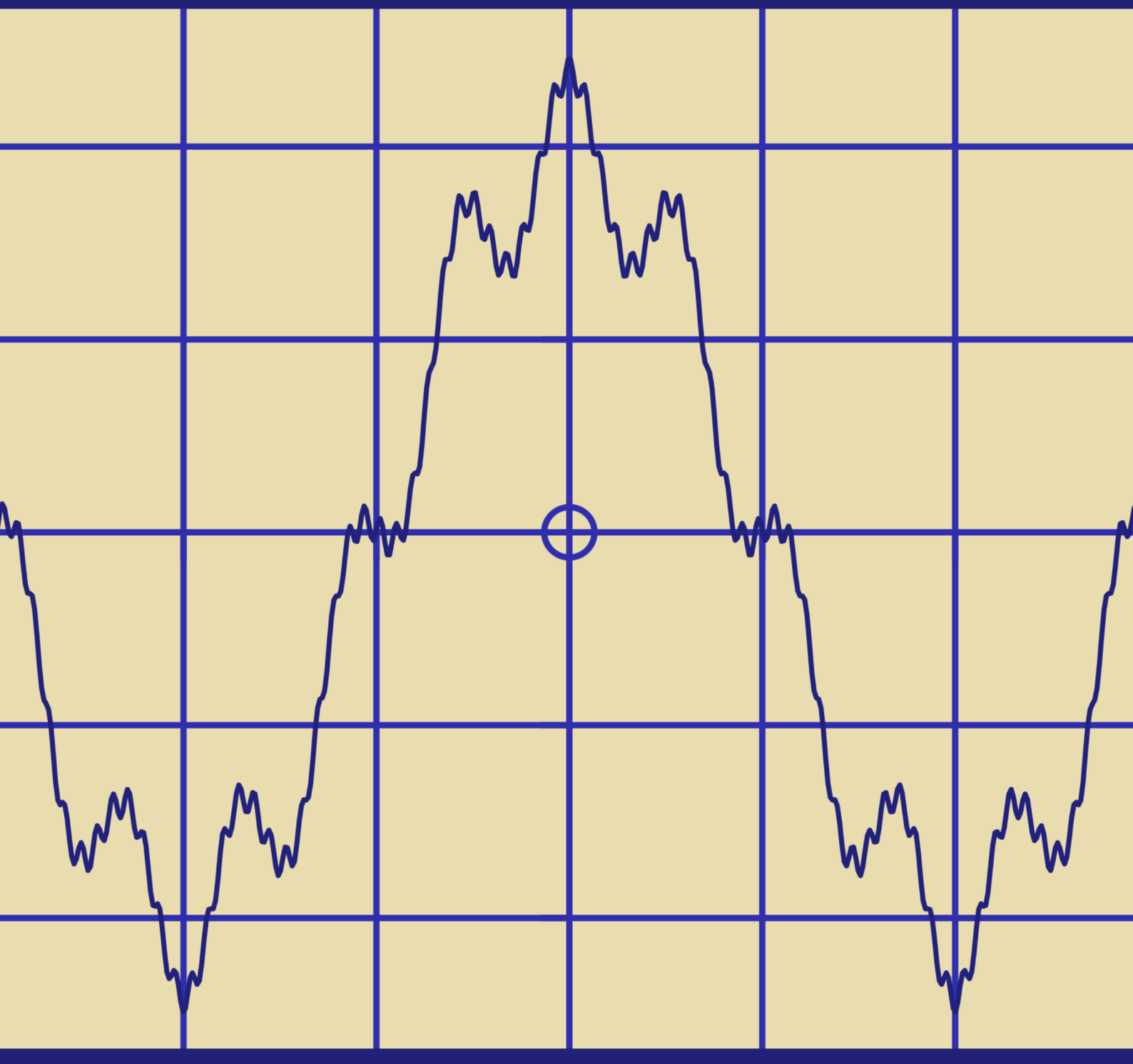
Calculus



Martin McBride

Calculus
by Martin McBride

Published by Axlesoft Ltd
info@axlesoft.com

About the author

Martin McBride is a software engineer with forty years of experience developing software for many applications including medical imaging, maths visualisation, image processing, computer graphics, data compression, real-time data acquisition, and machine control systems. Much of his work has been rooted in mathematics.

Martin has a BA in Physics from Oxford University. He has written many articles on maths and software engineering (on medium.com and other websites) as well as several other books, including *Functional Programming in Python* and *Computer Graphics in Python*.

Preface

I've been writing about maths for several years, mainly on my website graphicmaths.com, and on other sites including medium.com. I have also worked on maths visualisation software, most recently generativepy (see below), a Python maths visualisation library.

Writing individual articles can be quite limiting. Each article can only assume a certain level of knowledge on the reader's part, and so has to include a lot of groundwork. And since an article naturally has a limited length, there is only so far it can go.

This book has allowed me to combine many of my previous articles into a, hopefully, coherent work where each chapter builds on the earlier chapters, allowing me to cover each topic in more depth.

Approach

My aim in this book is to provide an intuitive understanding of the concepts it covers. Each topic includes plain language explanations, examples, and explanatory graphs and diagrams. I have tried to write the book in a conversational style.

Of course, formal proofs are important too. There are often several different proofs of any given theorem. I have attempted to present the most intuitive and accessible proof in each case. In some cases, I have included more than one proof if they each offer important insights.

Technical details

This book was written in LaTeX, using TeXstudio[1], an open-source LaTeX authoring system.

All the diagrams in the book were created in Python, mainly using the generativepy[2], an open-source maths visualisation library.

[1] https://www.texstudio.org/

[2] https://github.com/martinmcbride/generativepy, https://pypi.org/project/generativepy/

Contact

If you would like to be updated when I publish other books and articles, please join my Substack newsletter. I regularly post free articles on there, as well as news about other projects.

My YouTube channel contains lots of animated videos covering various maths topics, including calculus.

Substack newsletter:	**graphicmaths.substack.com**
YouTube channel:	**www.youtube.com/@graphicmaths7677**
LinkedIn:	**www.linkedin.com/in/martin-mcbride-0014b5257**
Books:	**www.amazon.co.uk/stores/Martin-McBride/author/B07XSF9NFZ**
	leanpub.com/u/martinmcbride
Articles:	**medium.com/@mcbride-martin**
	graphicmaths.com

Contents

Chapter 1

Introduction

This book provides an introduction to calculus, suitable for science and engineering students at undergraduate level, and anyone else who wishes to learn about calculus. It assumes an understanding of High School maths (UK KS4). The book mainly covers calculus of a single variable.

At its heart, calculus is concerned with the properties of mathematical functions. In particular, differentiation allows us to calculate the instantaneous slope of a function curve at some point on the curve, and integration allows us to calculate the area under a function curve between two points.

We often find the slope of a curve at point x by:

- Finding the slope of a line between points on the function curve between x and $x + \Delta x$.
- Finding the limit of that slope as Δx tends to zero.

We can find the area under the curve between points x_a and x_b by:

- Dividing the area under the curve into rectangles of width Δx.
- Finding the limit of the sum of the areas of the rectangles as Δx tends to zero.

These definitions are illustrated in figure 1.1.

According to the *fundamental theorem of calculus*, we can also find the area under a curve using the *antiderivative* of the function. If function f is the derivative of function g, we say that g is the antiderivative of g, ie it is the function you must differentiate to obtain f.

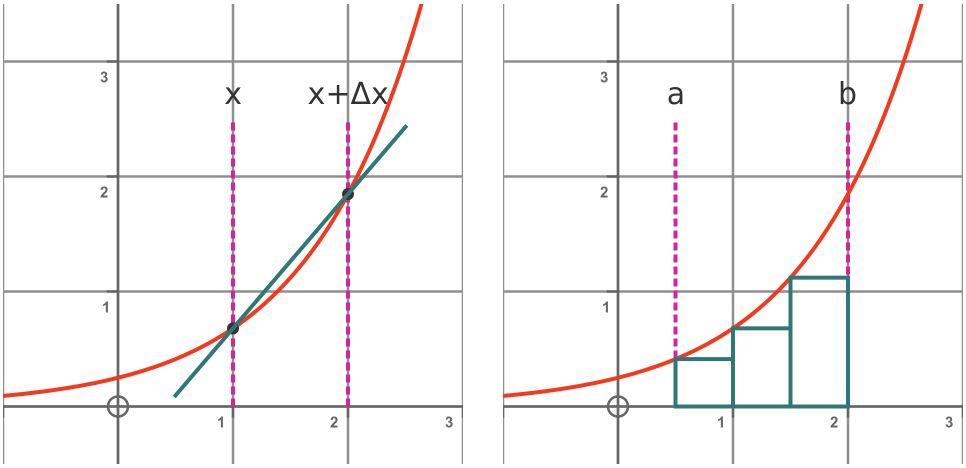


Figure 1.1: Derivative (left) and integral (right) as limits

1.1 Content summary

Functions and limits discusses the properties of functions of a single variable, including the ideas of domains, codomains, images, and inverse functions. It also covers limits, including the squeeze theorem. Finally, it introduces the idea of a tangent and normal to a curve and the concepts of smoothness and differentiability.

Differentiation covers the concepts of rate of change and differentiation, finding the rate of change as a limit, and the various notations used for derivatives. It also offers intuitive explanations, together with formal proofs, for the derivatives of many standard functions including polynomials, exponentials, and trig functions.

Integration introduces the antiderivative, or indefinite integral, as the inverse of a derivative, and gives the antiderivatives of many standard functions. It explains and proves the fundamental theorem of calculus, defining the definite integral as the area under a curve. It also discusses improper integrals.

Differentiation techniques covers more advanced techniques for differentiating more complex functions. These include the product rule, chain rule, quotient rule, reciprocal rule, and inverse function rule. Intuitive explanations and proofs are given for each rule.

Integration techniques covers more advanced techniques for integrating complex functions. These include integration by substitution, integration by parts, and the LIATE rule for integration by parts. Again, intuitive explanations and proofs are provided in each case, including how the rules relate to the differentiation techniques from the previous chapter.

Chapter 2

Functions and limits

In this chapter, we will lay the groundwork for our study of calculus with a discussion of functions, limits, and some other core concepts.

We will start with a deep dive into the topic of functions, specifically real-valued functions of a single variable, including:

- The definitions of various sets and ranges of numbers, including the concept of open and closed intervals.
- A naive approach to defining the set of possible input and output values of a function, using the domain and range.
- A more robust approach to sets of input and output values, using domain, codomains and images.
- Injective, surjective and bijective functions.

We will then discuss the concept of limits, including:

- The definition of a limit.
- Types of limit, including asymptotes and left/right limits.
- The laws for combining limits.
- Indeterminate forms of limits.
- The conditions where a limit does not exist.

We will look at Big O notation that can be used to describe the general behaviour of a function at extreme values of x, and the squeeze theorem that allows us to find limits in various difficult situations.

Tangents are an important concept in calculus, so we will look at:

- The definition of slope and tangent, and the microstraightness property.
- Stationary points and points of inflection.
- Equations of tangents.
- Normals to a curve.

We will finish by looking at some other properties of functions, such as continuous, smooth, differentiable, and analytic functions.

2.1 Functions

We are all familiar with mathematical functions such as x^2, $\sin x$, and $\ln x$. These functions accept an input value and return an output value. We say they are functions of one variable. It is possible to define functions of more than one variable, but in this book, we will mainly be dealing with functions of one variable.

We usually call the input variable x, and we often assign the output value to a variable y, for example $y = x^2$. We can then plot the function on xy axes. But of course, we can use any names we like for the input and output variables, and we do this quite often in calculus, for example when we use *variable subtitution* in more advanced integration techniques.

More formally, we can say that a function from a set X to a set Y maps each element of set X onto exactly one element of set Y. In this book we will be dealing (almost) exclusively with real-valued functions, so X will be the set of real numbers $\mathbb{R}$ or some subset of it. Likewise, Y will be the set of real numbers or some (potentially different) subset of it.

These are the main types of functions we will be using in this book, but not the only type. We will look at some other types later in this chapter.

2.1.1 Sets of numbers

We often need to refer to specific sets of numbers. We use a standard notation for describing these sets, shown in table 2.1.

For the main sets of numbers, we use stylised letters, such as $\mathbb{Z}$ to represent all integers or $\mathbb{R}$ to represent all real numbers.

We can further restrict this by adding conditions as a subscript, for example, $\mathbb{R}_{>0}$ represents all the real numbers that are greater than zero (ie all the positive real numbers).

Notation	Set of numbers
$\mathbb{R}$	The set of all real numbers.
$\mathbb{Z}$	The set of all integers.
$\mathbb{N}$	The set of natural numbers (the counting numbers 1, 2, 3 ...).
$\mathbb{Q}$	The set of rational numbers.
$\mathbb{R} - \mathbb{Q}$	The set of irrational numbers.[1]
$\mathbb{I}$	The set of all imaginary numbers.
$\mathbb{C}$	The set of all complex numbers.
$\mathbb{R}_{>0}$	The set of all positive real numbers.[2]
$\mathbb{R}^+$	Alternative notation all positive real numbers.[3]
$\{1, 2, 3\}$	A set of discrete possible values, eg 1, 2 and 3.
7	Only one possible value, eg 7.
$\mathbb{R}_{\leq -1} \cup \mathbb{R}_{\geq 1}$	Union of two or more sets, eg all the numbers with $\lvert x \rvert \geq 1$.
Note 1	This notation means the set of numbers that are in $\mathbb{R}$ but not in $\mathbb{Q}$.
Note 2	We can use other conditions, eg $\mathbb{R}_{\neq 0}$, and other base sets, eg $\mathbb{Z}_{\geq 0}$.
Note 3	This is an alternative notation for $\mathbb{R}_{\geq 0}$ (not used in this book).

Table 2.1: Sets of numbers

Notation	Set of numbers
$[-1, 1]$	A closed interval $-1 \leq x \leq 1$.
$(-1, 1)$	A open interval $-1 < x < 1$.
$[-1, 1)$	A semi-open[1] interval $-1 \leq x < 1$.
$(-1, 1]$	A semi-open[1] interval $-1 < x \leq 1$.
Note 1	A semi-open interval can also be called a semi-closed interval.

Table 2.2: Types of intervals

For more specific sets of numbers, we can use intervals (below). We can also use normal set notation, for example, $\{1, 2, 3\}$ represents the set of three numbers 1, 2 and 3.

2.1.2 Intervals

We can define a set of all numbers between two endpoints as an *interval*. For example, the set of every real number between 0 and 10 would be an interval.

When we define an interval, it is important to be clear whether or not that interval includes its endpoints. If an interval includes its endpoints it is called a *closed interval* and is written a $[0, 10]$. If it doesn't include its endpoints it is called an *open interval* and is written a $(0, 10)$.

Intervals can be open at one end and closed at the other. All of these cases are shown in table 2.2.

Where do the terms open and closed come from? This comes from the observation that an open interval doesn't have a largest or smallest value.

This might seem counter-intuitive at first. Looking at the open interval $(0, 1)$ for example, then clearly all the elements of that interval must be greater than zero and less than one.

But what is the smallest element of the interval? It isn't zero, because an open interval doesn't include its endpoints. So zero isn't a member of the interval.

The value 0.1 is an element of the interval, but it certainly isn't the smallest element of the interval, because 0.01 is smaller. And 0.001 is smaller still. In fact, it doesn't matter how many zeros we add before the one, we can always make the number even smaller by adding an extra zero. It is impossible to name the smallest element of the interval because there is no smallest element.

We can state this more formally. For any value x that is in an open interval, the values $x + \delta$ and $x - \delta$ will also be in the interval, for some values of $\delta > 0$. We might need to choose a very small value of δ, but it will always be possible.

An open interval has a limit on how small any element of the set can be, but it doesn't have a

smallest element. And the same is true of the largest element. So we say the interval is open.

We don't have the same problem with a closed interval. The interval $[0, 1]$ has a smallest interval of 0 and a largest interval of 1.

One interesting thing to notice is that the definitions of open and closed are not opposites of each other. Rather, they are complements of each other:

- Let I be the closed interval $[-1, 1]$. The complement $I^{\complement}$ of that interval contains all numbers that are either > 1 or < -1, which are both open sets.
- Conversely, let J be the open interval $(-1, 1)$. The complement $J^{\complement}$ of that set contains all numbers that are either ≤ 1 or ≤ -1, which are both closed sets.

Since the definitions are not mutually exclusive, an interval can be both open and closed. You can call them *clopen* intervals if you must. For example:

> The set of all real numbers, $\mathbb{R}$ is both open and closed.

How is this so? Well, $\mathbb{R}$ is unbounded, so it has no limit points. This means it contains all of its limit points (because there are none) so it is a closed set.

But, for any value x in $\mathbb{R}$, you will always be able to find a value in $\mathbb{R}$ that is larger than x, and you will always be able to find a value in $\mathbb{R}$ that is smaller than x. $\mathbb{R}$ does not contain a largest or smallest value. So $\mathbb{R}$ is also open.

2.1.3 Domain and range of a function

In elementary maths, we learn about the *domain* and *range* of a function, which are viewed as properties of the function itself.

> The slightly naive concept of domain and range works reasonably well for simple, real-valued functions, but it has limitations. We will briefly review this here, before moving on to a better alternative.

The domain of a function is X, the set of all permitted input values. For many functions, such as $\sin x$ for example, the domain is simply $\mathbb{R}$, that is to say, any real number is a valid input value.

Other functions might restrict the permitted input values to a particular interval. For example, the natural log function $\ln x$ is only valid for $x > 0$, therefore its domain is $\mathbb{R}_{>0}$

Some functions have far more specific domains. Consider the function:

$$\frac{x}{|x|}$$

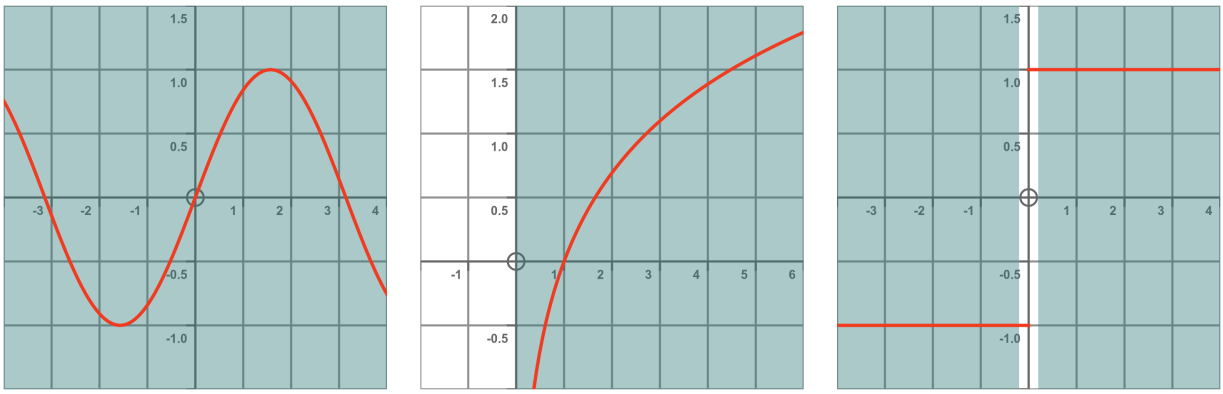


Figure 2.1: Domains of $\sin x$, $\ln x$ and $\frac{x}{|x|}$

There are three possible cases:

- When $x > 0$, the top and bottom of the fraction are both positive with the same magnitude, so the function value is 1.
- When $x < 0$, the top and bottom are opposite signs but with the same magnitude, so the function value is -1.
- When $x = 0$, the top and bottom are both zero the value is undefined (because $0/0$ is undefined).

So the domain of this function is any value other than zero, ie $\mathbb{R}_{\neq 0}$. All three functions are shown in figure 2.1. Notice that for $x/|x|$ the gap in the domain at zero is infinitesimally small, but it has been exaggerated on the graph to make it visible.

The range of a function is the set, Y, of possible outputs of a function.

For example, the function $\sin x$ only ever gives a value in the interval $[-1, 1]$ for any value of x. We would therefore say that the range of $\sin$ is the closed interval $[-1, 1]$ (which is a subset of $\mathbb{R}$).

The $\ln x$ function can return any value in the range $-\infty$ to ∞, so its range is $\mathbb{R}$.

The function $x/|x|$ is, once again, quite unusual. It can only have two possible values, 1 (if $x > 0$) or -1 (if $x < 0$). So its range is simply the two-member set $\{-1, 1\}$ All three ranges are shown in figure 2.2.

2.1.4 Domain, codomain and image

In reality, functions are a little more complicated than the simple model above. In particular, we might sometimes wish to use a different domain depending on the context (ie what we are using the function for).

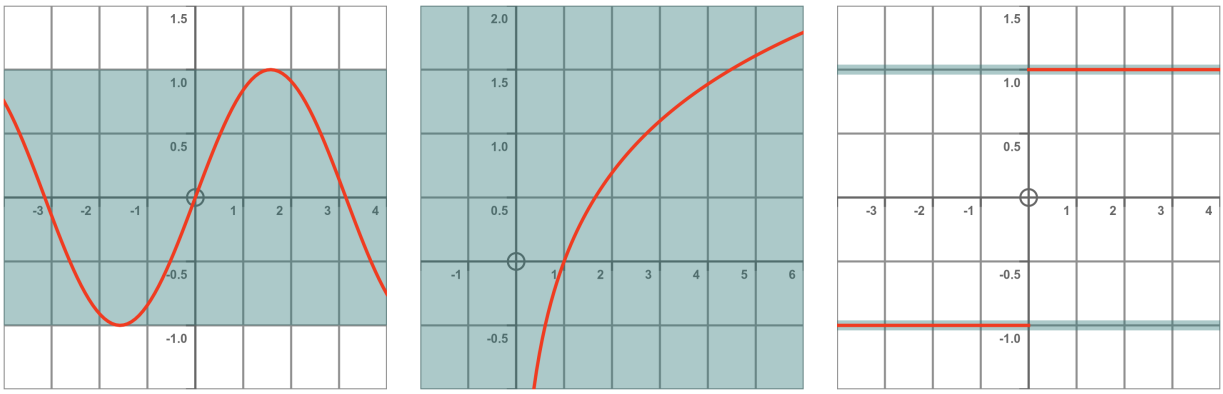


Figure 2.2: Ranges of $\sin x$, $\ln x$ and $\frac{x}{|x|}$

Let's consider the sine function. If we are working with angles and elementary trigonometry, then, of course, we treat $\sin$ as a function of a real variable, with an input interval $[0, \pi/2]$ (we would most likely call this interval 0 to 90 degrees when studying basic geometry) and an output interval of $[0, 1]$.

If we are working with slightly more advanced trigonometry, we might allow angles in the interval $(-\pi, \pi]$ (between -180 and 180 degrees), and we would expect output values in the interval $[-1, 1]$.

In this book, we will mainly be treating $\sin x$ purely as a mathematical function, so our input domain is $\mathbb{R}$ and we would expect output values in the interval $[-1, 1]$.

But you might also be aware that we can apply the sine function to a complex variable z. And if we do that, the result can be any complex number. So the sine function can have a domain of $\mathbb{C}$ and a range of $\mathbb{C}$.

The complex form of the sine function includes all the other interpretations as subsets, but that certainly doesn't mean that we should use that form under all circumstances. If we are just trying to solve a triangle we don't want to be dealing with complex number angles and side lengths!

So rather than thinking of the domain as a property of the function, we should think of the domain as being a choice we make when using the function. We can use $\sin$ over $[0, \pi/2]$ if we are doing elementary geometry, and $\sin$ over $\mathbb{R}$ in a book like this. We don't always have to explicitly state the domain, but it is a choice we are making, perhaps by default, every time.

When we choose a different domain, we often get a different set of possible output values from the function. We call this the *codomain* rather than the range, and it is also a choice we make. If we use $\sin$ over $\mathbb{R}$, the result will be real so we would normally use a codomain of $\mathbb{R}$ as well.

In this system, we are never using the just sine function, we are using the sine function over a particular domain and codomain. We write the complete real-valued function as:

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

The complex version would be written as:

$$\sin : \mathbb{C} \to \mathbb{C}$$

These are, in effect, two different functions both based on the sine function. The complete function definition includes the domain, the codomain, and the function itself. However, as we noted before, the domain and codomain often don't need to be stated as they will be obvious from the context. In this book, for example, the domain and codomain can be assumed to be $\mathbb{R}$ unless otherwise stated.

Looking again at the real version of the sine function, we have chosen to map $\mathbb{R}$ to $\mathbb{R}$, but it is still the case that the output of the function is limited to values in the interval $[-1, 1]$. We call this the *image* of the function:

> The image of a function f is the set of all output values the function can produce.

This is similar to the range in the simple scheme, except that the image depends on the domain and codomain. In summary:

- The codomain is the general set of values that function outputs, such as $\mathbb{R}$ or $\mathbb{C}$. We can choose the codomain.
- The image is the set of output values that the function actually produces. It is a property of the function plus the chosen domain and codomain.

In the case of $\sin : \mathbb{C} \to \mathbb{C}$, the image is $\mathbb{C}$.

The $\ln$ function can also be used in the complex domain $\mathbb{C}_{\neq 0}$, and like the $\sin$ function its codomain and image will both be $\mathbb{C}$ in that case.

What about $x/|x|$? Well $|x|$ is defined for a complex value of x, so we can define:

$$\frac{x}{|x|} : \mathbb{C}_{\neq 0} \to \mathbb{C}$$

This isn't a book on complex analysis so we won't go into detail, but if we use the modulus argument form of complex numbers, then $z = r\angle\theta$ will map onto $1\angle\theta$. The complex plane will be mapped onto the unit circle on an argand diagram, so the image of the function is $r = 1$.

This is shown in figure 2.3. Every point in the complex plane is mapped onto the unit circle, preserving its angle. For example, points b and c are mapped onto the unit circle at a.

2.1.5 Example - the $\sqrt{x}$ function

The square root function is an interesting example showing that the domain and codomain do not have to be identical.

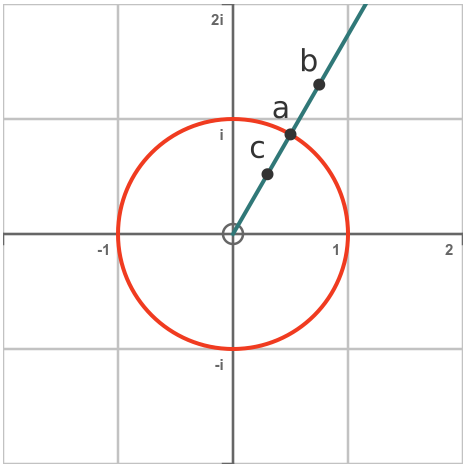


Figure 2.3: Image of function $\frac{x}{|x|} : \mathbb{C}_{\neq 0} \to \mathbb{C}$

Often when we work with the square root function we will use a domain of $\mathbb{R}_{\geq 0}$ (non-negative real numbers). We normally take the principal square root (for example $\sqrt{4}$ is 2 not -2).

With this domain, the codomain will be $\mathbb{R}$, so we write the function as $\sqrt{x} : \mathbb{R}_{\geq 0} \to \mathbb{R}$, The image will be $\mathbb{R}_{\geq 0}$ because we are always taking the positive square root. This is shown on the LHS of figure 2.4. Since the image relates to the y values, only the area where $y \geq 0$ is shaded.

We could choose a different domain, for example, $\mathbb{R}$. This means we will be taking the square root of negative x values as well as positive values. The square root of a negative number is an imaginary number, for example, $\sqrt{-4}$ is $2i$. This means that despite the domain being real, the codomain is $\mathbb{C}$. But notice that not all complex numbers are possible, the square root of a real can only be a real or imaginary number, and must be positive. So the image of $\sqrt{x} : \mathbb{R} \to \mathbb{C}$ is $\mathbb{R}_{\geq 0} \cup \mathbb{I}_{\geq 0}$ (ie all numbers that are positive and either real or imaginary).

This is shown in the centre graph of figure 2.4. An important thing to notice here is that, since the domain and codomain are both $\mathbb{C}$, the graph shown is an Argand diagram of the codomain, so we can mark the possible output values on the real and imaginary axes. This is different from most of the previous graphs where the x-axis represents the real domain and the y-axis represents the real codomain.

Finally, if we chose a domain of $\mathbb{C}$ the codomain will again be $\mathbb{C}$. The image is slightly more involved, but we won't go into it in detail. Every complex number has two square roots, and the primary root is usually defined as the root that has a positive real part (or a positive imaginary part if the real part is zero). This is shown on the RHS of figure 2.4.

2.1.6 Injective and bijective functions

We say that a function $y = f(x)$ is *injective* (or *one-to-one*) if every element in the codomain is mapped to at most one element of the domain. In other words, there are no two values of x that have the same y value.

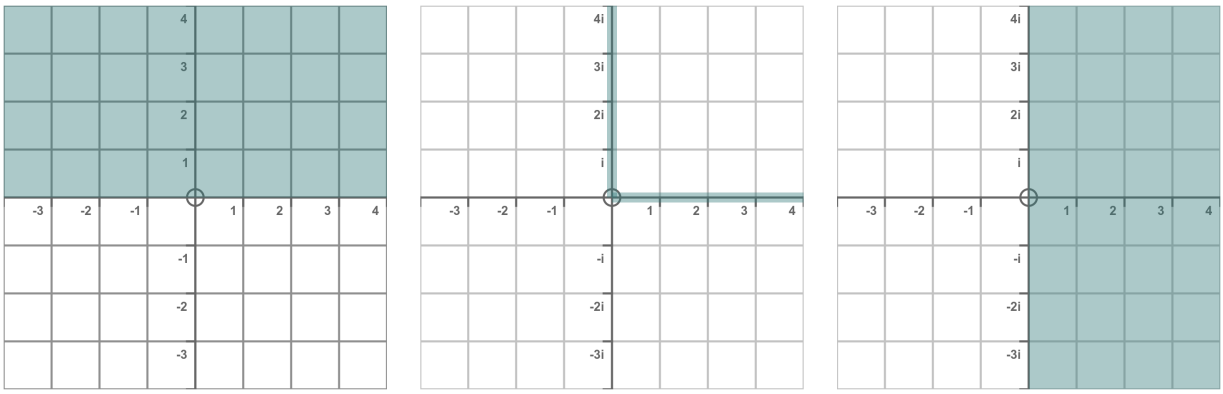


Figure 2.4: Image of $\sqrt{x}$ for $\mathbb{R}_{\geq 0} \to \mathbb{R}$, $\mathbb{R} \to \mathbb{C}$, and $\mathbb{C} \to \mathbb{C}$

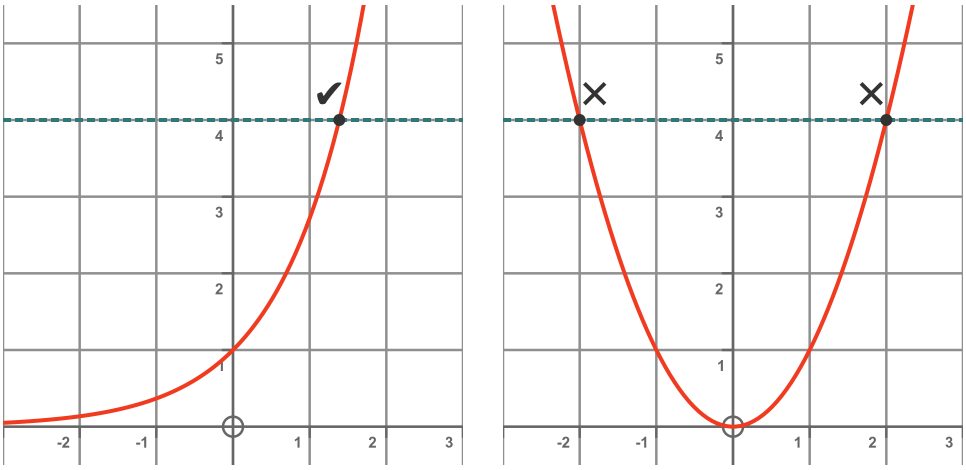


Figure 2.5: Function e^x (left) is injective, function x^2 (right) is not

Figure 2.5 shows that the exponential function e^x is injective. We can draw a horizontal line anywhere on the graph and it will never cross the curve in more than one place. But the function x^2 is not injective because there are values of y corresponding to two different x values - for example, $y = 4$ corresponds to x values 2 or -2 (because 2^2 and $(-2)^2$ are both equal to 4).

We say that a function is *surjective* if every element in the codomain is mapped to at least one element of the domain. In other words, for every possible y value, then there is at least one x value such that $f(x)$ returns that y value.

The function $\ln x$ is surjective because for any y value there is a corresponding x value. But e^x is not surjective, because there is no x value giving a result $y \leq 0$. This is shown in figure 2.6.

Finally, a function is *bijective* if it is both injective and surjective. This means that every possible y value corresponds to exactly one x value, and vice versa.

A good example of a bijective function is the cube function, x^3. This is shown on the LHS of figure 2.7. We can find the cube of any real number, so any x value in the domain will map onto a unique y value. Similarly, any y in the codomain will map onto a unique x value.

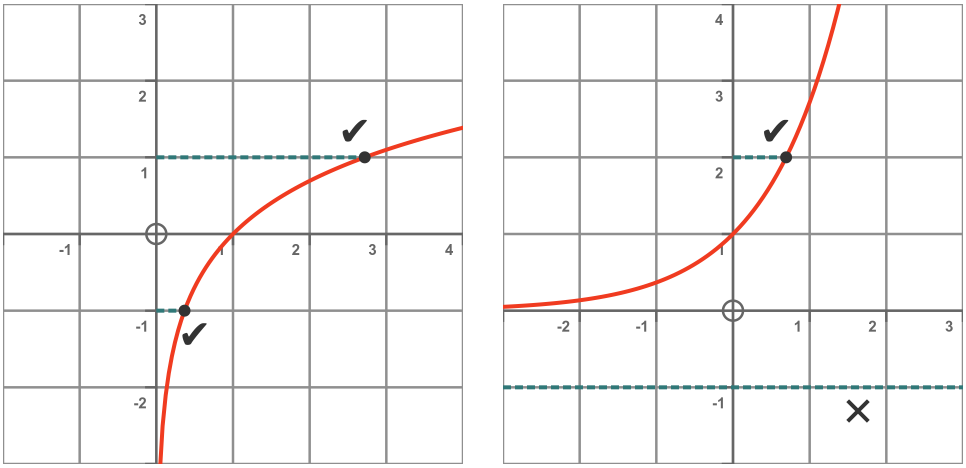


Figure 2.6: Function $\ln x$ (left) is surjective, function e^x (right) is not

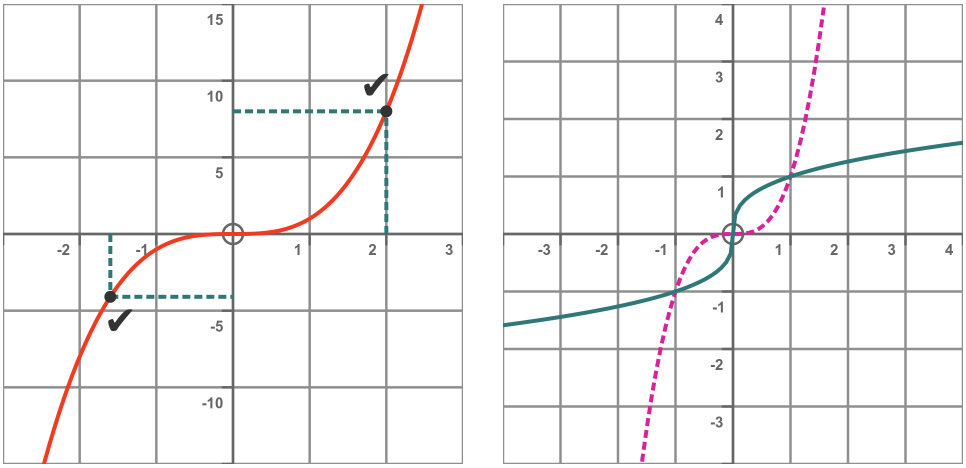


Figure 2.7: Function x^3 (left) and its inverse $\sqrt[3]{x}$ (right)

But we can go further than that. Bijective functions are invertible. Any x value x_a will map uniquely onto some value y_a, and the value y_a will map uniquely onto x_a.

The RHS of figure 2.7 illustrates this. The function $\sqrt[3]{x}$ (shown in cyan) inverts the function x^3 because $\sqrt[3]{x^3}$ is equal to x for all values of x. The function x^3 is shown on the same axes, as a dashed magenta line. This illustrates that a bijective function and its inverse are reflections of each other over the positive diagonal line $y = x$.

As a counterexample, x^2 is not bijective, because $\sqrt{x^2}$ is not always equal to x (eg -2 squared is 4, but the square root of 4 is 2).

2.1.7 Inverse functions over restricted domain

As we have seen, the function x^3 is bijective over $\mathbb{R} \to \mathbb{R}$. But not all functions that we think of as being inverses follow this pattern.

Take, for example, the two functions e^x and $\ln x$. These functions appear to be inverses - if we raise some real number to the power e then take the natural log of the result. But, in the real domain, we can't say the same of the log function because we can't take the log of a negative number.

We can avoid this problem by restricting the domain and codomain. We know that e^x has a domain $\mathbb{R}$ and always produces a positive result, so its codomain is $\mathbb{R}_{\geq 0}$. We also know that $\ln x$ has a domain of $\mathbb{R}_{\geq 0}$ but can create any real value as output, so its codomain is $\mathbb{R}$. Therefore, if we give full definitions of the two functions:

$$e^x : \mathbb{R} \to \mathbb{R}_{\geq 0}$$

$$\ln x : \mathbb{R}_{\geq 0} \to \mathbb{R}$$

We can see that they are both bijective and inverses, over the restricted domains and codomains. This is shown in figure 2.8. The figure shows the codomain of e^x (which is $\mathbb{R}_{\geq 0}$) and the domain of $\ln x$ (which is also $\mathbb{R}_{\geq 0}$).

Two other things to notice, that are generally true:

- The domain of each function is the same as the codomain of the other function.
- The two functions are reflected over the leading diagonal (the line $y = x$). This is shown on the RHS of figure 2.8.

This is because a function and its inverse effectively swap the roles of x and y.

Another interesting case is the $\sin$ function and its inverse $\arcsin$. The $\arcsin$ function has a domain of $[-1, 1]$ and a codomain of $\mathbb{R}$. But the inverse function is cyclical along the y axis, so for any value of x in $[-1, 1]$, there are infinitely many possible values of y. This is shown on the left in figure 2.9.

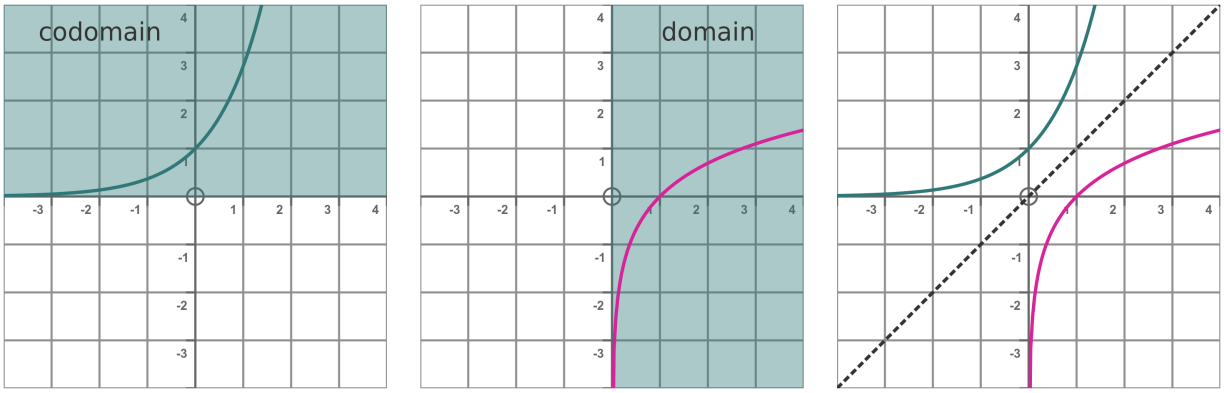


Figure 2.8: e^x and $\ln x$ are inverse functions if the domain of $\ln x$ is restricted to $\mathbb{R}_{\geq 0}$

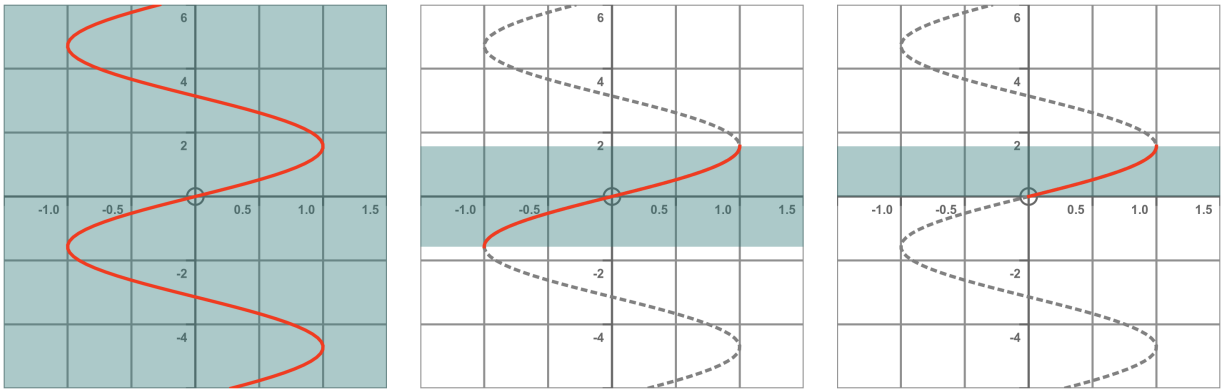


Figure 2.9: $arcsinx$ with a codomain of $\mathbb{R}$, $(-\pi/2, \pi/2]$, and $[0, \pi/2]$

This means that the version of $\arcsin$ shown in the graph isn't a function, because a function is only allowed to return one y value for any value of x.

We can turn this into a function by restricting its codomain. If we choose a codomain of $(-\pi/2, \pi/2]$ then every value of x in $[-1, 1]$ will have a unique value within the codomain. This is shown in the centre in figure 2.9.

Sometimes when using the sine function in geometry, we are only concerned with angles in the range 0 to 90 degrees (ie $[0, \pi/2]$). If we restrict our codomain to this range, then $\arcsin$ will only be valid for x values in $[0, 1]$. This is shown on the right in figure 2.9.

2.1.8 The domain preimage

We previously defined the image of a function as being the set of all possible output values of the function for a given domain and codomain.

We can also define the *preimage* (sometimes called the *inverse image*) as the set of all possible input values that correspond to the values in the image of the function.

In most cases, the preimage is the same as the domain, so the term preimage is not used often. However, in the case above for $\arcsin$, we restricted the codomain to $[0, \pi/2]$ we found that it restricted the permitted x values. It is sometimes helpful to express this in terms of a preimage. We can say that the domain is still $[-1, 1]$ but with a preimage of $[0, 1]$.

2.1.9 Choice of codomain and image

As discussed, the codomain is a choice we make to specify the set of function output values we are considering, while the image is a property that tells us the set of output values that the function can possibly produce.

There is a degree of flexibility in how we choose the codomain. For example, when we looked at the exponential function earlier, we chose to define its codomain as the set of positive reals $\mathbb{R}_{\geq 0}$. Using that definition, the image is also $\mathbb{R}_{\geq 0}$, because every positive real is a potential output of the function. The definition suited us because it matches the domain of the $\ln x$ function.

But we could also have defined the codomain as being the set of all reals, $\mathbb{R}$. That would have been similar to the approach we took with other functions. The image of the function would still be $\mathbb{R}_{\geq 0}$, of course.

This is largely a matter of preference or convenience, although we would not usually use a very specific set for the codomain, it is usually better to use a reasonably general set for the codomain and a more specific set for the image.

2.2 Limits

Suppose we have some function $f(x)$ and we wish to find the value of that function when $x = a$.

The obvious way to do that is to simply calculate $f(a)$.

But there is another way to do it. We could calculate the value of $f(x)$ for some value x_0 that is very close to a but not equal to a. We might then expect $f(x_0)$ to be quite close to $f(a)$.

We could then repeat that calculation with another value x_1 that is even closer to a, so $f(x_1)$ would be even closer to $f(a)$.

If we repeat this for $x_2, x_3 \ldots x_n$, with each successive value getting ever closer to a, we would see $f(x_n)$ getting closer and closer to $f(a)$.

This is called the *limit* L of $f(x)$ as x tends to a. We write this as:

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

But why would we go to all this trouble, rather than simply calculating $f(a)$? Well, normally we wouldn't. But there are some situations where we can't calculate the value of $f(a)$ directly for some reason, but we can find the limit as x tends to a. We will look at some examples of this below, but the most significant example is used in the process of differentiation from first principles, which we will study extensively in chapter **??**.

2.2.1 Simple example of a limit

We will start with a simple example of the function x^2, shown in figure 2.10.

The LHS simply shows x^2 evaluated at 2. Of course, $f(2) = 4$.

In the centre graph, we start with an x value of 1, then 1.5,1.75, and so on. As the value of x gets closer and closer to 2 from below, $f(x)$ gets closer and closer to 4, exactly as we would expect.

The RHS graph starts with an x value of 3, then 2.5, 2.25, etc. As the value of x gets closer and closer to 2 from above, $f(x)$ also gets closer and closer to 4.

This is no surprise. If we take a continuous, smoothly varying function like x^2, then of course if x is very close to 2, so $f(x)$ will be very close to $f(2)$. This becomes more useful when we look at functions that aren't quite so well-behaved.

2.2.2 The function x/x

One example of how limits can be useful is the function x/x. At first glance, you might think that we can simply cancel the fraction and get the result 1. And that is almost correct.

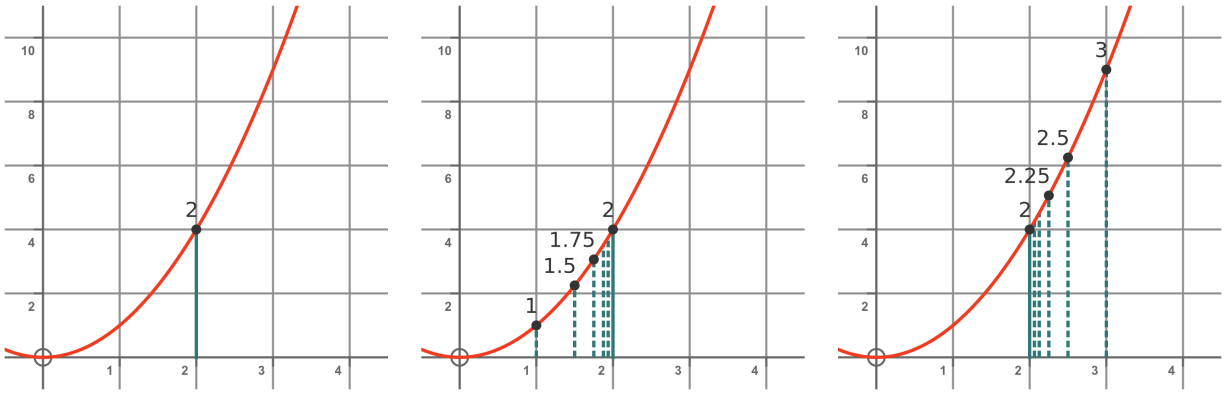


Figure 2.10: Value of x^2 when $x = 2$ (left) and $\lim_{x \to 2} x^2$ (centre, right)

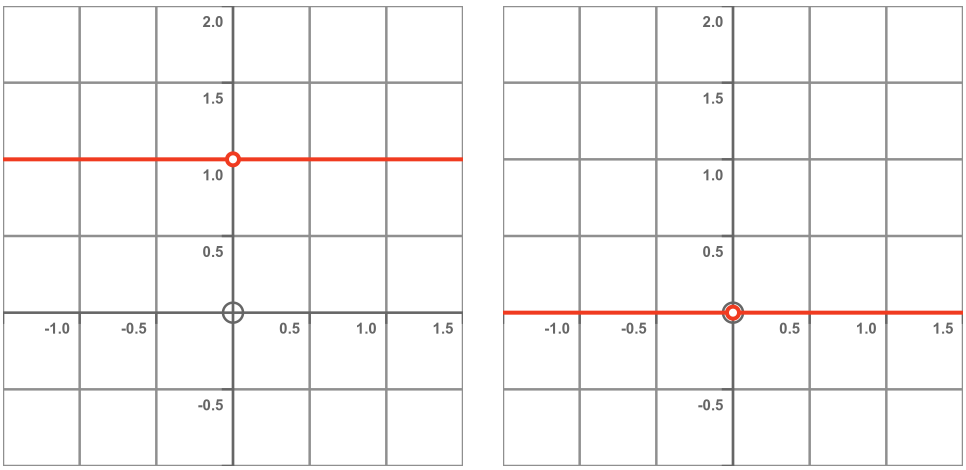


Figure 2.11: Functions x/x (left) and $0/x$ (right) are undefined when $x = 0$

A problem occurs when $x = 0$ because $0/0$ is undefined. We can't calculate $f(0)$. This is shown on the LHS of figure 2.11. The small empty circle where the function crosses the y-axis indicates that it is undefined at that exact point, despite being defined at every other point.

What we can do is calculate the limit of the function as x tends to 0. Let's try that with some small values of x:

x	$f(x) = x/x$
1	1
0.1	1
0.01	1
0.001	1

It is pretty clear from this that no matter how small x gets, provided it is not 0, $f(x)$ will always be

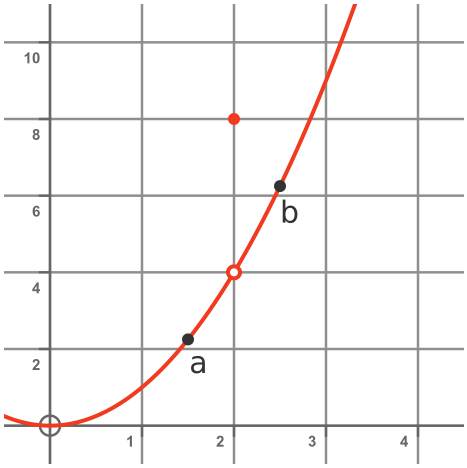


Figure 2.12: Function $f(x)$ is x^2 when $x \neq 2$ and 8 when $x = 2$

1. So we can say that

$$\lim_{x \to 0} \frac{x}{x} = 1$$

It is important to note that this *does not* mean that $0/0$ is 1. It just means that the limit of the particular function x/x, as x tends to 0, is 1. No matter how close we get to 0 (without actually reaching 0), the value of the function will be 1.

We can contrast this with the function $0/x$, shown on the RHS of figure 2.11. It is clear that this function has the value 0 for every $x \neq 0$, so it follows that its limit as x tends to 0 is 0.

So, in summary, the value of $0/0$ is undefined, but the limit of some function that tends towards $0/0$ might be defined, and that limit can take different values for different functions.

2.2.3 The limit might not equal the value of the function

In some cases, $f(a)$ might have a value, but the limit of $f(x)$ as x tends to a could have a different value. To see how this might work, consider this slightly contrived example (shown in figure 2.12):

$$f(x) = \begin{cases} x^2 & \text{if } x \neq 2 \\ 8 & \text{if } x = 2 \end{cases}$$

Again the graph uses an empty circle to indicate that the value of the function is different when $x = 2$. The function normally follows the curve x^2, but the solitary point at $(2, 8)$ indicates that the value is different when $x = 2$.

The value of $f(2)$ is 8, according to its definition. But if we start with a value less than 2 (point a for example), and find the limit as x tends to 2, that limit will be 2^2, which is 4. And if we start with a value greater than 2 (such as point b), and find the limit as x tends to 2, that limit will also be 4.

In this case, the limit of the function when $x = 2$ is different to the value of $f(2)$.

2.2.4 Formal definition of a limit

We can formally define a limit as follows:

Let $f(x)$ be a function defined over an interval that contains a. We say that:

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

if for any $\epsilon > 0$ there exists a $\delta > 0$ such that:

$$|f(x) - L| < \epsilon \qquad \text{whenever} \qquad |x - a| < \delta$$

What does this mean? Well, we know that the basic idea of a limit is that when x gets very close to a, then $f(x)$ gets very close to L.

But this definition goes further than that. First, let's rewrite the two conditions. For ϵ:

$$|f(x) - L| < \epsilon \quad \Longrightarrow \quad L - \epsilon < f(x) < L - \epsilon$$

In other words, $f(x)$ is within the range $L \pm \epsilon$.

For δ:

$$|x - a| < \delta \quad \Longrightarrow \quad a - \delta < x < a - \delta$$

In other words, x is within the range $a \pm \delta$.

So we can express the definition slightly in a different, but equivalent, way:

For any ϵ we can find a value δ such that, if x is in the range $a \pm \delta$ then $f(x)$ will be in the range $L \pm \epsilon$.

So if L is the limit f at a, then for a very small value δ we expect $f(a + \delta)$ to be quite close to L. But this definition tells us that we can get *as close as we like* to L simply by choosing a small enough δ.

Let's take the example of $f(x) = x^2$ and $a = 2$, therefor $f(a)$ = 4. We will choose $\epsilon = 2$ as an example, so we need $f(x)$ to be in the interval $4 \pm 2 = (2, 6)$.

To acheive this we must choose an x value in the interval $(\sqrt{2}, \sqrt{6})$. This is shown on the LHS of figure 2.13.

To express this in terms of a δ value, we must have an x value in the interval $2 \pm (\sqrt{6} - 2)$. Therefore $\delta = \sqrt{6} - 2$.Notice that this interval is smaller than $(\sqrt{2}, \sqrt{6})$ because we need an interval that

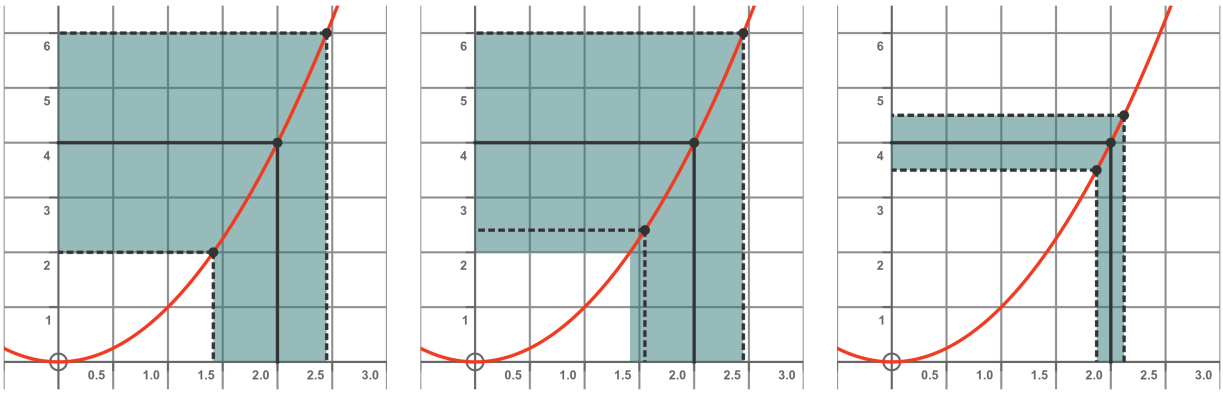


Figure 2.13: Limit of x^2 for $a = 2$ with $\epsilon = 2$ (left) $\delta = \sqrt{5} - 2$ (centre) and $\epsilon = 1/2$ (right)

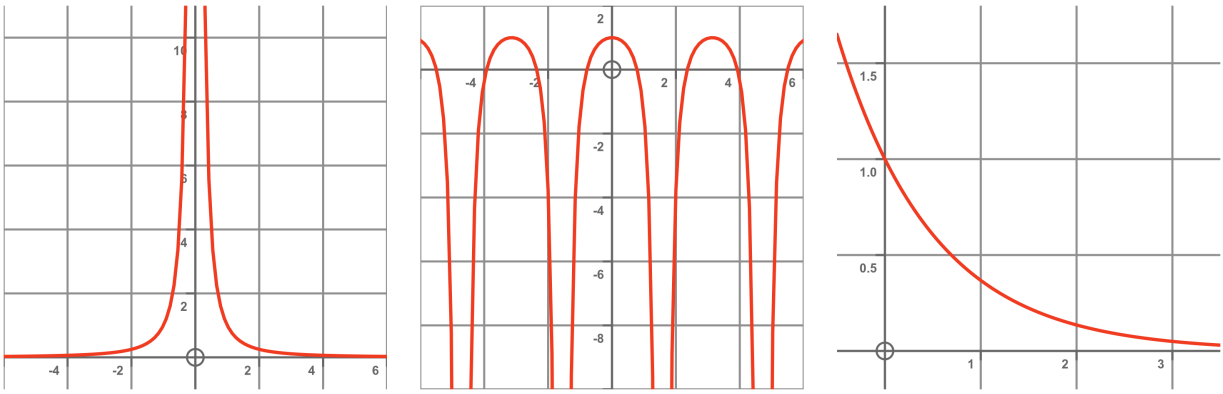


Figure 2.14: Functions $1/x^2$ (left) and $1 - tan^2x$ (centre) and e^{-x} (right) have asymptotes

is symmetrical about 2 and guarantees that $f(x)$ will be in $(2, 6)$. This is shown in the centre of figure 2.13.

So the upshot of all this is that, for a value of $a = 2$ and $\epsilon = 2$, then the condition stated in the limit will be met when $\delta = \sqrt{6} - 2$. But what if we chose a different value for ϵ, say $1/2$? Then there would be a range of x values that gave an $f(x)$ within $2 \pm 1/2$. The δ value would be smaller, but it would still exist. The graph on the RHS of figure 2.13 shows this.

We won't prove it here, but it is hopefully obvious that however small we make ϵ (provided it is > 0) there will be a δ that satisfies the condition.

2.2.5 Asymptotes

Let's look at a different function, $1/x^2$ (LHS of figure 2.14). We can see that firstly it is always positive, and secondly, the value gets bigger as x approaches zero from either side. In fact, it grows without bounds as we get closer to 0.

This is called an asymptote (or a vertical asymptote, since it heads towards infinity in the y direc-

tion). We might say that this function has a limit of ∞, ie:

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

This isn't a limit in the normal sense, because ∞ isn't a number. The main problem is that, no matter how big $1/x^2$ becomes, it will never be "close" to ∞ - it is still infinitely far away. So we need a slightly different definition when the limit is infinite:

A function $f(x)$ has a **limit of infinity** as x tends to a if $f(x)$ grows without bounds as x approaches a. This is written as:

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

More formally, the limit is infinity if for any $M > 0$ there exists a $\delta > 0$ such that:

$$|f(x)| > M \quad \text{whenever} \quad |x - a| < \delta$$

The centre graph in figure 2.14 also shows the function:

$$1 - \tan^2 x$$

This also has vertical asymptotes, but in that case, the function goes to $-\infty$. The definition above applies to these, because we are considering the modulus, $|f(x)| > M$, so very large negative values will also meet the criterion.

The RHS graph shows the function e^{-x}. This function tends to zero as x tends to ∞. This is called a horizontal asymptote. Again it is not quite like a normal limit, and it isn't quite like a vertical asymptote either, so we need another definition:

A function $f(x)$ has a **limit as** x **approaches** ∞ if $f(x)$ tends to some value L as x gets larger. This is written as:

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

More formally, the limit exists if, for any positive number ϵ, no matter how small, there exists a positive number N such that:

$$|f(x) - L| < \epsilon \quad \text{for any } x \text{ satisfying} \quad x > N$$

We can define a limit at $-\infty$ in a similar way, using the condition that $x < -N$.

2.2.6 Left and right limits

The function $\text{floor}(x)$, shown on the LHS of figure 2.15, is defined below:

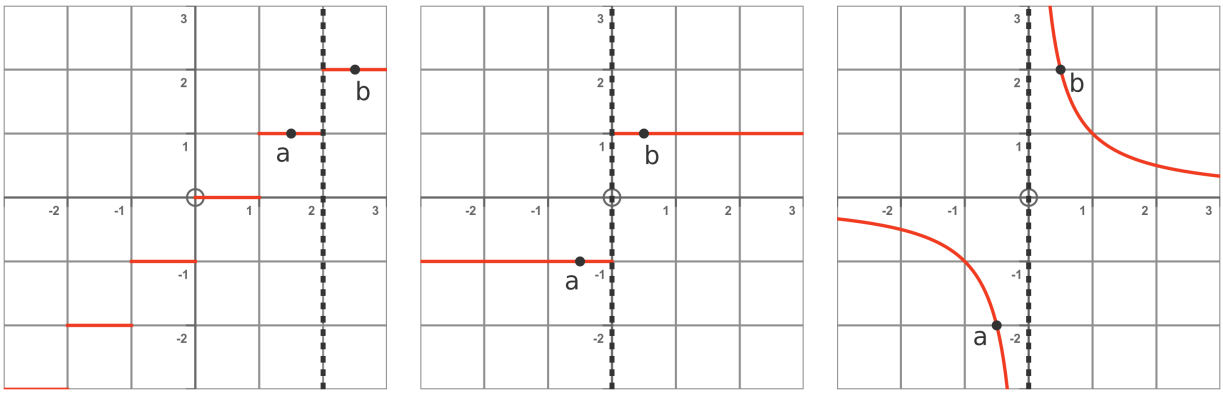


Figure 2.15: Functions $\lfloor x \rfloor$ (left) and $x/|x|$ (centre) and $1/x$ (right)

The $\text{floor}(x)$ function, written as $\lfloor x \rfloor$ returns the integer part of x, ie the largest integer that is $\leq x$.

For example, for x in the interval $[1, 2)$ (that is, $1 \leq x < 2$) $\lfloor x \rfloor$ has value 1, for x in the interval $[2, 3)$ $\lfloor x \rfloor$ has value 2, and so on.

This function is discontinuous at every point where x is an integer. This causes a problem if we try to find the limit at some integer value of x. For example, if we start at point a, where $x = 1.5$, then $\lfloor x \rfloor$ will be 1. If we move x closer to 2, without actually reaching 2, the value of $\lfloor x \rfloor$ will remain equal to 1. So the limit as x tends to 2 from the left is:

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

We call this the *left limit*. Notice the small raised minus sign in the limit $x \to 2^-$, which indicates this is a limit as x approached 2 from below (ie the left) only.

Alternatively, if we start at point b (where $x = 2.5$), then $\lfloor x \rfloor$ will be 2. If we move x closer to 2, without actually reaching 2, the value of $\lfloor x \rfloor$ will remain equal to 2. So the limit as x tends to 2 from the right is:

$$\lim_{x \to 2^+} \lfloor x \rfloor = 2$$

This is the *right limit* and has a small raised plus sign next to the 2.

The value of the function when $x = 2$ is 2, which is also equal to the right limit in this particular case. But for this function, the left limit at 2 and the right limit at 2 are not equal, so we say that the function does not have a limit at 2 (and in this case, the same is true for any integer value of x).

If $f(x)$ has a left limit and a right limit at a but those limits are not equal, $f(x)$ does not have a limit at a.

This follows from the definition of a limit. For a limit L to exist, we must be able to choose a value

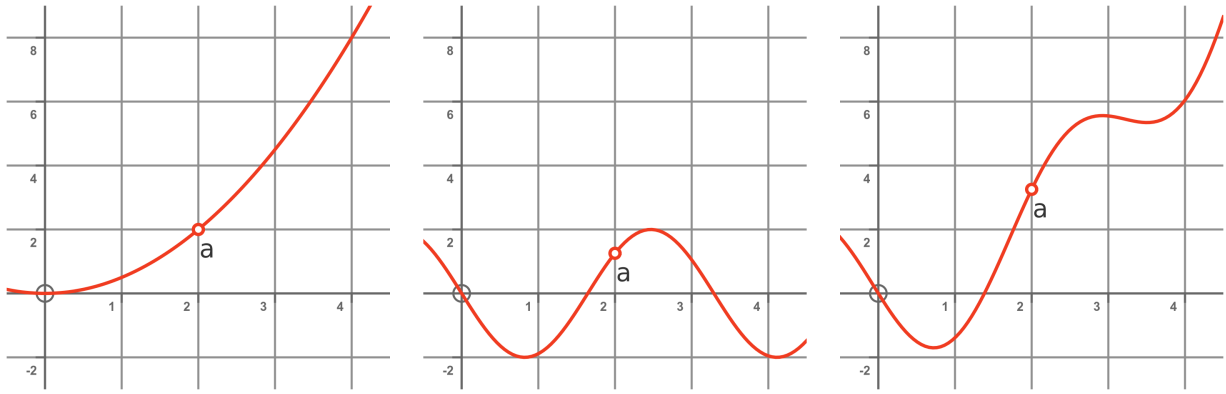


Figure 2.16: Limit at a of $f(x)$ (left), $g(x)$ (centre) and $f(x) + g(x)$ (right)

δ such that $f(a - \delta)$ and $f(a + \delta)$ are both as close as we want to L, and this is not possible if the left and right limits are not equal.

The centre graph in figure 2.15 shows the function $x/|x|$, which we have looked at before. This again has unequal left and right limits at zero:

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

Again say the function does not have a limit there.

Finally, the RHS of the figure shows the function $1/x$. At zero, there is a left limit of $-\infty$ and a right limit of ∞, so the function has no limit there. This contrasts with the function $1/x^2$, which we saw in figure 2.14, where the left and right limits are both ∞, so the function does have a limit of infinity as x tends to zero.

2.2.7 Limit laws

Let's assume we have two functions, $f(x)$ and $g(x)$, and an open region I that contains the value a. Both functions are defined over I, except for $x = a$. The functions have limits at a:

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

$$\lim_{x\to a} g(x) = L_g$$

Example functions $f(x)$ and $g(x)$ are shown in figure 2.16. The graph on the RHS shows the function $f(x) + g(x)$.

What is the limit of the sum of these two functions at a? Well, it is:

$$\lim_{x\to a}[f(x) + g(x)] = \lim_{x\to a} f(x) + \lim_{x\to a} g(x)$$

Law	Equation	Symbol
Sum rule	$\lim_{x\to a}[f(x)+g(x)] = \lim_{x\to a} f(x) + \lim_{x\to a} g(x)$	$L_f + L_g$
Difference rule	$\lim_{x\to a}[f(x)-g(x)] = \lim_{x\to a} f(x) - \lim_{x\to a} g(x)$	$L_f - L_g$
Constant multiply rule	$\lim_{x\to a} cf(x) = c\lim_{x\to a} f(x)$	cL_f
Product rule	$\lim_{x\to a}[f(x)\cdot g(x)] = \lim_{x\to a} f(x) \cdot \lim_{x\to a} g(x)$	$L_f \cdot L_g$
Quotient rule	$\lim_{x\to a}[f(x)/g(x)] = \lim_{x\to a} f(x) / \lim_{x\to a} g(x)$	L_f/L_g
Power rule	$\lim_{x\to a} f(x)^n = (\lim_{x\to a} f(x))^n$	$(L_f)^n$
Root rule	$\lim_{x\to a} \sqrt[n]{f(x)} = \sqrt[n]{\lim_{x\to a} f(x)}$	$\sqrt[n]{L_f}$

Table 2.3: The limit laws

We will call this L_{f+g}. In terms of the limits defined earlier this is:

$$L_{f+g} = \lim_{x\to a}[f(x)+g(x)] = L_f + L_g$$

We can derive similar results for other combinations of $f(x)$ and $g(x)$, known as the limit laws. These are shown in table 2.3.

2.2.8 Indeterminate forms

When applying the limit laws, we need to watch out for indeterminate forms. We need to take care wherever we are combining limits with values of 0, ∞ and $-\infty$. This includes:

$$\infty - \infty \qquad 0 \cdot \infty \qquad \frac{\infty}{\infty} \qquad \frac{0}{0} \qquad 0^1 \qquad 0^0 \qquad \infty^0$$

For example, if $L_f = \infty$ and $L_g = -\infty$, we can't assume the value $L_f + L_g$ is zero. We need to do a bit more work to find the actual limit.

As an example, consider the two functions (see figure 2.17):

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

$$g(x) = -x$$

These are familiar functions. We know that x^2 tends to ∞ as x tends to ∞. And $-x$ tends to $-\infty$ because of the minus sign.

But does this mean that the limit of $f(x) + g(x)$, or $\infty - \infty$, is zero? Well, no. The sum is indeterminate, but it is quite trivial to evaluate the limit of the sum of the functions:

$$\lim_{x\to\infty}(x^2 - x)$$

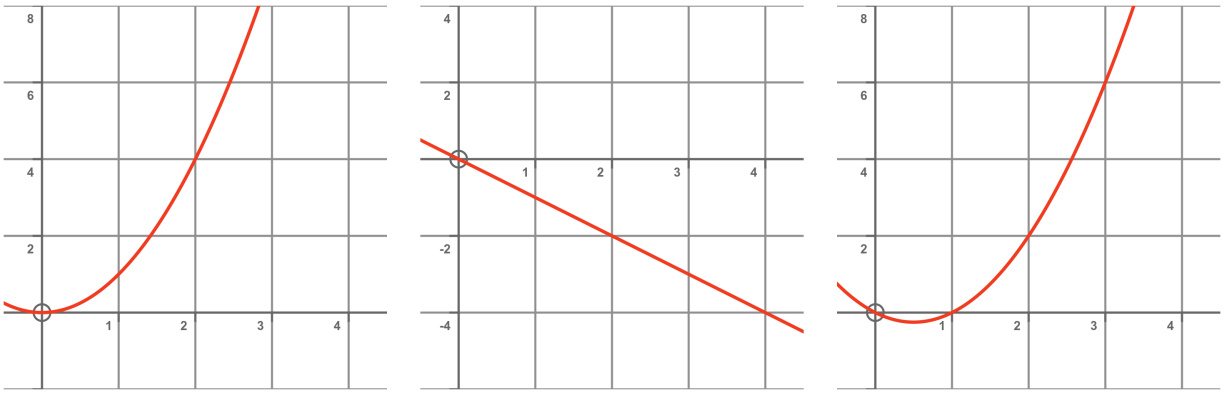


Figure 2.17: $f(x) = x^2$ (left), $g(x) = -x$ (centre) and $f(x) + g(x)$ (right)

For very large values of x, the x^2 term becomes much larger than the x term, so the positive part of the formula wins and the sum tends to ∞. The RHS of the graph shows this. A more formal way to deal with indeterminate forms of limits is L'Hôpital's rule, which we will meet in section **??**.

2.2.9 Limit doesn't always exist

There are several common cases where limits don't exist based on the limit definition given previously.

As we saw earlier, if a function has left and right limits at a but they are not equal, then the function does not have a limit as x tends to a. Examples of this are shown in figure 2.15. The problem there was that if we let x tend to a from the left we get one value for the limit, if we let x tend to a from the right we get a different value, so there is no single value we can give for the limit.

Another case is a function whose domain does not cover the entirety of $\mathbb{R}$. For example, the function $\sqrt{x}$ is only defined for non-negative numbers. This is shown on the LHS of figure 2.18.

You might think that this means there is a limit at zero because we can't go past that point. In reality, we can find the right limit as x tends to zero from the positive side, but there is no left limit because the function isn't defined for values less than zero.

However, in some cases, the right limit might be sufficient, for example, if we are only considering approaching zero from above (as might be the case since the function isn't defined for $x < 0$). Care is needed in cases like that.

Finally, consider a function such as e^x, shown on the RHS of figure 2.18. The value of this function, like many functions, heads towards infinity as x tends to infinity.

We previously looked at vertical and horizontal asymptotes, but e^x does not fit either definition. For a vertical asymptote, $f(x)$ must grow without bounds as we approach a particular finite value of x. For a horizontal asymptote, $f(x)$ must approach a particular finite value as x tends towards $\pm\infty$. But neither of those applies to the case where $f(x)$ grows without bounds as x tends to

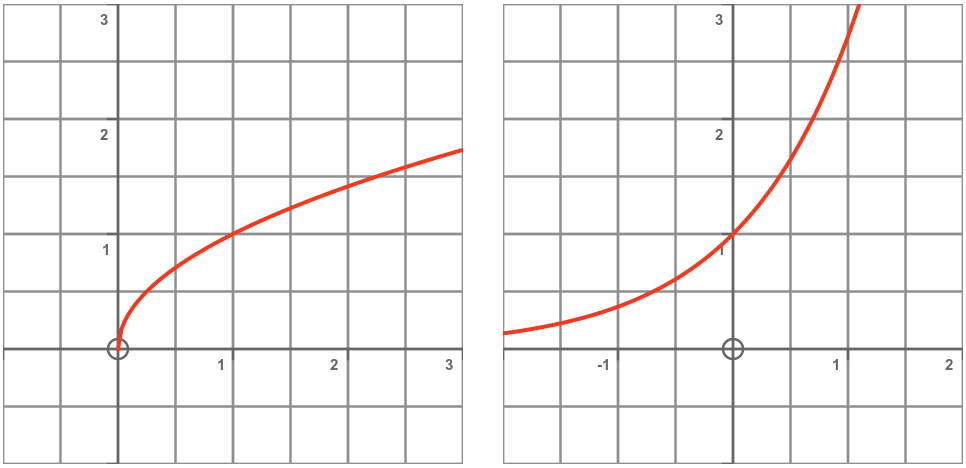


Figure 2.18: $\sqrt{x}$ has no limit at zero (left) and e^{-x} has no limit at ∞ (right)

infinity.

We might informally say that the limit of $f(x)$ as x tends to ∞ is ∞, but it isn't a true limit.

These aren't the only situations where a limit doesn't exist. We will look at a couple of pathological cases next.

2.2.10 Pathological cases

Some functions have strange properties when we examine their limits.

The first example is the function:

$$\sin\left(\frac{1}{x}\right)$$

This is defined over the domain $\mathbb{R}_{\neq 0}$, but we will only consider it for values of $x > 0$. It is shown on the LHS of figure 2.19.

As x tends to zero, $1/x$ gets bigger and bigger, so the sine function oscillates faster and faster, which means the oscillations get more and more compressed along the x-axis.

For any $x > 0$, there will be infinitely many cycles between x and zero. This means that no matter how small x becomes, the function will take every value between 1 and -1 over the interval $(0, x]$. So the function has no limit as x tends to zero.

One way to visualise this function is to imagine the equivalent function of a new variable u, where $u = 1/x$. The function is now $\sin u$, but rather than finding the limit as x tends to 0, we need to find the limit as u tends to ∞. This is shown on the RHS of figure 2.19. Imagine if u starts at 1 and increases to ∞. $\sin u$ has no limit - however big u gets, the sine function just keeps oscillating between 1 and -1.

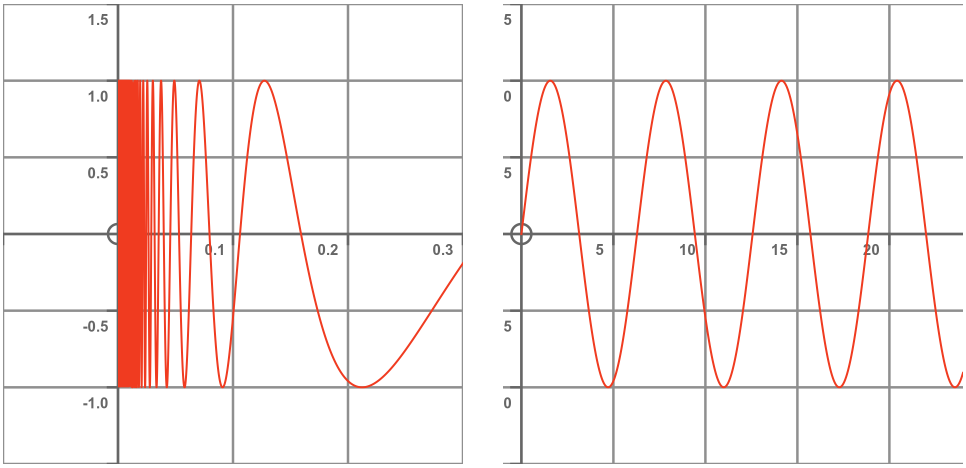


Figure 2.19: $\sin(1/x)$ as x approaches zero (left), and $\sin u$ (right)

The $\sin 1/x$ function is similar. As x starts at 1 and moves to 0, it also oscillates infinitely many times, but all those oscillations are crammed into a finite space, and oscillate faster and faster as x gets smaller. In either case, there is no limiting value.

Another strange function is the Dirichlet function, defined as:

$$f(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q} \\ 0 & \text{if } x \notin \mathbb{Q} \end{cases}$$

If you recall, $\mathbb{Q}$ is the set of rational numbers, so this function takes the value 1 if x is rational, and 0 if x is irrational. It is sometimes called the *indicator function* of the set of rational numbers, written as $\mathbf{1}_{\mathbb{Q}}$.

The set of rational numbers has some interesting properties:

- Any interval in $\mathbb{R}$ of length $\epsilon > 0$ will contain infinitely many rational numbers, no matter how small ϵ is.
- Any interval in $\mathbb{R}$ of length $\epsilon > 0$ will contain infinitely many irrational numbers, no matter how small ϵ is.

These two properties also imply the following:

- If we choose any two rational numbers, no matter how close they are, there will always be an irrational number between them.
- If we choose any two irrational numbers, no matter how close they are, there will always be a rational number between them.

The result of these properties is that the Dirichlet function contains no limits anywhere. That is $\lim_{x \to a} f(x)$ does not exist for any a.

To prove this, we will use the definition of a limit from section 2.2.4. Let's choose a value of $\epsilon = 1/2$. If there is a limit at a, it must be true that (for some value of $\delta > 0$) $f(x)$ is in the interval $f(a) \pm 1/2$ for all x in the interval $a \pm \delta$. If we can show that this is not true for any x, we have proved the result.

First, let's choose any a such that $f(a) = 1$ (ie any rational value of a), then for any value of $\delta > 0$, however small, there will be an irrational number within $a \pm \delta$. This number will have $f(x) = 0$, so $f(x)$ will be 1, which is not within the interval $f(a) \pm 1/2$. Therefore no rational value a can be a limit.

If we choose any irrational value for a, similar reasoning proves that a still cannot be a limit. Since a must either be rational or irrational, this means that no value a can be a limit of the Dirichlet function.

2.3 Big O notation

If we look at the function x^2, we know that as x gets very large, x^2 also gets very large. Many other functions behave similarly, for example, e^x, x factorial, x^x and so on.

But there is an important difference between these functions - they all increase at different rates. As we will see, these rates can be fundamentally different for different functions.

If we have any function that increases without bounds as x increases, we can ask how it increases. Does it increase in a similar way to x^2, or does it increase in a similar way to e^x, or maybe some other function? We call this the *order* of the function, and we use *Big O notation* to represent it. For example, if a function increases like x^2 we say it is an $O(n^2)$ function, or that it has order n^2. Notice we use n rather than x in the notation - that is a common convention.

2.3.1 Rates of increase of different functions

Figure 2.20 shows a variety of functions that increase without bounds as x increases.

The LHS shows various powers of x. The function x^0, which of course is just 1, doesn't increase with x, but it is shown as a starting point. x^1, (ie x), increases linearly. Higher powers, such as x^2 and x^3, curve upwards. The graph also shows the $\ln x$ function.

An important thing to realise is that each function isn't just a bit steeper than the previous function, it increases in a different, and fundamentally faster, way.

Consider the two functions $1000x^2$ and x^3. When x is 1, or 10, or 100, then the function $1000x^2$ is bigger than x^3. But eventually, x^3 catches up and overtakes $1000x^2$. We can say that there is a

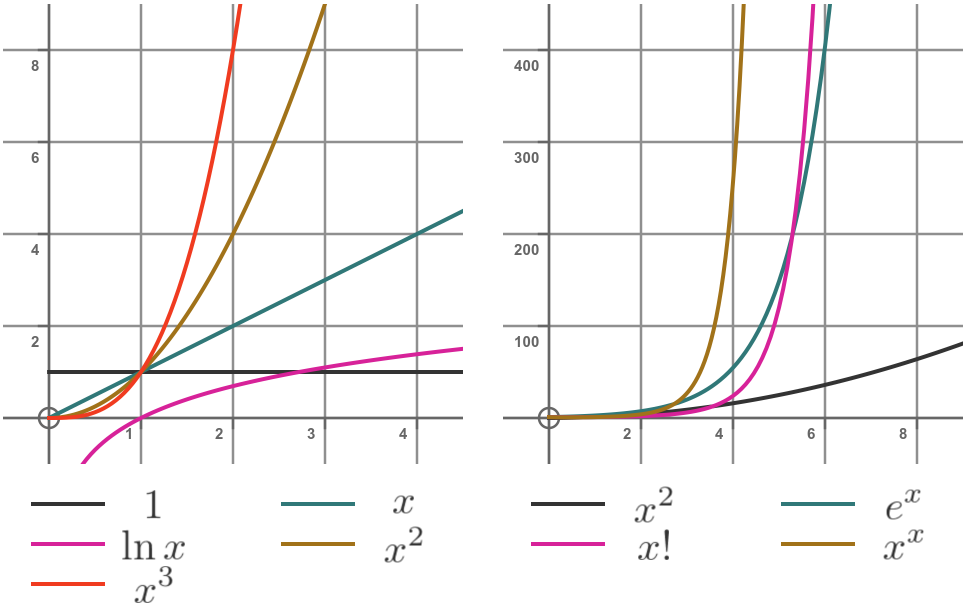


Figure 2.20: Various functions the increase without bounds as x increases

value v such that:

$$x^{n+1} > 1000x^n \quad \text{when} \quad x > v$$

In fact, if we multiply x^n by any finite, positive number M, no matter how big M is, then eventually x^{n+1} will become larger than Mx^n. The rate of increase of x^{n+1} is fundamentally different to the rate of increase of x^n.

This doesn't only apply to powers of x. The RHS of figure 2.20 shows some more functions – notice that this graph has a much larger scaling of the y-axis. In order of rate of increase, the functions from both graphs are 1, $\ln x$, x^n (for $n > 0$), e^x, x factorial (written as $x!$), and x^x.

As we noted before x^2 is a higher order than x^3, but e^x is a higher order than x^2, or x^3, or x^n for any finite n. The exponential function, ultimately, rises faster than any power function. $x!$ eventually rises faster than e^x, and x^x rises faster still.

2.3.2 The order of a function

The functions above represent some well-known orders. There are also some others. For example, $\ln(\ln x)$ is a function that increases even slower than the $\ln x$ function, while $x \ln x$ increases faster than $\ln x$ but slower than x^2.

Some functions increase even faster than x^x, for example x^{x^x}.

There are, in principle, infinitely many orders we could invent, but the ones mentioned in section 2.3.1 represent most of those that are commonly used.

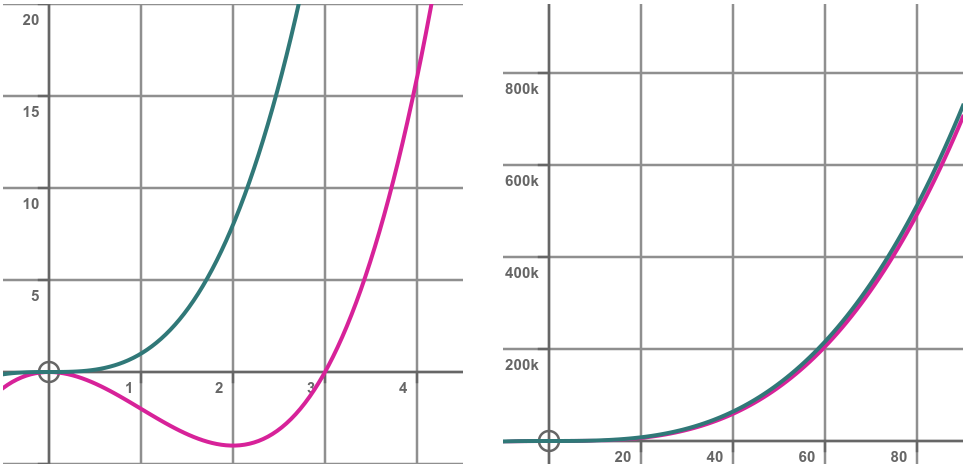


Figure 2.21: $g(x) = x^3$ (cyan) and $f(x) = x^3 - 3x^2$ (magenta) over different ranges

It is interesting to look at how this applies to the sum of two functions. For example:

$$f(x) = x^3 - 3x^2$$

Is there yet another order, something like $x^3 + x^2$? Well, fortunately not. Figure 2.21 shows why. Here, $f(x)$ is plotted in magenta. The function x^3, which we will call $g(x)$, is plotted in cyan.

The LHS shows the two graphs over the range 0 to 4. The graphs look quite different over that range.

The RHS shows the same graphs over the range 0 to 80. This time the graphs look a lot more similar.

What is happening here? Well, for small values of x, the cube and square terms are similar in size. Since $f(x)$ is the sum of these terms it looks quite different to $g(x)$. For larger x values, it is a different story. As x increases, both terms increase in magnitude (one is positive, one is negative). But the cubed term gets bigger much faster than the squared term.

Table 2.4 shows the value of x^3, the difference between the two terms for various values of x, and also shows the ratio of the two.

There are several points to notice here:

- The two curves never become equal as x gets larger. $f(x)$ is always less than $g(x)$ for $x > 0$.
- The absolute difference between the two curves, (which is equal to $3x^2$), always increases as x gets larger.
- BUT, the difference between the two curves as a proportion of $g(x)$ gets smaller and smaller as x increases.

x	$g(x) = x^3$	$f(x) - g(x) = -3x^2$	Ratio $\frac{f(x)-g(x)}{g(x)}$
1	1	-3	-3.0
10	1000	-300	-0.3
100	1000000	-30000	-0.03
1000	1000000000	-3000000	-0.003
10000	10000000000000	-300000000	-0.0003
100000	1000000000000000	-30000000000	-0.00003

Table 2.4: Order of $f(x) = x^3 - 3x^2$

In other words, $f(x)$ and $g(x)$ become more and more alike as x increases. In both functions, the fastest growing term, x^3 ultimately takes over. We say that both functions are of order x^3 because that is what they tend towards for large x. This can be written as $O(n^3)$ in Big O notation.

We can apply the same reasoning to any function that is the sum or difference of two or more other functions. For example:

- $x^x + x^4 + \ln x$ is an $O(n^n)$ function because x^x is higher order than x^4 and $\ln x$.
- $3e^x + 10$ is exponential, written as $O(c^n)$[1]. We ignore the multiplier, the function grows in the *same way* as e^x.
- $e^x - x^{1000}$ is $O(c^n)$.

Let's look at the final item in a bit more detail. x^{1000} is a function that gets very big, very quickly. For example. when $x = 2$, the value 2^{1000} is approximately equal to 10^{301}, that is a one followed by 301 zeros! Whereas e^2 is approximately equal to seven.

But for very large x, e^x will overtake x^{1000}, and will eventually grow far larger.

2.3.3 Formal definition of order of a function

We can give a formal definition of the order of a function:

A function $f(x)$ is $O(g(n))$ if there are two positive constants M and x_0 such that:

$$|f(x)| \leq Mg(x) \quad \text{for all} \quad x > x_0$$

[1]By convention, we use c^n (where c represents any positive constant) rather than e^n. The two functions have the same basic exponential shape since $e^n = c^{kn}$ for some k.

We can look at a few examples to make sense of this definition. First our example function:

$$f(x) = x^3 - 3x^2$$

We have already seen graphically that this function is $O(n^3)$, but now we can demonstrate it formally.

An important thing to realise is that we do not need to find the exact M and x_0 when the condition first becomes true, we just need to demonstrate that the condition is true for *some* value of M and x_0. In this case, we can use $M = 1$, and pick an arbitrary value for x_0, say 10. With these values, we have:

$$f(10) = 10^3 - 3 \cdot 10^2 = 700$$

$$g(10) = 10^3 = 1000$$

So $f(x)$ is less than $Mg(x)$ (where $M = 1$) at this point, and as x gets bigger the difference will only increase (because x^3 grows faster that x^2) so we have shown that:

$$|x^3 - 3x^2| \leq 1 \cdot x^3 \quad \text{for all} \quad x > 10$$

So our function is $O(n^3)$.

Looking at our another example:

$$3e^x + 10$$

We already made the observation that this function is going to look more and more like $3e^x$ as x gets larger. But our formal definition requires us to show that:

$$|3e^x + 10| \leq Me^x \quad \text{for all} \quad x > x_0$$

We might expect the value of M to be 3, to match the equation. But this condition will never be true:

$$3e^x + 10 \leq 3e^x$$

The solution is to use a larger value for M. But remember, we don't need to find the smallest values of M and x_0, we just need to find a pair of values that work. So we could choose $M = 4$. As x gets large, and the term 10 becomes insignificant, there will definitely be a value of x_0 where:

$$3e^x + 10 \leq 4e^x \quad \text{for all} \quad x > x_0$$

To satisfy this inequality we need $e^{x_0} > 10$. The value 3 works, but of course so would any larger value. We now have:

$$3e^x + 10 \leq 4e^x \quad \text{for all} \quad x > 3$$

This shows that the function is the same order as e^n, written as $O(c^n)$.

One more point to mention. According to the definition above, if a function meets the conditions to be of order $O(n^2)$ it will also meet the conditions to $O(n^3)$, $O(n^4)$ and so on. When we talk about the order of a function we normally mean the lowest order that matches.

2.3.4 Applications in computer science

If you have studied computer science you will probably have heard of Big O notation in that context too. It has a similar meaning, but it usually relates to the worst-case performance of an algorithm as a function of the amount of input data. Typically the performance being measured is the time taken to perform the algorithm, or sometimes the amount of memory required.

As an example, consider a search algorithm. We have a list containing n numbers, and we wish to check if a particular number is in the list. For example, we might want to know if the following list contains the number 3:

$$[1, 7, 8, 3, 5, 2, 9, 6]$$

One way to do that is a *linear search*. This means we start at the beginning of the list, and check every element in turn to see if it matches the value we are looking for.

The time taken to do this will vary. The first element might match, so we can stop searching after that - we know that the list contains the number. If no elements match, we will need to search right to the end of the list before we can be certain there is no match. In our example, we would need to check four elements before we find a match with the value 3.

But we are interested in the worst case. This is the case when there is no match, so we have to search right to the end of the list. This will require n checks (the length of the list) so we say this algorithm is $O(n)$.

A more efficient algorithm is the binary search. This requires a sorted list of numbers, ie:

$$[1, 2, 3, 5, 6, 7, 8, 9]$$

This is more efficient. We can pick the central element. Let's choose the fourth element, value 5. Since the elements are ordered, and 5 is greater than 3, we know that if 3 is present it must be in the first half of the list.

We can repeat this process, halving the length of the list each time. Since there are eight elements in the list, if we halve it three times, we will either have found the matching value or we will know it isn't present.

If we had started with sixteen elements, we would have needed four operations to search the list. Each time we double the length of the list, we only add one extra operation. If n is the length of the list, and t is the maximum number of tries required, we have:

$$n = 2^t$$

So we have:

$$t = \log_2 n$$

Of course, t has to be an integer, but when n and t are quite large the rounding error will be small. This tells us that binary search is an $O(\log n)$ process, which is a massive improvement over $O(n)$ for large n.

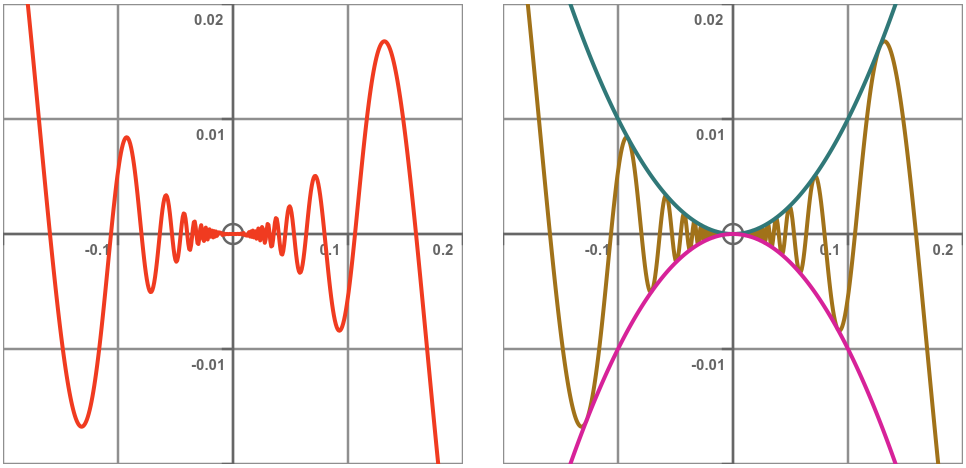


Figure 2.22: $x^2 \sin \frac{1}{x}$ (left) and same function with x^2 and $-x^2$ (right)

2.4 Squeeze theorem

The squeeze theorem is a useful way to find a limit in some quite specific situations. It is best explained with a couple of examples.

2.4.1 Example – $x^2 \sin(1/x)$

As a first example, we will use the squeeze theorem to find:

$$\lim_{x \to 0} x^2 \sin \frac{1}{x}$$

The function is shown on the LHS of figure 2.22.

The problem here is that we cannot evaluate or find the limit of $\sin(1/x)$ at zero because the argument $1/x$ goes to infinity, so the function oscillates infinitely many times as it approaches 0 (see section 2.2.10).

What can we do? Well, we can observe that the value of $\sin(1/x)$ is always in the range $[-1, 1]$ for any value of x. Even though its value oscillates infinitely many times as we move towards zero, it can never go outside that range. In other words:

$$-1 \leq \sin \frac{1}{x} \leq 1$$

This alone doesn't help us find the limit, because although the function is bounded, it is the oscillation that causes the problem. But we can multiply through by x^2 giving this inequality:

$$-x^2 \leq x^2 \sin \frac{1}{x} \leq x^2$$

The central term is now our original function, so we have:

$$-x^2 \leq f(x) \leq x^2$$

We are allowed to perform this multiplication because we know that $x^2 \geq 0$ for any x. We can multiply an inequality by a non-negative value and it still holds (if we multiplied by a negative value it would reverse the sense of the $\leq$ condition).

So as x tends to zero, the $\sin(1/x)$ might go a little crazy but it will always be finite, and x^2 will tend to zero. So the value of $f(x)$ must tend to zero because any finite number multiplied by zero is zero. So we have:

$$\lim_{x \to 0} f(x) = \lim_{x \to 0} x^2 \sin \frac{1}{x} = 0$$

The squeeze theorem is a generalisation of this example. We won't prove it here, but it states:

Let $f(x)$, $g(x)$ and $h(x)$ be three functions such that:

$$g(x) \leq f(x) \leq h(x) \quad \text{for all } x \text{ close to } a \text{ but not equal to } a$$

Suppose:

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

Then:

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

2.4.2 Example – $(\sin x)/x$

As a second example, we will consider the limit:

$$\lim_{x \to 0} \frac{\sin x}{x}$$

Although this expression appears simpler than the previous example, it is actually slightly more difficult to solve and has quite a surprising result. The limit turns out to be 1.

Solving this requires us to use a geometric argument combined with the squeeze theorem. Figure 2.23 shows the construction we will use.

We start with a circle with centre P and radius 1. We then construct a triangle PQR where Q and R are two points on the radius of the circle that make an angle x at the centre. Since we will be finding the limit as x tends to zero we can assume x is in $[0, \pi/2]$.

We will also extend the line PQ to a point S such that the angle PQS is a right angle.

This creates three different areas, shown in figure 2.23:

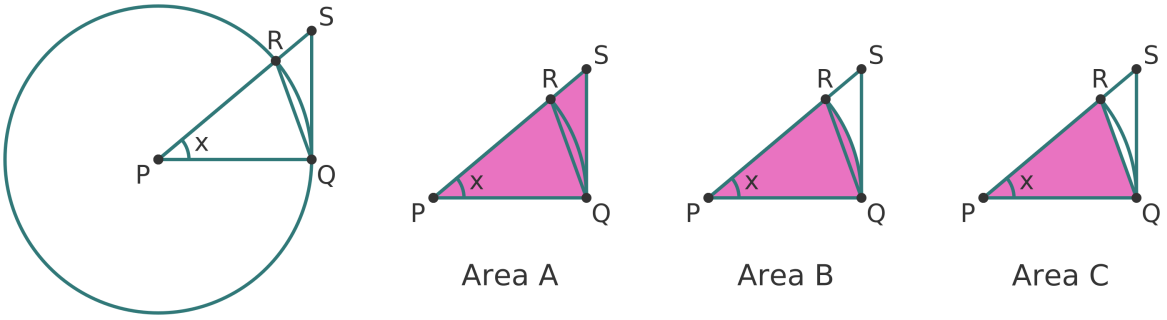


Figure 2.23: Geometric proof of limit of $(\sin x)/x$

- A is the area of the triangle PQS.
- B is the area of the minor sector of the circle between points Q and R.
- C is the area of the triangle PQR.

Now we can make an important observation. Area A completely encloses area B, which in turn completely encloses area C. Therefore:

$$A \geq B \geq C$$

Now let's calculate these areas, in terms of the angle x.

Area A is a right-angled triangle with base $PQ = 1$ (because the circle has a radius of 1). It's height is QS. Since it is a right-angled triangle, basic trigonometry tells us that $QS = PQ \tan x$.

The area of the triangle is therefore:

$$\begin{aligned} A &= \frac{1}{2}\, base \times height \\ &= \frac{1}{2} PQ \times PQ \tan x \\ &= \frac{\tan x}{2} && \text{since } PQ = 1 \end{aligned}$$

Area B is a sector of a circle with radius 1 and angle x. This is a standard formula:

$$\begin{aligned} B &= \frac{1}{2}\theta r^2 && \text{general formula} \\ &= \frac{x}{2} && \text{since } r = 1 \text{ and } \theta = x \end{aligned}$$

Finally area C is a triangle with base PQ. Its height is the perpendicular distance of R from the line PQ, which has a length $PR\sin x$. Its area is:

$$\begin{aligned} C &= \frac{1}{2}\, base \times height \\ &= \frac{1}{2} PQ \times PR \sin x \\ &= \frac{\sin x}{2} && \text{since } PQ = PR = 1 \end{aligned}$$

We can plug these values into our previous area inequality:

$$\begin{aligned} & A \geq B \geq C \\ & \frac{\tan x}{2} \geq \frac{x}{2} \geq \frac{\sin x}{2} \\ & \tan x \geq x \geq \sin x && \text{multiply through by 2} \end{aligned}$$

Let's divide through by $\sin x$:

$$\frac{1}{\cos x} \geq \frac{x}{\sin x} \geq 1$$

This is starting to look quite promising! We can invert the fractions, remembering that this will also reverse the inequality (if $a \geq b$ then $1/a \leq 1/b$), which gives:

$$\cos x \leq \frac{\sin x}{x} \leq 1$$

Now as x tends to 0, $\cos x$ tends to 1, so we have:

$$1 \leq \lim_{x \to 0} \frac{\sin x}{x} \leq 1$$

So by the squeeze theorem:

$$\lim_{x \to 0} \frac{\sin x}{x} = 1$$

Just one final point. We previously assumed x was small but positive, so the proof so far only applies to the right limit (the limit as x tends to 0 from above). But the limit depends on $\cos x$, which is an even function, so if we repeat the proof assuming small negative x we will find that the left limit is also 1. We won't repeat the calculation here, as it is almost identical to the one we have just done.

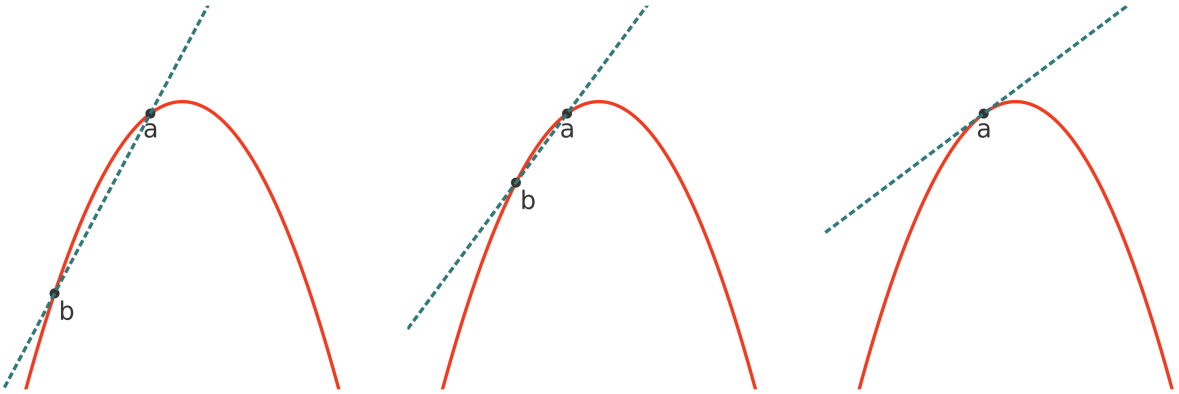


Figure 2.24: Tangent is limit of secant from a to b as $b \to a$

2.5 Tangents to a curve

Calculus, and differentiation in particular, frequently uses the concept of tangents. A tangent to a curve is a line that touches a curve and has the same slope as that curve at the point of contact.

It is sometimes said that a tangent touches a curve without crossing it, but that is not true in all cases as we will see.

2.5.1 Tangent as the limit of a secant

A *secant* to a curve is a line that crosses the curve at two points. We can view a tangent as a limit of that secant as the points get closer and closer together. Figure 2.24 shows this.

The tangent at point a can be approximated by a secant between point a and some nearby point b. As b gets closer to a, the line becomes a better approximation. In the limit, as b tends to a, the line tends towards the true tangent. This means that the slope of the line is equal to the slope of the curve at a.

2.5.2 Points of inflection

In figure 2.25 we see that the curve is horizontal at point a, with the curve increasing to the left of a and decreasing to the right of a. This means that the curve at point a crosses the tangent at a (while the curve at any other point p only touches the tangent at p).

We call this a point of inflection. There are several ways to define the point of inflection but the simplest is:

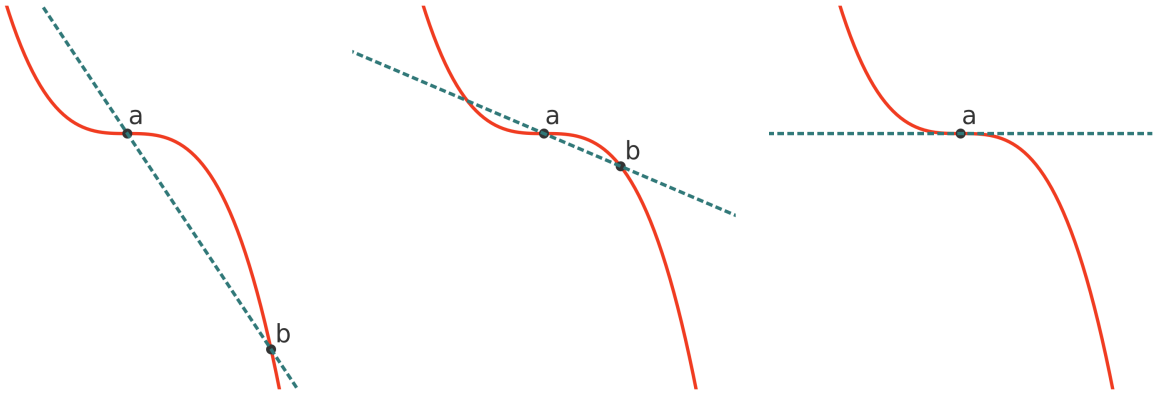


Figure 2.25: Tangent at a point of inflection will intersect the curve

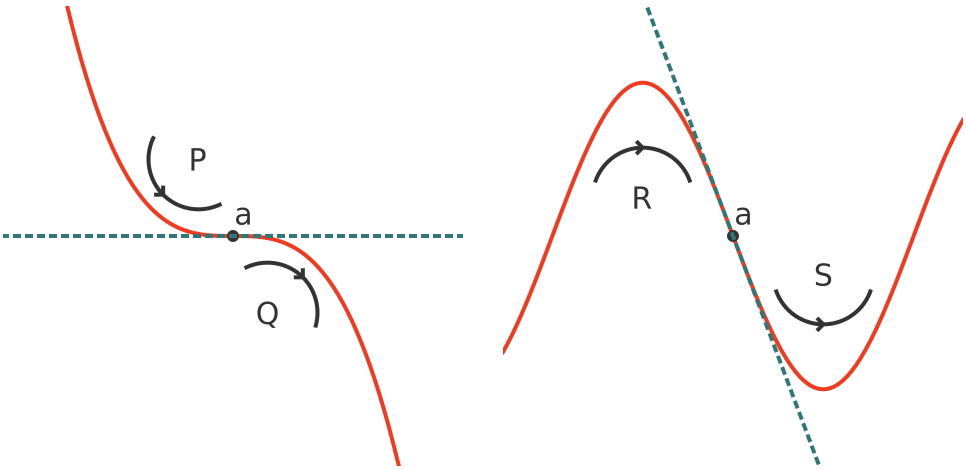


Figure 2.26: Curvature changes sign at a point of inflection

A point of inflection (or inflection point), is a point on a curve where the curve crosses its tangent at that point.

A point of inflection occurs when the *curvature* of a curve changes sign. This is shown in 2.26. The LHS shows the curvature of the same function from figure 2.25:

- To the left of point a, as we move along the x-axis the curvature of the function is counter-clockwise (shown by the arc P).
- To the right of point a, as we move along the x-axis the curvature of the function is clockwise (shown by the arc Q).

At point a the curvature flips from ccw to cw, and because the two sides are curving in opposite directions the curve crosses the tangent rather than just touching it.

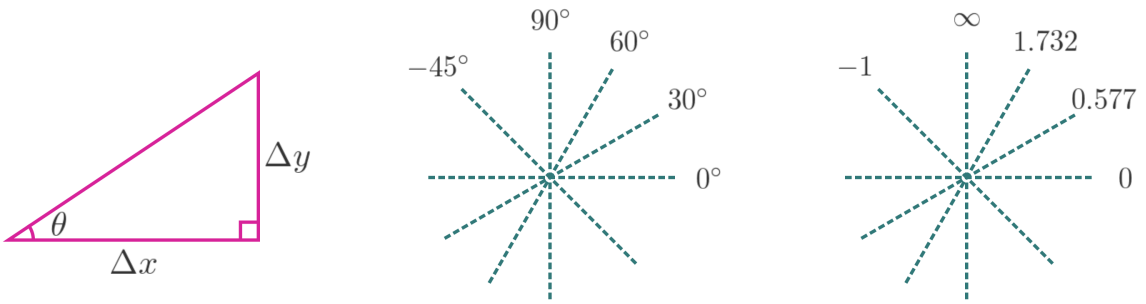


Figure 2.27: Angle θ in terms of Δx and Δy (left, centre), and slope $\Delta y/\Delta x$ (right)

This *often* happens because the second derivative of the curve (ie the rate of change of the rate of change of the curve) changes sign, but there are other situations where it can occur.

Notice that the slope at a point of inflection doesn't have to be horizontal. The RHS of 2.26 shows a sine function, with a point of inflection where the slope is diagonal. To the left of a, the curvature is clockwise (arc R) and to the right of a, it is counterclockwise (arc S).

2.5.3 What is slope?

We have talked about the slope of a line but what do we mean by slope?

> If we choose two points on the line that are separated by a horizontal distance Δx and a vertical distance Δy, then the slope is $\Delta y/\Delta x$

This is sometimes called the rise:run ratio. Figure 2.27 shows various examples of different slopes. Notable points are:

- If y increases as x increases, the slope is positive. If y decreases as x increases, the slope is negative
- A horizontal line has a slope of 0, and a vertical line has a slope of $+\infty$.
- A $45°$ line has a slope of magnitude 1, and smaller or larger angles have smaller or larger magnitudes of slope.

Figure 2.27 also shows the angle that each tangent makes with the x-axis. The general formula for this is:

$$\theta = \arctan \frac{\Delta y}{\Delta x}$$

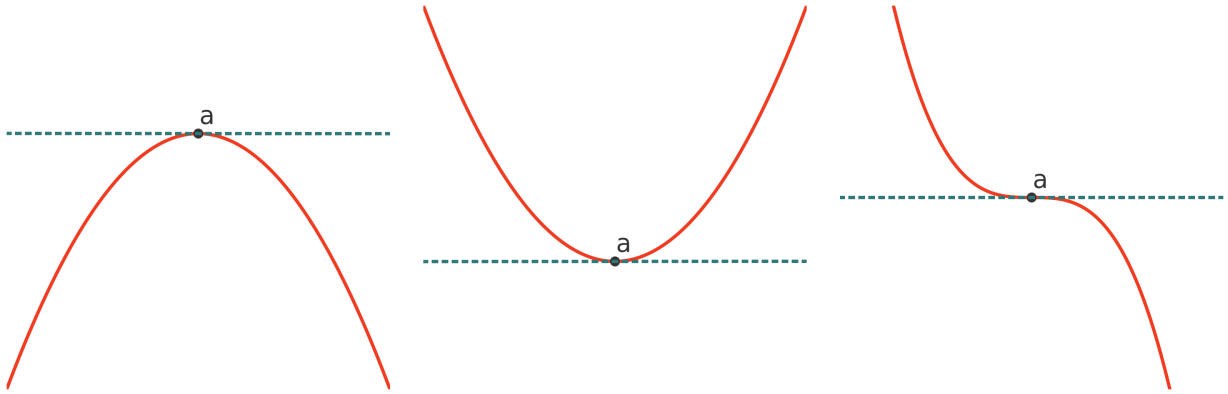


Figure 2.28: Stationary points have a horizontal tangent

θ can be interpreted as an angle between $-90°$ and $90°$ (so negative slopes have a negative angle), or as an angle between $0°$ and $180°$ (so negative slopes have an angle $> 90°$).

2.5.4 Stationary points

A *stationary point* is a point on a curve where the slope is zero. In other words, the tangent is horizontal. Examples are shown in figure 2.28.

Since the tangent is horizontal at a stationary point, it means that an infinitesimal change in x at that point will cause no change in y, which is why they are called stationary points.

Stationary points are often associated with local minima and maxima, as shown in the first two graphs of the figure. In fact *Fermat's theorem of stationary points* (section **??**) tells us that any local maxima and minima on a smooth, continuous curve must be a stationary point.

However, not every stationary point is a maximum or minimum. Some stationary points are points of inflection (such as the RHS of the figure).

2.5.5 Tangent to a straight line

If the curve in question happens to be a straight line, then the tangent is the line itself. This may seem obvious, but it is worth stating because it will be very important later on.

One of the intuitive principles of differentiation is that, if we zoom in on a small section of the curve of a well-behaved function, it eventually starts to look like a straight line. The more we zoom in, the more it resembles a straight line. This is sometimes called the *microstraightness property*. It is this property that allows us to use limits to find the tangent to a curve.

We will define "well-behaved" later when we look at this in more detail.

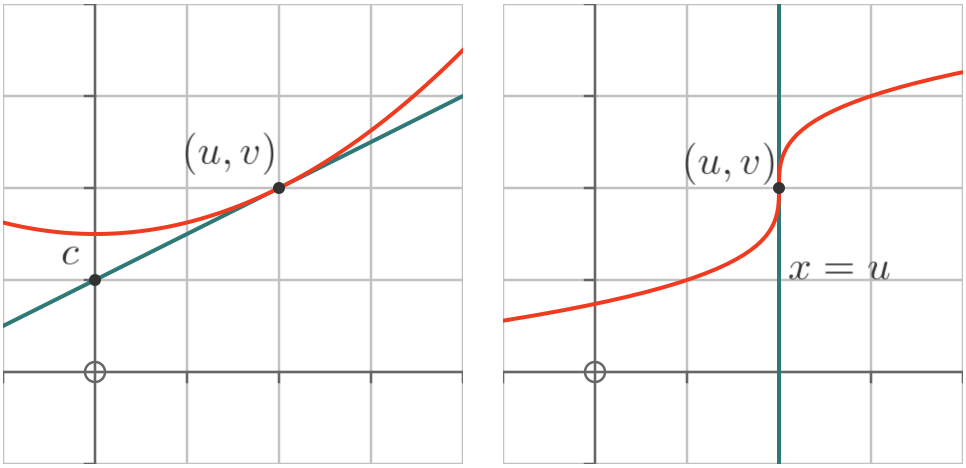


Figure 2.29: Tangent line $y = mx + c$ passing through (u, v), with vertical case shown on right

2.5.6 Equation of a tangent

The general equation of a straight line of slope m intercepting the y-axis at point $(0, c)$ is:

$$y = mx + c$$

To find the tangent to a curve at the point (u, v), where the slope of the function at that point is m, we need to solve the equation above to find c. Substituting u and v for x and y gives:

$$v = mu + c$$

So:

$$c = v - mu$$

Giving:

$$y = mx + v - mu$$

Or:

$$y = m(x - u) + v$$

This is shown in figure 2.29.

This equation is valid in all cases except where the straight line is vertical. In that case, m is infinite and x is constant, so the line has the equation $x = u$. The line doesn't intersect the y-axis, but if u is zero the line is coincident with the y-axis.

2.5.7 Normal to a curve

We sometimes need to find the normal to a curve. The normal to a curve at (u, v) is a line that is perpendicular to the tangent at that point. This is shown in figure 2.30.

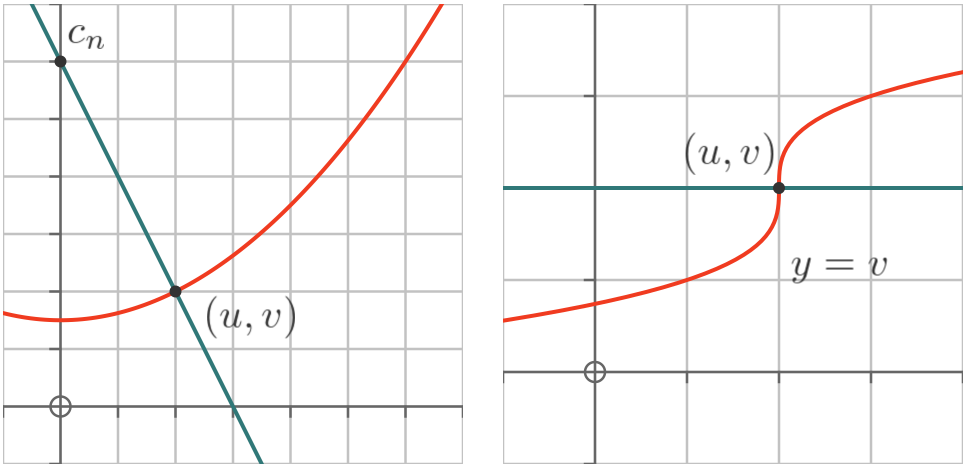


Figure 2.30: Normal line $y = -x/m + c_n$ passing through (u, v), with horizontal case shown on right

If the gradient of the tangent at that point is m, then the gradient of the normal is $-1/m$. If the normal crosses the y-axis at $(0, c_n)$, the equation of the line is:

$$y = -\frac{x}{m} + c_n$$

We can find the equation of this line in terms of m, u, and v, just like we did for the tangent. First we find c_n by substituting u and v in the equation above:

$$v = -\frac{u}{m} + c_n$$

So:

$$c_n = v + \frac{u}{m}$$

Replacing c_n in the original equation gives us:

$$y = -\frac{x}{m} + v + \frac{u}{m}$$

$$= \frac{u - x}{m} + v$$

If the tangent is vertical, this means that the normal will be horizontal, and since it passes through (u, v) the line will have the equation $y = v$.

2.6 More about functions

We will end this chapter by looking at a few other aspects of functions.

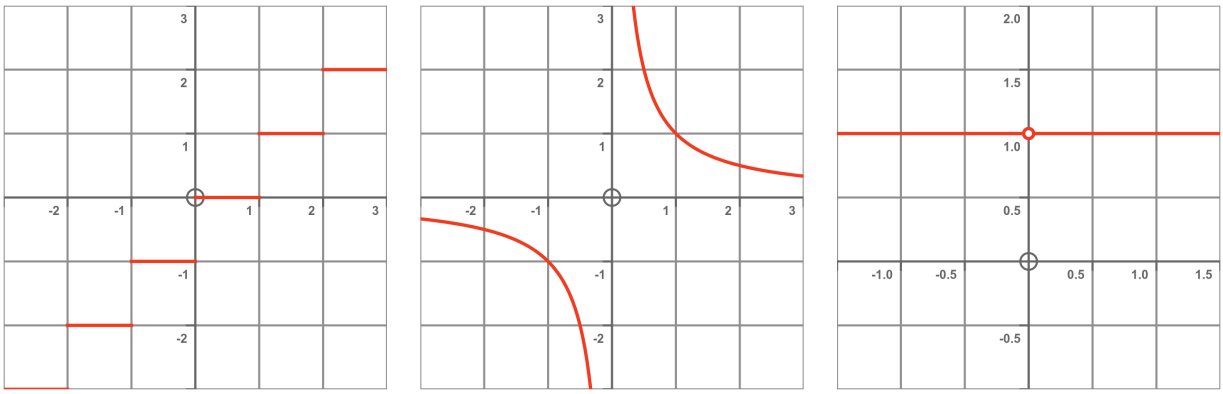


Figure 2.31: Functions $\lfloor x \rfloor$, $1/x$ and x/x have discontinuities

2.6.1 Continuous functions

Roughly speaking, a function is *continuous* if the function curve can be drawn as a single line, without lifting the pencil off the paper.

Another way to say this is that, at any point on the curve, if we change x by a very small amount, the value of the function will only change by a very small amount.

Many common functions are continuous, for example, x^2, $\sin x$, and $\ln x$. These are all functions where the function is a continuous, unbroken line.

Some examples of non-continuous functions are shown in figure 2.31. The LHS shows the floor function $\lfloor x \rfloor$, which is discussed in section 2.2.6. This function takes the value of the integer part of x and has a discontinuity for every integer value of x. For example at $x = 1$ the function jumps from 0 to 1.

The centre graph shows the function $1/x$. This function has a discontinuity at zero. That is because the function tends to $-\infty$ as we approach zero from the left, but it tends to ∞ as we approach zero from the right.

The RHS function is x/x, which is discussed in section 2.2.2. This function is equal to 1 everywhere except at zero where it is undefined. This means it is discontinuous at zero because there is an infinitesimally small region where the function is undefined.

One way to define a continuous function is as follows:

A function $f(x)$ is **continuous** if for any a in its domain, and for any $\epsilon > 0$ there exists a $\delta > 0$ such that:

$$|f(x) - f(a)| < \epsilon \qquad \text{whenever} \qquad |x - a| < \delta$$

You might recognise this as being very similar to the definition of a limit. In plain words, this is

saying that for any point on the function, if we choose a point that is very close to a, then the value of the function at that point will be very close to $f(a)$.

We can also express the definition in terms of limits:

A function $f(x)$ is **continuous** if for any a in its domain, the limit:

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

exists and is equal to $f(a)$.

This has exactly the same meaning as the previous definition. Looking at our example functions, $\lfloor x \rfloor$ fails because there is no limit at $x = 1$ (or any other integer value). The left and right limits exist but are different. The same is true of the function $1/x$ because at $x = 0$ the left and right limits are different.

For the function x/x, the test fails for a different reason. The function has a limit as x tends to zero. The limit is 1 as we approach from either the left or right side. However the function itself is undefined at zero, so the function and its limit are not equal at that point (because $f(0)$ is undefined).

2.6.2 Types of discontinuity

There are three different types of discontinuity, with different characteristics. They are removable discontinuities, jump discontinuities, and essential discontinuities. The classification is based on the behaviour of the left and right limits at the discontinuity (section 2.2.6).

Here is the definition of a removable discontinuity:

A function has a **removeable discontinuity** at x_0 if

- It is continuous in the region of x_0, everywhere except x_0.
- It has both a left limit L^- and a right limit L^+.
- L^- and L^+ are equal. We will call this value L
- But $f(x_0)$ either doesn't exist or is not equal to L.

For example the function x/x (figure 2.11) has a value of 1 everywhere except $x = 0$ where it is undefined. But it has a limit of 1 as we approach 0 from either direction. So we say it has a removable discontinuity at that point.

If we have a function f that has a removable discontinuity it is quite easy to create a similar function

f_c that is continuous. For example:

$$f(x) = \frac{x}{x}$$

$$f_c(x) = \begin{cases} f(x) & \text{if } x \neq 0 \\ 1 & \text{if } x = 0 \end{cases}$$

$f_c(x)$ has the same value as $f(x)$ for all values of x except x_0 (0 in this case), where it has the value L (1 in this case). It can be thought of as a continuous version of $f(x)$.

Here is the definition of a jump discontinuity:

A function has a **jump discontinuity** at x_0 if

- It is continuous in the region of x_0, everywhere except x_0.
- It has both a left limit L^- and a right limit L^+ at x_0.
- But L^- and L^+ are not equal.

An example of a jump discontinuity is the function $x/|x|$, shown in figure 2.15. This function is undefined at 0, but it has a limit of -1 as we approach 0 from the left, and a limit of +1 as we approach 0 from the right.

Finally, here is the definition of an essential discontinuity:

A function has an **essential discontinuity** at x_0 if

- It is not continuous at x_0.
- One of both of the left and right limits do not exist as we approach x_0

There are several ways that an essential discontinuity can occur:

- If a function has one or more vertical asymptotes (that is, if its value goes to infinity for some finite value x_0). For example, the function $1/x$ (figure 2.15) becomes infinite at 0.
- If a function is undefined over some region. For example, the function $\sqrt{x}$ (figure 2.18) has no left limit at 0 because it is undefined for $x < 0$.
- Certain pathological functions, for example $\sin 1/x$ has no limit at 0 (figure 2.19).

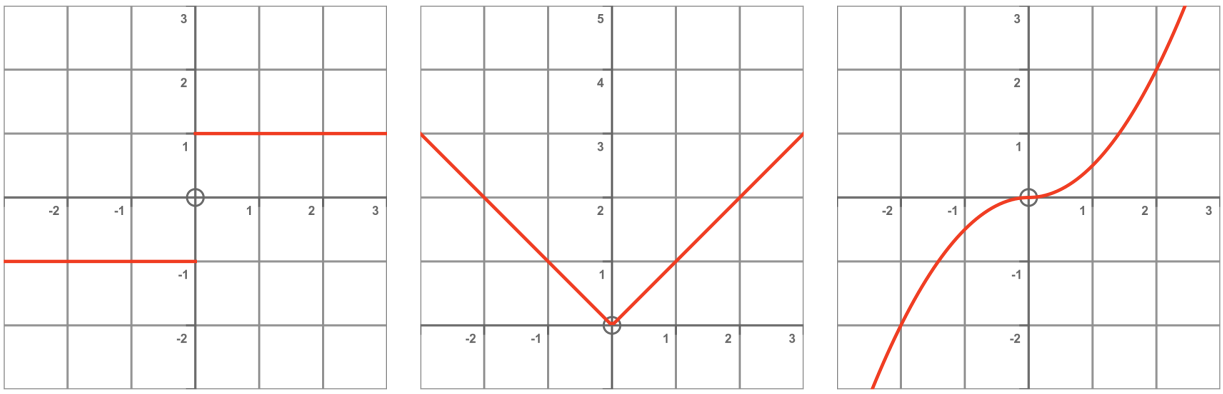


Figure 2.32: Function $\operatorname{sgn}(x)$ (left) $x \operatorname{sgn}(x)$ (centre) and $x^2 \operatorname{sgn}(x)/2$ (right)

2.6.3 Differentiable functions

We won't discuss *differentiation* fully until the next chapter, but we will quickly introduce it here so we can discuss smooth functions. In simple terms when we differentiate a function we get a new function called the *derivative* of the original function. The derivative tells us the slope of the curve at any point, which also represents the rate of change of the function.

We say that a function is *differentiable* if we can find its derivative at every point. The slope of a function, of course, is related to its tangent. A function is differentiable if it has a tangent at every point, and also if that tangent is not vertical at any point.

2.6.4 Smooth functions

A smooth function, as you might guess, is one where the curve representing the function is smooth. A smooth function changes gradually, with no discontinuities or sudden changes in its direction.

An intuitive test for smoothness might be: if the function represented the position of a car over time, would a passenger experience a smooth ride? The exact definition of this is slightly more involved than that, as we will see later. For now, we will just look at a few examples.

We have met many smooth functions already: x^2, e^x, $\ln x$, and $\sin x$ are all smooth functions. If we were to imagine a car travelling according to any of these functions, it would feel smooth, although e^x would reach a terrifying speed pretty quickly.

Figure 2.32 shows a couple of functions that are not smooth.

The function $\operatorname{sgn}(x)$ has the following definition:

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

This function is quite similar to a function we have seen before, the function $x/|x|$. The only difference is that $\mathrm{sgn}(x)$ is zero when x is zero, whereas $x/|x|$ is undefined at that point.

Since the function is discontinuous at $x = 0$ it is certainly not smooth at that point. In the car analogy, the vehicle would be stationary at a particular position, then it would instantaneously jump to a different position. That would be anything but a smooth ride!

The second curve shows the function $x\,\mathrm{sgn}(x)$ (which is equivalent to $|x|$).

We can see that when $x < 0$ the slope is -1 and when $x > 0$ the slope is 1. In our analogy, the slope of the curve represents the velocity of the car. This means that the velocity of the car is the same as the left-hand graph, $\mathrm{sgn}(x)$.

For a smooth ride, we would want to speed of the car to remain constant or change gradually. But In this case, when $x = 0$ the curve forms a sharp point. The slope (ie velocity) changes abruptly at that point - the car is reversing (negative velocity) and then instantaneously switches to travelling forwards at the same velocity. Again this abrupt change in velocity, which represents an infinite acceleration, is not smooth.

The right-hand function is $x^2\,\mathrm{sgn}(x)/2$. This curve looks pretty smooth to the naked eye, so would this give us a smooth ride? Unfortunately not.

As we will see in the next chapter, the function $x^2/2$ has slope x, so in this case the velocity of the car would look like the centre graph, and the acceleration would look like the left-hand graph. There would be an instantaneous change in acceleration.

Would that be a smooth ride? Probably not. In physics, the rate of change of acceleration is called *jolt* (or sometimes *jerk*). These names are quite apt. When a car brakes suddenly, it isn't the change in velocity that you feel, it is the change in acceleration.

For a function to be considered smooth, the function must be continuous, its derivative must be continuous, the derivative of its derivative must be continuous, the derivative of the derivative of its derivative must be continuous, and so on, all the way down. That isn't as onerous as it might sound, there are many functions (including x^2, e^x, $\ln x$, and $\sin x$ mentioned above) that satisfy the condition for smoothness.

2.6.5 Analytic functions

Certain functions can be expressed as infinite polynomials of the form:

$$f(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + a_3(x - x_0)^3 + \cdots = \sum_{n=0}^{\infty} a_n(x - x_0)^n$$

This is called a *Taylor series* (or a Taylor expansion). x_0 is a constant, and a_n are a set of constants that can be derived from the function $f(x)$.

We can attempt to calculate an approximate value for $f(x)$ by calculating a finite number of terms,

that is:

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

We say that the series is convergent if the approximation above gets closer and closer to the true version of $f(x)$ as we increase p. In other words, if we can get as close as we like to the true value of $f(x)$ by choosing a sufficiently large value of p.

Loosely speaking, a function is classed as *analytic* if it has a series that converges (we will give a more precise definition below).

As an example, the exponential function can be written as:

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

In this case, $x_0 = 0$ and $a_n = 1/n!$ (remembering, of course, that 0! is 1). Let's use this to calculate e^1, which is just $e \approx 2.718281828$. Here is the calculation for 2, 4 and 8 terms (result to 8 dp):

$$e^1 = 1 + 1 = 2.00000000$$

$$e^1 = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} = 2.66666666$$

$$e^1 = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \frac{1}{5!} + \frac{1}{6!} + \frac{1}{7!} = 2.71825396$$

After 8 terms, the result is accurate to 4 places, and calculating more terms would give an even more accurate result.

For the special case of $x_0 = 0$, a Taylor series is sometimes called a Maclaurin series.

In the case of the exponential function, the series will converge for any value of x. This is not true for all functions. For some functions:

- The Taylor series will only converge if x is in the neighbourhood of x_0 (ie it will not converge if x is too far away from x_0).
- Convergence also requires x_0 to be within a certain interval D.

We can define an analytic function as:

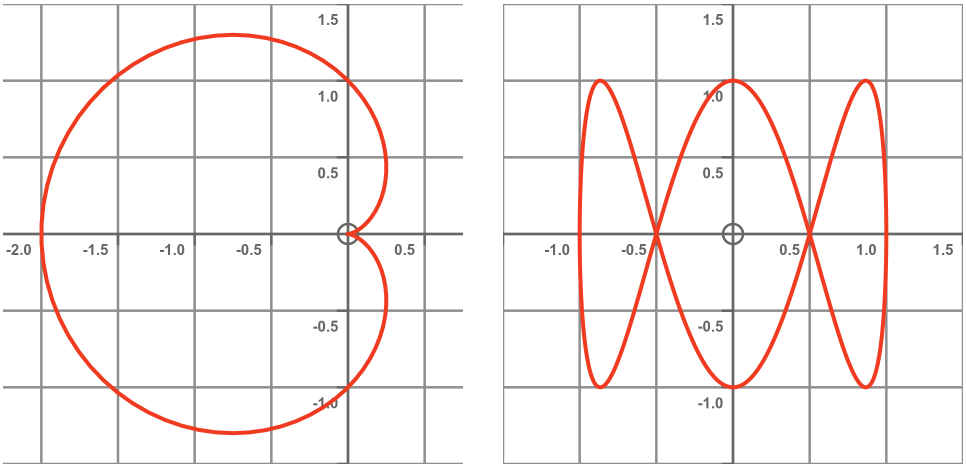


Figure 2.33: Polar cardioid curve (left) parametric Lissajous figure (right)

A function $f(x)$ is analytic if over the interval D if, for any x_0 in D, the series:

$$a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + a_3(x - x_0)^3 + \cdots = \sum_{n=0}^{\infty} a_n(x - x_0)^n$$

converges to $f(x)$ when x is in the neightbourhood of x_0.

2.6.6 Other types of functions

In this book, we will deal almost exclusively with real functions of a single real variable, and mainly functions of the form $y = f(x)$ that are normally drawn using xy coordinates. But there are other types of functions that we will mention here.

Polar functions take the form $r = f(\theta)$. For a polar function, each point on the curve is defined by two values, r and θ. r is the straight-line distance of the point from the origin, and θ is the angle the point makes at the origin, measured counterclockwise from the positive going x-axis.

The LHS of figure 2.33 shows a polar plot of a *cardioid curve* given by:

$$r = 1 - \cos\theta$$

Because the curve is defined in terms of r and θ, it is possible for there to be multiple points on the curve with the same x value. In the curve shown, for example, there are three points with $x = 0$, but each of those points has a different value of θ.

Parametric curves use two separate functions, f and g to calculate x and y values, based on some parameter, often called t. Each value of t defines a point (x, y) using $x = f(t)$ and $y = g(t)$. We might think of t representing time, and the functions representing the position of a moving point.

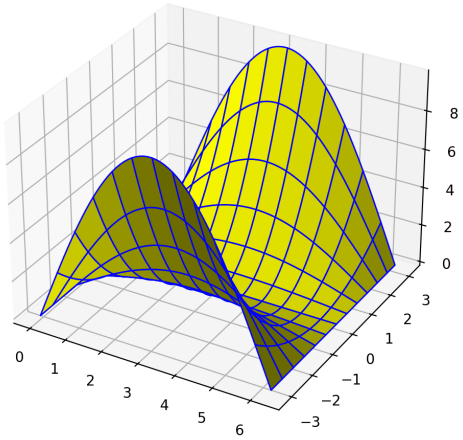


Figure 2.34: $z = y^2 \sin x$, a function of two variables

The RHS of figure 2.33 shows a parametric plot of a *Lissajous figure* given by the pair of functions as t moves from 0 to 2π:

$$x = \sin t$$

$$y = \cos 3t$$

Functions of more than one variable are functions where the value depends on two or more variables, for example, $z = f(x, y)$. A function of two variables can be plotted on 3D axes, for example, figure 2.34 shows the function $z = y^2 \sin x$.

Functions of more than two variables cannot be easily plotted on a single set of axes, and visualising them can be difficult.

There are other types of functions that we won't consider here:

- Line graphs in higher dimensions. For example, if $y = f(x)$ and $z = g(x)$, this defines a line curve in 3 dimensions.
- 3D polar plots, where we define a point in terms of a length r and two angles θ and ϕ.
- Parametric plots higher dimensions. For example, in 3 dimensions we can define a line curve using three functions $x = f(t)$, $y = g(t)$ and $z = h(t)$. We can define a surface using two parameters t and u with $x = f(t, u)$, $y = g(t, u)$ and $z = h(t, u)$. In n dimensions we can define surfaces of any dimension from 1 to $n - 1$

Complex valued functions are functions that accept a complex number and return a complex number, for example, $w = f(z)$. Since a complex number is a 2-dimensional quantity (a real and imaginary part), complex-valued functions effectively exist in 4 dimensions.

Typically, when plotting such a function, we split it into two 3D graphs. Usually, the first graph shows the real part of the output value as a function of the complex input value $z = x + iy$. A second graph shows the imaginary part of the output value.

2.7 Summary

In this chapter, we have looked at functions and their properties:

- Domain, codomain and image.
- Injective, surjective and bijective functions.
- Tangents and normals to a curve.
- The order of a function using Big O notation.
- Conditions for functions to be continuous, smooth, differentiable, and analytic.

We also looked at different sets of numbers, and open/closed intervals.

The chapter also covered the definition of a limit, the various types of limits, and the conditions for a limit to exist.