MATHEMATICS IA: Calculus

Edward Green
School of Mathematical Sciences
University of Adelaide

July 7, 2018

Preface

These lecture notes contain all the material you will be required to know for Mathematics IA (Calculus). They are an aid to your learning, and will *complement* the lectures. Whilst there will be considerable overlap, the material presented here will not be identical to what you see in lectures, and you will put yourself at a disadvantage if you do not attend them. In particular, lectures will feature active learning experiences which will help you develop and test your understanding, and provide opportunities to ask questions of the lecturer or discuss important concepts with your fellow students.

Contents

Chapter 1

Functions

Lecture 1

1.1 Functions

In science, technology, finance and many other fields, we need to be able to express relationships between quantities in a precise way. For example, we might be interested in the force required to make a body accelerate at particular rate, or the amount of energy that could be released from a certain mass. Mathematics provides a natural way of expressing these relationships *e.g.* $F = ma$, $E = mc^2$. The quantities which are related to each other are called the *variables*, and the relationship itself is encapsulated by a *function*, a rule which relates the values of the variables. For the simplest case where there are just two quantities which are related to each other, you will already be familiar with writing these kinds of relationships in the form:

$$y = f(x),$$

where, of course, the choice of symbols is arbitrary, and the meaning of the quantities x and y, and the nature of the function, f, will depend on the situation being considered.

We can think of x (which is usually called the *independent variable*) as the *input*. Then, the function, f, tells us what to do to x to get the value of the *dependent variable*, y. In this course, we will deal only with relationships where x and y are real numbers. In precise mathematical language, we say that f is a real-valued function of a single real variable.

You will have already met a wide variety of functions of this type, such as:

- **Polynomials**

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

  is a polynomial function of degree n (if $a_n \neq 0$; each of the $a_0, a_1, \ldots, a_n$ is a (constant) real number).

- **Rational Functions**

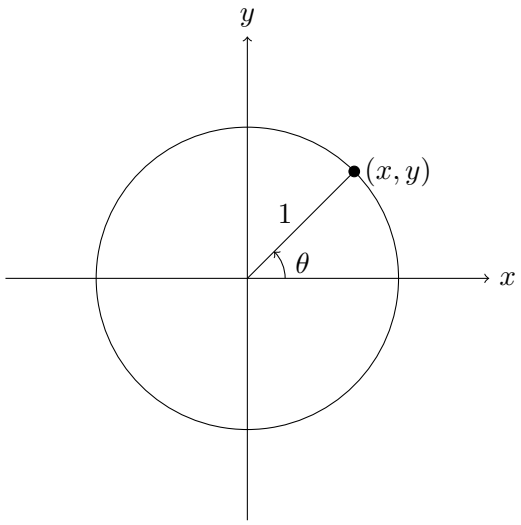
  If $p(x)$ and $q(x)$ are polynomial functions

  $$r(x) = \frac{p(x)}{q(x)} \quad \text{is a } \textit{rational function}.$$

Note that when $q(x) = 0$, $r(x)$ is undefined.

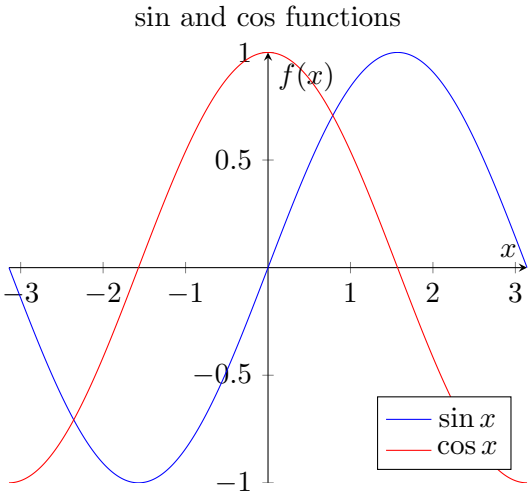
- **Trigonometric Functions**

  Recall that the basic definitions of the trigonometric functions are based on the unit circle



$$\cos\theta = x \quad \sin\theta = y \quad \tan\theta = \frac{\sin\theta}{\cos\theta} = \frac{y}{x}$$

Remember: In mathematics, angles are always radians.



There are three other trigonometric functions, which may be a little less familiar:

$$\sec\theta = \frac{1}{\cos\theta}$$
$$\csc\theta = \frac{1}{\sin\theta}$$
$$\cot\theta = \frac{1}{\tan\theta} = \frac{\cos\theta}{\sin\theta}$$

- **Exponential functions**

  These are functions of the form $f(x) = a^x$, where a is a positive real number.

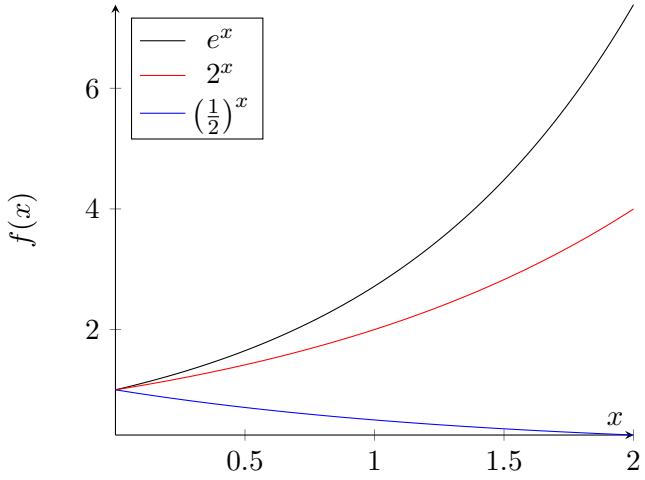
  A particularly important example is known as *the* exponential function, $f(x) = e^x = \exp x$ (where $e \approx 2.72$). This function has a number of properties that make it useful in calculus, as we will see later in the course.

  You should ensure you are familiar with the basic properties of exponential functions, such as:

$$a^0 = 1, \quad (a^x)(a^y) = a^{x+y}, \quad (a^x)^y = a^{xy},$$

  where x and y are any real numbers. You will need to use these properties in assignments and tutorials, and in the final examination.

Examples of exponential functions



- **Logarithmic functions**

  These are closely related to exponential functions. If $x = a^y$ ($a > 0$, $a \neq 1$), then the logarithm to the base a, $\log_a x = y$ (*i.e.* $\log_a x$ is the power to which we must raise a to get x).

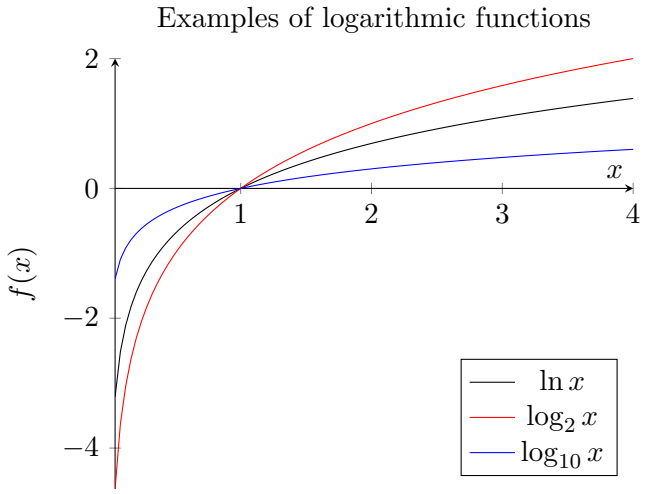
  The logarithm to the base e (≈ 2.72) is called the *natural logarithm* (or *natural log* for short) and has the special notation $\ln x$. Like its counterpart, the exponential function, it also plays an important role in calculus.

  You should ensure you are familiar with the following properties

$$\log_a (xy) = \log_a x + \log_a y, \quad \log_a \left(\frac{x}{y}\right) = \log_a x - \log_a y,$$

$$\log_a x^y = y \log_a x, \quad \log_b x = \frac{\log_a x}{\log_a b},$$

where again x and y can be any real numbers, and $b > 0$, $b \neq 1$.



1.2 Definition of a function

Whilst these basic types of functions (and combinations of them) will come up many times throughout the course (and in future courses), they do not cover every possible type of relationship that we might want to describe. What, then, are the essential features that define a function?

In fact, the mathematical concept of a function is very broad. In earlier work, you have probably become used to seeing functions which can described by a single, straightforward equation (*e.g.* $f(x) = x^2$), and which have graphs that are nice smooth curves. But functions are allowed to have jumps or kinks, can be specified using words or tables rather than formulae, or be described by different formulae for different values of x. The precise mathematical definition of a function below makes this clear.

Definition 1.1 (Definition of a function)**.** Let $\mathcal{D}$ be a set of real numbers. A *function* f with *domain* $\mathcal{D}$ is a rule (or a set of rules) that assigns to each number x in $\mathcal{D}$ a unique real number called $f(x)$.

The *range* of f is the set of all values of $f(x)$, that is, $\mathcal{R} = \{f(x) \mid x \text{ is a number in } \mathcal{D}\}$.

Remarks:

- We will sometimes use the notation $\mathcal{D}(f)$ and $\mathcal{R}(f)$ for the domain and range of f, respectively, particularly if we need to distinguish between the domains (or ranges) of different functions (*e.g.* the domain of g versus the domain of f).
- Note the significance of the phrase '... *assigns to each number* $x \in \mathcal{D}$ *a* **unique** *real number called* $f(x)$.' This means that, for example $f(x) = \pm\sqrt{x}$ $(x \geq 0)$ **does not** satisfy the

definition of a function, because it assigns two values of $f(x)$ to each nonzero value of x. (For $x = 4$, it would give the values of $+2$ and -2 for $f(x)$.)

- The definition does not mean that there are no values x_1, x_2, $(x_1 \neq x_2)$ such that $f(x_1) = f(x_2)$. For example, $f(x) = x^2$ (where x can be any real number), obeys the definition of a function, even though $f(2) = f(-2) = 4$; each value of x is only assigned a single value, $f(x)$.

- Formally, in order to specify a function completely we must give the domain $\mathcal{D}$ and range $\mathcal{R}$. However, for brevity, often we do not write them down explicitly. In these cases we adopt the *convention* that the domain is the largest possible subset of the real numbers for which the function can be defined. The range is then the set of points which are the images under the function of points in the domain.

1.2.1 How to denote the domain and range of a function

The functions we consider in this course will have domains and ranges which are subsets of the real numbers (for brevity, we denote the set of real numbers by $\mathbb{R}$). The most common of these subsets are *intervals*.

We use square brackets to denote intervals which include endpoints, and round parentheses for intervals not including endpoints.

- $(1, 2) = \{x \mid 1 < x < 2\}$ is an *open* interval (excludes end points)

- $[1, 2] = \{x \mid 1 \leq x \leq 2\}$ is a *closed* interval (includes end points).

- We can have mixed combinations such as $(1, 2]$ or $[1, 2)$.

- Whereas $[2, 1)$ is the empty set, denoted by $\emptyset$.

  All of these examples of intervals are *bounded*. This means we can find real numbers m and M such that every number x in the interval satisfies $m \leq x \leq M$. If an interval is not bounded, we say it is *unbounded*.

- $(1, \infty) = \{x \mid 1 < x\}$ is an example of an unbounded interval, since there is no real number M such that $x \leq M$ for every x in the interval.

  Note: '∞' is *not* a real number; it is a symbol to denote that the interval extends forever. We can never write $[a, \infty]$ or $[-\infty, b)$ etc.

- $(-\infty, \infty) = \mathbb{R}$, the set of real numbers.

Sometimes, we may want to exclude certain points from an interval, or join two (or more) intervals together to give the complete domain or range. Intervals are examples of sets, and the following items of set notation can be useful for expressing these kinds of relationships concisely. If A and B are sets (*e.g.* intervals), then:

- $x \in A$ means that the object x is an element of the set A (*e.g.* $x \in [0, 1]$ means $0 \leq x \leq 1$)

- $A \cup B$ means the *union* of the sets A and B: $A \cup B = \{x \mid x \in A \text{ or } x \in B\}$, so $(0, 2] \cup [2, 3) = (0, 3)$. (Note that the word 'or' is inclusive in mathematics, so if $x \in A \cup B$, then x can be in A or B or in both A and B.)

- $A \cap B$ the *intersection* of the sets A and B: $A \cap B = \{x \mid x \in A \text{ and } x \in B\}$, so $[0,3] \cap [1,4] = [1,3]$

- $A \backslash B$ means the *difference* of the sets A and B: $A \backslash B = \{x \mid x \in A \text{ but } x \notin B\}$. For example, $(0,2) \cup (2,3) = (0,3) \backslash \{2\}$.

Example 1.1. If $f(x) = \sin x$, $x \in \mathbb{R}$ (that is, f has domain $\mathbb{R}$) and if $g(x) = \sin x$, $-\frac{\pi}{2} \leq x \leq \frac{\pi}{2}$, then f and g are *different* functions.

Usually we consider $\sin x$, $x \in \mathbb{R}$ (that is, function f), but sometimes we will want to consider $\sin x$ for some (proper) interval in $\mathbb{R}$; after a few lectures we return to this example.

Example 1.2. If we write $f(x) = \sqrt{x+2}$, and do not specify the domain, then implicitly we mean

$$f(x) = \sqrt{x+2}, \quad x \geq -2;$$

that is, the domain is the set $\{x | x \geq -2\}$ - the largest subset of the real numbers for which $f(x) = \sqrt{x+2}$ is defined (as a real number).

1.3 Other examples of functions

Lecture 2

Some important functions which you might not be familiar with include the following.

- The **hyperbolic functions** are a collection of functions defined in terms of the exponential function. The most important are sinh, cosh and tanh which are defined to be:

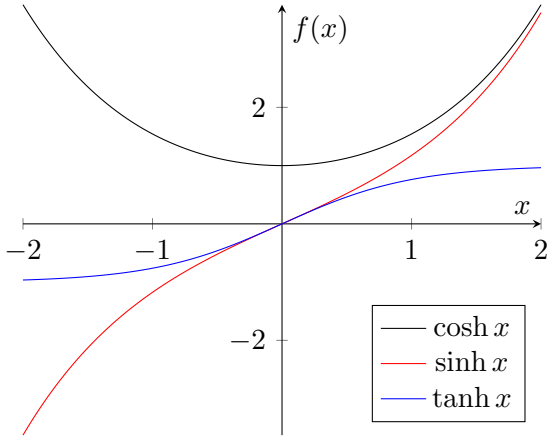
$$\cosh x = \frac{e^x + e^{-x}}{2}, \quad \sinh x = \frac{e^x - e^{-x}}{2}, \quad \tanh x = \frac{\sinh x}{\cosh x}.$$

Derived from these are three further functions

$$\operatorname{sech} x = \frac{1}{\cosh x}, \quad \operatorname{cosech} x = \frac{1}{\sinh x}, \quad \coth x = \frac{1}{\tanh x}.$$

The naming of the hyperbolic functions deliberately echoes that of the trigonometric functions. As we will see during the course, they have a number of similar properties: for example $\cosh^2 x - \sinh^2 x = 1$ (compare the trigonometric identity $\cos^2 x + \sin^2 x = 1$). (In later courses, you will learn how the two types are closely connected through the theory of complex functions.) However, unlike the trigonometric functions, the hyperbolic functions are not periodic. Graphs of $\cosh x$, $\sinh x$ and $\tanh x$ are shown below.

The hyperbolic functions



Many people encounter difficulty with pronouncing the names of some of the hyperbolic functions. Although there are some variations between people and countries (particularly US vs. UK) the following pronunciations are widely used:

$$\cosh - \text{'kosh'}, \quad \sinh - \text{'shine'}, \quad \tanh - \text{'fan'}$$

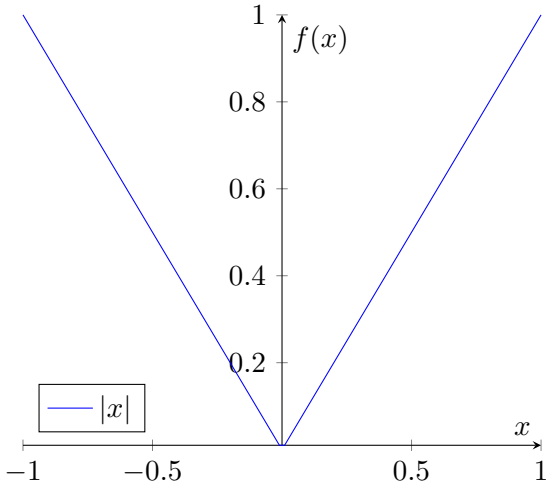
$$\text{sech} - \text{'shek'}, \quad \text{cosech} - \text{'co-shek'}, \quad \coth - \text{'koff'}.$$

- The **absolute value function** (also known as the modulus function), $f(x) = |x|$ is defined by a two part rule:

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

Domain $\mathcal{D} = \mathbb{R}$; the range is $\{x \mid x \geq 0\} = [0, \infty)$.

The absolute value function



- The **Heaviside function**, $H(x)$, which is defined to be

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

  This function is also called the **unit step function**. It is frequently encountered in electronics applications, since, if we let the independent variable be time, t, $H(t)$ represents a signal which switches on at $t = 0$.

- The **Dirichlet function**, $D(x)$, is defined as

$$D(x) = \begin{cases} 1 & \text{if } x \text{ is a rational number}, \\ 0 & \text{otherwise} \end{cases}$$

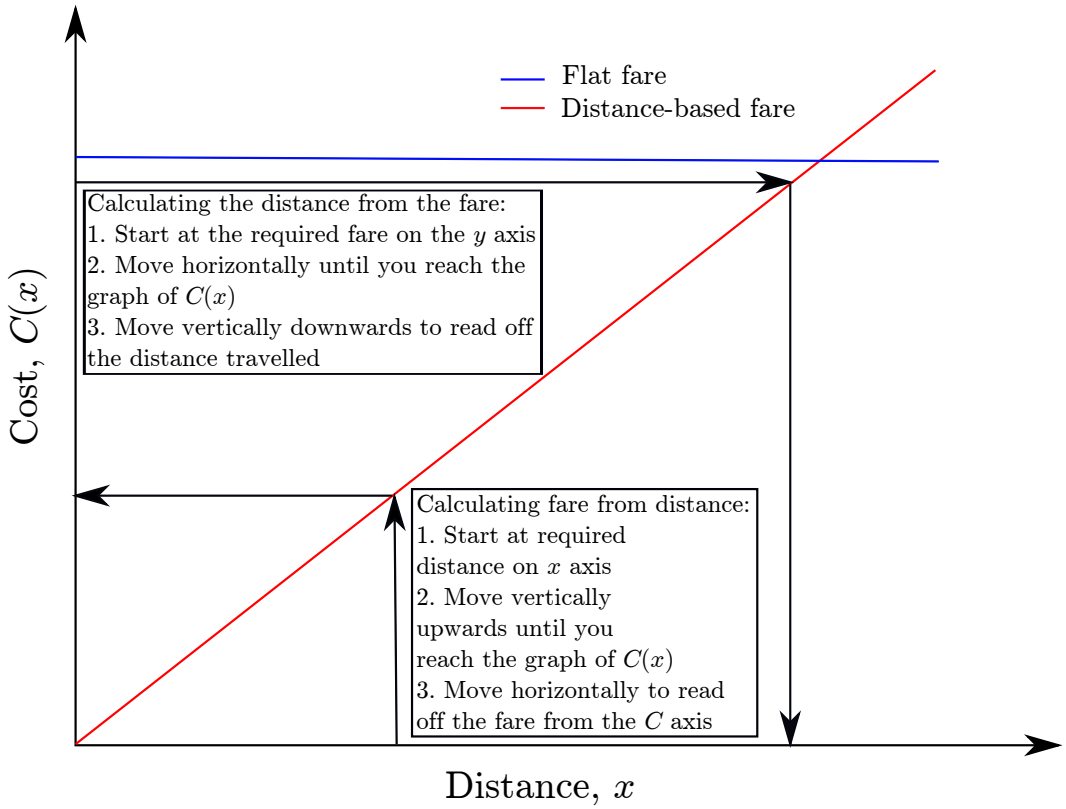
  The Dirichlet function has many strange properties - for example, what would its graph look like? However, it clearly *is* a function, as it satisfies Definition 1.1 above!

1.4 Inverse functions

Consider the problem of pricing a ticket for a bus or train journey. Since many of the costs involved scale with the distance travelled, in the past, it was common for companies to calculate fares using a fixed rate per kilometre (this method is still used in some places). Mathematically, we would say the cost C (in dollars) of your ticket would be a function of the distance x that you would be travelling - *i.e.* $C = f(x)$. For example, if the transport company set a rate of \$0.30 per km, we would have $C = f(x) = 0.3x$ (a linear function).

For this type of pricing scheme, as well as being able to calculate the cost of the ticket when the distance to be travelled is known, we can reverse the process, and find the distance travelled if we know the cost. For example, if a ticket from my home to the city is \$3.60, then the distance between these two points must be 12 km (3.60/0.3). In mathematical terms, what we have done here is calculate the inverse of the function, f, which we denote by f^{-1}. If f tells us the price of the ticket (output) given the distance (input), the inverse f^{-1} reverses the process; it tells us the distance given the cost. We would write $C = f(x) = 0.3x \Leftrightarrow x = f^{-1}(C) = C/0.3 = 10C/3$.

Nowadays, most cities have introduced a more streamlined pricing system, where fares are constant within certain zones, or simply charged at a flat rate. For example, Adelaide Metro charges around \$3.60 irrespective of distance. In this case, the cost of the ticket can still be thought of as a function of distance $C = f(x) = 3.60$ (a constant function). Note that now, although it is straightforward to calculate the fare when the distance to be travelled is known (the result is always \$ 3.60), we can no longer reverse the process to find out the distance travelled from the ticket cost.



Finding the cost given the distance travelled (and vice versa)

When can we find the inverse of a function f? What property of f does it rely on?

If we draw the graphs of the two functions, what is going on becomes clearer. In both cases, given a value of x, we can find the value of C by moving from x vertically upwards to the graph of the function, then moving leftwards to read off the value of C. As both functions obey Definition 1.1, we get one corresponding value of C for each x (though the value of C we get may be the same for many values of x). Now, let's try to reverse the process, beginning with the first pricing scheme (by distance). We start by reading off the cost on the vertical axis, and move rightwards until we hit the line, then move vertically down until we hit the x-axis, where we can read off the corresponding distance. If we try this with the flat fare pricing structure, we cannot determine a value of x (it could be any distance x).

The examples illustrate the property we need to be able to invert a function. Whilst for any function (by definition) there will only be one output value $f(x)$ for any input value x, in order to reverse the process, we need to certain that for every output value $f(x)$ there is only one input value x. If we have two values $x_1 \neq x_2$, with $f(x_1) = f(x_2)$ (which is permitted by the definition of a function), then knowing the value of f does not allow us to determine the value of x; it could be x_1 or x_2. Functions that have the property that each value of $f(x)$ corresponds to only one value of x are called *one-to-one* (1-1) or *injective*. The definition below makes this idea precise.

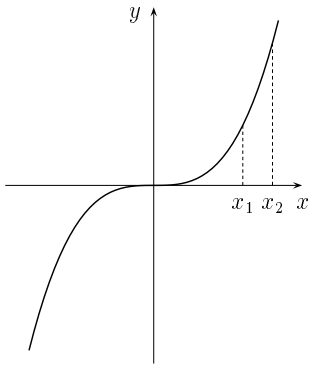
Definition 1.2 (One-to-one functions)**.** A function $f : \mathcal{D} \to \mathcal{R}$ ($\mathcal{D}, \mathcal{R}$, the domain and range of f are subsets of $\mathbb{R}$) is *one-to-one* (*1–1*) if

$$\text{for any } x_1, x_2 \in \mathcal{D}, \text{ if } x_1 \neq x_2 \text{ then } f(x_1) \neq f(x_2).$$

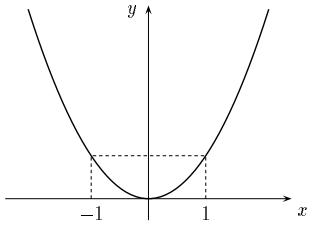
Equivalently, f is 1–1 provided

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

Example 1.3. 1. $f(x) = x^3$ is 1–1

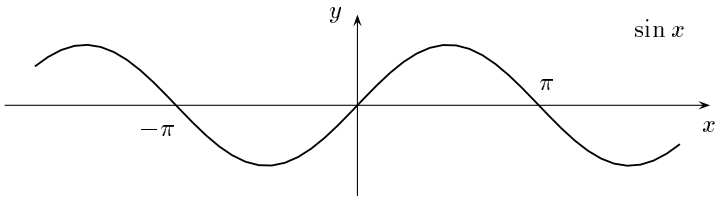


2. $f(x) = x^2$ is *not* 1–1



for example, $f(1) = f(-1) = 1$

3. $f(x) = \sin x$ is not 1–1



as $0 = \sin 0 = \sin \pi = \sin(-\pi) = \cdots$.

Going back to the graphs, we can observe that, if we require every value of $f(x)$ (*i.e.* every y value) to correspond to just one x value, then if we draw **any** horizontal line, it must meet the graph of $f(x)$ at most once. (If there was a least one horizontal line which touched the graph twice

or more, then we would have a y value that corresponds to two or more x values. Then the function could not be one-to one). This gives us a useful way of testing if a particular functions is one-to-one or not.

Horizontal Line Test A function f is 1–1 if and only if any horizontal line meets the graph of f *at most once.*

Increasing, decreasing and monotonic functions Functions which are either *always increasing* or *always decreasing* are called monotonic. Mathematically, we define them as follows:

Definition 1.3 (Monotonic functions)**.** Let $f : \mathcal{D} \to \mathcal{R}$ be a function, and let x_1, x_2 be real numbers in $\mathcal{D}$. Then, we say that f is:

- *increasing* if $f(x_1) < f(x_2)$ for any $x_1 < x_2$;
- *decreasing* if $f(x_1) > f(x_2)$ for any $x_1 < x_2$.

Functions which are increasing or decreasing are called *monotonic.*

If $f(x)$ is monotonic (either always increasing, or always decreasing), any horizontal line can cross the graph of $f(x)$ at most once. (If the graph crosses a horizontal line twice, then there must be $x_1 < x_2$ with $f(x_1) = f(x_2)$, violating monotonicity.) Thus, using the horizontal line test, we can see that monotonic functions are one-to-one.

1.4.1 Calculating the inverse of a function

Lecture 3

We have now established that, provided our function is one-to-one, it will have an inverse. But what exactly do we mean by this? As we have discussed, whilst the function f takes an input value, x, and gives an output $y = f(x)$, the inverse function, f^{-1} takes the output value $y = f(x)$, and tells you the input value x. Thus, if $y = f(x)$, then $f^{-1}(y) = x$. This means $f^{-1}(f(x)) = x$ *i.e.* , applying the function f to x and then applying the inverse function, f^{-1}, to the result takes you back to where you started. Similarly, if $x = f^{-1}(y)$ then $f(x) = f(f^{-1}(y)) = y$. Note that since f takes values in the domain, $\mathcal{D}$ and maps them to values in the range, $\mathcal{R}$, and f^{-1} reverses the process, the domain of f^{-1} is the range of f, and the range of f^{-1} is the domain of f. We now use these intuitive ideas to define precisely what we mean by the inverse of a function.

Definition 1.4 (Inverse functions)**.** Let $f : \mathcal{D} \to \mathcal{R}$ be a one-to-one function with domain $\mathcal{D}$ and range $\mathcal{R}$. Then f has an inverse, $f^{-1} : \mathcal{R} \to \mathcal{D}$, such that

$$f^{-1}(f(x)) = x \quad \text{for all } x \in \mathcal{D}$$
$$f\left(f^{-1}(y)\right) = y \quad \text{for all } y \in \mathcal{R}$$

Now we need to find a method for calculating the inverse. Consider the process we use when reading off the value of $f(x)$ given x from a graph. We find the required value of x on the horizontal axis, move vertically upward until we meet the graph of the functions, and then move horizontally to read off the value of $y = f(x)$. If we want to find x given $y = f(x)$, we do this by starting with the y value, tracing across to the graph of $f(x)$, and then reading off the value of x (this is equivalent to finding $x = f^{-1}(y)$). Note that the only difference is that we have exchanged the roles of the x and y axes. Swapping the x and y axes is equivalent to performing a reflection in the line $y = x$. Hence, if we take the graph $y = f(x)$, and reflect it in the line $y = x$, we end up with the graph of $x = f^{-1}(y)$ (where the y axis is now the horizontal axis, the the x axis is the vertical one). Finally, if we swap the symbols x and y to get a horizontal x axis and vertical y axis, then our graph will be $y = f^{-1}(x)$. This gives us a method by which we can find the inverse function $f^{-1}(x)$ for a given $f(x)$.

Method for finding f^{-1}

1. Check that $f(x)$ is a one-to-one function (only one-to-one functions have an inverse).
2. Write $y = f(x)$.
3. Solve this equation for x to obtain $x = f^{-1}(y)$.

   We now have the functional form for the inverse function, specified in terms of the variable y. There is nothing particularly special about the symbols x and y; we could simply change them to, say, α and β, giving us $\alpha = f^{-1}(\beta)$. However, we often want to have the inverse function specified as a function of x (for example, this can be convenient for plotting purposes). To obtain it in that form, there is one final step.

4. *Interchange* x and y; then $y = f^{-1}(x)$.

Note: from the explanation above it is clear that the graph of $y = f^{-1}(x)$ is obtained by reflecting the graph of $y = f(x)$ in the line $y = x$.

Example 1.4. A spherical balloon is being inflated in such a way that its volume, V, (in m^3) at time, t, is given by the increasing function $V = f(t)$. At what time does it attain a radius of 1m?

Using the formula $V = f(t)$, we can find the volume of the balloon at a given time, t. However, we in fact want to find the time, given the volume. This is given by $t = f^{-1}(V)$ (we can think of this as taking the relationship $V = f(t)$ and applying f^{-1} on both sides, using the fact $f^{-1}(f(t)) = t$). Note the we know f^{-1} exists, since we are told f is increasing. Now, we know that, for a sphere, the relationship between the volume V and radius r is $V = \frac{4}{3}\pi r^3$. Thus, when the radius is 1m, the volume is $\frac{4}{3}\pi$ m^3. So, the time when the balloon attains a radius of 1m is $t = f^{-1}\left(\frac{4}{3}\pi\right)$.

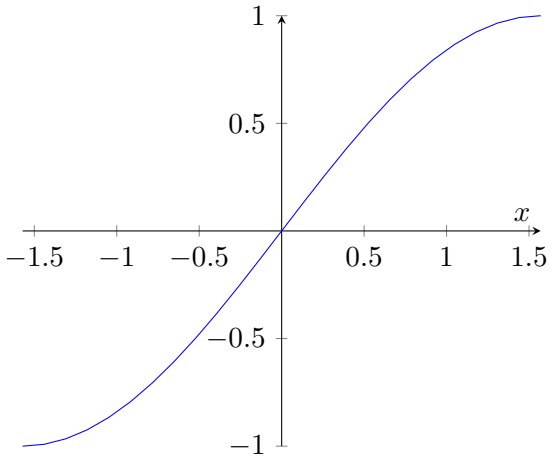
1.5 Inverse trigonometric functions

We have already seen that $\sin x$, $x \in \mathbb{R}$, is not 1–1. However, in your earlier studies in mathematics, you will have solved problems where you needed to find one of the angles in a right-angled triangle,

given the lengths of two sides. In order to do this, you will at some point have used the button on your calculator marked $\sin^{-1}$ or arcsin - which gives you the angle, θ, when $\sin\theta$ is known. Hence, this button calculates the inverse function of $\sin\theta$. How is this possible?

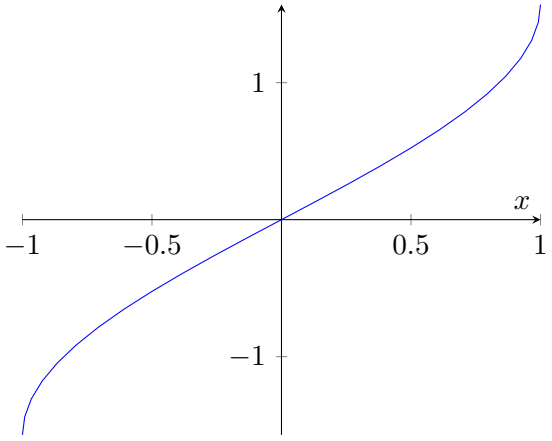
Recall that we said that, formally, $f(x) = \sin x,\ x \in \mathbb{R}$ and $g(x) = \sin x,\ x \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ are *different* functions, since their domains are different. If we look at the plot of $g(x)$ below, we can see that it is increasing on $[\frac{-\pi}{2}, \frac{\pi}{2}]$ and so it is one-to-one. Hence, it has an inverse function, $g^{-1}(x)$; this inverse function is written as $\arcsin x$, or $\sin^{-1} x$. Essentially, in order to be able to find the inverse of the sin function, we 'restrict the domain' to make the function one-to-one. The choice to restrict the domain to $[-\frac{\pi}{2}, \frac{\pi}{2}]$ is conventional, but arbitrary; any other domain on which the sin function is one-to-one would be equally valid.

$$g(x) = \sin x, \quad -\tfrac{\pi}{2} \le x \le \tfrac{\pi}{2}$$



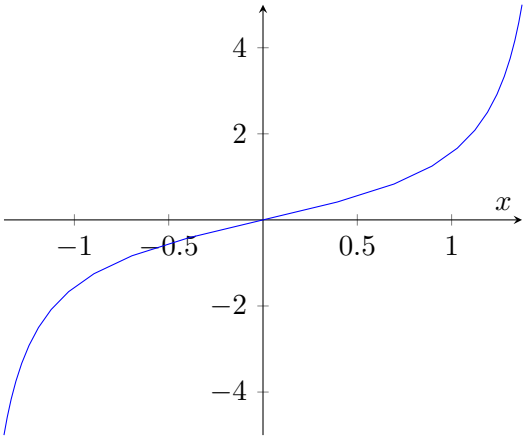
As the range of g, $\mathcal{R}(g) = [-1, 1]$, the domain $\mathcal{D}(\arcsin) = [-1, 1]$. Also the domain of g is $\mathcal{D}(g) = [-\frac{\pi}{2}, \frac{\pi}{2}]$, so the range $\mathcal{R}(\arcsin) = [-\frac{\pi}{2}, \frac{\pi}{2}]$.

$$g^{-1}(x) = \arcsin x = \sin^{-1} x, \quad -1 \le x \le 1$$



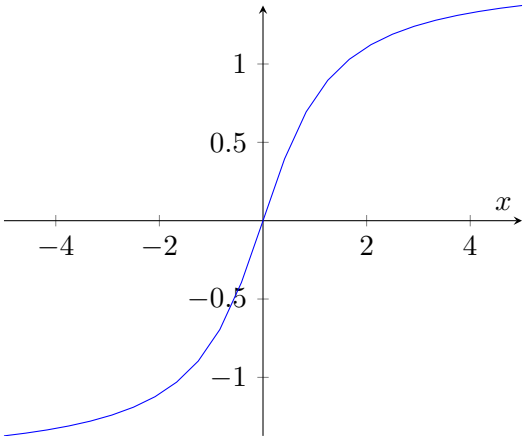
Inverse tangent We similarly restrict the domain of tan so we can define the inverse tangent function arctan or $\tan^{-1}$. Let $g(x) = \tan x,\ -\frac{\pi}{2} < x < \frac{\pi}{2}$. As we can see from the graph below, on the interval $(-\frac{\pi}{2}, \frac{\pi}{2})$ $\tan x$ is increasing and therefore 1–1.

$$g(x) = \tan x\,, \quad -\tfrac{\pi}{2} < x < \tfrac{\pi}{2}$$



The inverse tangent function is

$$g^{-1}(x) = \arctan x \; = \tan^{-1}(x)$$

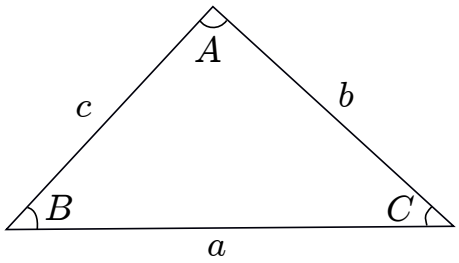


Domain of arctan is $\mathbb{R} = (-\infty, \infty)$ as $\mathcal{R}(\tan) = \mathbb{R}$

Range of arctan is $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ as $\mathcal{D}(\tan) = \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$

Example 1.5. Although we can define inverse sin, cos and tan functions by restricting their domains in this way, we need to be aware that we have done this when we are using them to solve problems. Recall the Sine Rule, which relates the angles and lengths of the sides of any triangle (not necessarily right-angled):

$$\frac{\sin A}{a} = \frac{\sin B}{b} = \frac{\sin C}{c}.$$



Definition sketch for the Sine Rule and Cosine Rule

Consider a triangle such that $a = 11$ cm, $b = 15.5$ cm and the angle $A = 0.75$. What is the angle B?

Rearranging the Sine Rule equation, we have

$$\sin B = \frac{b \sin A}{a} = \frac{15.5 \sin 0.75}{11} \approx 0.961$$

Now, using the definition of the inverse sine function from earlier, we have $\sin^{-1}(0.961) = 1.29$. Does this mean $B = 1.29$ is the only solution?

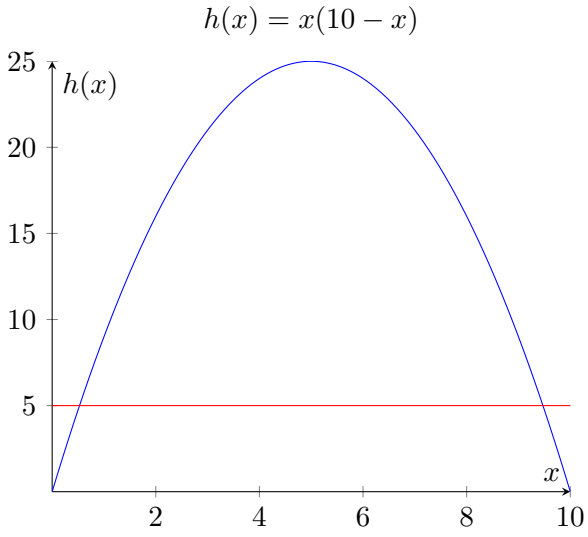
In fact, it **does not**. Recall that, for any angle, θ, $\sin\theta = \sin\pi - \theta$. Hence, we need to check whether $\pi - 1.29 \approx 1.85$ is also a solution for B. Since the angles in a triangle must add up to π and $0.75 + 1.85 = 2.60 < \pi$, this is indeed another possible solution.

Question: The Cosine Rule states that for *any* triangle (not necessarily right-angled)

$$a^2 = b^2 + c^2 - 2bc\cos A$$

(where we have used the same notation as in the Sine Rule above). Can we ever have two possible solutions if we use the Cosine Rule to calculate an angle? If not, why not?

1.6 Zeros of functions



In practical applications, we frequently need to be able to find the value of x, such that $f(x)$ takes a particular value. For example, let the height above ground of a ball (in metres) when it is a horizontal distance x from the point where it was thrown be given by $h(x) = x(10-x)$. How far has the ball travelled horizontally when it is five metres above the ground? Mathematically, we need to find the values of x which satisfy the equation $x(10-x) = 5$. These values occur where the blue curve meets the red line in the graph above. If we define a new function $f(x) = h(x) - 5 = x(10-x) - 5$, our problem reduces to finding the values of x satisfying $f(x) = 0$. These values of x are called the *zeros* of the function, $f(x)$.

We can re-write $f(x) = 0$ as

$$f(x) = -x^2 + 10x - 5 = 0.$$

This is a quadratic equation, which we can easily solve using the quadratic formula. The zeros of $f(x)$ are

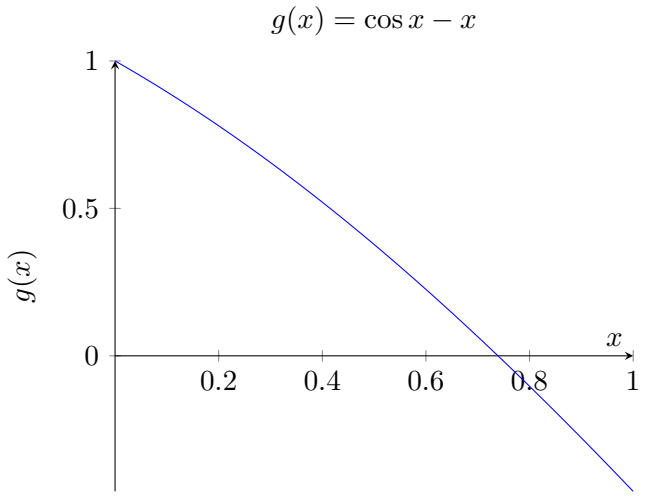
$$x = \frac{-10 \pm \sqrt{100 - 20}}{-2} = 5 \pm \sqrt{20} = 5 \pm 2\sqrt{5}.$$

The two values are approximately $x_1 = 0.53$ and $x_2 = 9.5$.

As this example illustrates, many problems reduce to finding the zeros of a particular function.

Now consider the general problem of finding the zeros of a function, $f(x)$. If we can write down the inverse function f^{-1}, this is very straightforward, since if $f(x) = 0$, then $x = f^{-1}(0)$. However, there are many situations, including the example of the thrown ball above, where we want to find the zeros of a function which is not one-to-one and hence has no inverse. In other cases, it may not be possible to write down the inverse in an explicit form. Then what should we do?

In the case of the thrown ball, although the function is not one-to-one, the problem is still straightforward, because we can use the quadratic formula to calculate the zeros. But for most functions that you could think of, no such simple formula would exist. Suppose we want to find solutions of $\cos x = x$. This is equivalent to finding the zeros of $g(x) = \cos x - x$. Although this function is one-to-one on $x \in [0, 1]$ (as can be seen from the graph of $g(x)$), we cannot write down an expression for the inverse function explicitly. However, the graph does suggest we might proceed by noting that $g(0.8) \approx -0.103 < 0$ and $g(0.7) \approx 0.065 > 0$. Since the value of the function changes sign between $x = 0.7$ and $x = 0.8$, we could reason that a solution will lie somewhere in the interval (0.7, 0.8). What property of the function are we relying on here when make that deduction? Can you think of a function where $f(a) < 0$ and $f(b) > 0$, but there is no value of $x \in (a, b)$ where $f(x) = 0$?



Consider the following function:

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

This f obeys the definition of a function we gave earlier, and $f(-1) < 0$ and $f(1) > 0$, but since $f(x)$ is nowhere equal to zero, there cannot be a solution of $f(x) = 0$ for $x \in (-1, 1)$. How has the reasoning that worked for $g(x) = \cos x - x$ gone wrong here?

The answer is obviously because the function $f(x)$ 'jumps' from the value -1 to $+1$ 'skipping out' zero, whereas $g(x)$ goes through every value between -0.103 and 0.065. The graph of $g(x)$ is an unbroken line, whilst that for $f(x)$ has a 'gap' near $x = 0$. Clearly, our method of tracking down the zero by finding places where the function changes sign will only work for functions like $g(x)$ where the graph is an unbroken curve, which excludes some functions.

The property that a function must have in order for our zero-finding method to work is called *continuity*. Roughly speaking, it means that we can draw the graph of the function without ever having to take our pen off the paper. However, in order to define more precisely what we mean by continuity, we must first revisit the idea of a *limit*, which you have probably already come across in connection with the definition of a derivative (a topic we will return to in Chapter 2).

1.7 Limits

Lecture 4

In order to define continuity, we first need to think carefully about what happens to the value of $f(x)$ as x approaches some value of interest. We intuitively define the limit of $f(x)$ as x approaches c as follows.

Definition 1.5 (Limits)**.** Let f be a function defined on a domain that includes all values close to c, but need not include c itself. For example, the domain could be a set of the form $(a, c) \cup (c, b) = (a, b) \setminus \{c\}$ where $a < c < b$. We say that the limit of $f(x)$ as x approaches c is

the real number L if the values of $f(x)$ get closer and closer to L as x gets closer and closer to c (with $x \neq c$). We write

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

Example 1.6. Consider $f(x) = x^2 - 3$ and $c = -2$.
Intuitively, as x gets closer and closer to -2, $f(x)$ gets closer and closer to $f(-2) = (-2)^2 - 3 = 1$. Hence $L = 1$.

The example above is straightforward; in this case, $c = -2$ is actually in the domain of f, and the value of $f(x)$ gets closer and closer to $f(c)$ as $x \to c$.

It is important realise that not all examples of limits are like the preceding example where we can substitute for the value of $f(x)$ at c. In fact, the number $L = \lim_{x \to c} f(x)$ depends on the values of $f(x)$ for x close to c but not at $x = c$. The definition makes this clear, since the function f does not even need to be defined at c. However, if f *is* defined at c, we can arbitrarily redefine its value at c without affecting L.

Example 1.7. Consider

$$f(x) = \begin{cases} x^2 - 3 & \text{if } x > -2, \\ -7 & \text{if } x = -2 \\ x^2 - 3 & \text{if } x < -2. \end{cases}$$

and $c = -2$.

Now, suppose we let x get closer and closer to -2, but making sure we always choose values such that $x > -2$. Then $f(x)$ approaches 1. If we do the same thing, but this time always choosing values of $x < -2$, $f(x)$ again approaches 1. Thus, once again, $L = 1$ – despite the fact that $f(-2) = -7$.

In the next example, we illustrate that the limit L can exist even if $f(c)$ is undefined.

Example 1.8. Consider

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

which is not defined at $x = 0$. However if we compute some values we find:

x	$f(x)$
0.1	0.9983341665
0.01	0.9999833334
0.001	0.9999998333
0.0001	0.9999999983

So we should not be surprised to find that $\lim_{x \to 0} \sin x/x = 1$.

It is important to note that, in determining $\lim_{x\to c} f(x)$, we must consider values on both sides of c. Although we did not explicitly check negative values of x in our table above, if we were to do so, we would get the same result, $L = 1$, since for this function, $f(x) = f(-x)$.

Thus far we have seen examples where the limit exists. However limits often do not exist. For example consider

$$f(x) = \sin\frac{1}{x}.$$

As x approaches 0 this oscillates wildly between -1 and 1 and is definitely not getting closer and closer to any particular value L.

Similarly, if we consider the Heaviside function,

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

with $c = 0$, then we can see that if we always take negative values of x, as x gets closer and closer to zero, $H(x) = 0$. However, if we let x take only positive values, then, as we get closer and closer to zero, $H(x) = 1$. Here, the limit again does not exist, since we cannot get a consistent value for L when we consider x values on both sides of zero. An alternative way to show this is to think about what happens to $H(x_n)$ for $x_n = \left(\frac{-1}{2}\right)^n$ for $n = 1, 2, 3 \ldots$. Although x_n gets closer and closer to zero as n increases, the value of the function oscillates between the values of 0 and 1; it does not get any closer to a fixed value, L.

The definition of limit we have given above is 'intuitive' in that it suffers from a number of ambiguities due to the lack of precision of the everyday language employed. For example what does 'closer and closer to' mean ? It could be read as meaning that the values of $f(x)$ are monotonically approaching L as x approaches c. This is definitely not the case as we can see by considering the example

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

Like the preceding example this also oscillates up and down but the amplitude of the oscillation decreases as we approach 0 because of the x factor. It has limit $L = 0$ as x approaches 0.

In order to avoid these kinds of issues, which result from the imprecision of everyday language, mathematicians use a more formal definition of limits. Many of you will meet this definition in future, when taking more advanced mathematics courses. However, for the purposes of this course, the intuitive definition we have presented above will suffice.

Later in the course, we will need to be able to calculate limits. Hence, in the next two sections, we present some basic facts about them, together with some useful properties that are known as the *limit laws*.

1.7.1 Basic properties of limits

Property 1.1 (Limits are unique)**.** If $\lim_{x\to c} f(x)$ exists then it is unique.

That is, if $\lim_{x\to c} f(x) = L$ and $\lim_{x\to c} f(x) = M$ then $L = M$.

Note that this makes sense intuitively, since if L and M are not equal they must be some distance apart and if $f(x)$ is very close to one of them it cannot also be very close to the other.

Property 1.2. The following statements of limits are equivalent:

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

(ii) $\lim_{x\to c}(f(x) - L) = 0$

(iii) $\lim_{x\to c} |f(x) - L| = 0$

(iv) $\lim_{h\to 0} f(c+h) = L$ (follows from putting $x = c + h$ in (i))

1.7.2 The limit laws

It would clearly be cumbersome if we had to start from the definition every time we wanted to calculate a limit. (For example, finding $\lim_{x\to 0}(x^{34} + 3x^7 + 2) = 2$ would be horrendous!) Instead, we can rely on the following rules, which make things much more straightforward in practice.

Property 1.3 (Elementary limits)**.** (i) $\lim_{x\to c} x = c$; (ii) $\lim_{x\to c} K = K$, if K is a constant.

Again these are intuitive: if x gets closer and closer to c then x gets closer and closer to c! Also, if K is a constant then as x gets close to c, K does not change and so can only be close to itself.

To calculate results such as $\lim_{x\to 0}(x^{34} + 3x^7 + 2) = 2$ we use these elementary limits and the limit laws.

Property 1.4 (Limit Laws)**.** Let f and g be functions both defined in $(a, b)\backslash\{c\}$ (a neighbourhood of c, not including c itself) and suppose that $\lim_{x\to c} f(x) = L$ and $\lim_{x\to c} g(x) = M$.

Then for any real numbers α and β we have

(i) $\lim_{x\to c}\{\alpha f(x) + \beta g(x)\} = \alpha L + \beta M$

(ii) $\lim_{x\to c}\{f(x)g(x)\} = LM$

(iii) $\lim_{x\to c} \dfrac{f(x)}{g(x)} = \begin{cases} \dfrac{L}{M} & \text{if } M \neq 0 \\ \text{does not exist} & \text{if } M = 0, L \neq 0 \\ \text{may or may not exist} & \text{if } M = 0, L = 0 \end{cases}$

(iv) Composition of functions: if $\lim_{x\to c} g(x) = M$ and $\lim_{x\to M} f(x) = f(M)$ then, if f is defined in an open interval containing M, $\lim_{x\to c} f(g(x)) = f(M)$.

Example 1.9.

$$\begin{aligned}\lim_{x\to c} x^2 &= \lim_{x\to c} x . \lim_{x\to c} x \quad \text{using limit law for products}\\ &= c \times c\\ &= c^2\end{aligned}$$

Similarly $\lim_{x\to c} x^n = c^n$ for any integer n.

More generally if $P(x)$ is any polynomial, *i.e.*

$$P(x) = a_0 + a_1 x + \cdots + a_n x^n,$$

then

$$\lim_{x\to c} P(x) = P(c).$$

Example 1.10. Calculate $\lim_{x\to 0}(x^{34} + 3x^7 + 2)$

$$\begin{aligned}\lim_{x\to 0}(x^{34} + 3x^7 + 2) &= \lim_{x\to 0} x^{34} + 3(\lim_{x\to 0} x^7) + \lim_{x\to 0} 2\\ &= 0 + 3.0 + 2\\ &= 2\end{aligned}$$

Similarly for any rational function $r(x) = \dfrac{P(x)}{Q(x)}$ where $P(x)$ and $Q(x)$ are polynomials, we can use the limit laws to show that

$$\lim_{x\to c} r(x) = \frac{P(c)}{Q(c)} \quad \text{provided } Q(c) \neq 0.$$

Example 1.11.

$$\lim_{x\to 2} \frac{3x^2 + 5x + 8}{x^3 + 7} = \frac{3.2^2 + 5.2 + 8}{2^3 + 7} = \frac{12 + 10 + 8}{8 + 7} = \frac{30}{15} = 2$$

Example 1.12.

$$\begin{aligned}\lim_{x\to 1}(x^2 + 3)^6 &= \Big[\lim_{x\to 1}(x^2 + 3)\Big]^6 \quad \text{(using limit law for composition of functions)}\\ &= [4]^6 = 4096.\end{aligned}$$

Sometimes a limit does not exist because as x approaches c, $f(x)$ becomes arbitrarily large so that no matter what number L you choose $f(x)$ will not get closer and closer to it. If $f(x)$ is always positive as $x \to c$, we denote this limit by ∞ but be aware that ∞ is not a number and writing

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

is just a shorthand way of saying that $f(x)$ gets larger and larger as x gets closer and closer to c.

Similarly, if $f(x)$ is negative, but can become arbitrarily large in magnitude as $x \to c$, then we would write

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

Example of unbounded limit

Based on our discussion above, we can see that, for example,

$$\lim_{x\to 0} \frac{1}{x^2} = \infty, \qquad \lim_{x\to 1} -\frac{1}{(x-1)^4} = -\infty.$$

However, it is not true that $\lim_{x\to 0} \frac{1}{x} = \infty$ since it is large and negative for negative values of x close to zero and large and positive for positive values of x close to zero (on the other hand $\lim_{x\to 0} \left|\frac{1}{x}\right| = \infty$.)

Recall also $\lim_{x\to 0} \sin\frac{1}{x}$ which does not exist even though it is not unbounded (the values always lie between 1 and -1).

1.8 Continuous functions

Lecture 5

Now that we have revisited the idea of a limit, and discussed how they can be calculated, we are in a position to move forward with our definition of continuity. Recall, that this is the property we will exploit in order to track down the zeros of functions.

In discussing $\lim_{x\to c} f(x) = L$ we noted that the value of L did not depend on the value of $f(c)$, and in fact, f did not even have to be defined at c for L to exist. However, for our first example, we considered the function $f(x) = x^2 - 3$ and $c = -2$, and saw that $\lim_{x\to c} f(x) = f(c)$. Functions that have this property are precisely those that we call continuous, as the following definition makes clear.

Definition 1.6 (Continuity at a point)**.** Let $f(x)$ be defined on (a, b) and let c be a point in (a, b). We say that f is *continuous at* c if $\lim_{x\to c} f(x) = f(c)$.

Note that there are three essential requirements for f to be continuous at c:

1. the function f is defined at c ($c \in \mathcal{D}(f)$),

2. the limit $\lim_{x \to c} f(x)$ exists, and

3. the value of the limit is $f(c)$.

If the function is defined at c but either of conditions (2) and (3) are not met, we say that f has a *discontinuity* at c, or equivalently, that f is *discontinuous* at c.

Definition 1.7 (Continuity on an interval)**.** We say that f is continuous on the open interval (a, b) if f is continuous at every point of (a, b).

From this definition, we note that polynomial functions, $|x|$, $\cos x$ and $\sin x$ are all continuous on $(-\infty, \infty)$. Rational functions $f(x) = p(x)/q(x)$ are not defined at points where $q(x) = 0$; hence they cannot be continuous at such points. For example x^{-1} is undefined at $x = 0$. Similarly, $\tan x$ is undefined at $x = \pm\frac{(2n+1)\pi}{2}$ (where $n = 0, 1, 2, \ldots$).

The Heaviside function is discontinuous at $x = 0$, since, as we saw earlier, the limit $\lim_{x \to 0} H(x)$ does not exist. However, it is continuous at all other points. The Dirichlet function $D(x)$ is discontinuous everywhere, despite being bounded and defined for all real numbers. We can see this is true if we consider what happens to $D(x)$ as $x \to c$ for both rational and irrational values of c. If c is irrational, $D(c) = 0$, but there are rational values of x arbitrarily close to c, so $\lim_{x \to c} D(x) \neq D(c) = 0$. Similarly, if c is rational, there are irrational values of x arbitrarily close to it, and so $\lim_{x \to c} D(x) \neq D(c) = 1$.

The following properties of continuous functions are often useful in applications where we need to demonstrate that a function is continuous, and avoid the need to return to the definition of continuity every time. They follow from the Limit Laws (Property 1.4).

Property 1.5 (Properties of continuous functions)**.** If f and g are continuous at c, and α and β are any real numbers, then

(i) $\alpha f(x) + \beta g(x)$ is continuous at c.

(ii) $f(x)g(x)$ is continuous at c.

(iii) $\dfrac{f(x)}{g(x)}$ is continuous at c, provided $g(c) \neq 0$.

(iv) If g is continuous at c and f is continuous at $g(c)$ then $f(g(x))$ is continuous at c.

Remarks on applications of continuous functions

Most of the functions you will have met previously, and indeed, most that you will meet in this course, are continuous. This is perhaps not surprising, as continuous functions arise naturally in many applications: the distance travelled by a vehicle, its velocity or its acceleration, for example,

would all usually be continuous functions of time. Less familiar examples would include the variation in gravitational field strength with the distance from a planet, or the variation in the strength of the electric field with distance from a charged body. However, you must not let this blind you to the fact that there are many applications which give rise to functions with discontinuities: for example, the change over time of the population of a city or the amount of money in a bank account. Since the values of these functions must change by multiples of a fixed amount (one person for the population, or one cent for the bank account), we can expect their graphs to have 'jumps' when they are plotted as functions of time.

In a large number of cases, whether a relationship is represented by a continuous or discontinuous function will depend on the situation we are interested in. Consider, for example, volumes of air or water, which are composed of large (but finite) numbers of individual molecules. If we think about the mass of water in a container as a function of time, at a microscopic level it must change only in discrete 'jumps', equal to the mass of a single water molecule. However, in most everyday applications, like filling up a fish tank, these jumps in mass, and the time intervals between them, are so small that it makes more sense to treat the mass of water as a continuous function of time. Conversely, we may need to approximate a continuous function by a discontinuous one: for example, many computer programmes approximate continuous functions by constant values over very short intervals.

1.9 The interval bisection method

We can find the zeros of a *continuous function* $f(x)$ defined on an interval $[a, b]$ to any degree of accuracy required by using the *interval bisection method*. The method relies on the fact that if the function is continuous, then if we sketch the graph of $f(x)$, we must be able to connect $f(a)$ and $f(b)$ without taking our pen off the paper. Hence, we have the following property:

Property 1.6 (The Intermediate Value Theorem)**.** Let $f(x)$ be a continuous function defined on an interval $[a, b]$. Then $f(x)$ takes every value between $f(a)$ and $f(b)$ at least once.

A formal proof of this property is beyond the scope of this course.

An obvious consequence of this result is that if we have a continuous function, f and we know two values of x, a and b (with $a < b$) in the domain of f, such that $\text{sign} f(a) \neq \text{sign} f(b)$, then a zero of f must be somewhere in the interval (a, b). We progressively narrow down the interval by considering the sign of $f(m)$ where $m = (a + b)/2$ is the midpoint of the interval. If $f(m)$ has the same sign as $f(a)$ then a zero must lie in the interval (m, b); conversely, if $f(m)$ has the same sign as $f(b)$, then a root lies in (a, m). We continue this process until we have narrowed the interval sufficiently to achieve the required accuracy.

We demonstrate the method using the function $f(x) = \cos x - x$ which we discussed earlier. Suppose we want to find the zero of this function to an accuracy of 0.01. To begin the procedure, we need to find the two values for a and b. Often this can be done by looking at a graph of the relevant function. Earlier, we saw that $a = 0.7$ and $b = 0.8$ are suitable values, since $f(0.7) > 0$ and $f(0.8) < 0$. The steps are now as follows:

1. We calculate the mid-point of the interval (0.7, 0.8), $m = (0.7 + 0.8)/2 = 0.75$.

2. We find $f(m) = f(0.75) = -0.018 < 0$. This has the same sign as $f(0.8)$. Hence, the root must lie in the smaller interval (0.7, 0.75).

3. Now we go through the procedure again, calculating the mid-point of the new interval, $m = (0.7 + 0.75)/2 = 0.725$.

4. Since $f(0.725) = 0.023 > 0$, we now know our root must line in the even-smaller interval (0.725, 0.75).

5. The mid-point of this new interval is $m = 0.7375$, and $f(0.7375) = 0.0027 > 0$. Hence, we now know the zero lies in the interval (0.7375, 0.75).

6. The midpoint of this interval $m = 0.74375$ and $f(0.74375) = -0.0078 < 0$. Thus, our zero lies in the interval (0.7375, 0.74375).

7. Since the lower and upper end points of the interval now differ by less that the required accuracy (0.01), we can say that the zero of f occurs at $x \approx 0.74$ (to two decimal places).

1.10 Summary of learning outcomes

Now that we have reached the end of this chapter, you should be able to:

- Define what is meant by a function
- Specify the *domain* and *range* of a function using interval notation and simple set notation
- Determine if a function is one-to-one
- Define the inverse of a function, and determine if a given function has an inverse
- Calculate the inverse $f^{-1}(x)$, given an invertible function $f(x)$
- Define the inverse trigonometric functions by appropriately restricting the domain of the function
- Explain intuitively what is meant by a limit
- Compute limiting values of functions, where they exist
- Define what it means for a function f to be continuous at a point, c.
- Use the interval bisection method to find zeros of functions which are continuous on an interval.

Chapter 2

Differentiation and its applications

Lecture 6

In the previous chapter, we have revisited ideas about functions, and defined them in precise mathematical terms. In practical applications, the function giving the relationship between the variables we are interested in encapsulates in a compact form a great deal of information about the situation. However, this information may not always be in the form that is most useful to us. For example, we might know the position of an aeroplane as a function of time, but want to find out how its speed depends on time. Understanding relationships involving rates of change is the goal of differential calculus.

2.1 The derivative

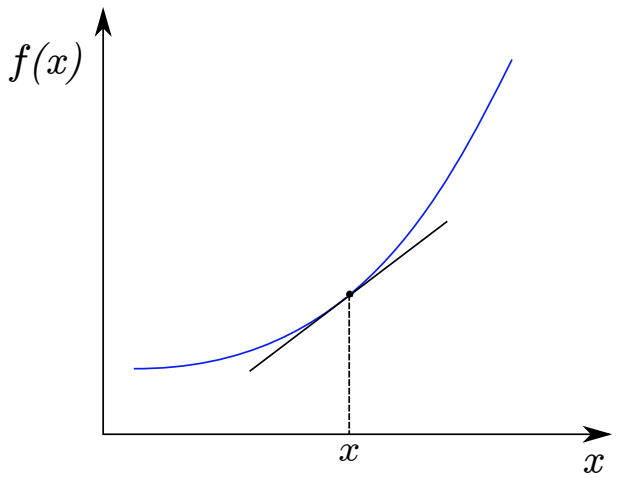
From your previous studies, you will already know that the rate at which a function $f(x)$ changes with x is called the **derivative of** f. Common notations for the derivative include

$$\frac{df}{dx} = f'(x).$$

For example, velocity v is the rate of change of position x with time. Hence, if we know a vehicle is moving such that its distance (in m) from some start point at time t (in s) is given by $x = 15t$, then

$$v = \frac{dx}{dt} = x'(t) = 15\text{ms}^{-1}.$$

Note that when the independent variable represents time, as here, you will sometimes see the notation $\dot{x}(t)$ for the derivative of x.



The derivative $f'(x)$ gives the gradient of the tangent to the graph of f at x.

Geometrically, the derivative of a function $f(x)$ at a point x is the gradient of the tangent to the graph of the function $f(x)$ at x (or, equivalently, the slope of the graph of $f(x)$ at x).

You already know how to calculate the derivatives of a variety of different functions. For example:

$$\begin{aligned} \tfrac{dx^r}{dx} &= rx^{r-1} \\ \tfrac{d}{dx}\sin x &= \cos x \\ \tfrac{d}{dx}\cos x &= -\sin x \\ \tfrac{d}{dx}e^x &= e^x \end{aligned}$$

Question: In theory, can I always calculate the derivative $\dfrac{df}{dx}$ if I know $f(x)$?

Consider the Heaviside function, $H(x)$. What is the derivative of $H(x)$ at $x = 0$? In fact, it is undefined. We note that, at $x = 0$, the function jumps by one unit in the y direction. Calculating the slope (= change in y / change in x) would involve a division by zero; hence the slope (or derivative) is not defined there. We have already seen that the function is not continuous at $x = 0$, so perhaps it is not so surprising that we cannot differentiate it there. But what about the absolute value function $|x|$ - can we calculate its derivative at $x = 0$? Note that $|x|$ is continuous at $x = 0$, as its graph is an unbroken curve. However, if we try to calculate the slope of the graph at $x = 0$ using our usual methods, we might get one of three different answers:

$$\frac{|0.01| - |0|}{0.01} = 1, \qquad \frac{|0.01| - |-0.01|}{0.02} = 0, \qquad \frac{|0| - |-0.01|}{0.01} = -1.$$

This is not acceptable; we require that the definition of the derivative should give us a well-defined, unique answer. Hence the derivative of $|x|$ is undefined at $x = 0$. This means that being *differentiable* is a stronger condition on a function than merely being continuous. All differentiable functions are continuous, but not all continuous functions are differentiable.

We can see that the problem with $|x|$ at $x = 0$ is due to the 'kink' in the graph of the function there. If we look at the slope of $|x|$, we see it is -1 for $x < 0$ and 1 for $x > 0$. Hence, the value

of the slope jumps by two at $x = 0$. Informally, we can think of a function being differentiable at x as meaning that the function is continuous at x, and does not have a 'kink' or 'twist' (where the slope suddenly changes) there.

2.1.1 Definition of the derivative

Definition 2.1 (The derivative)**.** The derivative of f at x is defined as

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

if this limit exists. If the limit exists, then we say the function f is *differentiable* at x.

The limit laws tell us that limits are unique. Thus, if f is differentiable at x, then the derivative is unique. Note that the derivative $f'(x)$ is itself a function of the variable x.

Above, we have demonstrated that not all continuous functions are differentiable. However, if a function *is* differentiable, then it is also continuous. This fact is straightforward to demonstrate using the limit laws. If f is differentiable at x, then setting $y = x + h$, we know that the following limit exists

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

Now, $\lim_{y \to x} y - x = 0$, and then, using the limit law for a product we know the following

$$0 = 0 \times f'(x) = \lim_{y \to x}(y - x) \lim_{y \to x} \frac{f(y) - f(x)}{y - x} = \lim_{y \to x} \frac{f(y) - f(x)}{y - x}(y - x) = \lim_{y \to x} f(y) - f(x).$$

Thus $\lim_{y \to x} f(y) = f(x)$, which tells us that f is continuous at x.

2.1.2 Interpreting the meaning of the derivative

In defining the derivative, we considered the graph $y = f(x)$. Then, $\frac{dy}{dx} = f'(x)$ is the steepness of the graph at the point x (*i.e.* it tells us how much y would increase if x increased by one unit). If $f'(x) > 0$, then y is increasing with x; if $f'(x) < 0$, y decreases as x increases. In applications, the meaning of the derivative will depend on what the quantities y and x represent (and, of course, we may use different symbols from y and x where this is more natural for the problem being considered). For example, suppose $f(t)$ is the height (in metres) of a ball above ground level at time t (measured in seconds). Then we can interpret the derivative as the rate of change of height, *i.e.* the vertical velocity of the ball. We note that the definition of the derivative involves a fraction, where the function f appears in the numerator, and an increment of the independent variable (in this case, t) appears in the denominator. Thus

$$\text{Units of } \frac{df}{dt} = \frac{\text{Units of } f}{\text{Units of } t}.$$

In our case, the units of f are m, and the units of t are s; so the units of $\frac{df}{dt}$ are ms^{-1}. This makes sense, as these are units of velocity.

Example 2.1 (Housebuilding costs)**.** Suppose the cost, C, (in dollars) of building a house of area, A m^2, is given by $C = f(A)$. What is the practical interpretation of $\frac{dC}{dA} = f'(A)$? What are its units?

$\frac{dC}{dA}$ is a cost divided by an area, so its units are dollars per m^2. Recalling the definition of the derivative, we can think of it as the change in cost, ΔC, when there is a small increase, ΔA in the area of the house, so $\frac{dC}{dA}$ is the additional cost per square metre. Hence, if you are planning to build a house of roughly A m^2, then $f'(A)$ is the approximate cost per m^2 of the additional expense involved in building a slightly larger house. It is called the *marginal cost*. We would expect the marginal cost to be lower than the average cost per square metre, since once you have already hired the necessary equipment, builders, *etc.*to build a fairly large house, the cost of adding a little extra space is likely to be small.

2.1.3 Concavity and the second derivative

We have defined the derivative of a function $f(x)$, and we note that the derivative $\frac{df}{dx} = f'(x)$ is also a function of x. Provided that $f'(x)$ is a differentiable function, we can calculate its derivative in the same way. We call this *the second derivative of* f (with respect to x), and write it as

$$\frac{d^2f}{dx^2} = f''(x) = \frac{d}{dx}\left(\frac{df}{dx}\right).$$

Third and higher derivatives are similarly defined by repeated differentiation.

Note that the phrase *'Provided that $f'(x)$ is a differentiable function'* is very important: $f'(x)$ may not be, even if $f(x)$ was itself differentiable. For example, suppose $f(x) = x^{\frac{3}{2}}$. Then $f'(x) = \frac{3}{2}\sqrt{x}$ and $f''(x) = \frac{3}{4\sqrt{x}}$. Hence we can see that although $f(x)$ is differentiable at 0, $f'(x)$ is not, since $f''(x)$ is undefined there.

What information does the second derivative convey? As usual, that depends on the problem being studied. Suppose that the position of a vehicle at time t is given by a function $x(t)$. The the first derivative is the rate of change of position of the vehicle - *i.e.* its velocity, $v(t)$. Hence we would write

$$\frac{dx}{dt} = x'(t) = v(t).$$

The second derivative of x, $\frac{d^2x}{dt^2} = x''(t) = \frac{dv}{dt} = v'(t)$, is the rate of change of velocity. But this is just the acceleration of the vehicle, $a(t)$.

What does the second derivative tell us about the shape of the graph of a function, $y = f(x)$?

Remember that if $\frac{df}{dx} = f'(x) > 0$ on an interval, then $f(x)$ is increasing on that interval. If $\frac{df}{dx} = f'(x) < 0$ on an interval, then $f(x)$ is decreasing on that interval. Since $\frac{d^2f}{dx^2} = f''(x)$ is the derivative of f, then:

- $\frac{d^2f}{dx^2} = f''(x) > 0$ on an interval, then $f'(x)$ is increasing on that interval;
- $\frac{d^2f}{dx^2} = f''(x) < 0$ on an interval, then $f'(x)$ is decreasing on that interval.

What precisely does this tell us? We illustrate the two types of behaviour in the diagram below.

- If $f''(x) > 0$, then $f'(x)$ is increasing; this means the curve is bending upwards.
- If $f''(x) < 0$, then $f'(x)$ is decreasing; this means the curve is bending downwards.

In fact, we can usefully describe what the second derivative tells us in terms of the *concavity* of the graph, which is defined as follows.

Definition 2.2 (Concavity)**.** We say a function $f(x)$ is *concave up* at a point x if the tangent line to the graph at that point lies below the graph in the region close to the point.
A function $f(x)$ is *concave down* at a point x if the tangent line to the graph at that point lies above the graph in the region close to the point.
Where concavity changes from up to down or *vice versa*, the tangent line must cross the graph. Such a point is called a *point of inflection*

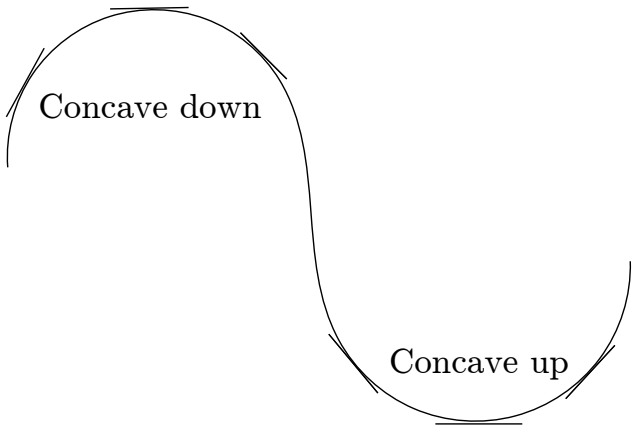


Illustration of a curve with concave up and concave down regions

Now we can say that:

- if $f''(x) > 0$, then $f'(x)$ is increasing and the curve is *concave up*;
- if $f''(x) < 0$, then $f'(x)$ is decreasing and the curve is *concave down.*

We have to be a little careful, as the converse is not true. A curve can be concave up (or down) everywhere and have $f''(x) = 0$ at a point. For example, the curve $f(x) = x^4$ is concave up everywhere, but has $f''(0) = 0$. Thus, if we know a function is concave up on an interval, the most we can say is that $f''(x) \geq 0$ on that interval. Similarly, if a function is concave down, the most we can say is that $f''(x) \leq 0$.

2.2 Rules for differentiation

Lecture 7 In principle, we could use the definition to calculate the derivative of any function of interest. However, the procedure would be long-winded and tiresome. In order to save time in calculations, we need to know the derivatives of a small repertoire of common functions. You will probably have already learned a number of these at school. Henceforth, it will be assumed that you are familiar with the following:

$$\begin{aligned}
\frac{dx^r}{dx} &= rx^{r-1} \quad (r \neq 0) \\
\frac{d}{dx}\sin x &= \cos x \\
\frac{d}{dx}\cos x &= -\sin x \\
\frac{d}{dx}\tan x &= \sec^2 x \\
\frac{d}{dx}e^x &= e^x \\
\frac{d}{dx}\ln x &= \frac{1}{x}.
\end{aligned}$$

Using these derivatives and the general rules below, we are able to calculate the derivatives of many more complex functions.

2.2.1 Differentiating linear combinations, products and quotients

Let u and v be differentiable functions, and c_1 and c_2 be fixed real numbers. Then:

Property 2.1 (Derivative of a linear combination)**.**

$$\frac{d}{dx}(c_1u + c_2v) = c_1\frac{du}{dx} + c_2\frac{dv}{dx}$$

Property 2.2 (The Product Rule)**.**

$$\frac{d}{dx}(uv) = v\frac{du}{dx} + u\frac{dv}{dx}$$

We can derive this result from the definition of the derivative as follows:

$$\begin{aligned}
\frac{d}{dx}\left(u(x)v(x)\right) &= \lim_{h\to 0}\frac{u(x+h)v(x+h)-u(x)v(x)}{h}\\
&= \lim_{h\to 0}\frac{u(x+h)v(x+h)-u(x)v(x+h)+u(x)v(x+h)-u(x)v(x)}{h}\\
&= \lim_{h\to 0}\frac{u(x+h)v(x+h)-u(x)v(x+h)}{h}+\lim_{h\to 0}\frac{u(x)v(x+h)-u(x)v(x)}{h}\\
&= \lim_{h\to 0}v(x+h)\left(\frac{u(x+h)-u(x)}{h}\right)+\lim_{h\to 0}u(x)\left(\frac{v(x+h)-v(x)}{h}\right)\\
&= \lim_{h\to 0}v(x+h)\lim_{h\to 0}\left(\frac{u(x+h)-u(x)}{h}\right)+\lim_{h\to 0}u(x)\lim_{h\to 0}\left(\frac{v(x+h)-v(x)}{h}\right)\\
&= v(x)\frac{du}{dx}+u(x)\frac{dv}{dx}
\end{aligned}$$

Property 2.3 (The Quotient Rule)**.** Provided $v(x)\neq 0$

$$\frac{d}{dx}\frac{u}{v}=\frac{v\dfrac{du}{dx}-u\dfrac{dv}{dx}}{v^2}$$

We require $v(x)\neq 0$ to ensure the quotient is differentiable. We can then derive this result using the Product Rule. Let $q(x)=\dfrac{u(x)}{v(x)}$; then $u(x)=q(x)v(x)$. Applying the Product Rule to $u(x)$ we have

$$\frac{du}{dx}=q\frac{dv}{dx}+v\frac{dq}{dx}.$$

Rearranging gives

$$\begin{aligned}
\frac{dq}{dx}&=\frac{1}{v}\left(\frac{du}{dx}-q\frac{dv}{dx}\right)=\frac{1}{v}\left(\frac{du}{dx}-\frac{u}{v}\frac{dv}{dx}\right)\\
&=\frac{1}{v^2}\left(v\frac{du}{dx}-u\frac{dv}{dx}\right).
\end{aligned}$$

Example 2.2. Using the linear combination rule (2.1)

$$\frac{d}{dx}\left(-5x^3+3x^2-x+2\right)=-15x^2+6x-1.$$

Example 2.3. Newton's Second Law of motion states that the rate of change of momentum of a body is equal to force, F, acting on it. The momentum of a body of constant mass m, moving in a straight line with velocity $v(t)$ is given by mv. Hence, for such a body, combining Newton's Second Law with the linear combination rule for derivatives tells us that

$$F=\frac{d}{dt}(mv(t))=m\frac{dv}{dt}=ma,$$

since the mass (m) is constant, and the rate of change of velocity, $\frac{dv}{dt}$, is the acceleration, a. Note that this familiar formula is only applicable to bodies of constant mass. This assumption does not apply to *e.g.* rockets, where the fuel is rapidly burned up, changing the mass of the vehicle. We shall consider this example in more detail later in the course.

Example 2.4. Using the product rule (2.2)

$$\frac{d}{dx}\left((\sin x)\sqrt{x}\right) = \cos x\sqrt{x} + \sin x\frac{1}{2\sqrt{x}}.$$

Example 2.5. Using the quotient rule (2.3)

$$\begin{aligned}\frac{d}{dx}\tan x &= \frac{d}{dx}\frac{\sin x}{\cos x}\\ \frac{d}{dx}\tan x &= \frac{\cos x\cdot\cos x - \sin x(-\sin x)}{\cos^2 x}\\ &= \frac{\cos^2 x + \sin^2 x}{\cos^2 x}\\ &= \frac{1}{\cos^2 x} = \sec^2 x.\end{aligned}$$

Example 2.6. Using a combination of several rules:

$$\begin{aligned}&\frac{d}{dx}\frac{5x^4+x^2}{x^3-4x+3}\\ &= \frac{(20x^3+2x)(x^3-4x+3)-(5x^4+x^2)(3x^2-4)}{(x^3-4x+3)^2}.\end{aligned}$$

2.3 Differentiating compositions of functions

Suppose that a spherical balloon is being inflated in such a way that the radius of the ballon, r, at time t is given by $r = 3 + \frac{t}{2}$ cm. How fast is the volume, V, of the balloon (in cm3) increasing?

We know that, since the balloon is spherical, $V(r) = \frac{4}{3}\pi r^3$. Since $r(t) = 3 + \frac{t}{2}$, we have

$$V(r(t)) = \frac{4}{3}\pi\left(3+\frac{t}{2}\right)^3.$$

The rules we have learned so far are not so helpful in this case. One approach is to multiply out the brackets, which gives

$$V = \frac{4}{3}\pi\left(27 + \frac{27t}{2} + \frac{9t^2}{4} + \frac{t^3}{8}\right).$$

Now, we can find the derivative of V with respect to t using the linear combination rule

$$\frac{dV}{dt} = \pi\left(18 + 6t + \frac{t^2}{2}\right).$$

However, multiplying out brackets is a rather tedious process. Luckily, here we only had a cubic to contend with; in some problems the power might be much higher. What we would do if we had, say, $\left(3+\frac{t}{2}\right)^{13}$? The calculation then would be much more tedious, and the likelihood of making

a mistake, greater. Fortunately, there is a very useful property of derivatives which helps us in cases like these, where we are interested in differentiating a function which can be written as a composition of functions.

Property 2.4 (The Chain Rule)**.** Suppose that y is a function of u, which is itself a function of x - *i.e.* $y = y(u)$, where $u = u(x)$. Then, the derivative of y with respect to x is given by

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}.$$

In alternative notation, let f and g be differentiable functions. Then

$$\frac{d}{dx} f\left(g(x)\right) = f'\left(g(x)\right) g'(x).$$

For the case of our balloon example, we have $V = V(r) = \frac{4}{3}\pi r^3$ and $r = r(t) = 3 + \frac{t}{2}$, so

$$\frac{dV}{dr} = 4\pi r^2, \qquad \frac{dr}{dt} = \frac{1}{2}.$$

Hence, on using the Chain Rule, we find

$$\frac{dV}{dt} = \frac{dV}{dr}\frac{dr}{dt} = 4\pi r^2 . \frac{1}{2} = 2\pi \left(3 + \frac{t}{2}\right)^2 = 2\pi \left(9 + 3t + \frac{t^2}{4}\right) = \pi \left(18 + 6t + \frac{t^2}{2}\right).$$

This is, of course, the same answer as we previously obtained. However, here there was no need for us to multiply out the brackets (we only did so to demonstrate the fact that the two answers were the same).

Examples:

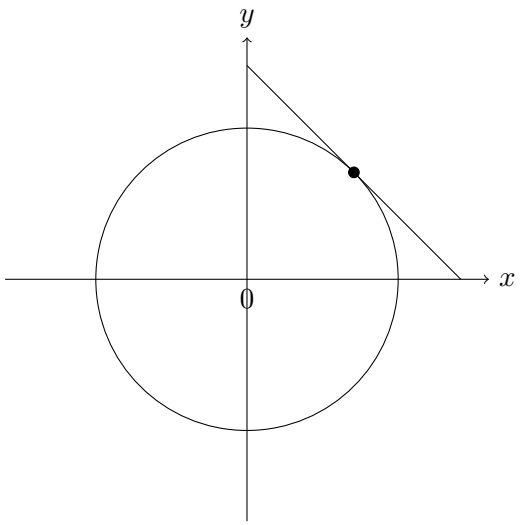
1. $\dfrac{d}{dx} \sin \sqrt{x} = (\cos \sqrt{x}) \cdot \dfrac{1}{2\sqrt{x}}.$

2. $\dfrac{d}{dx} \sqrt{1 + x^3} = \dfrac{1}{2\sqrt{1 + x^3}} \cdot 3x^2.$

3. $$\begin{aligned} \frac{d}{dx} \sqrt[3]{\frac{1+x}{1-x}} &= \frac{d}{dx} \left(\frac{1+x}{1-x}\right)^{1/3} \\ &= \frac{1}{3} \left(\frac{1+x}{1-x}\right)^{-2/3} \cdot \frac{1 \cdot (1-x) - (1+x)(-1)}{(1-x)^2} \\ &= \frac{1}{3} \left(\frac{1+x}{1-x}\right)^{-2/3} \cdot \frac{2}{(1-x)^2}. \end{aligned}$$

2.4 Implicit differentiation

Lecture 8



How steep is the tangent to a circle at some point (x, y) on the circle?

As a starting point, we could begin by considering how we would answer this question for another sort of curve, rather than a circle. For example, what about the curve $y = x^2$? In that case, the solution would be easy. We know that the gradient of the tangent to the curve is given by $\frac{dy}{dx} = 2x$; the steepness is just $\left|\frac{dy}{dx}\right| = |2x|$. Hence, the key part of the solution is calculating $\frac{dy}{dx}$, which is straightforward provided we know the equation of the curve $y = f(x)$. However, in the case of a circle of radius a (centred at the origin) the equation of the curve is given by

$$x^2 + y^2 = a^2,$$

rather than specifying y explicitly as a function of x. One way we might try to get around the problem is by rearranging the equation, to get

$$y = \sqrt{a^2 - x^2},$$

but this introduces another problem. For a circle, y can take values between $-a$ and a, but by choosing to take the positive square root, I can only obtain the positive values. Hence, I only have the upper half of the circle (I would need two equations to get the complete curve).

Instead, our lives are much easier if we simply differentiate the whole of the original equation with respect to x. We can do this using the Chain Rule - we just need to remember that y depends on x (*i.e.* $y = y(x)$). Hence,

$$\frac{d}{dx}(x^2 + y^2) = 2x + 2y\frac{dy}{dx} = \frac{da^2}{dx} = 0.$$

Rearranging, we obtain

$$\frac{dy}{dx} = \frac{-2x}{2y} = -\frac{x}{y}.$$

Using this formula, we can find the gradient of the tangent to the circle at any point (x, y) on the circle, and hence we will know the steepness.

2.4.1 Derivatives of inverse functions

Recall that a function f which is 1–1 on its domain has an inverse f^{-1}. (A function f is 1–1 if it satisfies the horizontal line test; that is, each horizontal line intersects the graph of f at most once.)

Let f be a one-to-one function which is differentiable on some domain of interest, with inverse f^{-1}. Recall that if $y = f^{-1}(x)$, then we must have $x = f(y)$, and $f(y) = f(f^{-1}(x)) = x$. In order to ensure f^{-1} is differentiable at x, we require $f'(y) \neq 0$ (where $y = f^{-1}(x)$) or equivalently, $f'(f^{-1}(x)) \neq 0$. The reason for this will become clear shortly.

Using implicit differentiation on this last equation, we see that

$$\frac{df(y)}{dx} = \frac{df}{dy}\frac{dy}{dx} = \frac{dx}{dx} = 1$$

and so

$$\frac{dy}{dx} = \frac{1}{df/dy}.$$

But since $f(y) = x$, $\frac{df}{dy} = \frac{dx}{dy}$, and thus

$$\frac{dy}{dx} = \frac{1}{dx/dy}.$$

Now we can see why we needed to be a little careful about the differentiability of f^{-1} earlier. If $\dfrac{dx}{dy} = 0$, then $\dfrac{dy}{dx}$ will be undefined (and hence f^{-1} is not differentiable at that point).

Note that when we calculate $\dfrac{dx}{dy}$, we will usually obtain it as a function of y, and hence the right hand side of the equation above will be a function of y too. If we want to find $\dfrac{dy}{dx}$ as a function of x, we must use the fact that $y = f^{-1}(x)$ express the right-hand side in terms of x. This is perhaps clearer when we write the result in the alternative notation

$$\frac{df^{-1}}{dx} = \frac{1}{f'(y)} = \frac{1}{f'(f^{-1}(x))} \qquad (\text{for } f'(f^{-1}(x)) \neq 0).$$

Example 2.7. Consider $g(x) = x^2$, $x \geq 0$, find $(g^{-1})'(4)$.

Before we begin, we need to think carefully about what we are trying to calculate, and state it clearly: there is scope for us to get confused by a poor choice of notation here.

We want to calculate the derivative of g^{-1}. Hence, to get things set up in the same form as we have above, we need let $y = g^{-1}(x)$. Then, $x = g(y)$, and since $g(x) = x^2$, $x \geq 0$, we have $g(y) = y^2$, $y \geq 0$ (simply changing the symbol from x to y). Then

$$x = g(y) = y^2 \quad \Rightarrow \quad \frac{dx}{dy} = 2y,$$

so

$$\frac{d}{dx}g^{-1}(x) = \frac{dy}{dx} = \frac{1}{dx/dy} = \frac{1}{2y}.$$

Now we need to express the last term as a function of x. We do this using the inverse function $y = g^{-1}(x) = \sqrt{x}$, $(x \geq 0)$ and so obtain

$$\frac{dy}{dx} = \frac{1}{2\sqrt{x}} = \frac{1}{2}x^{-1/2}.$$

Hence

$$(g^{-1})'(4) = \frac{1}{2\sqrt{4}} = \frac{1}{4}.$$

Example 2.8. Let $f(x) = \dfrac{1+3x}{5-2x}$, find $f^{-1}(x)$. Hence find $\dfrac{d}{dx}f^{-1}(x)$ by using the formula for the derivative of an inverse, and also by direct differentiation.

Be careful to identify when the roles of x and y swap. From $y = \frac{1+3x}{5-2x}$

$$\begin{aligned} 5y - 2xy &= 1 + 3x \\ 5y - 1 &= x(3 + 2y) \\ x &= \frac{5y-1}{3+2y}, \end{aligned}$$

thus

$$f^{-1}(x) = \frac{5x-1}{2x+3}.$$

Derivative $y = f^{-1}(x)$ so $x = f(y) = \frac{1+3y}{5-2y}$

$$\begin{aligned} \frac{dy}{dx} &= \frac{1}{\frac{dx}{dy}} = \frac{1}{\frac{3(5-2y)+2(1+3y)}{(5-2y)^2}} \\ &= \frac{(5-2y)^2}{17} \\ &= \frac{\left(5 - 2\left(\frac{5x-1}{2x+3}\right)\right)^2}{17} \\ &= \frac{(5(2x+3) - 2(5x-1))^2}{17(2x+3)^2} \\ &= \frac{17^2}{17(2x+3)^2} = \frac{17}{(2x+3)^2} \end{aligned}$$

Check (Easier!)

$$\begin{aligned} \frac{df^{-1}(x)}{dx} &= \frac{d}{dx}\left(\frac{5x-1}{3+2x}\right) \\ &= \frac{5(3+2x) - 2(5x-1)}{(2x+3)^2} \\ &= \frac{17}{(2x+3)^2} \end{aligned}$$

2.4.2 Derivatives of inverse trigonometric functions

We can use the results we have just established to find the derivatives of inverse trigonometric functions.

Example 2.9. Show that

$$\frac{d}{dx}\arcsin x = \frac{1}{\sqrt{1-x^2}}\,.$$

Recall $y = \arcsin x \Leftrightarrow x = \sin y$

$$\frac{dy}{dx} = \frac{1}{\frac{dx}{dy}} = \frac{1}{\frac{d\sin y}{dy}} = \frac{1}{\cos y}.$$

We now need to express $\cos y$ in terms of x. Recall $\cos^2 y = 1 - \sin^2 y = 1 - x^2$ as $x = \sin y$, so $\cos y = \pm\sqrt{1-x^2}$. Which *sign* $\pm$ do we use?

Recall for $y = \arcsin x$, $-1 \le x \le 1$ and $\frac{-\pi}{2} \le y \le \frac{\pi}{2}$. For $\frac{-\pi}{2} \le y \le \frac{\pi}{2}$, $\cos y \ge 0$. Thus $\cos y = +\sqrt{1-x^2}$, and hence $\frac{d}{dx}\arcsin x = 1/\sqrt{1-x^2}$.

Similarly, we can show that $\frac{d}{dx}\arctan x = \frac{1}{1+x^2}$.

Recall $y = \arctan x$ iff $x = \tan y$ for which we know $dx/dy = \sec^2 y$, then

$$\frac{dy}{dx} = \frac{1}{\frac{dx}{dy}} = \frac{1}{\sec^2 y} = \cos^2 y\,.$$

We use $x = \tan y$ to get a formula for $\sec^2 y$:

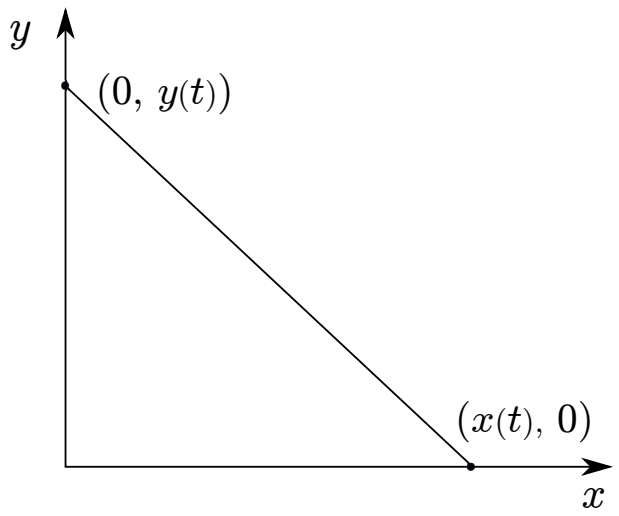
$$\begin{aligned}
&\cos^2 y + \sin^2 y = 1\\
&1 + \frac{\sin^2 y}{\cos^2 y} = \frac{1}{\cos^2 y}\\
&1 + \tan^2 y = \sec^2 y\\
&\frac{dy}{dx} = \frac{1}{1+x^2} \quad \text{as } x = \tan y.
\end{aligned}$$

2.5 Related rates

Lecture 9

Suppose we have a situation where we know two quantities are related in a certain way. Implicit differentiation can be a useful tool to establish the relationship between the rates of change of these quantities, known as *related rates*. The idea is illustrated by the following examples.

Example 2.10 (A falling ladder)**.** A ladder of length 3 m stands on flat ground, leaning against a vertical wall. The bottom of the ladder is at $(x(t), 0)$, and the top is at $(0, y(t))$. At $t = 0$ the ladder begins to slip, with the bottom moving horizontally outwards at 0.1 ms^{-1}. Assuming it moves only in the vertical direction, how fast is the top of the ladder slipping down when the foot is 1m from the wall? How does the speed at which the top falls change as the ladder slips further?



Sketch of the ladder problem.

Since the length of the ladder is constant, we must have

$$x^2 + y^2 = 9,$$

by Pythagoras' theorem. Differentiating with respect to time (using the chain rule) we find

$$2x\frac{dx}{dt} + 2y\frac{dy}{dt} = 0.$$

Re-arranging we find

$$\frac{dy}{dt} = -\frac{x}{y}\frac{dx}{dt} = -(0.1)\frac{x}{y}. \tag{2.1}$$

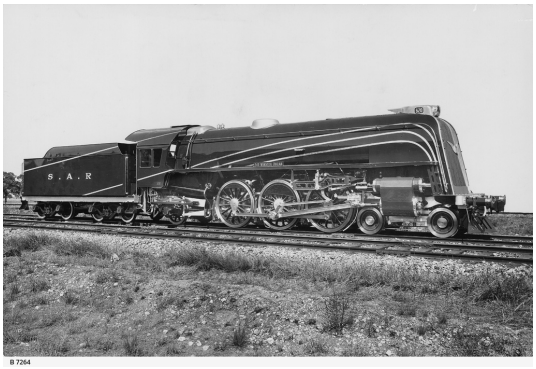
When $x = 1$, $y = \sqrt{9-1} = \sqrt{8}$ and so when the ladder is 1m from the wall, the top is moving at a speed

$$\frac{dy}{dt} = -(0.1)\frac{1}{\sqrt{8}}\text{ms}^{-1}.$$

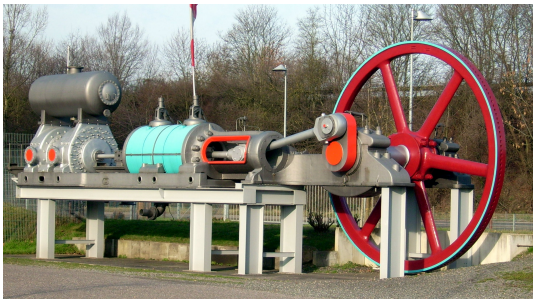
Note that the minus sign indicates movement in the downward (negative-y) direction.

As the ladder slips away from the wall, $x(t)$ will increase, and, correspondingly, $y(t)$ must decrease. Looking at equation (2.1), we can then see that, the further the ladder moves from the wall, the faster to top of it is dropping.

Example 2.11 (Piston motion)**.** In a steam engine, the high pressure steam drives a piston back and forth. This motion is translated into the rotation of a wheel by a connecting rod. (A similar principle operates in a petrol or diesel engine, except that in that case, the movement of the piston is caused by the expanding exhaust gases from the ignition of the fuel, rather than by steam.)

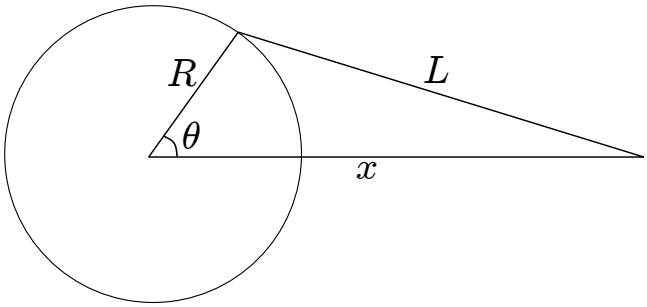


Sir Winston Dugan (SAR Class 620)



Model stationary steam engine at Speyer, Germany

We consider a simplified version of this situation. Suppose the piston is linked to a wheel of radius R by a straight, rigid rod of length, L. Let x be the horizontal distance between the centre of the wheel, and the point where the piston is joined to the rod, as shown in the sketch below. This changes with time as the piston moves back and forth.



How is the rate of rotation of the wheel related to how fast is the piston moving?

From the diagram, we can see that the speed of the piston is just the rate of change of $x(t)$. Using the cosine rule, we have

$$L^2 = R^2 + x^2 - 2Rx\cos\theta.$$

Implicit differentiation of the above, using the product rule gives:

$$2x\frac{dx}{dt} - 2R\cos\theta\frac{dx}{dt} + 2Rx\sin\theta\frac{d\theta}{dt} = 0.$$

Rearranging we have

$$(R\cos\theta - x)\frac{dx}{dt} = Rx\sin\theta\frac{d\theta}{dt},$$

so

$$\frac{dx}{dt} = \left(\frac{Rx\sin\theta}{R\cos\theta - x}\right)\frac{d\theta}{dt}.$$

Lecture 10

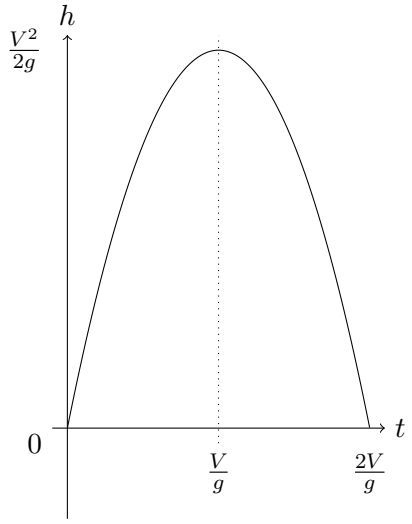
2.6 Maxima and minima of functions

Example 2.12 (Projectile motion)**.** Consider the height of a ball, or similar object, thrown upwards with speed, V, and acted upon only by gravity (we neglect air resistance, *etc.*). The height h of the ball at time t is then given by

$$h = h_0 + Vt - \frac{1}{2}gt^2,$$

where h_0 is the height above ground from which the object is thrown, and $g \approx 9.8$ ms^{-2} is the acceleration due to gravity. What is the maximum height reached by the ball?

There are various different ways we might approach this problem. We can see from the formula that the quantity we need to maximise is $Vt - \frac{1}{2}gt^2$ (since h_0 just adds a constant to this). Hence, let us take $h_0 = 0$ for the moment, to simplify things. For small values of t, the first term will dominate, but for larger t, the second term becomes significant, and will eventually cancel it out. Thus we expect a 'middling' value of t will give the greatest height. If we knew the value of V numerically, one 'low-tech' method we might try would simply be to try different values of t in the formula. We could then progressively 'narrow down' the possible value, until we find t to whatever accuracy we might require. But this would be extremely time consuming and tedious.



A better method would be to plot the graph of h against t, as above. Then we can see that the maximum height is achieved at a time halfway between when the ball is thrown up, and when it returns to the ground. The first of these times is $t = 0$; the second is $t = \frac{2V}{g}$. Hence the time at which the maximum height is reached is $t^* = \frac{V}{g}$. The height is

$$h_{max} = h\left(t^*\right) = \frac{V^2}{2g}.$$

This method worked even when we did not know the numerical value of V, but we relied on the symmetry of the parabola (the fact that the maximum occurred at a value of t half way between the two roots of $h(t) = 0$). In other problems, we might want to find the maximum of a function that does not have such a helpful symmetry; what would we do then?

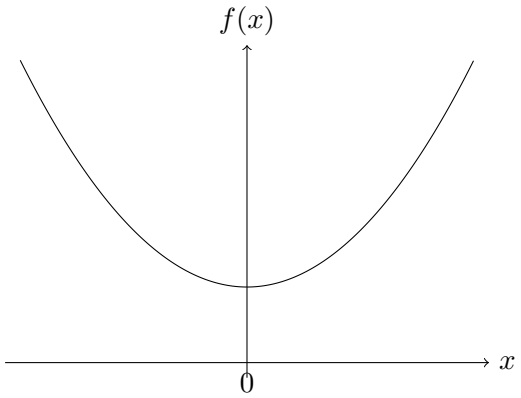
Briefly thinking about the physics of the problem, and the graph of the function $h(t)$ suggests another way forward. Initially, when the ball is thrown, it is moving upwards, so the function $h(t)$ is increasing. This means $\frac{dh}{dt} > 0$. Later, the ball will be moving downwards, so $h(t)$ will be decreasing, and $\frac{dh}{dt} < 0$. When the ball reaches its maximum height, it changes from moving upwards to moving downwards, so for an instant its velocity is zero. Hence $\frac{dh}{dt} = 0$. This observation provides us with another way of finding the maximum height. We can easily calculate

$$\frac{dh}{dt} = V - gt.$$

The derivative $\frac{dh}{dt}$ will be zero when $V - gt = 0$, which implies $t^* = \frac{V}{g}$, as we found previously. This method, however, did not require us to use any properties of the graph of $h(t)$ to determine t^*. Instead, it relies on the simple observation, that as the graph of a function goes through the maximum value, the function goes from increasing to decreasing. Hence the derivative of the function changes from being positive to being negative. The transition will occur at the maximum point itself, so the derivative there will be zero.

2.6.1 Local maxima and minima

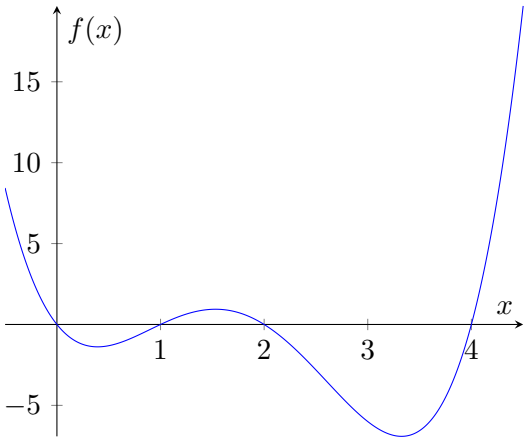
In the example we have just considered, the point where $h(t)$ took its maximum value was the transition point between the region where the function was increasing ($\frac{dh}{dt} > 0$), and the region where it was decreasing ($\frac{dh}{dt} < 0$). Hence, the maximum corresponded exactly to the unique point where $\frac{dh}{dt} = 0$. In other circumstances, we might want to find the minimum value of some function, $f(x)$, say. Now, if we think about the shape of such as graph (such as the one drawn below), we can see that as the function goes through the minimum value, it will transition from being a decreasing function ($\frac{df}{dx} < 0$) to being an increasing function ($\frac{df}{dx} > 0$); at the minimum value itself, we will have $\frac{df}{dx} = 0$.



A function $f(x)$ with a minimum at $x = 0$

Thus, at a maximum *or* minimum point, the derivative of the function will be zero. This suggests some interesting questions about the property we used to find the maximum in the last example. For instance, can we always find the maximum or minimum value of a function f simply by finding the roots of $f'(x) = 0$? What happens if $f'(x)$ has multiple zeros?

Graph of $f(x) = x(x-1)(x-2)(x-4)$



Consider the function $f(x) = x(x-1)(x-2)(x-4)$ for $-1 \leq x \leq 5$. We can see that the graph has three points where the tangent to the curve would be parallel to the x axis (*i.e.* where the derivative $\frac{df}{dx} = 0$). This is what we expect: since $f(x)$ is a quartic, its derivative will be a cubic, and a cubic equation has at most three roots. Now, there can be at most one minimum and one maximum value, so clearly at least one of the three roots cannot correspond to the maximum or minimum value. Equally, we see that the function takes its largest value at $x = 5$, where $\frac{df}{dx} \neq 0$. We need to be a little more precise with our definitions if we want to understand the relationship between points where the derivative is zero, and the maximum and minimum values of the function.

Firstly, we introduce the idea of local maxima and minima, where we consider the values of the function only in some small neighbourhood of the point of interest.

Definition 2.3. A function f has a *local maximum* at a point x if $f(x)$ is greater than or equal to the values of f for all points near x. A function f has a *local minimum* at a point x if $f(x)$ is less than or equal to the values of f for all points near x. If a point is either a *local maximum* or a *local minimum* we call it a *local extremum* of the function.

We can now see that the three points where the derivative of $f(x) = x(x-1)(x-2)(x-3)$ is zero are all local extrema. The end points $x = -1$ and $x = 5$ are also local extrema.

2.6.2 Critical points

We now need to introduce an important piece of terminology.

Definition 2.4. A *critical point* of a function, f, is a point x in the domain of f where either

$$\frac{df}{dx} = 0 \quad \text{or} \quad \frac{df}{dx} \text{ is undefined.}$$

Note: the definition tells us that for x to be a critical point, $f(x)$ must be defined even if $\dfrac{df}{dx}$ is undefined. This means that $x = 0$ *is* a critical point of the absolute value function, $|x|$. However, $x = 0$ *is not* a critical point of $f(x) = x^{-1}$, even though $f'(x)$ is undefined there, because $f(x)$ itself is undefined at $x = 0$.

Property 2.5 (Local extrema at interior points are critical points)**.** Suppose f that is a non-constant function defined on an interval, and that it has a local extremum at x (where x is not one of the end points of the interval). If f is differentiable at x, then $f'(x) = 0$. Thus, if x is a local extremum of f, it is also a critical point of f.

This property is straightforward to prove. First, consider the case where f is not differentiable at x. This means $f'(x)$ is undefined, and hence, from the definition, x is a critical point of f.

Now, consider the case where f is differentiable at x, and has a local maximum there. Then, for any sufficiently small $|h|$ the existence of a local maximum at x means $f(x+h) \leq f(x)$. If $h > 0$ this means that

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

and if $h < 0$ this means that

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

Using these inequalities and the fact that f is differentiable at x we have, for $h > 0$

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

and, for $h < 0$

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

Since the limit is unique, we must obtain the same answer whether we approach x from the left or the right (*i.e.* with h positive or negative). Hence $f'(x) = 0$. The proof for a minimum point follows in the same manner. Thus, if f has a local extremum at x, then x is a critical point.

Note that the converse of this result is *not* true: a point x may be a critical point of f which is not a local extremum. For example, if $f(x) = x^3$, then $f'(0) = 0$, but zero is not a local extremum.

2.6.3 The second derivative test

Lecture 11

Now that we have established that all local extrema correspond to critical points (though not *vice versa*), we can see that one way to find local extrema is to find all the critical points, and then check each one to see if it is a maximum, minimum or neither. Looking at whether the function is concave up or concave down at the critical point helps us to do this. If $f'(x) = 0$ and the function is concave down, that means the curve of the graph is bending downwards, and the point must be a local maximum. Conversely, if $f'(x) = 0$ and the function is concave up, the curve is bending up, so we must be at a local minimum. Since the second derivative of f tells us about the concavity, we can use it to test for local maxima and minima.

Property 2.6 (**The second derivative test for local maxima and minima).** Let f be a twice differentiable function defined on some interval containing the point x.

If $\frac{df}{dx} = 0$ and $\frac{d^2f}{dx^2} < 0$ then f has a local maximum at x.
If $\frac{df}{dx} = 0$ and $\frac{d^2f}{dx^2} > 0$ then f has a local minimum at x.
If $\frac{df}{dx} = 0$ and $\frac{d^2f}{dx^2} = 0$ then the test provides no information.

In the last case, we would need to use some other method *e.g.* considering the graph of the function, to decide what type of point we were dealing with. One possibility is that x is a *point of inflection.*

2.6.4 Global extrema

At this stage, we have established how to find local extrema of functions: they are among the critical points of the function, or the end points of the interval on which the function is defined. The maximum or minimum value of the function over the entire domain is called the *global maximum* or *global minimum.* In order to find the global maximum or global minimum, intuitively, we would just check the values of the function at all the local maxima or minima, and choose whichever of those that gives the greatest or least value of the function. This idea is essentially sound, but there are two ways things could go wrong.

The global maximum or minimum function f may not exist if, for example:

- The function is defined over an open interval (a, b). Consider $f(x) = x$ defined on $x \in (0, 1)$; the maximum value of f would occur for the largest value x, but there is no largest value of x in $(0, 1)$.
- The function is defined over an unbounded interval such as $[0, \infty)$. Again, consider $f(x) = x$. Although there is now a global minimum at $x = 0$, there is no global maximum, as $f(x)$ increases without bound as $x \to \infty$.

For continuous functions defined on closed and bounded intervals, our intuitive idea is correct.

Property 2.7 (The Extreme Value Theorem)**.** Let f be a continuous function defined on a closed, bounded interval $[a, b]$. Then, f has a global maximum and a global minimum on $[a, b]$.

The proof of this theorem is beyond the scope of this course.

We can use the Extreme Value Theorem to find the global maximum and minimum of continuous functions on closed, bounded intervals by exploiting the connections we have established between critical points and local extrema. The steps are:

- Find the critical points of f by solving $\frac{df}{dx} = 0$ and finding any points where $f'(x)$ is undefined.
- Evaluate f at every critical point and the end points of the interval, a and b.

- Compare the values of $f(x)$ thus obtained; the largest value gives the global maximum, the smallest the global minimum.

2.7 Optimisation

We have now demonstrated the way to find the smallest and largest values a function can take. In practical problems, such values are often important, *e.g.* to minimise the amount of fuel used by a vehicle, or to maximise the amount of a product of a chemical reaction. This type of problem is illustrated in the example below.

Example 2.13. A soft drink can is approximately cylindrical, and is required to hold 330 cm^3 (equivalently, ml) of liquid. As a can manufacturer, you aim to minimise costs by using the smallest amount of metal possible. Assuming the cost of materials used in each can is proportional to the can's surface area, what dimensions should the can be made to, in order to minimise the cost of producing it?

A cylindrical can basically consists of two circular pieces for the top and bottom, and a large rectangular piece which is curved around to form the sides. Briefly thinking about the problem, we can see that if we try to save material on the sides, by making it short, we will have to make the discs for the top and bottom very large to accommodate the liquid. Similarly, if we try to save material from the top and bottom by making them smaller, the can will need to be tall to hold the drink. So, there is a trade-off that we need to make.

Let us now consider a cylindrical can of radius r and height h. The volume of the can is given by $V = \pi r^2 h$, and its surface area is the sum of the two end pieces and the curved piece that forms the sides. The area of the top and bottom are each πr^2. The piece that forms the sides of the can is a rectangle, with one side of length h, and the other side of length $2\pi r$ (the circumference of the circles forming the top and bottom, around which it is wrapped); hence its area is $2\pi rh$. Therefore, the total surface area of the can is $A = 2\pi rh + 2\pi r^2$. We wish to minimise this value, subject to the constraint that $V = \pi r^2 h = 330$ cm^3 (so the can contains the required amount of liquid). However, we only know how to find the maxima and minima of functions of a single variable; here A depends on both r and h. How can we simplify things?

Fortunately, the constraint that we require the can to contain a certain amount of liquid comes to our rescue. We can rearrange the volume equation to get h in terms of r. We find

$$h = \frac{330}{\pi r^2}.$$

Then, we substitute this expression for h into our equation for A, which yields

$$A = \frac{660}{r} + 2\pi r^2.$$

We can find the value of r which minimises A by computing the derivative if A and setting it equal to zero. Thus

$$\frac{dA}{dr} = -\frac{660}{r^2} + 4\pi r = 0.$$

We rearrange this equation to get the value of r:

$$r = \left(\frac{330}{2\pi}\right)^{\frac{1}{3}} \approx 3.7\text{cm}.$$

Note that the corresponding height is

$$h = \frac{330}{\pi r^2} = 2\left(\frac{330}{2\pi}\right) = 2r \approx 7.5\,\text{cm}.$$

The final stage is to check our answer really does give a minimum of the area, A, rather than a maximum. The second derivative of A is

$$\frac{d^2A}{dr^2} = \frac{1320}{r^3} + 4\pi,$$

which is positive for all $r > 0$. Hence, $r \approx 3.7$ corresponds to a minimum.

In fact, if we look up product information on the web, it appears 330ml cans (the European standard size) are made to a diameter of 6.6cm and height 11.5cm which means they are taller and thinner than the dimensions we have calculated here. Why might there be such a difference?

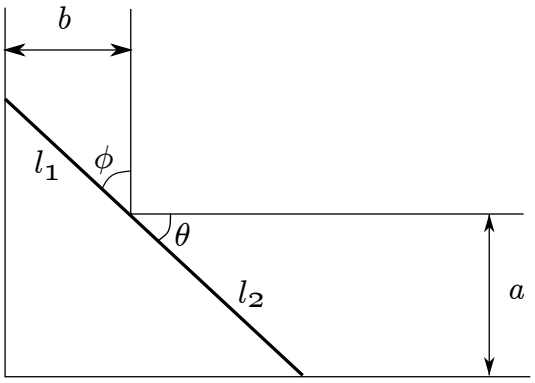
Lecture 12 The above example was fairly straightforward. However, practical situations can lead to a wide variety of optimisation problems. Some will require more careful though before we can simply apply our procedure for finding extrema of differentiable functions. The following fairly general guidelines, although they may not all be relevant in every problem, will often help.

The steps are:

1. Draw a picture (if possible). Label the variables.
2. Identify the quantity to be maximised or minimised. This is the dependent variable.
3. Identify the quantities on which the dependent variable depends. Write down the relations between these variables.
4. Select *one* of the quantities from step 3 and express the dependent variable as a function of this variable – the independent variable – alone. Use the physical constraints of the problem to fix the domain of the function. (Note: frequently one variable may be a better choice than others, so a little thought should be put into this step).
5. The problem should now be converted into a mathematical one of finding the global extremum a certain function over an interval.
6. Answer the original question.

The next example illustrates the more complicated situations we might encounter in real life problems.

Example 2.14. Two corridors, which meet at right angles, have widths a and b m. Find the length of the longest pole which will go around the corner, assuming the pole remains horizontal at all times.



Sketch of the pole in the corridor problem

Consider the diagram above. For a given angle θ between the pole and the wall, the longest pole which can fit in the angle between the corridors will be just touching the walls at each side, and at the corner, as illustrated.

The length of this longest pole is

$$l = l_1 + l_2.$$

Since $\phi = \frac{\pi}{2} - \theta$, simple trigonometry gives

$$l_1 \sin\phi = l_1 \cos\theta = b, \qquad l_2 \sin\theta = a.$$

Hence, as a function of the angle θ, the length of the longest pole that can fit in the corner is

$$l(\theta) = \frac{a}{\sin\theta} + \frac{b}{\cos\theta}.$$

Now, as we turn the pole through the corner, the angle θ will go from 0 to $\pi/2$. In order to find the *maximum* length of pole that will be able to make the turn, we need to find the *minimum* value of $l(\theta)$ (the maximum length of pole that fits in the corridor for a given θ) for $\theta \in [0, \frac{\pi}{2}]$. This pole will be able to fit in the corner for all angles between 0 and $\frac{\pi}{2}$, and hence can traverse the corner successfully.

Differentiating l with respect to θ we obtain

$$l'(\theta) = -\frac{a\cos\theta}{\sin^2\theta} + \frac{b\sin\theta}{\cos^2\theta}.$$

l is minimised when $l'(\theta) = 0$ *i.e.*

$$a\cos^3\theta = b\sin^3\theta \quad \Rightarrow \quad \tan^3\theta = \frac{a}{b}.$$

Equivalently, we can write this as

$$\theta = \tan^{-1}\left(\frac{a}{b}\right)^{\frac{1}{3}}.$$

Let us call this value θ^*. Then, the length of the pole is given by

$$l = \frac{a}{\sin\theta^*} + \frac{b}{\cos\theta^*}.$$

Since l is large and positive for θ close to 0 and $\frac{\pi}{2}$, and there are no other critical points, we can see this is indeed a global minimum. (Alternatively, the second derivative test can be used.) To express l more compactly, let $\alpha = \tan\theta^* = \left(\frac{a}{b}\right)^{\frac{1}{3}}$. Then, by drawing the appropriate triangle, we see that $\sin\theta^* = \frac{\alpha}{\sqrt{1+\alpha^2}}$ and $\cos\theta^* = \frac{1}{\sqrt{1+\alpha^2}}$. Substituting this into the formula for l we get

$$l = \frac{a\sqrt{1+\alpha^2}}{\alpha} + b\sqrt{1+\alpha^2}.$$

Since $\sqrt{1+\alpha^2} = \sqrt{1+\left(\frac{a}{b}\right)^{\frac{2}{3}}} = \frac{1}{b^{\frac{1}{3}}}(a^{\frac{2}{3}}+b^{\frac{2}{3}})^{\frac{1}{2}}$ we have

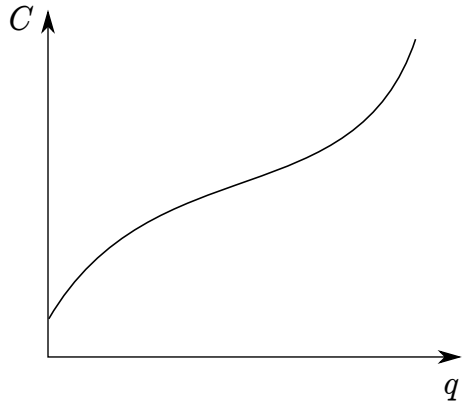
$$l = \frac{ab^{\frac{1}{3}}}{a^{\frac{1}{3}}b^{\frac{1}{3}}}(a^{\frac{2}{3}}+b^{\frac{2}{3}})^{\frac{1}{2}} + \frac{b}{b^{\frac{1}{3}}}(a^{\frac{2}{3}}+b^{\frac{2}{3}})^{\frac{1}{2}}.$$

Hence finally,

$$l = (a^{\frac{2}{3}}+b^{\frac{2}{3}})^{\frac{1}{2}}(a^{\frac{2}{3}}+b^{\frac{2}{3}}) = (a^{\frac{2}{3}}+b^{\frac{2}{3}})^{\frac{3}{2}}$$

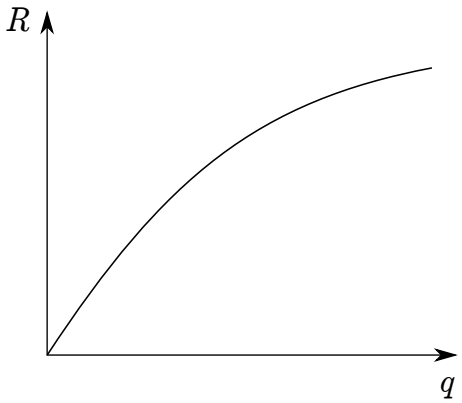
2.8 Applications to marginality

As well as the well-known applications in physics and engineering, another important area where optimisation can be applied is to business and economic decision-making.



Sketch of the typical shape of the cost function, $C(q)$

At the simplest level, to continue and expand, a business needs to bring in more revenue from selling its wares than it expends in producing them. Consider a very simple case where a company makes only a single product. We introduce the *cost function*, $C(q)$, which gives the total cost of producing a quantity q of some good. (For simplicity, we can think of q as being the number of 'widgets' the company produces, though of course, what the company produces need not be a physical object.) We expect the cost function to look something like the sketch above. In particular, since the more goods that are made, the greater the cost, we expect C to be an increasing function of q. The intercept on the vertical axis represents the *fixed costs*, which are incurred before anything can be produced. These might include the costs of buildings and machinery. The cost function increases rapidly at first, but then more slowly, as producing larger quantities of something is usually more efficient than producing small quantities (so-called *economies of scale*). However, at very high levels of production, the costs start increasing more rapidly again as shortages of resources like staff or raw materials begin to push up their costs.



Sketch of the typical shape of the revenue function, $R(q)$

The income that comes from selling the goods produced by the business is given by the *revenue function*, $R(q)$, where q is again the quantity of goods. If the price per item is p, and the quantity of goods sold is q then, $R = pq$. If the price of the goods does not change with quantity, then the graph of R against q will be a straight line. However, in reality we would expect that, when the quantity of goods sold becomes very large, the market for that particular product will become saturated, causing the price to drop. This would give a graph somewhat like that illustrated above.

The profit, $\pi(q)$ that the business makes is the difference between the revenue and the cost, *i.e.*

$$\pi(q) = R(q) - C(q).$$

If this is negative, the business makes a loss. (Note that the profit is often denoted $\pi(q)$ since π is the Greek letter corresponding to p, and p has already be used to denote price; it is nothing to do with the constant $\pi \approx 3.14$.) Hence, if we plot R and C on a graph, the business makes a profit where R is above C.

Marginal analysis

As the owner of a profitable business, you might be tempted to go one stage further and ask, 'Could I refine what I am doing to make the business more profitable?'. Many economic decisions are based on an analysis of how costs and revenues would change, if a small change was made to what the business is currently doing. This is called 'marginal analysis'; the additional costs involved are the *marginal costs*, and the associated revenues are *marginal revenues*.

Suppose we are running a large bakery, and need to decide if we should increase our production of bread from the current 1000 loaves per day. Assuming we make our decisions solely on financial grounds, the logical way to do this would be to consider the extra costs of baking more bread and compare this to the additional revenue we get from selling more. If the revenue increase is greater than the cost increase, we should increase production.

If we were to increase our production by one loaf, then the increase in costs would be

$$\text{Increase in cost} = C(1001) - C(1000) = \frac{C(1001) - C(1000)}{1001 - 1000}.$$

Note from the final term on the RHS that this is just the average rate of change of cost between 1000 and 1001 loaves, or approximately the derivative of $C(q)$ at $q = 1000$, which is the instantaneous rate of change of cost at $q = 1000$. Since we are interested in the effect of 'small' changes to the

business, many economists define the *marginal cost*, M_C, as the instantaneous rate of change of costs - *i.e.*

$$M_C = \frac{dC}{dq}.$$

Similarly, the additional revenue we would receive from increasing our bread production by one loaf would be

$$\text{Increase in revenue} = R(1001) - R(1000) = \frac{R(1001) - R(1000)}{1001 - 1000}.$$

Hence we similarly define the marginal revenue, M_R as

$$M_R = \frac{dR}{dq}. \tag{2.2}$$

If the marginal revenue exceeds the marginal cost, then producing the one extra loaf increases our profits, and we should make the change. If the marginal cost exceeds the marginal revenue, baking the extra loaf would reduce our profits.

Of course, what we really care about as company directors is profit: this is what we really want to maximise. From our previous discussions of extrema, we know the maximum profit will occur either at an end point ($q = 0$ or $q = q_{max}$, the maximum amount of goods that can be produced) or at a critical point of $\pi(q)$ - *i.e.* a point where

$$\frac{d\pi}{dq} = 0.$$

But

$$\pi(q) = R(q) - C(q),$$

so

$$\frac{d\pi}{dq} = \frac{dR}{dq} - \frac{dC}{dq} = M_R - M_C = 0.$$

Thus, the maximum profit can occur when the marginal cost equals the marginal revenue ($M_C = M_R$).

2.9 Summary of learning outcomes

Now that we have reached the end of this chapter, you should be able to:

- Explain the concept of the derivative of a function in terms of the rate of change, or geometrically, the slope of the tangent line
- State the definition of the derivative of a function
- Give examples of functions which are, and which are not, differentiable
- Interpret the meaning of the derivative in the context of a practical problem, including specifying its units
- Differentiate common functions, such as polynomials, exponentials, logarithms and trigonometric functions

- Recall the rules of differentiation (linear combination rule, Product Rule, Quotient Rule and Chain Rule)
- Use the rules of differentiation (possibly in combination) to differentiate complicated functions
- Differentiate implicitly, and thus differentiate inverse functions (including inverse trigonometric functions)
- Use implicit differentiation to find relationships between related rates of change (related rates) and apply this knowledge to solve practical problems
- Define local and global extrema, and critical points.
- Use the properties of critical points and local extrema to find global maxima and minima of differentiable functions.
- Apply your knowledge to solve practical optimisation problems

Chapter 3

Integration

We started this course by considering functional relationships between variables. In the last chapter we saw how, once we have such a functional relationship, we are also able to obtain information about the rates of change of the variables by differentiating. For example, if we know the velocity of a vehicle, v, as a function of time, t, we can determine its acceleration $a = \frac{dv}{dt}$ by differentiation. However, there are many applications where we would want to reverse this process - *e.g.* we might know the acceleration of a vehicle from Newton's Second Law ($F = ma$), but need to know its velocity. In such a case, we need to use integration. **Lecture 13**

You will have already come across the concept of integration at school, or in earlier courses. It will probably have been explained to you that integration is the inverse operation of differentiation (*i.e.* if we take a function $f(x)$ integrate it with respect of x and then differentiate the result, we get back our original function, $f(x)$). You might also have learned that integration gives the area under a curve $y = f(x)$. Some of you may have been told that 'integration is a process of summation'. All of these things are true, but at first sight it is not obvious that they should correspond to the same mathematical operation. Thus, to begin this section on integration, we look at some examples to revisit these ideas, and define precisely what we mean by terms which you may already have heard, such as definite and indefinite integral.

3.1 Finding the displacement of a vehicle from the velocity

Example 3.1. Suppose a car is moving in a straight line with increasing velocity, and its velocity is recorded each second for five seconds:

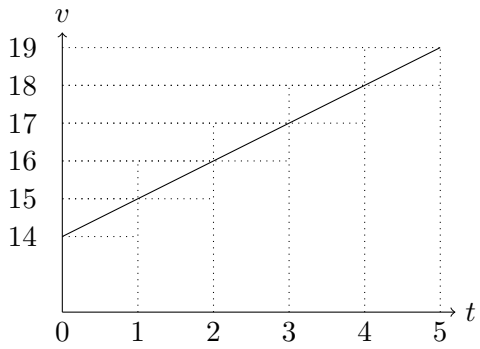
Time t (sec)	0	1	2	3	4	5
Velocity v (m/sec)	14	15	16	17	18	19

If the car starts at $x = 0$, where is it at the end of these five seconds?

We know velocity = displacement / time

or Displacement = velocity $\times$ time.

If the velocity was constant over the five seconds, then we would easily calculate the displacement as the product of velocity and time. However, the velocity is increasing.



Intuitively, we can work out an estimate of the displacement by assuming that over a short interval (say, one second), the velocity will be roughly constant. Then, by multiplying the velocity by that time, we have the distance moved by the car over that one second. Adding up all the contributions from each of the first five seconds will give an estimate of how far the car travelled in that time.

Taking the maximum possible velocity over the five one-second time intervals gives an upper bound (overestimate) for the displacement:

$$\begin{aligned} U_5 &= 15 \cdot 1 + 16 \cdot 1 + 17 \cdot 1 + 18 \cdot 1 + 19 \cdot 1 \\ &= 85\,\text{m} \end{aligned}$$

(The distance travelled at 15 m/sec for 1 sec is the area of the rectangle and is $15 \times 1 = 15$ m.)

Similarly, taking the minimum velocity on each time interval gives a lower bound (underestimate):

$$\begin{aligned} L_5 &= 14 \cdot 1 + 15 \cdot 1 + 16 \cdot 1 + 17 \cdot 1 + 18 \cdot 1 \\ &= 80\,\text{m}. \end{aligned}$$

We know the car travelled between 80 m and 85 m from the starting point, a difference of 5 m. If we estimate the displacement as the midpoint between $U_5 = 85$ and $L_5 = 80$, namely 82.5 m, then we know that:

$$x = 82.5 \underbrace{\pm 2.5}_{\text{possible error}}\,\text{m}.$$

Now, suppose that in this case, we know the functional relationship between v and t for any value of t in $[0, 5]$, rather than just knowing the velocity at one second intervals. We can quickly check that the function $v(t) = 14 + t$ is consistent with the values in the table. Now that we know the value of $v(t)$ for any $t \in [0, 5]$, we could produce more accurate estimates of the distance travelled by dividing the time into more than 5 intervals, since as the time intervals get smaller, the speed of the car will change less over that time, and our approximation that the speed is constant over the interval comes closer to being true.

If we plot the function on a graph, we can see that, geometrically, what we have done when we were estimating the distance travelled was to estimate the area under the line $v(t)$ by adding up the area of rectangles. Since the function is a straight line, we can calculate this area under the graph (which represents the distance) precisely, as we know how to calculate areas of rectangles and triangles. We obtain: $x = 14 \times 5 + 0.5 \times 5 \times 5 = 82.5$. Note that this is consistent with our previous estimates.

Now let us consider the function $G(t) = 14t + \frac{1}{2}t^2$; differentiating gives $\frac{dG}{dt} = 14 + t = v(t)$. Hence,

$$x = \int_0^5 v(t)\,dt = [14t + \frac{1}{2}t^2]_0^5 = G(5) - G(0) = 14 \times 5 + 0.5 \times 5 \times 5 = 82.5.$$

This simple example is important because it demonstrates the connection we discussed between sums (we added up the distances travelled by the car over each small time interval to estimate the displacement), finding the area under a graph (in this case, a velocity-time graph) and the 'inverse operation' of differentiation (which gave us the function $G(t)$). In the final stage, we used the notation for a definite integral, which you are probably already familiar with, although we have not yet precisely defined what it means. However, having demonstrated the connection between the three concepts, we now need to decide how to define precisely what integration is.

In some ways, it would be convenient to define $G(t)$ as 'the integral' of $v(t)$. In the past, you have probably been used to thinking of integration in this way - the problem of finding the integral of a certain function $v(t)$ simply came down to finding some function $G(t)$, such that $G'(t) = v(t)$. (We call $G(t)$ an *antiderivative* for $v(t)$; we will define this term precisely later.) For our example, it was easy to compute a suitable $G(t)$, since $v(t)$ was a linear function. But what if $v(t)$ were a more complicated function - say, $v(t) = \sqrt{1 + t^3 + t^4}$? Could you find a suitable $G(t)$ in that case?* If you cannot, does that mean we are unable to calculate the distance travelled from $t = 0$ to $t = 5$ for this velocity function? Of course not! We can use the summation procedure we used earlier to find upper and lower bounds on the distance travelled, and by making the time intervals we consider small enough, we can determine the answer to any desired degree of accuracy.

As there are many more functions on which we can use the summation procedure than for which we can find antiderivatives it makes sense to use this procedure to define integration, since it gives a more widely-applicable definition. This is important, because there are many practical problems where we need to be able to calculate integrals (like the distance in this example), but for which there is no known antiderivative. In the next section, we will use the summation procedure to give a general definition of a definite integral. Then, we will show that the familiar method you have used for computing them (*i.e.* finding an antiderivative) gives a result consistent with this definition. But, before we can do all that, we need to introduce some notation.

3.2 Summation notation

In the previous example, we estimated the areas under the curve using sums. For example, we had

$$\begin{aligned} U_5 &= 15 \cdot 1 + 16 \cdot 1 + 17 \cdot 1 + 18 \cdot 1 + 19 \cdot 1 \\ &= 85\,\text{m} \end{aligned}$$

If we want to calculate areas using this method we potentially need to evaluate sums with large numbers of terms, which it would be tedious to write out in full. Summation notation provides a convenient and concise way to express sums of this type. You have probably already met this notation at school when you studied statistics (*e.g.* in calculations of mean and variance). However, we will briefly revise it here for the sake of completeness.

*Before you waste too much effort trying to find one, it is worth noting that an explicit formula for $G(t)$ does not exist for this $v(t)$!

If $a_1, a_2, \ldots, a_n$ are real numbers, the sum of $a_1, \ldots, a_n$ is written

$$\sum_{i=1}^{n} a_i = a_1 + a_2 + \cdots + a_n.$$

"i" is called the "index" of the sum. The symbol $\sum$ is a capital sigma (a Greek letter approximating to s); this tells us we need to form a sum. The terms of the sum appear to its right. The value of i at the bottom of the $\sum$ symbol tells us what value of i to use in the first term of the sum. Then, we add one to i to get the second term , and so on, until i reaches the value at the top of the $\sum$ symbol, which will give the last term in the sum. The letter i is a "dummy index" as we can change it without altering the sum:

$$\sum_{i=1}^{n} a_i = \sum_{j=1}^{n} a_j = \sum_{k=1}^{n} a_k = a_1 + a_2 + \cdots + a_n.$$

The value of a_i will be a function of i, $a_i = f(i)$ so the sum is $\sum_{i=1}^{n} f(i) = f(1)+f(2)+\cdots+f(n)$. Hence, we can see that using summation notation, our sum from earlier can be written

$$U_5 = \sum_{i=1}^{5} v(i),$$

where $v(i) = 14 + i$.

Examples

1. (a) $\sum_{i=1}^{4} i^2 = 1^2 + 2^2 + 3^2 + 4^2 = 30$
   (b) $\sum_{i=3}^{6} i = 3 + 4 + 5 + 6 = 18$
   (c) $\sum_{j=0}^{3} 2^j = 2^0 + 2^1 + 2^2 + 2^3 = 1 + 2 + 4 + 8 = 15$
   (d) $\sum_{i=1}^{4} 2 = 2 + 2 + 2 + 2 = 8$

2. Write $2^3 + 3^3 + \cdots + n^3$ in sigma notation.

   There are many ways to write a sum:

$$\begin{aligned} 2^3 + 3^3 + \cdots + n^3 &= \sum_{i=2}^{n} (i)^3 \\ &= \sum_{j=1}^{n-1} (j+1)^3 \\ &= \sum_{k=3}^{n+1} (k-1)^3 \end{aligned}$$

Properties of $\sum$:

1. $\sum_{i=m}^{n} ca_i = c\sum_{i=m}^{n} a_i$, for c a constant;
2. $\sum_{i=m}^{n}(a_i + b_i) = \sum_{i=m}^{n} a_i + \sum_{i=m}^{n} b_i$;
3. $\sum_{i=m}^{n}(a_i - b_i) = \sum_{i=m}^{n} a_i - \sum_{i=m}^{n} b_i$.
4. But $\sum_{i=m}^{n}(a_i \times b_i) \quad \neq \quad (\sum_{i=m}^{n} a_i) \times (\sum_{i=m}^{n} b_i)$ in general;
5. and $\sum_{i=m}^{n} \frac{a_i}{b_i} \quad \neq \quad \frac{\sum_{i=m}^{n} a_i}{\sum_{i=m}^{n} b_i}$ in general.

These can each be verified by writing out the sums on each side

Some important sums are those of constants, linear and squares.

1. $\sum_{i=1}^{n} 1 = \underbrace{1 + \cdots + 1}_{n} = n.$
2. $\sum_{j=0}^{n} ar^j = a + ar + ar^2 + \cdots + ar^n = a\frac{1-r^{n+1}}{1-r}$ is the geometric sum.
3. $\sum_{i=1}^{n} i = 1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2}$

   Derivation of formula

$$\begin{aligned}\sum_{i=1}^{n} i &= 1 + 2 + \cdots + n \\ &= n + n - 1 + \cdots + 1\end{aligned}$$

   Adding:

$$2\sum_{i=1}^{n} i = \underbrace{(n+1) + (n+1) + \cdots + (n+1)}_{n} = n(n+1)$$

$$\sum_{i=1}^{n} i = n\frac{(n+1)}{2}.$$

4. $\sum_{i=1}^{n} i^2 = \frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6} = \frac{n}{6}(2n^2 + 3n + 1) = \frac{n}{6}(2n+1)(n+1)$

   Proof

   We can show this formula is correct using proof by induction. We begin by checking it is correct for $n = 1$: in this case, the sum is 1, and the formula gives

$$\frac{1}{6}(2+1)(2) = \frac{6}{6} = 1.$$

Hence, the formula is true for $n = 1$. We now assume that it is correct for $n = k$, and consider what happens when $n = k + 1$. We have

$$\begin{aligned}\sum_{i=1}^{k+1} i^2 &= \sum_{i=1}^{k} i^2 + (k+1)^2 = \frac{k}{6}(2k+1)(k+1) + (k+1)^2 \\ &= \frac{(k+1)}{6}[k(2k+1) + 6(k+1)] = \frac{(k+1)}{6}[2k^2 + k + 6k + 6] = \frac{(k+1)}{6}[2k^2 + 7k + 6] \\ &= \frac{(k+1)}{6}(2k+3)(k+2) = \frac{(k+1)}{6}(2[k+1]+1)([k+1]+1)\end{aligned}$$

Hence, if the formula is true for $n = k$ it is also true for $n = k + 1$. Since we have already shown it is true for $n = 1$, this completes our proof by induction.

Example 3.2. Evaluate $\sum_{i=3}^{10}(i+2)^2$.

$$\begin{aligned}\sum_{i=3}^{10}(i+2)^2 &= \sum_{j=5}^{12} j^2 \quad \text{where } j = i+2 \\ &= \sum_{j=1}^{12} j^2 - \sum_{j=1}^{4} j^2 \\ &= \frac{12}{6} \cdot 25 \cdot 13 - \frac{4}{6} \cdot 9 \cdot 5 = 620.\end{aligned}$$

Lecture 14

3.3 Defining the definite integral

Now that we have introduced summation notation, we are in a position to move towards defining integration more precisely. Specifically, we will begin by considering **definite integrals** - these are integrals over a stated range, $[a, b]$ (where a, $b \in \mathbb{R}$). Let f be a real valued function defined on the interval $[a, b]$. We denote a definite integral, I as

$$I = \int_a^b f(x)\,dx.$$

I is a unique real number which depends on the function f being integrated (called the *integrand*), and on the values of a and b.

Can we find I given any function f and interval $[a, b]$?

The answer to this question is, 'No'; not every function is *integrable*. Some functions are not integrable for any values of a and b; in other cases, a function might be integrable on some intervals, but not others. If the function is continuous on $[a, b]$, it can be shown to be integrable. However, continuity of f is not a requirement; for example, we can integrate the Heaviside function $H(x)$ between -1 and 1, and obtain

$$\int_{-1}^{1} H(x)\,dx = 1.$$

We will see why this is true shortly.

Note that, for a definite integral, the variable, x, with respect to which we are integrating, is a 'dummy variable', much like the index i was a 'dummy index' when we used summation notation; we can replace x with any other symbol without changing the result, I. Thus

$$I = \int_a^b f(x)\,dx = \int_a^b f(t)\,dt = \int_a^b f(\theta)\,d\theta.$$

In order to define the definite integral using sums, we need to follow these steps:

1. We divide the interval $[a, b]$ into n equal subintervals. Each subinterval will have width $\Delta x = (b-a)/n$. We denote the end points of the subintervals by

$$a = x_0 < x_1 < x_2 < \ldots < x_n = b.$$

2. On each subinterval $[x_{i-1}, x_i]$ let

$$m_i = \text{minimum value of } f \text{ on } [x_{i-1}, x_i],$$
$$M_i = \text{maximum value of } f \text{ on } [x_{i-1}, x_i].$$

   If the function f is continuous on $[a, b]$, then m_i and M_i are guaranteed to exist on each subinterval. If f is not continuous, either or both of them may fail to exist on one or more subintervals.

3. Then, we define the lower and upper sums, as we did for the distance-finding example:

$$\begin{aligned} \text{Lower sum} \quad L_n &= m_1\Delta x + m_2\Delta x + \cdots + m_n\Delta x \\ &= \sum_{i=1}^{n} m_i\Delta x = \left(\sum_{i=1}^{n} m_i\right)\Delta x \\ \text{and Upper sum} \quad U_n &= M_1\Delta x + M_2\Delta x + \cdots + M_n\Delta x \\ &= \sum_{i=1}^{n} M_i\Delta x = \left(\sum_{i=1}^{n} M_i\right)\Delta x. \end{aligned}$$

   Note that $L_n \le U_n$ for all values of n (since each term in the lower sum is less than or equal to the equivalent term in the upper sum). In fact, $L_m \le U_n$ for any numbers m and n. We can understand this intuitively if we imagine the graph of a continuous function, $f(x)$, and visualise the rectangles representing each of the terms in the sum (as we drew for the distance travelled example). The lower sum is always less than the area below the curve, no matter how many subintervals are used, because all the rectangles lie below the curve, whereas the upper sum is always greater than the area below the curve.

Definition 3.1. The *definite integral*, I, of a function, f, from a to b is defined to be the unique real number which satisfies

$$L_n \le I \le U_n \quad \text{for all } n = 1, 2, 3, \ldots$$

where such a number exists. We write this as:

$$I = \int_a^b f(x)\,dx.$$

If I exists, then we say the function f is *integrable* over $[a, b]$.

Note that the symbol we use to denote an integral, $\int$ is actually an elongated letter s, chosen because of the close connection between integrals and sums.

Riemann sums An alternative way of defining the definite integral, which you might meet in textbooks, is based on *Riemann sums* (named after German mathematician Bernhard Riemann). Instead of 'trapping' the value of I between the upper and lower sums, we approximate it by introducing

$$S_n = \sum_{i=1}^{n} f(x_i^*)\Delta x.$$

where x_i^* is any value in the i^{th} interval (*i.e.* $x_{i-1} < x_i^* < x_i$). We then define

$$I_{Riemann} = \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i^*)\Delta x.$$

Now, if we assume that f is a continuous function on $[a, b]$, we can show that this definition is equivalent to Definition 3.1. Note that for any n, $L_n \leq S_n \leq U_n$. As $n \to \infty$, $x_i \to x_{i+1}$, so, $\lim_{n\to\infty} f(x_i^*) = f(x_i)$; similarly, $\lim_{n\to\infty} m_i = \lim_{n\to\infty} M_i = f(x_i)$ (the latter two equalities follow from the fact that M_i and m_i are the images under f of two points in the i^{th} interval). Hence $\lim_{n\to\infty} L_n - S_n = \lim_{n\to\infty} U_n - S_n = 0$ and so, using the inequality in Definition 3.1 $\lim_{n\to\infty} I - S_n = 0$. Hence,

$$I = \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i^*)\Delta x = I_{Riemann}.$$

What functions can we integrate? We have not clearly stated what types of functions are 'integrable', except for saying that they are those functions for which the definite integral exists (a rather circular definition!). A precise definition is outside the scope of this course. However, it can be shown from the definition that if $f(x)$ is a continuous function on $[a, b]$, then it is integrable. Continuity of f on $[a, b]$ ensures that the values m_i and M_i exist for every sub-interval in $[a, b]$ (by the Extreme Value Theorem), which is necessary for our definition of I to make sense. However, continuity is not a requirement. We say that a function f is *bounded* on the interval $[a, b]$ if there is some real number M such that for all $x \in [a, b]$, $-M < f(x) < M$. As long as the function is bounded on $[a, b]$, it can have *finitely many* jump discontinuities and the definite integral will still exist.

For example, the Heaviside function, $H(x)$, is integrable over any finite interval $[a, b]$ where $a\, b \in \mathbb{R}$, with

$$I = \int_a^b H(x)dx = \begin{cases} (b-a) \text{ if } a, b \geq 0 \\ b \text{ if } a < 0 \leq b \\ 0 \text{ if } a, b < 0 \end{cases}$$

If $a, b \geq 0$, then we have $m_i = M_i = 1$ on every subinterval, and so $L_n = U_n = \dfrac{(b-a)}{n} \displaystyle\sum_{i=1}^{n} 1 = b - a$. Hence we must have $I = b - a$. Similarly, if $a, b < 0$, $m_i = M_i = 0$ on every subinterval, and $L_n = U_n = \dfrac{(b-a)}{n} \displaystyle\sum_{i=1}^{n} 0 = 0$, so $I = 0$. If $a < 0 \leq b$, we split the interval $[a, b]$ into $[a, 0)$ and $[0, b]$. Let $[a, 0)$ be divided into n_1 subintervals. On each subinterval, we have $m_i = M_i = 0$. We similarly subdivide on $[0, b]$ into n_2 subintervals, on each of which we have $m_i = M_i = 1$. Set $n = n_1 + n_2$; then, adding the contributions from both $[a, 0)$ and $[0, b]$, we have $L_n = U_n = \dfrac{a}{n_1} \displaystyle\sum_{i=1}^{n_1} 0 + \dfrac{b}{n_2} \sum_{i=1}^{n_2} 1 = b$, so $I = b$. Hence the class of integrable functions is larger than the class of continuous functions (which in turn is larger than the class of differentiable functions).

However, not all functions are integrable. For example, the function $f(x) = x^{-1}$ is not integrable over $[0, 1]$ because it is undefined at 0. The Dirichlet function is not integrable over any finite interval, $[a, b]$, despite the fact that it is defined for all real numbers, and $0 \leq D(x) \leq 1$. We can show this as follows. Suppose that we divide $[a, b]$ into n subintervals. Now, $m_i = 0$ for all i, since there will be an irrational number in $[x_{i-1}, x_i]$. Similarly, $M_i = 1$ for all i, since there will be an rational number in $[x_{i-1}, x_i]$. Thus $L_n = 0$ and $U_n = \Delta x \displaystyle\sum_{i=1}^{n} 1 = \dfrac{(b-a)}{n} \sum_{i=1}^{n} 1 = (b - a)$ for all values of n. This means that there is no *unique* real number I with $L_n \leq I \leq U_n$ for all n (any number between 0 and $b - a$ would be equally good!).

Properties of definite integrals

1. $\displaystyle\int_a^a f(x)\, dx = 0$. This follows immediately from the definition of the definite integral.

2. $\displaystyle\int_b^a f(x)\, dx = -\int_a^b f(x)\, dx$.

3. If f is integrable on $[a, b]$ and $a < c < b$, then

$$\int_a^b f(x)\, dx = \int_a^c f(x)\, dx + \int_c^b f(x)\, dx.$$

4. The average value, $\bar{f}$, of a function f over the interval $[a, b]$ is defined to be

$$\bar{f} = \frac{1}{(b-a)} \int_a^b f(x)\, dx.$$

To see why this definition makes sense, consider the following. We know that the average (mean) of n numbers is simply the sum of those numbers, divided by n. Suppose we divide our interval $[a, b]$ into n subintervals, each of length $\Delta x = (b-a)/n$. Now, let x_i^* be a number in the i^{th} subinterval (we could choose the mid-point, for example). Then, we could calculate the (approximate) average value of f over $[a, b]$ as:

$$\bar{f} \approx \frac{f(x_1^*) + f(x_2^*) + f(x_3^*) + \cdots + f(x_n^*)}{n}.$$

But, we have $n = (b-a)/\Delta x$. Substituting this in, we get

$$\begin{aligned}
\frac{f(x_1^*) + f(x_2^*) + f(x_3^*) + \cdots + f(x_n^*)}{n} &= \frac{f(x_1^*) + f(x_2^*) + f(x_3^*) + \cdots + f(x_n^*)}{\frac{(b-a)}{\Delta x}} \\
&= \frac{(f(x_1^*) + f(x_2^*) + f(x_3^*) + \cdots + f(x_n^*))\Delta x}{(b-a)} \\
&= \frac{f(x_1^*)\Delta x + f(x_2^*)\Delta x + f(x_3^*)\Delta x + \cdots + f(x_n^*)\Delta x}{(b-a)} \\
&= \frac{1}{(b-a)} \sum_{i=1}^{n} f(x_i^*)\Delta x
\end{aligned}$$

Now, as n gets larger and larger (*i.e.* we are evaluating f at more and and more points in the interval), our approximation of the average value will improve. Thus, we let

$$\bar{f} = \frac{1}{(b-a)} \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i^*)\Delta x,$$

But, comparing the RHS of the expression above to the definition of the definite integral using Riemann sums, we can see that

$$\bar{f} = \frac{1}{(b-a)} \int_a^b f(x)\, dx.$$

3.3.1 Using summation to calculate definite integrals

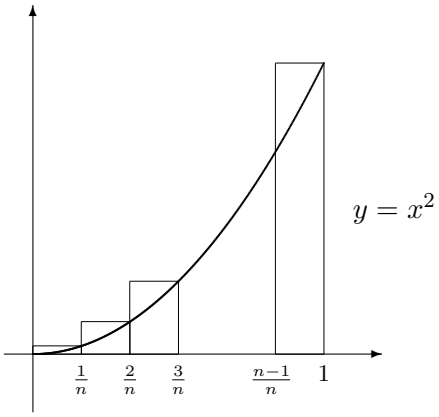
Lecture 15 An important advantage of defining integration in terms of summation is that it provides a way to calculate definite integrals numerically for any integrable function. In order to demonstrate that the definition works in the way we expect, we consider the problem of finding the area between $y = x^2$ and the x-axis over the interval $x = 0$ to $x = 1$.

Numerically, we can see the integral takes the value ≈ 0.333. In fact, we can find this definite integral exactly using the definition, and our knowledge of summation notation from earlier.

Following the steps from the definition of the definite integral, we begin by partitioning the interval $[0, 1]$ into n equal subintervals:

$$\left[0, \frac{1}{n}\right], \left[\frac{1}{n}, \frac{2}{n}\right], \left[\frac{2}{n}, \frac{3}{n}\right], \ldots, \left[\frac{n-1}{n}, 1\right].$$

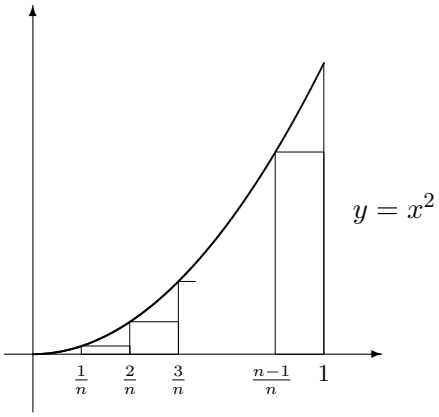
On each subinterval $[\frac{i-1}{n}, \frac{i}{n}]$ let m_i and M_i respectively denote the minimum value and maximum value of $y = f(x)$ on that interval.



Adding the areas of all rectangles of height M_i and base $\frac{1}{n}$ gives the upper sum.

Calculating the upper sum

$$\begin{aligned} U_n &= \frac{1}{n} \cdot \left(\frac{1}{n}\right)^2 + \left(\frac{2}{n}\right)^2 \cdot \frac{1}{n} + \cdots + (1)^2 \frac{1}{n} \\ &= \left[\left(\frac{1}{n}\right)^2 + \left(\frac{2}{n}\right)^2 + \left(\frac{3}{n}\right)^2 + \cdots + \left(\frac{n}{n}\right)^2\right] \frac{1}{n} \\ &= [1^2 + 2^2 + 3^2 + \cdots + n^2] \frac{1}{n^3} \end{aligned}$$



Similarly, adding the areas of all rectangles of height m_i and base $\frac{1}{n}$ gives the lower sum.

Calculating the lower sum

$$\begin{aligned} L_n &= \frac{1}{n}\cdot 0^2 + \frac{1}{n}\left(\frac{1}{n}\right)^2 + \cdots + \frac{1}{n}\left(\frac{n-1}{n}\right)^2 \\ &= \left[0^2 + \left(\frac{1}{n}\right)^2 + \cdots + \left(\frac{n-1}{n}\right)^2\right]\frac{1}{n} \\ &= \left[0^2 + 1^2 + \cdots + (n-1)^2\right]\frac{1}{n^3} \end{aligned}$$

$$\begin{array}{ccccc} L_n & \leq & A & \leq & U_n \\ \text{area of rectangles} & & \text{area under} & & \text{area of rectangles} \\ \text{below the curve} & & \text{curve} & & \text{above curve} \end{array}$$

As $n \to \infty$, that is, the width of the rectangles gets smaller and smaller, we expect the upper and lower sums to get closer together and better approximate the area.

What is the area A?

We saw earlier that

$$\sum_{i=1}^{n} i^2 = 1^2 + 2^2 + 3^3 + \cdots + n^2 = \frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}$$

$$\begin{aligned} \text{Thus} \quad U_n &= \left(\frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}\right)\frac{1}{n^3} \\ &= \frac{1}{3} + \frac{1}{2n} + \frac{1}{6n^2}. \end{aligned}$$

We apply the same formula to L_n:

$$\begin{aligned} &0^2 + 1^2 + 2^2 + \cdots + (n-1)^2 \\ &= 1^2 + 2^2 + \cdots + (n-1)^2 + n^2 - n^2 \\ &= \frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6} - n^2 = \frac{n^3}{3} - \frac{n^2}{2} + \frac{n}{6} \\ L_n &= \left[\frac{n^3}{3} - \frac{n^2}{2} + \frac{n}{6}\right]\frac{1}{n^3} \\ &= \frac{1}{3} - \frac{1}{2n} + \frac{1}{6n^2}. \end{aligned}$$

$$\begin{array}{rl} L_n & \leq A \leq \quad U_n \\ \dfrac{1}{3} - \dfrac{1}{2n} + \dfrac{1}{6n^2} & \leq A \leq \quad \dfrac{1}{3} + \dfrac{1}{2n} + \dfrac{1}{6n^2} \end{array}$$

As $n \to \infty$, $\quad L_n \to \frac{1}{3}, \quad U_n \to \frac{1}{3}, \quad$ so $\quad A = \frac{1}{3}$.

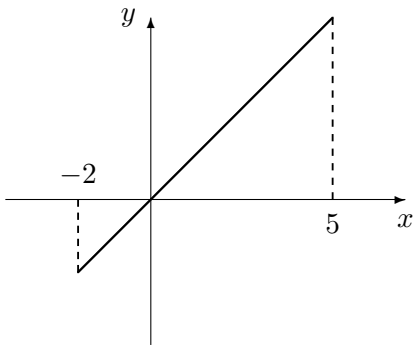
This result is consistent with what we have found numerically; as we use larger and larger values of n our numerical approximation of the integral becomes closer and closer to the exact value we have just found.

3.3.2 Definite integrals and areas

In our earlier examples, we considered functions which take only positive values. The definition of the definite integral above does not require the values of $f(x)$ to be positive; nothing is changed if either or both of the maximum and minimum values (M_i and m_i) are negative on any sub-interval. In many applications, such as finding the displacement of a vehicle from the velocity, the fact that some terms in the sum would be negative makes sense: if the velocity was negative for some part of the time, the vehicle would be moving backwards, and hence it would not end up so far from the starting point. The negative terms in the sum would cancel out some of the positive ones, so the final answer would be smaller. However, we do need to think a little more carefully about our interpretation of the integral as the area under the relevant graph. If some of the M_i or m_i are negative, then the corresponding value $M_i\Delta x$ or $m_i\Delta x$ must be also; but areas cannot be negative.

We can resolve this problem by interpreting the definite integral as representing the 'signed area': *i.e.* over regions where $f(x) > 0$, the definite integral gives the area between the curve and the x-axis; where $f(x) < 0$, the definite integral gives minus the area. Thus, if we wish to use the integral to compute areas when $f(x) \leq 0$, $a \leq x \leq b$, then the (positive) area *between* $y = f(x)$ and the x-axis is given by $-\int_a^b f(x)\,dx$.

Example 3.3. Find the area between $y = x$ and the x-axis for $-2 \leq x \leq 5$.



$$\text{Area} = -\int_{-2}^{0} x\,dx + \int_{0}^{5} x\,dx = \left(\frac{1}{2}\times 2\times 2\right) + \frac{1}{2}(5\times 5)$$
$$= \frac{29}{2} \quad \text{(using areas of triangles).}$$

However, this difficulty only really arises if we become too wedded to the idea that a definite integral gives 'the area under the curve'. Whilst these area-finding problems are common in school textbooks, in most real-life applications, calculating the definite integral of a function does not yield an area. For example, if the integrand is a velocity (in ms^{-1}, the integral with respect to time (measured in seconds) gives a displacement (in m). If the integrand $f(x)$ represent the force (in N) acting in the x-direction on a body at position x (measured in metres from an origin), then $\int_a^b f(x)\,dx$ gives the work done (in J) in moving the body from $x = a$ to $x = b$. We can represent quantities like $v(t)$ or $f(x)$ using graphs, and then the area between the curve and the horizontal axis gives a helpful visual representation of the integral. But they are not areas, as is clear from the units.

You will find it much easier to understand some of the concepts in later courses if you keep in mind how integrals are really defined: as a limit of a sum, the terms of which consist of the value

of a function over some small region, multiplied by the size of that region. In the examples we have seen so far the region is an interval, and so the 'size' of the region is simply its length. However, this quite general way of thinking about integration allows us to extend the idea of integration on part of the real line, as we have been doing here, to integrating along an arbitrary curve, or over an area or volume.

For example, consider a cube with sides of length a m, composed of a material that has density ρ kg m^{-3}. If the density, ρ is constant, then we can calculate the mass of the cube, M, by multiplying the density by the cube's volume: $M = \rho a^3$ kg. But what if the density of the cube varies with position (*e.g.* the material is heavier near the bottom than near the top)? In that case, we could find the mass of the cube by subdividing it into smaller cubes, and finding the minimum and maximum values of the density in each small cube. Then, upper and lower bounds for the mass of each small cube could be calculated as the product of the maximum (or minimum) density and the volume of the small cube. On summing up the masses of each small cube, we would obtain upper and lower bounds for the mass of the large cube, M. The smaller we make the little cubes, the better our estimate of M would become. Notice that our process here is basically the same as that we used to define the definite integral of a function $f(x)$ on the interval $[a, b]$. The difference here is that our function, ρ (the density) could potentially depend on x, y and z, and instead of sub-intervals of $[a, b]$, we sum up the contribution from sub-volumes (small cubes) of our region $0 \leq x \leq a$, $0 \leq y \leq a$, $0 \leq z \leq a$. In the limit that the volume of the small cubes tends to zero, we would obtain the mass, M as a **volume integral**, written as

$$M = \int_0^a \int_0^a \int_0^a \rho(x, y, z)\, dx\, dy\, dz.$$

These kinds of extensions are outside the scope of this course. However, they help to illustrate why we have defined integration using sums, which may seem rather convoluted at first.

3.4 Antiderivatives and The Fundamental Theorem of Calculus

Lecture 16 We have now seen that we can calculate integrals (at least approximately) using the definition for a very wide range of functions. But, no matter whether we do it numerically using the computer (by taking increasingly large numbers of subdivisions, n, until we reach the level of accuracy required) or analytically by working out the sums L_n and U_n, and taking the limit $n \to \infty$ (which is only possible for a small class of relatively simple integrals), the procedure is long-winded and tedious. It would be nice to avoid it if possible.

Fortunately, it very often can be avoided. In the earlier example, to find the displacement x of the car, we calculated the definite integral

$$x = \int_0^5 v(t)\, dt = \int_0^5 (14 + t)\, dt.$$

We did this by finding a function $G(t)$ that satisfied $G'(t) = v(t)$; for this example, $G(t) = 14t + \frac{1}{2}t^2$ is such a function. Then,

$$x = \int_0^5 v(t)\, dt = [G(t)]_0^5 = G(5) - G(0).$$

This method is how you have been used to calculating integrals at school. Similarly, if you had been asked to find the area in the last example without being told which method to use, you would probably have gone straight ahead and calculated it using

$$A = \int_0^1 x^2\,dx = \left[\frac{x^3}{3}\right]_0^1 = \frac{1}{3}(1^3 - 0^3) = \frac{1}{3}.$$

We will now show that this method of finding a function whose derivative is the integrand gives an answer consistent with the summation method we have used to define integration. But first, we need to introduce the following definition.

Definition 3.2. A function G is an **antiderivative** of f on (a, b) if $G'(x) = f(x)$, $a < x < b$.

Note: the term *primitive* can also be used instead of antiderivative.

If $G(x)$ is an antiderivative for $f(x)$, we note from the rules of differentiation that $G(x) + c$ (where c is a constant - *i.e.* any real number) is also an antiderivative for $f(x)$, because

$$\frac{d}{dx}(G + c) = \frac{dG}{dx} = f(x).$$

In fact, *all antiderivatives for a function $f(x)$ on an interval can differ from each other only by a constant.* We can demonstrate this fact as follows. Suppose that we have two antiderivatives for $f(x)$ on our interval $[a, b]$, $G_1(x)$ and $G_2(x)$. Then

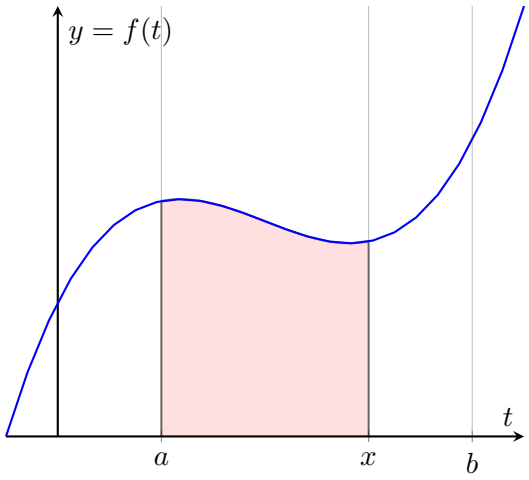
$$\frac{d}{dx}(G_1 - G_2) = \frac{dG_1}{dx} - \frac{dG_2}{dx} = f(x) - f(x) = 0.$$

Since the derivative of $G_1 - G_2$ is zero, we must have $G_1(x) - G_2(x) = c$, where c is a constant, and thus $G_1(x) = G_2 + c$.

Now, consider a function $f(t)$ which is *continuous* on the interval $t \in [a, b]$, and let x be some value in this interval. For any x, we can calculate the definite integral of f from a to x using Definition 3.1. The answer will depend on the particular value of x chosen. Hence we can define a function, $G(x)$, given by

$$G(x) = \int_a^x f(t)\,dt.$$

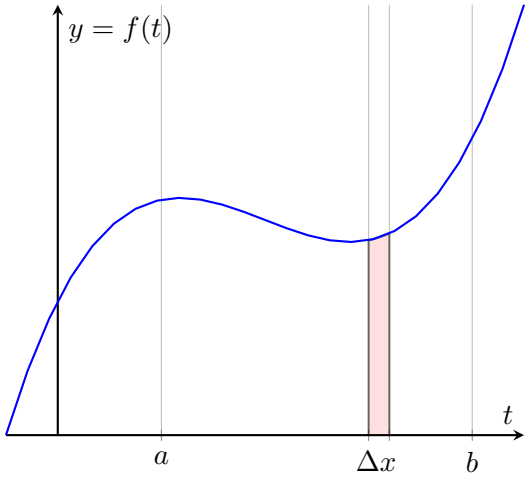
As we discussed earlier, the definite integral $G(x)$ can represented by the (signed) area between the graph $y = f(t)$ and the horizontal axis, for t between a and x, as shaded in pink on the diagram below.



Now let Δx be a small real number. We have

$$G(x + \Delta x) = \int_a^{x+\Delta x} f(t)\, dt.$$

Then, the change in G as t increases from x to $x + \Delta x$ is represented by the area of the region shown in pink in the diagram below. Since the function f is continuous, it is approximately constant over the small interval $[x, x + \Delta x]$. Thus the area of the shaded region can be approximated as a rectangle of height $f(x)$ and width Δx.



This means that

$$G(x + \Delta x) - G(x) \approx f(x)\Delta x$$

or

$$\frac{G(x + \Delta x) - G(x)}{\Delta x} \approx f(x).$$

As Δx gets smaller, the approximation will get better and better, so that in the limit $\Delta x \to 0$, we will have

$$\lim_{\Delta x \to 0} \frac{G(x + \Delta x) - G(x)}{\Delta x} = f(x).$$

But, if we look back at Definition 2.1, we can see that the left hand side of this equation is precisely the definition of the derivative of $G(x)$. Hence

$$\lim_{\Delta x \to 0} \frac{G(x + \Delta x) - G(x)}{\Delta x} = \frac{dG}{dx} = f(x).$$

Since $\frac{dG}{dx} = f(x)$, the function $G(x)$, satisfies the definition of an antiderivative of $f(x)$. The following theorem summarises what we have just shown.

Theorem 3.1 (The First Fundamental Theorem of Calculus)**.** Let $f(t)$ be a continuous function on $[a, b]$, and let $a < x < b$. Then

$$G(x) = \int_a^x f(t)\, dt$$

is an antiderivative for f on (a, b) - *i.e.*

$$\frac{dG}{dx} = f(x).$$

We now need to show that antiderivatives provide a handy way to calculate definite integrals. Let

$$G_1(x) = \int_a^x f(t)\, dt.$$

where $a < x < b$. Then, as we have just demonstrated, $\frac{dG_1}{dx} = f(x)$. We note that from the way we have defined G_1,

$$\int_a^b f(t)\, dt = G_1(b).$$

Now, as we saw earlier

$$\int_a^a f(t)\, dt = G_1(a) = 0.$$

But suppose we have another antiderivative for $f(x)$, $G_2(x)$ - can we also use this to calculate the required definite integral?

As we stated earlier, antiderivatives for the same function can differ only by a constant. Hence

$$G_1(x) = G_2(x) + c,$$

where c is a constant. Then,

$$\int_a^b f(t)\, dt = G_1(b) = G_1(b) - G_1(a) = (G_2(b) + c) - (G_2(a) + c) = G_2(b) - G_2(a) = [G_2(x)]_a^b.$$

We have thus shown that we can calculate definite integrals using **any** antiderivative for our function f, which is summarised by the following theorem.

Theorem 3.2 (The Second Fundamental Theorem of Calculus)**.** Let f be a continuous function on $[a, b]$ and G any antiderivative of f on $[a, b]$. Then

$$\int_a^b f(t)\,dt = G(b) - G(a) = [G(x)]_a^b\,.$$

This extremely powerful result gives us a much quicker and easier way of calculating integrals *when an algebraic antiderivative exists.* It is clear from our manipulations with sums earlier in this section that this is the method we would ideally use whenever possible. But it is not always possible to find such an antiderivative: for example, the integrands $f(t) = \sqrt{1 + t^3 + t^4}$ and $f(t) = \sin\left(t + \frac{1}{t}\right)$ do not have one. In that case, we would use numerical methods, which are based on the definition of the integral. These will be covered in more detail in later courses.

We can use the Fundamental Theorems of Calculus, possibly in combination with the rules of differentiation we learnt in the previous chapter to calculate the derivatives of some integrals.

Example 3.4. Find $\dfrac{dF}{dx}$ if $F(x) = \displaystyle\int_x^0 \cos t^2\,dt$.

If the limits on the integral defining $F(x)$ were the opposite way around, this would be a trivial application of the First Fundamental Theorem of Calculus (FFTC), since if $G(x) = \displaystyle\int_a^x f(t)\,dt$, then $\dfrac{dG}{dx} = f(x)$.

Here, we need to swap the order of the limits, and recalling the properties of definite integrals, we have

$$F(x) = \int_x^0 \cos t^2\,dt = -\int_0^x \cos t^2\,dt = \int_0^x (-\cos t^2)\,dt.$$

Then, the integrand is $f(t) = -\cos t^2$, and using the FFTC, we have

$$\frac{dF}{dx} = f(x) = -\cos x^2.$$

Example 3.5. Find $\dfrac{d}{dx}\displaystyle\int_{\frac{1}{2}}^{x^4} \sec t\,dt$.

In this example, we again need to find the derivative of an integral, so we expect to use the FFTC. However, the upper limit on the integral causes us a problem: if it was x, things would be straightforward, but instead we have x^4 here. Let us define

$$G(x) = \int_{\frac{1}{2}}^{x^4} \sec t\,dt.$$

Hence we are trying to find $\dfrac{dG}{dx}$.

We need to get the integral into a form where we can use the FFTC. We can do this by making a substitution, $u = x^4$, so

$$G(u) = \int_{\frac{1}{2}}^{u} \sec t\,dt.$$

Then, we know that

$$\frac{dG}{du} = \sec u,$$

by the FFTC.

In order to obtain the derivative with respect to x, we now use the Chain Rule

$$\frac{dG}{dx} = \frac{dG}{du}\frac{du}{dx} = \sec u \frac{d}{dx}(x^4) = 4x^3 \sec x^4,$$

where we substituted $u = x^4$ in the final step.

3.4.1 Indefinite integrals

So far, we have focused our attention on *definite integrals* where the range of values over which we are integrating is stated. However, you are probably more used to being asked to find *indefinite integrals*, for which no range is specified. The indefinite integral of a function $f(x)$ is written as **Lecture 17**

$$\int f(x)\, dx.$$

This denotes the collection of *all* antiderivatives of $f(x)$ (whether known algebraically or not). Thus if $G'(x) = f(x)$, that is, G is an antiderivative, then

$$\int f(x)\, dx = G(x) + c,$$

where c a constant, known as the *constant of integration*.

You will already have learned the antiderivatives of many common mathematical functions. For the rest of this course, it will be assumed you are familiar with the following indefinite integrals:

$$\int x^r\, dx = \frac{x^{r+1}}{r+1} + c, \quad r \neq -1$$

$$\int \frac{1}{x}\, dx = \ln|x| + c$$

$$\int e^x\, dx = e^x + c$$

$$\int \sin x\, dx = -\cos x + c$$

$$\int \cos x\, dx = \sin x + c$$

$$\int \sec^2 x\, dx = \tan x + c$$

$$\int \sinh x\, dx = \cosh x + c$$

$$\int \cosh x\, dx = \sinh x + c$$

$$\int \frac{dx}{\sqrt{1-x^2}} = \sin^{-1} x + c, \quad |x| < 1$$

$$\int \frac{dx}{1+x^2} = \tan^{-1} x + c.$$

Knowing the indefinite integrals of these basic functions is essential to being able to calculate the more complicated integrals you will encounter in future courses, and your professional life after university. *You should ensure you learn them.*

Now we know the antiderivatives of some basic functions, we can compute integrals of them quickly and easily. For example, if we needed to calculate $\int_0^1 x^{23}\,dx$, we can use the fact that we know $\int x^{23}\,dx = \frac{x^{24}}{24} + C$. Then, putting in the limits on the integral we obtain

$$\int_0^1 x^{23}\,dx = \left[\frac{x^{24}}{24}\right]_0^1 = \frac{1}{24} - 0 = \frac{1}{24}.$$

Notice that in the above we did not bother to include the constant of integration, since when we evaluate a definite integral it cancels between the two terms.

3.5 Integration by substitution

Although some of the integrands we encounter in applications will be of a type we can recognise, often we will find ourselves needing to integrate a function which looks more complex than those we have seen above. For example, what would we do if asked to find $\int_0^1 (2x+7)^{23}\,dx$?
A 'low-tech' solution would be to expand to get a polynomial of degree 23, but this would be very tedious, prone to error, and time-consuming. There is a better way.

Suppose that G is an antiderivative for a function, f. Then, by definition, $G'(x) = \dfrac{dG}{dx} = f(x)$. Now, suppose we have a situation where G is a function of another function - say, for example, $G(u(x))$. By the Chain Rule, we know that

$$\frac{dG}{dx} = G'(u)u'(x) = \frac{dG}{du}\frac{du}{dx}.$$

Now, using the fact that G is an antiderivative of f we have

$$\frac{d}{dx}\left(G(u(x))\right) = f(u(x))u'(x) = f(u(x))\frac{du}{dx}.$$

If we integrate this expression with respect to x we have

$$G(u(x)) + c = \int \frac{d}{dx}\left(G(u(x))\right)\,dx = \int f(u(x))\frac{du}{dx}\,dx.$$

But, we also have

$$G(u(x)) + c = G(u) + c = \int G'(u)\,du = \int f(u)\,du,$$

since, by the definition of an antiderivative, $G'(u) = f(u)$. Hence,

$$\int f(u(x))u'(x)\,dx = \int f(u)\,du. \tag{3.1}$$

Thus, if we can 'spot' the function $u(x)$, we can turn the complicated looking integral $\int f(u(x))\frac{du}{dx}\,dx$ into the much simpler-looking one $\int f(u)\,du$. Essentially, we are using the Chain Rule in reverse to calculate the antiderivative we are looking for.

The notation in equation (3.1) makes it look as if we can treat dx and du as entities separate from the rest of the notation for an integral (although we have not given them any independent meaning) and 'cancel' the factor of dx, so that

$$\frac{du}{dx}\,dx = u'(x)\,dx = du.$$

This expression provides a convenient way of 'converting' an integral with respect to x into one with respect to u in calculations, as we will illustrate with the example below.

Now let us consider again the problem of finding $\int_0^1 (2x+7)^{23}\,dx$. A first step would be to find the indefinite integral $\int (2x+7)^{23}\,dx$. The integrand can be written as a function, $f(u) = u^{23}$, where $u = 2x+7$. Thus, we have $\frac{du}{dx} = u'(x) = 2$ so, using the expression above, $du = 2dx$ or $dx = \frac{1}{2}du$. Hence we can re-write our integral as

$$\int (2x+7)^{23}\,dx = \int u^{23}\,dx = \int u^{23}\cdot\frac{1}{2}\,du = \frac{1}{2}\int u^{23}\,du = \frac{1}{2}\left(\frac{1}{24}u^{24}\right) + c = \frac{1}{48}u^{24} + c.$$

Now, we can substitute $u = 2x+7$ back into our answer, to get the integral in terms of x if we want to do so:

$$\int (2x+7)^{23}\,dx = \frac{1}{48}(2x+7)^{24} + c.$$

In this particular example, we wanted to calculate a definite integral, so we need to include the limits, which are $x = 0$ and $x = 1$. There are two ways we can handle this.

1. Change the limits in x to the corresponding values of u; so $x = 0$ becomes $u = 2(0)+7 = 7$ and $x = 1$ becomes $u = 2(1)+7 = 9$. Then calculate the integral in terms of u. Thus we have

$$\int_0^1 (2x+7)^{23}\,dx = \int_7^9 u^{23}\cdot\frac{1}{2}\,du = \frac{1}{2}\int_7^9 u^{23}\,du = \frac{1}{48}\left[u^{24}\right]_7^9 = \frac{1}{48}\left(9^{24} - 7^{24}\right).$$

2. Alternatively, we can find the indefinite integral, and substitute for u to get the answer in terms of x (as we did above). Then we evaluate the result at the relevant values of x - *i.e.*

$$\int (2x+7)^{23}\,dx = \frac{1}{48}(2x+7)^{24}+c \quad\Rightarrow\quad \int_0^1 (2x+7)^{23}\,dx = \left[\frac{1}{48}(2x+7)^{24}\right]_0^1 = \frac{1}{48}\left(9^{24} - 7^{24}\right).$$

The technique we have just demonstrated is called *integration by substitution*. It allows us to calculate integrals of the form

$$\int f(u(x))u'(x)\,dx = \int f(u(x))\frac{du}{dx}\,dx = \int f(u)\,du.$$

3.6 Integration by parts

As we have just seen, integration by substitution allows us to calculate integrals of some quite complicated-looking functions essentially by reversing the Chain Rule of differentiation. We will **Lecture 18**

now look at a useful technique for computing integrals of functions which can be broken down into a product of two terms - *integration by parts*. Integration by parts can be seen as analogous to reversing the Product Rule of differentiation. The idea is to break up a complicated function into two parts, at least one of which we are able to integrate using the rules we have learned earlier.

Let $u(x)$ and $v(x)$ be two functions. Then, the **product rule** tells us that

$$\frac{d}{dx}(uv) = u\frac{dv}{dx} + v\frac{du}{dx}.$$

Integrating both sides with respect to x gives

$$uv = \int u\frac{dv}{dx}\,dx + \int v\frac{du}{dx}\,dx.$$

Rearrangement of the above equation gives us the standard formula for integration by parts:

$$\int u\frac{dv}{dx}\,dx = uv - \int v\frac{du}{dx}\,dx.$$

Note that when using this formula in calculations, you only add the constant of integration when the very last indefinite integral disappears.

Integration by parts is very useful for integrals where the integrand can be written as the product of two functions. In particular, for integrals of the form

$$\int x^n \cdot F(x)\,dx,$$

where $F(x)$ is a trigonometric, exponential function, etc. In this case take $u = x^n$, $\frac{dv}{dx} = F(x)$. However, these are not the only kind of integrals for which it can be helpful.

We will now look at some examples of how technique can be used in practice.

Example 3.6. Evaluate the integral $I_1 = \int xe^x\,dx$.

Our first task is to get our integral into the same form as the LHS of the formula. We can see that the integrand can be broken down into a product of x and e^x, but we need to choose which term to identify with u, and which with $\frac{dv}{dx}$. There is no rule which will always tell us the best way to do this. We could use trial and error, but often a little thought can save us unnecessary work. If there was one term which we could not integrate that would be the obvious candidate for u, but in this case it is equally easy to integrate x and e^x. However, we want the integral on the RHS of the formula to be one we are able to recognise. Hence we choose

$$u(x) = x \quad \Rightarrow \quad \frac{du}{dx} = 1, \qquad\qquad \frac{dv}{dx} = e^x \quad \Rightarrow \quad v = e^x.$$

Substituting the above into the formula then gives

$$I_1 = uv - \int v\frac{du}{dx}\,dx = xe^x - \int e^x(1)\,dx = xe^x - e^x + c = (x-1)e^x + c,$$

where we have included the constant of integration, c, as the final step.

- N.B. It is always good practice to check your answer by differentiating.
- *Exercise:* Why would it have been a bad idea to choose $u = e^x$, $\frac{dv}{dx} = x$?

For some integrals, a useful trick is to choose one of the terms in the product to be 1. This allows us to evaluate integrals we might not be able to do using other methods.

Example 3.7. Evaluate the integral $I_2 = \int \ln x \, dx$.

The first step is to re-write the integral as $I_2 = \int (1) \ln x \, dx$. Since we obviously do not know the integral of $\ln x$, we set $u = \ln x$, in which case we must choose $\frac{dv}{dx} = 1$. We thus have

$$\frac{du}{dx} = \frac{1}{x}, \qquad v = x.$$

Then, applying the integration by parts formula we obtain

$$I_2 = uv - \int v \frac{du}{dx} \, dx = x \ln x - \int \frac{x}{x} \, dx = x \ln x - x + c.$$

Sometimes, we might need to integrate by parts more than once.

Example 3.8. Evaluate the integral $I_3 = \int x^2 e^x \, dx$.

This is similar to the first example, and since differentiating x^2 leads to a simpler function, we choose

$$u(x) = x^2, \quad \Rightarrow \quad \frac{du}{dx} = 2x,$$

$$\frac{dv}{dx} = e^x \quad \Rightarrow \quad v = e^x.$$

Substituting the above into the formula then gives

$$I_3 = uv - \int v \frac{du}{dx} \, dx = x^2 e^x - \int e^x (2x) \, dx = x^2 e^x - 2 \int x e^x \, dx.$$

We can use integration by parts again to find $\int x e^x \, dx$ (we did this in Example 1). Hence

$$I_3 = x^2 e^x - 2 \int x e^x \, dx = x^2 e^x - 2(x-1)e^x + c = (x^2 - 2x + 2)e^x + c.$$

Sometimes by repeating the integration by parts procedure you end up with the original expression again, and after a little algebra, this enables us to evaluate the integral.

Example 3.9. For the integral $\int \cosh x \, \sin x \, dx$, let

$$\begin{aligned}
u &= \cosh x, \quad \frac{dv}{dx} = \sin x \\
\frac{du}{dx} &= \sinh x, \quad v = -\cos x \\
I &= \int \cosh x \, \sin x \, dx = uv - \int v \frac{du}{dx} \, dx \\
&= \cosh x(-\cos x) \, - \int (-\cos x) \sinh x \, dx \\
&= -\cosh x \, \cos x + \int \cos x \, \sinh x \, dx
\end{aligned}$$

Now repeat the process, pushing on in the same 'direction': we differentiated cosh to sinh, so must differentiate this sinh; conversely, we integrated sin to get cos, so we must integrate cos.

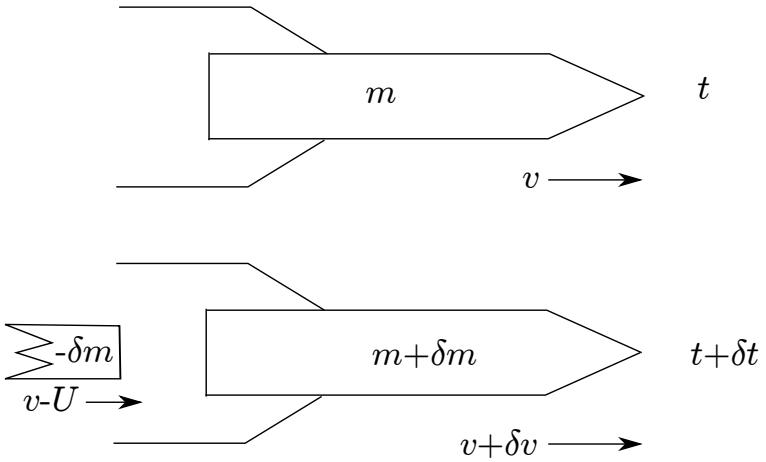
$$\begin{aligned}
u &= \sinh x, \quad \frac{dv}{dx} = \cos x \\
\frac{du}{dx} &= \cosh x, \quad v = \sin x \\
I &= -\cos x \, \cosh x + \sinh x \, \sin x - \int \cosh x \, \sin x \, dx \\
&= -\cos x \, \cosh x + \sinh x \, \sin x - I + c \\
\implies 2I &= -\cos x \, \cosh x + \sinh x \, \sin x + c \\
I &= \frac{1}{2}(-\cos x \, \cosh x + \sinh x \, \sin x + c)
\end{aligned}$$

Remember the rule: only add the integration constant when the last integral sign disappears.

3.7 Application: rocket flight

Lecture 19

Earlier in the course, we noted that Newton's Second Law (the rate of change of momentum of a body is equal to the force acting on it), when applied to a body of constant mass, m, leads to the well-known equation $F = ma$. Whilst this equation can be used to solve a wide variety of problems, there are important examples where the mass of the body in motion is not constant. One of these is the motion of a rocket, which emits hot gases (from burning fuel) to produce thrust. The fuel represents a significant proportion of the initial mass of the rocket (*e.g.* for the space shuttle, around 80%), almost all of which is used in reaching orbit.



Definition sketch for the derivation of the rocket equation

We begin by deriving an equation of motion for the rocket. We let the rocket's mass be $m(t)$ and its velocity $v(t)$. For simplicity, we assume it will move in a straight line, which we take as the positive x direction. The force acting on the rocket is $F(t)$. The rocket emits exhaust gases at a rate $|\dot{m}|$, (where, because this reduces the rocket's mass, we expect $\dot{m} < 0$), with constant velocity U relative to the rocket. Now let us consider what happens to the rocket in the small interval of time between t and $t + \delta t$. There will be a small change in mass of the rocket, δm (where we expect $\delta m < 0$), and similarly, a small change in the rocket's velocity, δv. A small amount of exhaust gas (of mass $-\delta m$) will also be emitted. Then, the change in momentum of the system (rocket plus gas) is

$$\begin{aligned} \underbrace{[(m+\delta m)(v+\delta v) - mv]}_{\text{change in rocket momentum}} + \underbrace{[-\delta m(v-U)]}_{\text{momentum of gas}} &= mv + m\delta v + v\delta m + \delta m\delta v - mv - v\delta m + U\delta m \\ &= F\delta t \end{aligned}$$

After a little tidying up, we obtain

$$m\delta v + \delta m\delta v + U\delta m = F\delta t$$

On dividing through by δt and taking the limit $\delta t \to 0$ we find

$$m\frac{dv}{dt} = F - U\frac{dm}{dt}. \tag{3.2}$$

Note that the term involving a product of the two small quantities δm and δv vanishes as $\delta t \to 0$.

Now consider the particular case of a rocket firework that lifts off vertically upwards, so the force acting upon it is gravity (for now, we will neglect other forces such as air resistance that may be present). Thus $F = m(t)g$ (directed vertically downwards), where g is the acceleration due to gravity (approximately 9.81 ms^{-2}) . The initial mass of the rocket is m_0 kg, but it burns fuel at a constant rate of α kg s^{-1}, so its mass after t seconds is $m(t) = m_0 - \alpha t$ kg. The exhaust gases are emitted at a constant speed, U ms^{-1}, relative to the rocket. Substituting this information into our rocket equation above, the upward velocity, v (in ms^{-1}) of the rocket obeys

$$(m_0 - \alpha t)\frac{dv}{dt} = -(m_0 - \alpha t)g - U(-\alpha),$$

where we have a minus sign in front of the force term because it is directed downwards (in the opposite direction to v). Dividing through by $m_0 - \alpha t$ yields

$$\frac{dv}{dt} = -g + \frac{\alpha U}{m_0 - \alpha t},$$

Note that this equation can only be valid for times $0 < t \leq \dfrac{m_0}{\alpha}$, at most. The reason for this is that when all the fuel is burned, the rocket will not longer experience a thrust, and the forces on it will change. This will happen before $m(t)$ reaches zero, which occurs at $t = \dfrac{m_0}{\alpha}$

1. Assuming that at $t = 0$ the rocket is stationary, what is the vertical velocity, v, of the rocket at time t?

   To answer this question, we need integrate the expression for $\frac{dv}{dt}$. The first term on the RHS poses no challenge, since g is a constant. The second term is more tricky. However, if we look at it for a moment or so we see it is a function of $m_0 - \alpha t$ multiplied by the derivative of $m_0 - \alpha t$. Hence, it is the type of integral that integration by substitution can be used to calculate. We make the substitution $w = m_0 - \alpha t$, and proceed as follows:

   $$\begin{aligned} v &= -gt + \int \frac{\alpha U}{m_0 - \alpha t}\, dt = -gt + \int \left(\frac{\alpha U}{w}\right) \frac{-1}{\alpha}\, dw = -gt - U \int \frac{1}{w}\, dw \\ &= -gt - U \ln w + c = -gt - U \ln (m_0 - \alpha t) + c. \end{aligned}$$

   Applying $v(0) = 0$ gives $c = U \ln m_0$ and hence

   $$v = -gt - U \left[\ln (m_0 - \alpha t) - \ln m_0\right] = -gt - U \ln \left(1 - \frac{\alpha t}{m_0}\right).$$

2. Now find the height, x, of the rocket at time, t. (Note: $\frac{dx}{dt} = v$.)

   We need to compute

   $$x(t) = \int -gt - U \ln \left(1 - \frac{\alpha t}{m_0}\right)\, dt = -\frac{gt^2}{2} - U \int \ln \left(1 - \frac{\alpha t}{m_0}\right)\, dt$$

   Once again, integrating the first term is straightforward. To calculate the integral of the second term, we start by making the substitution, $z = 1 - \frac{\alpha t}{m_0}$ in order to simplify the argument of the logarithm. We note that the substitution gives $\frac{dz}{dt} = -\frac{\alpha}{m_0}$. Hence

   $$\int \ln \left(1 - \frac{\alpha t}{m_0}\right)\, dt = \int \ln z \left(-\frac{m_0}{\alpha}\right)\, dz = -\frac{m_0}{\alpha} \int \ln z\, dz.$$

   We can now use integration by parts (as in the example from lectures) to calculate this integral. However, we need to be careful with our notation. Usually, we call the two functions in the integration by parts formula u and v but in this problem we have introduced U and v as velocities, so there is scope for confusion. Instead, we will call the functions in the integration

by parts formula f and g. We set $f(z) = \ln z$ and $g'(z) = 1$. Hence $f'(z) = \frac{1}{z}$ and $g(z) = z$, and we have

$$\begin{aligned}\int \ln z \, dz &= f(z)g(z) - \int g(z)f'(z) \, dz = z \ln z - \int z \left(\frac{1}{z}\right) dz = z \ln z - z + c \\ &= \left(1 - \frac{\alpha t}{m_0}\right) \ln \left(1 - \frac{\alpha t}{m_0}\right) - \left(1 - \frac{\alpha t}{m_0}\right) + c\end{aligned}$$

where c is a constant.

Putting all of the above together yields:

$$\begin{aligned}x(t) &= -\frac{gt^2}{2} - U \int \ln \left(1 - \frac{\alpha t}{m_0}\right) dt = -\frac{gt^2}{2} + \frac{Um_0}{\alpha} \int \ln z \, dz \\ &= -\frac{gt^2}{2} + \frac{Um_0}{\alpha} \left(1 - \frac{\alpha t}{m_0}\right) \left[\ln \left(1 - \frac{\alpha t}{m_0}\right) - 1\right] + K\end{aligned}$$

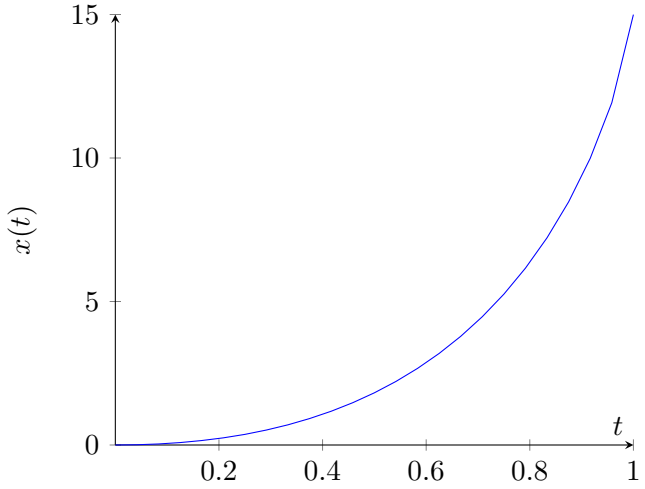
where K is a constant. If the firework starts at $x = 0$ when $t = 0$, we must have $K = \dfrac{Um_0}{\alpha}$

Now, recall that we set $m(t) = m_0 - \alpha t$ and hence the fraction of the original rocket mass remaining at time t is $\dfrac{m(t)}{m_0} = \dfrac{m_0 - \alpha t}{m_0} = 1 - \dfrac{\alpha t}{m_0}$. Hence, rewriting the equation above gives

$$\begin{aligned}x(t) &= -\frac{gt^2}{2} + \frac{Um(t)}{\alpha} \left[\ln \left(\frac{m(t)}{m_0}\right) - 1\right] + \frac{Um_0}{\alpha} \\ &= -\frac{gt^2}{2} + \frac{Um_0}{\alpha} \left[\frac{m(t)}{m_0} \left(\ln \left(\frac{m(t)}{m_0}\right) - 1\right) + 1\right]\end{aligned}$$

The plot below illustrates the rocket's trajectory for $g = 10$, $U = 20$, $\alpha = m_0 = 1$.

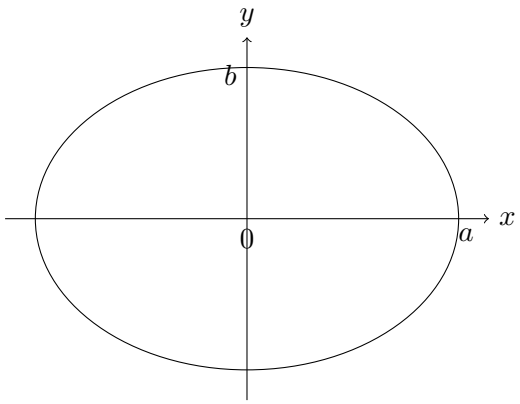
The rocket height $x(t)$, where $g = 10$, $U = 20$, $\alpha = m_0 = 1$.



If the firework consists almost entirely of fuel (so the mass of the containing shell is negligible), what height does the rocket attain when all the fuel is used up?

As this example very nicely demonstrates, in real-life problems we frequently need to use a combination of the techniques we have learned in order to calculate the quantities we are interested in. Unlike the problems you see at school, they almost never involve the application of just one part of your knowledge!

3.8 Trigonometric substitutions



Consider the problem of finding the area of an ellipse which is given by the equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

On drawing a diagram, we can see that the area of interest is four times that in the first quadrant (where x and y are both positive). Recalling that the area beneath a curve $y = f(x)$ (where $f(x) > 0$) is given by the integral, the area we wish to find is

$$A = 4\int_0^a y(x)\,dx$$

where $y(x)$ is the equation of the ellipse in the first quadrant.

Taking the original equation of the ellipse, and rearranging gives

$$y = b\sqrt{1 - \frac{x^2}{a^2}},$$

where we have chosen the positive square root, since in the first quadrant both x and y are positive. Thus

$$A = 4b\int_0^a \sqrt{1 - \frac{x^2}{a^2}}\,dx = \frac{4b}{a}\int_0^a \sqrt{a^2 - x^2}\,dx.$$

At this point, it would be easy to get disheartened; it is not immediately clear that there is a way to convert the integrand into something we recognise. The main difficulty is the square root sign, since this makes manipulations difficult. It would be nice is we could make a substitution such that we were taking the square root of a squared quantity. Then, the integrand would look much less intimidating.

If we think about the properties of trigonometric functions, then we can recall a property which will come to our rescue here: the fact that, for any θ

$$\sin^2\theta + \cos^2\theta = 1, \quad \Rightarrow \quad \cos^2\theta = 1 - \sin^2\theta.$$

On multiplying by a^2, we have

$$a^2\cos^2\theta = a^2 - a^2\sin^2\theta.$$

Comparison with our integrand suggests we try the substitution $x = a\sin\theta$. Then

$$a^2 - x^2 = a^2 - a^2\sin^2\theta = a^2(1 - \sin^2\theta) = a^2\cos^2\theta$$

Therefore

$$\sqrt{a^2 - x^2} = a\cos\theta,$$

where we have taken the positive square root, since $y(x)$ is positive over the given range. We note that the limits $x = 0$ and $x = a$ correspond to $\theta = 0$ and $\theta = \frac{\pi}{2}$ respectively, and that $\cos\theta > 0$ for $0 \leq \theta \leq \frac{\pi}{2}$ as required. Since $\frac{dx}{d\theta} = a\cos\theta$, our substitution gives

$$A = 4b\int_0^{\frac{\pi}{2}} \cos\theta \cdot (a\cos\theta)\, d\theta = 4ab\int_0^{\frac{\pi}{2}} \cos^2\theta\, d\theta.$$

Now, $\cos^2\theta$ is not something we can integrate straight away, but once again the properties of trigonometric functions come to the rescue. Recall the double angle identity

$$\cos 2\theta = \cos^2\theta - \sin^2\theta = 2\cos^2\theta - 1.$$

Thus

$$A = 4ab\int_0^{\frac{\pi}{2}} \cos^2\theta\, d\theta = 4ab\int_0^{\frac{\pi}{2}} \frac{1}{2}(\cos 2\theta + 1)\ d\theta = 2ab\int_0^{\frac{\pi}{2}} (\cos 2\theta + 1)\ d\theta.$$

Finally, we note that $\frac{d}{d\theta}\left(\frac{1}{2}\sin 2\theta\right) = \cos 2\theta$ and hence

$$A = 2ab\int_0^{\frac{\pi}{2}} (\cos 2\theta + 1)\ d\theta = 2ab\left[\frac{1}{2}\sin 2\theta + \theta\right]_0^{\frac{\pi}{2}} = 2ab\left(\frac{\pi}{2} - 0\right) = \pi ab.$$

This example is an illustration of the fact that integrands involving $\sqrt{a^2 \pm x^2}$, $\sqrt{x^2 \pm a^2}$ can often be evaluated by an appropriate trigonometric substitution. The strategy is to make a substitution of the type $x = a\sin\theta, x = a\sec\theta$ or something similar to simplify the expression using trigonometric identities such as

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

$$\sec^2\theta = 1 + \tan^2\theta.$$

Often the result can then be expressed as an integral of trigonometric functions of the type already dealt with above. We have already demonstrated the sin substitution; the next example uses a tan substitution.

Example 3.10. Find $I = \displaystyle\int \frac{1}{x^2\sqrt{x^2+9}}\,dx$

This time, we have a factor of the form $x^2 + a^2$ in our integrand; this suggests we should try to exploit the identity $1 + \tan^2\theta = \sec^2\theta$ to simplify it. If we set $x = 3\tan\theta$ then $\frac{dx}{d\theta} = 3\sec^2\theta$ and

$$x^2 + 9 = 9\tan^2\theta + 9 = 9(\tan^2\theta + 1) = 9\sec^2\theta.$$

Hence our integral becomes

$$I = \int \frac{3\sec^2\theta}{(9\tan^2\theta)(3\sec\theta)}\,d\theta = \frac{1}{9}\int \frac{\sec\theta}{\tan^2\theta}\,d\theta.$$

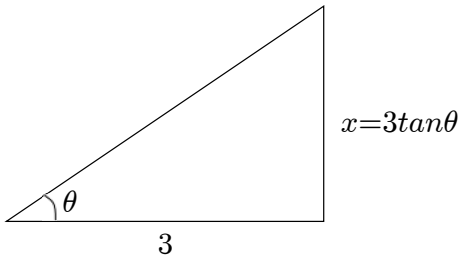
Recalling that $\sec\theta = \frac{1}{\cos\theta}$ and $\tan\theta = \frac{\sin\theta}{\cos\theta}$ we obtain

$$I = \frac{1}{9}\int \frac{\cos\theta}{\sin^2\theta}\,d\theta.$$

The integrand is now in the classic form for integration by substitution. We set $u = \sin\theta$ so $\frac{du}{d\theta} = \cos\theta$, which transforms the integral to

$$I = \frac{1}{9}\int \frac{\cos\theta}{u^2}\,\frac{1}{\cos\theta}\,du = \frac{1}{9}\int \frac{1}{u^2}\,du = -\frac{1}{9u} + c = -\frac{1}{9\sin\theta} + c.$$

Now, we need to substitute for $\sin\theta$ in terms of x.



Recall that in a right-angled triangle, if the adjacent side has length 3, the opposite side will have length $3\tan\theta = x$, and the hypotenuse will have length $\sqrt{3^2 + (3\tan^2\theta)} = \sqrt{9+x^2}$. Since $\sin\theta$ is the ratio of the lengths of the opposite side to the hypotenuse, $\sin\theta = \frac{x}{\sqrt{9+x^2}}$. Using this expression, we obtain:

$$I = \int \frac{1}{x^2\sqrt{(x^2+9)}}\,dx = -\frac{1}{9\sin\theta} + c = -\frac{\sqrt{9+x^2}}{9x} + c.$$

Note: The hyperbolic functions can also be useful for simplifying integrals of this form, as they have similar properties to the trigonometric functions. For example, we can use the identity $\cosh^2 t - \sinh^2 t = 1$ to find $I = \displaystyle\int \frac{1}{(x^2+4)^{\frac{3}{2}}}\,dx$ by setting $x = 2\sinh t$. Then we have $(x^2+4)^{\frac{3}{2}} = 8\cosh^3 t$, and $dx = 2\cosh t\,dt$. The rest of the calculation is left as an exercise.

Whilst both trigonometric or hyperbolic substitutions will work, sometimes one method will produce a solution more quickly than the other. Unfortunately, there is no clear rule for which will work best in any particular situation.

3.9 Application: population growth

Lecture 20

Suppose we are interested in knowing the population of bacteria being grown in a laboratory experiment. Let the number of bacteria at time t be given by $b(t)$. At time $t = 0$, b_0 bacteria are placed in a dish which contains a medium to supply them with necessary nutrients. Early on in the experiment, when there is only a small number of bacteria present, space and nutrients are plentiful and the cells reproduce rapidly by division. At this stage, we expect the rate of increase of the bacteria will be proportional to the population. However, when the bacteria become more numerous, overcrowding becomes an issue and nutrients begin to become more scarce, so the rate of population increase slows. Finally, when the bacteria are consuming the nutrients as fast as they are replenished, the population reaches the *carrying capacity* of the environment, and population growth stops. This scenario can be modelled by the *logistic equation*:

$$\frac{db}{dt} = rb\left(1 - \frac{b}{K}\right), \tag{3.3}$$

where r is the maximum reproduction rate of the bacteria, and K is the carrying capacity of the dish.

We can find the population $b(t)$ using the method of *separation of variables*. We rearrange equation (3.3) so that all the terms involving b are on one side, and all those involving t on the other

$$\frac{1}{rb\left(1 - \frac{b}{K}\right)}\,db = \frac{K}{rb\,(K - b)}\,db = dt.$$

We then integrate this equation

$$\int \frac{K}{rb\,(K - b)}\,db = \int 1\,dt.$$

The function on the right hand side is easy to compute. Using the fact that $b = b_0$ at $t = 0$ to specify the limits on the integral, we find that the time t at which the population reaches b bacteria is given by

$$t = \frac{K}{r}\int_{b_0}^{b} \frac{1}{x(K - x)}\,dx. \tag{3.4}$$

You will not need to recall the method of separation of variables for this course, as it will be covered in detail in Maths IB. For now, you can just accept that we need to calculate the integral in equation (3.4).

The integral on the RHS of equation (3.4) is not one that we can do by substitution; nor will integration by parts help. What can we do?

Consider adding two fractions, $\frac{A}{x}$ and $\frac{B}{K-x}$ where A and B are real numbers. Then

$$\frac{A}{x} + \frac{B}{K - x} = \frac{A(K - x)}{x(K - x)} + \frac{Bx}{x(K - x)} = \frac{A(K - x) + Bx}{x(K - x)} \tag{3.5}$$

The terms on the left hand side of the equation would be straightforward to integrate; what we have in the denominator on the far right hand side is the same as our integrand. If we can find

numbers A and B such that the numerator $A(K-x)+Bx=1$, then we have a way of calculating the integral in equation (3.4).

To find the numbers A and B such that $A(K-x)+Bx=1$, there are two ways we can proceed.

- We can compare coefficients of the same powers of x on both sides of the equation. These must be equal. In our example coefficient of x^0 (*i.e.* the constant term) is AK on the LHS, and 1 on the RHS. Hence $AK=1$, and so $A=1/K$. Comparing coefficients of x gives $B-A=0$, so $B=A=1/K$.
- Since the relationship $A(K-x)+Bx=1$ must hold for any value of x, we are free to substitute particular values of x which make the equation simpler. In this case, if we choose $x=0$, the second term in our equation vanishes, and we are left with $AK=1$. Hence $A=1/K$. If we set $x=K$, then the first term in the equation will vanish. Then, we have $BK=1$, so $B=1/K$.

Of course, both methods give the same answer.

We have now established that

$$\frac{1}{x(K-x)}=\frac{1}{K}\left(\frac{1}{x}+\frac{1}{K-x}\right).$$

Hence, we can re-write our integral as

$$t=\frac{K}{r}\int_{b_0}^{b}\frac{1}{x(K-x)}\,dx=\frac{1}{r}\int_{b_0}^{b}\left(\frac{1}{x}+\frac{1}{K-x}\right)\,dx. \tag{3.6}$$

The integral of the first terms is straightforward, and the second term can be integrated easily using substitution $u=K-x$

$$t=\frac{1}{r}\left[\ln x\right]_{b_0}^{b}+\frac{1}{r}\int_{K-b_0}^{K-b}\frac{-1}{u}\,du=\frac{1}{r}(\ln b-\ln b_0-[\ln u]_{K-b_0}^{K-b})=\frac{1}{r}(\ln b-\ln b_0-\ln(K-b)+\ln(K-b_0))$$

A little algebra can now be used to obtain an equations for the population b at time t:

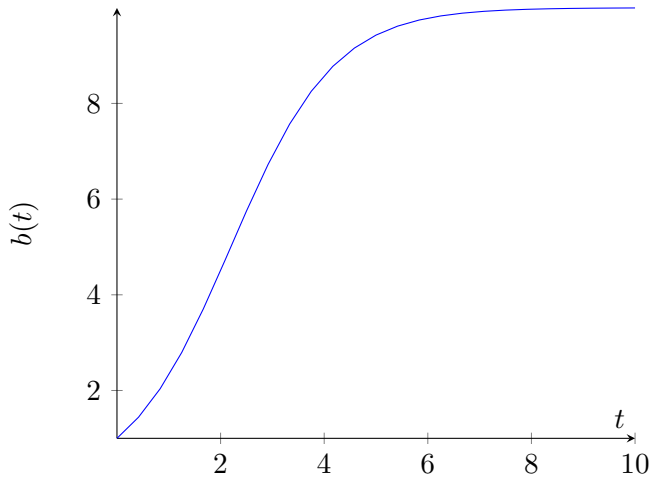
$$rt=\ln\frac{b(K-b_0)}{b_0(K-b)}\quad\Rightarrow\quad\frac{b}{(K-b)}=\frac{b_0}{(K-b_0)}e^{rt}.$$

One final rearrangement gives

$$\left(1+\frac{b_0}{(K-b_0)}e^{rt}\right)b=\frac{Kb_0}{(K-b_0)}e^{rt}\quad\Rightarrow\quad b=\frac{K}{1+\left(\frac{K}{b_0}-1\right)e^{-rt}}.$$

Now let us consider the behaviour of the function $b(t)$. We will assume that $K>b_0$ (so that the carrying capacity - the maximum number of bacteria the environment can sustain - is larger than the initial population) and note that $r>0$ (so bacteria increase in number). Then, the exponential term in the denominator is positive but decreasing, and hence the function itself is always increasing. As t increases, the exponential term decays rapidly at first, and so the population will initially grow quickly. However, at later times this term will be negligibly small, and hence the population will stop growing and approach the carrying capacity, K. This behaviour is illustrated by the graph below.

The bacteria population, $b(t)$, where $b_0 = 1$, $r = 1$, $K = 10$



3.10 Partial fractions

The method we used in the last example to break up the single, complicated fraction in the integrand into two simpler fractions which were easier to integrate is called the method of *partial fractions*. It is part of the procedure we need to use to integrate rational functions - functions which are of the form $f(x) = \frac{p(x)}{q(x)}$, where $p(x)$ and $q(x)$ are polynomials. We will assume that the degree of $p(x)$ is lower than that of $q(x)$. Then, there are three basic cases we need to consider (of which we have seen the simplest already).

1. If the denominator (bottom line) of the fraction is a product of distinct linear factors, then we write the function as a sum of fractions with the linear factors being the denominators of each fraction (this is what we did in the population growth example).

   Example 3.11. If we want to expand the function $\dfrac{1}{(x+1)(x+2)}$ into partial fractions, we write:

   $$\frac{1}{(x+1)(x+2)} = \frac{A}{x+1} + \frac{B}{x+2},$$

   where A and B are real numbers, which are as yet unknown. To find A and B we re-write the partial fractions, putting them back over a common denominator:

   $$\frac{A}{x+1} + \frac{B}{x+2} = \frac{A(x+2) + B(x+1)}{(x+1)(x+2)} = \frac{1}{(x+1)(x+2)}.$$

   Then, since the denominators of the middle and right-hand terms of the equation are equal, the numerators must be also, *i.e.* $1 = A(x+2) + B(x+1)$. As in our earlier example, there are two ways to proceed at this point. The first is to compare coefficients of the same power of x on both sides, which would give two simultaneous equations for A and B:

   $$1 = 2A + B \qquad A + B = 0.$$

solving these gives $A = 1$, $B = -1$. Alternatively, since $1 = A(x+2) + B(x+1)$ is an *identity* (it holds for *all* values of x) we can substitute appropriate values of x to determine A and B.

$$\begin{aligned} &1 = A(x+2) + B(x+1); \\ &x = -2 \implies B = -1; \\ &x = -1 \implies A = 1. \end{aligned}$$

You can use either method (or a combination of them).

Hence, we find that

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

2. If the denominator of the fraction contains one or more *repeated* linear factors, we must include terms of all powers up to the multiplicity of the factor in our partial fractions.

Example 3.12. If we want to expand the function $\dfrac{x^2+2x-1}{(x+1)^3}$ into partial fractions, we write

$$\frac{x^2+2x-1}{(x+1)^3} = \frac{A}{x+1} + \frac{B}{(x+1)^2} + \frac{C}{(x+1)^3} = \frac{A(x+1)^2 + B(x+1) + C}{(x+1)^3}.$$

Then, comparing the numerators we have

$$x^2 + 2x - 1 = A(x+1)^2 + B(x+1) + C.$$

We can quickly find A, B and C using a combination of the two methods described in the previous example:

- Substitute $x = -1 \implies C = -2$.
- Compare the coefficient of $x^2 \implies A = 1$.
- Compare the coefficient of $x^0 \implies -1 = A + B + C \implies B = 0$.

Hence

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

3. If the denominator of the fraction contains a irreducible quadratic (one which cannot be factorised into real linear factors) then we must allow the numerator of the corresponding partial fraction to be a linear function.

Example 3.13. To expand the function $\dfrac{x}{(x-1)(x^2+2x+2)}$ into partial fractions, we write:

$$\begin{aligned} \frac{x}{(x-1)(x^2+2x+2)} &= \frac{A}{x-1} + \frac{Bx+C}{x^2+2x+2} \\ &= \frac{A(x^2+2x+2) + (Bx+C)(x-1)}{(x-1)(x^2+2x+2)} \end{aligned}$$

As before, we can find the values of A, B and C by either equating coefficients, or using the substitution method. We choose the latter here. Setting $x = 1 \implies A = \frac{1}{5}$. Then, we choose two convenient values of x say $x = 0, -1$ to obtain two linear equations to determine B and C:

$$x = 0 \implies 0 = 2A - C \implies C = \frac{2}{5};$$
$$x = -1 \implies -1 = A + 2B - 2C \implies B = -\frac{1}{5}.$$

Thus

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

We will not consider cases involving repeated irreducible quadratic factors in this course.

(Note: a quadratic equation can always be factored into complex linear factors, and it is not uncommon to see students do this when tackling partial fraction problems in assignments or exams. Although technically correct, the problem is that, since the original function is clearly real-valued, its partial fraction decomposition (and any resulting integral) should be clearly real-valued too. This approach usually results in a function with complex terms (which, in fact, cancel out), but unless it is clearly written as a real-valued function it will not receive full marks.)

3.11 Integration of rational functions

Now that we can use partial fractions to simplify rational functions $\frac{p(x)}{q(x)}$ where the degree of $p(x)$ is less than that of $q(x)$, we are close to being able to integrate a general rational function. **Lecture 21**

The steps in the procedure are as follows.

1. If the degree of the polynomial $p(x)$ is greater than the degree of $q(x)$, use long division to simplify the integrand into a polynomial plus a (new) rational function.
2. Now we should be left with a rational function $\frac{p(x)}{q(x)}$ where the degree of $p(x)$ is less than that of $q(x)$. If necessary, factorise $q(x)$ into a product of *linear* and *irreducible quadratic* terms.
3. Use the method outlined above to decompose the rational function into partial fractions.
4. For the partial fractions with irreducible quadratic terms in the denominator, completing the square can be helpful (often the result is recognisable as the derivative of an inverse trigonometric function).

We will now illustrate how this works in practice with examples.

Example 3.14. Find $\displaystyle\int \frac{x^2+1}{x-3}\,dx$.

Long division:

$$\begin{array}{r} x+3 \\ x-3 \;\overline{)\; x^2+1} \\ \underline{x^2-3x} \\ 3x+1 \\ \underline{3x-9} \\ 10 \end{array}$$

$$\begin{aligned}\int \frac{x^2+1}{x-3}\,dx &= \int (x+3) + \frac{10}{x-3}\,dx \\ &= \frac{x^2}{2} + 3x + 10\ln|x-3| + C.\end{aligned}$$

Example 3.15. Find $\displaystyle\int \frac{2x^4-6x^3-9x^2-x-6}{x^5-3x^4+x^3-3x^2}\,dx$.

1. As the degree of the denominator is five and is greater than the degree of the numerator, no long division is necessary.

2. We next factorise the denominator:

$$\begin{aligned}x^5-3x^4+x^3-3x^2 &= x^2(x^3-3x^2+x-3) \\ &= x^3(x^2+1) - 3x^2(x^2+1) \\ &= x^2(x-3)(x^2+1).\end{aligned}$$

3. Express integrand as a sum of partial fractions:

$$\begin{aligned}&\frac{2x^4-6x^3-9x^2-x-6}{x^2(x-3)(x^2+1)} = \frac{A}{x-3} + \frac{B}{x} + \frac{C}{x^2} + \frac{Dx+E}{x^2+1}; \\ &2x^4-6x^3-9x^2-x-6 \\ &= Ax^2(x^2+1) + Bx(x-3)(x^2+1) \\ &\quad + C(x-3)(x^2+1) + (Dx+E)x^2(x-3)\end{aligned}$$

   - Substitute $x=0 \implies -6 = -3C \implies C = 2$.
   - Substitute $x=3 \implies 162-162-81-3-6 = 90A \implies -90 = 90A \implies A = -1$.
   - As there is nothing else to conveniently substitute, compare coefficients:

$$\begin{aligned}x^1 &\implies -1 = -3B + C \implies B = 1. \\ x^2 &\implies -9 = A + B - 3C - 3E = -1 + 1 - 6 - 3E \implies E = 1. \\ x^4 &\implies 2 = A + B + D = -1 + 1 + D \implies D = 2.\end{aligned}$$

4. Therefore

$$\begin{aligned}&\int \frac{2x^4-6x^3-9x^2-x-6}{x^2(x-3)(x^2+1)}\,dx \\ &= -\int \frac{1}{x-3}\,dx + \int \frac{1}{x}\,dx + 2\int \frac{1}{x^2}\,dx + \int \frac{2x+1}{x^2+1}\,dx \\ &= -\ln|x-3| + \ln|x| - 2x^{-1} + \int \frac{2x}{x^2+1}\,dx + \int \frac{1}{x^2+1}\,dx \\ &= \ln\left|\frac{x}{x-3}\right| - \frac{2}{x} + \ln(x^2+1) + \arctan x + K.\end{aligned}$$

3.12 Improper integrals

Present value of a continuous income stream **Lecture 22**

Suppose you place an amount of money, P dollars in a bank account, with a rate of interest $r\%$. Let t be the time in years since the money was deposited. Then, provided the interest is added at intervals much shorter than a year (*e.g.* daily or weekly), the amount of money in the bank, $A(t)$, is well-approximated by

$$A(t) = Pe^{rt}. \tag{3.7}$$

Suppose that, having enjoyed your time studying mathematics so much, you decide you would like to set up a mathematics PhD scholarship to help educate future generations of lecturers. You would like to be able to offer a scholarship of \$27,000 every year into the future. How much money do you need to give to the university?

We can re-arrange equation (3.7) to determine the amount we need to place in the account in order to have A dollars, after t years:

$$P = A(t)e^{-rt}. \tag{3.8}$$

We call this the *present value* of year t (we must pay P dollars now to get A dollars in a future year, t). If we want to award a scholarship worth A dollars (a constant amount) every year for n years we need to find the sum

$$P = \sum_{t=0}^{n} Ae^{-rt} \approx \int_0^n Ae^{-rt}\,dt.$$

(The approximation follows from the definition of the integral in terms of a sum: think of plotting the graph of Ae^{-rt} against t; the sum approximates the area between the curve and the horizontal axis.) But what if you wanted the scholarship to continue 'in perpetuity' (*i.e.* forever into the future)? This would mean taking the limit $n \to \infty$, and so we would need to calculate

$$\int_0^\infty Ae^{-rt}\,dt.$$

This kind of integral, which is performed over an unbounded domain, is called an *improper integral of the first kind*. It is improper, because when we defined definite integrals, we considered only bounded intervals. Thus we need to take care here; it is not obvious that a sensible answer will exist. It is quite plausible that if we wanted the scholarship to continue forever, we would require an infinite amount of money.

We proceed in the intuitive way, by calculating the amount of money required for the scholarship to run for n years, and then taking the limit $n \to \infty$. From earlier

$$P \approx \int_0^n Ae^{-rt}\,dt = -\frac{A}{r}\left[e^{-rt}\right]_0^n = -\frac{A}{r}(e^{-rn} - 1) = \frac{A}{r}(1 - e^{-rn}).$$

In the limit $n \to \infty$, we see that $e^{-rn} \to 0$, and so we have

$$\int_0^n Ae^{-rt}\,dt = \frac{A}{r}.$$

Thus, if the interest rate is 3% (assumed to remain constant) and we want to offer a scholarship of \$27,000 every year in perpetuity, we need to give a sum of

$$P = \frac{27,000}{0.03} = \$900,000.$$

3.12.1 Improper integrals of the first kind

More generally, we define improper integrals of the first kind using the limiting procedure we saw in the above example.

Definition 3.3. For a function $f(x)$, and with a, $b \in \mathbb{R}$

$$\int_a^\infty f(x)\,dx = \lim_{b\to\infty} \int_a^b f(x)\,dx,$$

where this limit exists. If the limit exists, we say the integral *converges.* Otherwise, we say it *diverges.*

We can define integrals over $(-\infty, b]$ and $(-\infty, \infty)$ in a similar way.

Improper integrals play an extremely important role in probability theory. You are already familiar with discrete random variables (*e.g.* the number of heads which appear when a coin is tossed ten times) where the outcome can only take certain discrete values. The concept of a continuous random variable is similar, but the outcome can be any real number within a certain range *e.g.* the number of hours a light bulb operates before failing. For a continuous random variable X the probability that X lies in the range $[a, b]$ is defined in terms of the integral of a *probability density function* $f(x)$. We write

$$P(a \le X \le b) = \int_a^b f(x)\,dx.$$

Note that we require

$$\int_{-\infty}^\infty f(x)\,dx = 1.$$

Why is this?

Of course, the functional form of the probability density $f(x)$ will depend on the application.

Example 3.16. The lifetime, T, of a certain type of light bulb is a continuous random variable with a probability density which follows the exponential distribution. The probability density function for this distribution is given by

$$f(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}} \quad x \ge 0,$$

where μ is the mean of the distribution. (Exercise: check that $P(0 \le T < \infty) = 1$.)

The mean lifetime of this type of bulb is found to be 1 year. What is the probability that a bulb will last for more than 5 years?

We need to calculate $P(T > 5) = P(5 \leq T < \infty)$. Using the formula above

$$P(T > 5) = P(5 \leq T < \infty) = \int_5^\infty e^{-x}\,dx = \left[-e^{-x}\right]_5^\infty = 0 - (-e^{-5}) = e^{-5} \approx 0.007$$

Note that

$$P(T > 5) = 1 - P(T < 5) = 1 - \int_0^5 e^{-x}\,dx = 1 - \left[-e^{-x}\right]_0^5 = 1 - (-e^{-5} - (-1)) = e^{-5}.$$

3.12.2 Improper integrals of the second kind

The fact that we have just discussed improper integrals of the *first kind* might suggest to you there is at least one other kind; indeed, there is. An *improper integral of the second kind* arises when we wish to calculate an integral on a bounded domain, but the integrand is undefined at some point within the interval, or at an end point. For example,

$$\int_0^1 \frac{1}{\sqrt{x}}\,dx$$

is an improper integral of the second kind, because $x^{-\frac{1}{2}}$ is undefined at $x = 0$.

To deal with this situation, we proceed in a similar way as before. We consider

$$\int_a^1 \frac{1}{\sqrt{x}}\,dx,$$

and then examine its behaviour as $a \to 0$. In this particular case

$$\int_a^1 \frac{1}{\sqrt{x}}\,dx = [2\sqrt{x}]_a^1 = 2(1 - \sqrt{a}).$$

Hence

$$\int_0^1 \frac{1}{\sqrt{x}}\,dx = \lim_{a\to 0}\int_a^1 \frac{1}{\sqrt{x}}\,dx = \lim_{a\to 0} 2(1 - \sqrt{a}) = 2.$$

By contrast $\displaystyle\int_0^1 \frac{1}{x^2}\,dx$ does not exist. If we consider

$$\lim_{a\to 0}\int_a^1 \frac{1}{x^2}\,dx = \lim_{a\to 0}\left[-\frac{1}{x}\right]_a^1 = \lim_{a\to 0} -1 + \frac{1}{a}$$

then we can see that the required limiting value is undefined.

In the above two cases, the value for which the function is undefined occurs at an end point. How would we deal with an integral like $\displaystyle\int_0^6 \frac{1}{(x-4)^{\frac{2}{3}}}\,dx$?

In this case, the problem is at $x = 4$, so we split the integral into two before taking the limit:

$$\int_0^6 \frac{1}{(x-4)^{\frac{2}{3}}}\,dx = \lim_{c\to 4}\int_0^c \frac{1}{(x-4)^{\frac{2}{3}}}\,dx + \lim_{c\to 4}\int_c^6 \frac{1}{(x-4)^{\frac{2}{3}}}\,dx.$$

Now, we have

$$\lim_{c\to 4}\int_0^c \frac{1}{(x-4)^{\frac{2}{3}}}\,dx = \lim_{c\to 4}[3(x-4)^{\frac{1}{3}}]_0^c = \lim_{c\to 4} 3(c-4)^{\frac{1}{3}} - 3(-4)^{\frac{1}{3}}$$
$$= 3(4)^{\frac{1}{3}},$$

and similarly,

$$\lim_{c\to 4}\int_c^6 \frac{1}{(x-4)^{\frac{2}{3}}}\,dx = \lim_{c\to 4}[3(x-4)^{\frac{1}{3}}]_c^6 = \lim_{c\to 4} 3(2)^{\frac{1}{3}} - 3(c-4)^{\frac{1}{3}}$$
$$= 3(2)^{\frac{1}{3}}$$

Since both of the two integrals exist in the limit $c \to 4$, we have

$$\int_0^6 \frac{1}{(x-4)^{\frac{2}{3}}}\,dx = 3(2)^{\frac{1}{3}} + 3(4)^{\frac{1}{3}}.$$

3.13 Summary of learning outcomes

Now that we have reached the end of this chapter, you should be able to:

- Use summation notation to express sums in compact form
- Compute simple sums expressed in summation notation
- Understand how and why the definite integral is defined in terms of a sum
- Recall and apply the properties of definite integrals
- Interpret the meaning of an integral geometrically, or in the context of a practical problem
- Define what is meant by an antiderivative
- State the First and Second Fundamental Theorems of Calculus, and apply them to solve problems
- Recall the antiderivatives of common functions, such as polynomials, exponentials, logarithms and trigonometric functions
- Define the indefinite integral, and recall its relation to antiderivatives
- Apply the techniques of integration by substitution and integration by parts to calculate the integrals of complicated functions
- Use a trigonometric substitution to calculate integrals
- Find the partial fractions expansion of a rational function
- Use long division and partial fractions expansions to calculate integrals of rational functions
- Calculate improper integrals of the first and second kind (or recognise that they do not exist)

You should also look back at the summaries of learning outcomes for the previous two chapters: this will help you prepare for the final written examination.

Chapter 4

Quick reference section

This section collects together some useful notation, for easy reference.

4.1 The Greek alphabet

A	α	alpha	B	β	beta	Γ	γ	gamma
Δ	δ	delta	E	ϵ	epsilon	Z	ζ	zeta
H	η	eta	Θ	θ	theta	I	ι	iota
K	κ	kappa	Λ	λ	lambda	M	μ	mu
N	ν	nu	Ξ	ξ	xi	O	o	omicron
Π	π	pi	P	ρ	rho	Σ	σ	sigma
T	τ	tau	Υ	υ	upsilon	Φ	ϕ	phi
X	χ	chi	Ψ	ψ	psi	Ω	ω	omega

4.2 Notation

iff "if and only if"
$\Longrightarrow$ "implies"
$\Longleftrightarrow$ "is equivalent to"
$\exists$, $\nexists$ "there exists", "there does not exist"
$\forall$ "for all" (or "for every")
$\equiv$ "is identically equal to"
$\therefore$ "therefore"

4.3 Sets

A set is a collection of objects called *elements*. The notation $\{x \mid \ldots\}$ or $\{x : \ldots\}$ is read "the set of objects x such that ...".

$x \in A$	the object x is an element of the set A
$\emptyset$	the *empty* set, that is, the set with no elements at all
$A \subset B$	the set A is contained in B, that is, every element of A is also an element of B. This does not exclude the possibility that $A = B$.
$A \cup B$	the *union* of the sets A and B: $A \cup B = \{x \mid x \in A \text{ or } x \in B\}$
$A \cap B$	the *intersection* of the sets A and B: $A \cap B = \{x \mid x \in A \text{ and } x \in B\}$
$A \backslash B$	the *difference* of the sets A and B: $A \backslash B = \{x \mid x \in A \text{ but } x \notin B\}$
$\mathbb{N}$	the set of natural numbers
$\mathbb{Q}$	the set of rational numbers
$\mathbb{R}$	the set of real numbers
$\mathbb{C}$	the set of complex numbers.
$\mathbb{R}^2$	$\{(x, y) \mid x, y \in \mathbb{R}\}$.

4.4 The Real Numbers

$$\mathbb{N} = \text{the natural numbers} = \{1, 2, 3, \ldots\}$$
$$\mathbb{Z} = \text{the integers} = \{0, \pm 1, \pm 2, \ldots\}$$
$$\mathbb{Q} = \text{the rational numbers} = \left\{ \frac{m}{n} \,\middle|\, n, m \in \mathbb{Z},\ n \neq 0 \right\}$$

Any rational number can be represented in infinitely many ways: for example,

$$\frac{1}{2} = \frac{2}{4} = \frac{3}{6} = \cdots = \frac{-1}{-2} \quad \text{etc.}$$