+Y^2-Y^3+\ldots \\ &= 1-(-\frac{x^2}{2!}+\frac{x^4}{4!}-\ldots)+(-\frac{x^2}{2!}+\frac{x^4}{4!}-\ldots)^2+\ldots \\ &= 1+\frac{x^2}{2!}-\frac{x^4}{4!}+\frac{x^4}{4}+\ldots \\ &= 1+\frac{x^2}{2!}+\frac{5x^4}{4!}+\ldots\end{aligned}$$

where we have limited $|Y|<1$, notice that in general $Y=\cos(x)-1\in[-2,0]$, so this restriction on Y corresponds to constraining $0\leq\cos(x)\leq 1$. By attempting other manipulations of functions where they are know to be analytic one can come up with other wonderful examples of Taylor series.

We conclude the course with a wonderful observation which is built upon the Taylor series and is fantastically useful for computing limits.

6.5 l'Hôpital's Rule

The rule is named after the 17^{th} century French mathematician Guillaume de l'Hôpital (1661-1704) and l'Hôpital's rule provides a method for evaluating limits which appear to involve indefinite forms (e.g the wretched $\frac{0}{0}$ or $\frac{\infty}{\infty}$). Should we wish to find the limit $\lim_{x\to 0}\left(\frac{f(x)}{g(x)}\right)$ for which $f(0)=0$ and $g(0)=0$ and we are able to find the functions $f(x)$ and $g(x)$ as power series expanded about $x=0$ then we have:

$$\begin{aligned}\lim_{x\to 0}\left(\frac{f(x)}{g(x)}\right) &= \lim_{x\to 0}\left(\frac{f(0)+xf^1(0)+\frac{1}{2}x^2f^2(0)+\ldots}{g(0)+xg^1(0)+\frac{1}{2}x^2g^2(0)+\ldots}\right) \\ &= \lim_{x\to 0}\left(\frac{xf^1(0)+\frac{1}{2}x^2f^2(0)+\ldots}{xg^1(0)+\frac{1}{2}x^2g^2(0)+\ldots}\right) \\ &= \lim_{x\to 0}\left(\frac{f^1(0)+\frac{1}{2}xf^2(0)+\ldots}{g^1(0)+\frac{1}{2}xg^2(0)+\ldots}\right) \\ &= \frac{f^1(0)}{g^1(0)}\end{aligned}$$

where we have substituted $f(0) = 0$ and $g(0) = 0$ before taking the limit. So long as $g^1(0) \neq 0$ this result gives a well-defined limit. Let us summarise the generalisation of this result which is known as l'Hôpital's Rule:

Theorem 6.1. *(l'Hôpital's Rule) Let $f(x)$ and $g(x)$ be two functions which are differentiable on the interval I (except possibly at the point $x = x_0$) such that $\lim_{x\to x_0}(f(x)) = 0$, $\lim_{x\to x_0}(g(x)) = 0$ and $g'(x) \neq 0$ for all $x \in I \setminus x_0$ then*

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

where $x_0 \in I$.

It is possible to further generalise this statement of the theorem. To see this we will consider again the example where $x_0 = 0$ but now consider functions such that $\lim_{x\to 0}(f(x)) = \pm\infty$ and $\lim_{x\to 0}(g(x)) = \pm\infty$ then we have,

$$\lim_{x\to 0}\left(\frac{f(x)}{g(x)}\right) = \lim_{x\to 0}\left(\frac{1/g(x)}{1/f(x)}\right) = \lim_{x\to 0}\left(-\frac{g'(x)/(g(x))^2}{-f'(x)/(f(x))^2}\right)$$

where we have been able to invoke l'Hôpital's rule as we have $\lim_{x\to 0}(1/f(x)) = 0$ and $\lim_{x\to 0}(1/g(x)) = 0$. Now if the limit $\lim_{x\to 0}\left(\frac{(g(x))^2}{(f(x))^2}\right)$ exists we can multiply the above by it to find

$$\lim_{x\to 0}\left(\frac{g(x)}{f(x)}\right) = \lim_{x\to 0}\left(\frac{g'(x)}{f'(x)}\right).$$

Example 6.10. *Use l'Hôpital's rule to determine the following limit:*

$$\lim_{x\to 0}\left(\frac{a^x - 1}{x}\right).$$

We note that $\lim_{x\to 0}(a^x - 1) = 0$ and $\lim_{x\to 0}(x) = 0$, so we have,

$$\lim_{x\to 0}\left(\frac{a^x - 1}{x}\right) = \lim_{x\to 0}\left(\frac{a^x \ln(a)}{1}\right) = \ln(a).$$

Where to take the derivative we note that if $y = a^x$ then $\ln(y) = x\ln(a)$ and on taking the derivative we have

$$\frac{1}{y}\frac{dy}{dx} = \ln(a)$$

and hence

$$\frac{dy}{dx} = a^x \ln(a).$$

Example 6.11. *Use l'Hôpital's rule to determine the following limit:*

$$\lim_{x\to 0}\left(\frac{\sin(x)}{x}\right).$$

We note that $\lim_{x\to 0}(\sin(x)) = 0$ *and* $\lim_{x\to 0}(x) = 0$*, so we have,*

$$\lim_{x\to 0}\left(\frac{\sin(x)}{x}\right) = \lim_{x\to 0}\left(\frac{\cos(x)}{1}\right) = 1.$$

Example 6.12. *Use l'Hôpital's rule to determine the following limit:*

$$\lim_{x\to 0}\left(\frac{2\sin(x)-\sin(2x)}{x-\sin(x)}\right).$$

We note that $\lim_{x\to 0}(2\sin(x)-\sin(2x)) = 0$ *and* $\lim_{x\to 0}(x-\sin(x)) = 0$*, so we have,*

$$\begin{aligned}\lim_{x\to 0}\left(\frac{2\sin(x)-\sin(2x)}{x-\sin(x)}\right) &= \lim_{x\to 0}\left(\frac{2\cos(x)-2\cos(2x)}{1-\cos(x)}\right)\\ &= \lim_{x\to 0}\left(\frac{-2\sin(x)+4\sin(2x)}{\sin(x)}\right)\\ &= \lim_{x\to 0}\left(\frac{-2\cos(x)+8\cos(2x)}{\cos(x)}\right)\\ &= 6.\end{aligned}$$

Example 6.13. *Use l'Hôpital's rule to determine the following limit:*

$$\lim_{x\to 0^+}\left(x\ln(x)\right).$$

We note that

$$x\ln(x) = \frac{\ln(x)}{1/x}$$

and as $\lim_{x\to 0^+}(\ln(x)) = \infty$ *and* $\lim_{x\to 0^+}(1/x) = \infty$*, so we have,*

$$\lim_{x\to 0^+}\left(x\ln(x)\right) = \lim_{x\to 0^+}\left(\frac{\ln(x)}{1/x}\right) = \lim_{x\to 0^+}\left(\frac{1/x}{-1/x^2}\right) = \lim_{x\to 0^+}\left(\frac{-x}{1}\right) = 0.$$

---

l'Hôpital's rule is constructed upon the assumptions that

- $f(x)$ and $g(x)$ have well-defined Taylor expansions around the limit point, and

- the limit $\lim_{x\to x_0}\left(\frac{f'(x)}{g'(x)}\right)$ exists while $f(x_0)=0$ and $g(x_0)=0$ or $f(x_0)=\pm\infty$ and $g(x_0)=\pm\infty$.

Although l'Hôpital's rule seems like a panacea for all limits, the assumptions around its construction means one must take care when using it. For example consider the limit

$$\lim_{x\to 2}\left(\frac{2x^2-3x+1}{5x+4}\right)=\frac{3}{14}$$

as the quotient is well-defined at the limit point, if we had mistakenly applied l'Hôpital's rule we would have evaluated:

$$\lim_{x\to 2}\left(\frac{\frac{d}{dx}(2x^2-3x+1)}{\frac{d}{dx}(5x+4)}\right)=\lim_{x\to 2}\left(\frac{4x-3}{5}\right)=1$$

which is not the correct answer. Another example where the rule is not valid is the limit

$$\lim_{x\to\infty}\left(\frac{x+\sin(x)}{x}\right)=1+\lim_{x\to\infty}\left(\frac{\sin(x)}{x}\right)=1.$$

If we had used l'Hôpital's rule we would have evaluated

$$\lim_{x\to\infty}\left(\frac{\frac{d}{dx}(x+\sin(x))}{\frac{d}{dx}(x)}\right)=\lim_{x\to\infty}\left(\frac{1+\cos(x)}{1}\right)$$

and this limit does not exist, so l'Hôpital's rule was not of use here and some other method is needed[3].

[3]We made use of $\lim_{x\to 0}(\frac{sin(1/x)}{1/x})=0$ to deduce the limit by using a change of the limit variable $y=\frac{1}{x}$.

Appendix

6.A The Ratio Test Implies the Root Test.

[4] Given a power series

$$S(x) \equiv \sum_{n=1}^{\infty} a_n x^n$$

we have two formulae for its radius of convergence[5]:

$$R_1 = \lim_{n\to\infty} (|a_n|^{-\frac{1}{n}}) \tag{6.1}$$

$$R_2 = \lim_{n\to\infty} \left|\frac{a_n}{a_{n+1}}\right|. \tag{6.2}$$

For different series it is useful to prefer one of these definitions for a radius of convergence over another. In this appendix we will show that if R_2 exists, then R_1 also exists and in that case $R_1 = R_2$ giving a single radius of convergence from the pair of definitions.

Let us commence by recalling the $\epsilon - \delta$ definition for the existence of a limit to equation 6.1 before putting the resulting expression in a helpful format.

If the limit defining R_2 exists then

$$|\left|\frac{a_n}{a_{n+1}}\right| - R_2| < \epsilon_2 \qquad \forall\, \epsilon_2 > 0, \quad n > N$$

where N is some sufficiently large value of $n \in \mathbb{Z}$ chosen such that the inequality is satisfied. Now it will be useful to consider the natural logarithm of R_2 in order to later compare it with R_1. We have,

$$\ln(R_2) \equiv \ln(\lim_{n\to\infty} \left|\frac{a_n}{a_{n+1}}\right|) = \lim_{n\to\infty} \Big(\ln|a_n| - \ln|a_{n+1}| \Big).$$

Now the definition of the existence of the limit give us:

$$|\ln|a_n| - \ln|a_{n+1}| - \ln(R_2)| < \epsilon \qquad \forall\, \epsilon > 0, \quad n > N. \tag{6.3}$$

Now to compare this with the limit R_1 we will also take the natural logarithm of R_1, so,

$$\ln(R_1) \equiv \ln(\lim_{n\to\infty} |a_n|^{-\frac{1}{n}}) = \lim_{n\to\infty} (\ln(|a_n|^{-\frac{1}{n}})) = \lim_{n\to\infty} (-\frac{1}{n} \ln|a_n|) \tag{6.4}$$

[4] The material in this appendix is provided for interest and is not an examinable part of the course.

[5] Recall that $|x| < R$ defines the values of x for which the power series is convergent and well-defined and R is the radius of convergence.

and by the definition of the existence of the limit we have,

$$|-\frac{1}{n}\ln|a_n| - \ln(R_1)| < \epsilon_1 \qquad \forall\,\epsilon_1 > 0, \quad n > N. \tag{6.5}$$

Our aim now is to show that equation (6.3) implies equation (6.5) and furthermore that $R_1 = R_2$.

Given equation (6.3) we may say that

$$\sum_{n=N}^{M} |\ln|a_n| - \ln|a_{n+1}| - \ln(R_2)| < \sum_{n=N}^{M} \epsilon = (M-N+1)\epsilon \tag{6.6}$$

where $M \in \mathbb{Z}$ and $M \geq N$. We may develop the left-hand-side of the equation above by using the triangle inequality (i.e. $|A+B| \leq |A| + |B|$), so that,

$$\begin{aligned}
\sum_{n=N}^{M} |\ln|a_n| - \ln|a_{n+1}| - \ln(R_2)| &\geq \left| \sum_{n=N}^{M} (\ln|a_n| - \ln|a_{n+1}| - \ln(R_2)) \right. \\
&= \Big| \ln|a_N| - \ln|a_{N+1}| + \ln|a_{N+1}| - \ln|a_{N+2}| + \ldots \\
&\qquad + \ln|a_M| - \ln|a_{M+1}| - (M-N+1)\ln(R_2) \Big| \\
&= \left| \ln|a_N| - \ln|a_{M+1}| - (M-N+1)\ln(R_2) \right|.
\end{aligned}$$

Hence from equation (6.6) we have,

$$\left| \ln|a_N| - \ln|a_{M+1}| - (M-N+1)\ln(R_2) \right| < (M-N+1)\epsilon.$$

Now we remark that as $M \geq N$ then $M - N \geq 0$ and $M - N + 1 > 0$. Also we note that $M - N + 1 = M - (N-1) \leq M$. It is also worth remarking that N is a fixed number, fixed by the definition in equation (6.3). Following these observations we see that if we divide the expression above by $M+1$ and take the limit as $M \to \infty$ we have:

$$\lim_{M\to\infty} \left| \frac{1}{M+1}\ln|a_N| - \frac{1}{M+1}\ln|a_{M+1}| - \frac{M-N+1}{M+1}\ln(R_2) \right| < \lim_{M\to\infty} \left(\frac{M-N+1}{M+1}\epsilon \right).$$

Since N is fixed, when we take the limit we find the expression above simplifies to

$$\lim_{M\to\infty} \left| -\frac{1}{M+1}\ln|a_{M+1}| \right| - \ln(R_2) < \epsilon.$$

In finding the above we have used $\lim_{M\to\infty} \left| \frac{1}{M+1}\ln|a_N| \right| = 0$ (N is fixed and so a_N is constant as $M \to \infty$) and $\lim_{M\to\infty} \left(\frac{M-N+1}{M+1} \right) = \lim_{M\to\infty} \left(1 - \frac{N}{M+1} \right) = 1$. Now from equation (6.4) we

have that,

$$\lim_{M\to\infty}\left|-\frac{1}{M+1}\ln|a_{M+1}|\right| = \ln(R_1),$$

hence,

$$\ln(R_1)-\ln(R_2) = \ln\left(\frac{R_1}{R_2}\right) < \epsilon \qquad \forall\,\epsilon.$$

This is true for all $\epsilon > 0$ which implies that $R_1 = R_2$. Implicitly our work is now done, as if R_2 exists then we have shown that it is equal to R_1 and hence R_1 exists. However we set out to show that equation (6.3) implies equation (6.5). This will be achieved in a simple step from the equation we have derived above:

$$\lim_{M\to\infty}\left|-\frac{1}{M+1}\ln|a_{M+1}|\right| - \ln(R_2) = \lim_{M\to\infty}\left|-\frac{1}{M+1}\ln|a_{M+1}| - \ln(R_2)\right| < \epsilon.$$

Therefore there exists $M_0 \in \mathbb{Z}^+$ such that for all $M > M_0$ then

$$\left|-\frac{1}{M+1}\ln|a_{M+1}| - \ln(R_2)\right| < \epsilon$$

which is equation (6.5) where $M+1 = n$, $M_0+1 = N$, $\epsilon = \epsilon_1$ and $R_2 = R_1$.

In this appendix we have shown that the ratio test for the convergence of a power series implies the root test for convergence.