Calculus MA181 (II)/MA186 : Lecture Notes

Dr Rachel Quinlan
School of Mathematics, Statistics and Applied Mathematics, NUI Galway

March 25, 2013

Contents

Chapter 1

The Real Numbers

1.1 The set $\mathbb{R}$ of real numbers

This section involves a consideration of properties of the set $\mathbb{R}$ of real numbers, the set $\mathbb{Q}$ of rational numbers, the set $\mathbb{Z}$ of integers and other related sets of numbers. In particular, we will be interested in what is special about $\mathbb{R}$, what distinguishes the real numbers from the rational numbers and why the set of real numbers is such an interesting and important thing that there is a whole branch of mathematics (real analysis) devoted to its study.

WHAT IS $\mathbb{R}$?

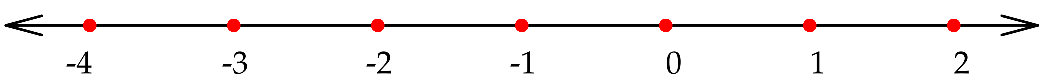
There are at least two useful ways to think about what real numbers are, and before considering them it is useful to first recall what *integers* are and what *rational numbers* are.

Integers are "whole numbers". The set of integers is denoted by $\mathbb{Z}$:

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

The notation "$\mathbb{Z}$" comes from the German word *Zahlen* (numbers).

On the number line, the integers appear as an infinite set of evenly spaced points. The integers are exactly those numbers whose decimal representations have all zeroes after the decimal point.



Note that the integers on the number line are separated by *gaps*. For example there are no integers in the chunk of the number line between $\frac{7}{5}$ and $\frac{63}{32}$.

The set of integers is *well-ordered*. This means (more or less) that given any integer, it makes sense to talk about *the next* integer after that one. For example, the next integer after 3 is 4. To see why this property (which might not seem very remarkable at first glance) is something worth bothering about and to understand what it says, it might be helpful to observe that the same property does not hold for the set $\mathbb{Q}$ of *rational numbers* described below.

A *rational number* is a number that can be expressed as a fraction with an integer as the numerator and a non-zero integer as the denominator. The set of all rational numbers is denoted by $\mathbb{Q}$ ($\mathbb{Q}$ is for *quotient*).

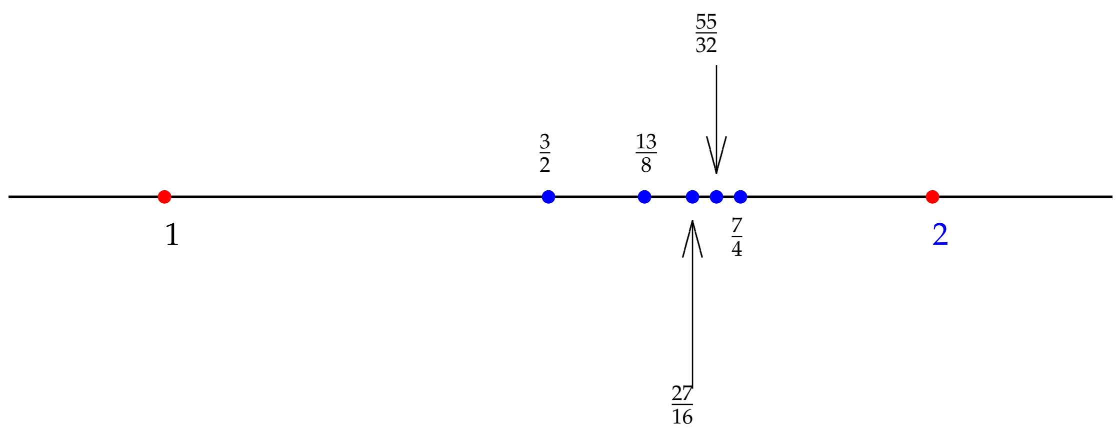
$$\mathbb{Q} = \left\{ \frac{a}{b} : a \in \mathbb{Z}, b \in \mathbb{Z}, \ b \neq 0 \right\}.$$

Note: The above statement can be *read* as "*Q is the set of all numbers that can be written in the form* $\frac{a}{b}$ *where* a *is an integer,* b *is an integer, and* b *is not zero*". In order to be able to make sense of written mathematics it is essential to be able to read all the parts of statements of this kind and form a clear mental impression of what is being said. In written mathematics, every mark on the page has meaning and you are expected to attend to that. This takes practice and care and it is not optional if you want to learn mathematics. Mathematical writing is and needs to be entirely unambiguous - this means it has to be technical and fussy, but there is no opportunity for misinterpretation once you are familiar with the relevant definitions and notation.

So $\mathbb{Q}$ includes such numbers as $\frac{5}{7}$, $\frac{-8}{16}$, $\frac{3141}{22445}$ and so on. It includes all the integers, since any integer n can be written in the form of a fraction as $\frac{n}{1}$. The rational numbers are exactly those numbers whose decimal representations either terminate (i.e. all digits are 0 from some point onwards) or repeat (i.e. from some point onwards the decimal part consists of some string of digits repeated over and over without end).

Note: The statement that $\mathbb{Q}$ includes all the integers can be written very concisely as $\mathbb{Z} \subset \mathbb{Q}$. This says : $\mathbb{Z}$ is a subset of $\mathbb{Q}$, i.e. every element of $\mathbb{Z}$ is an element of $\mathbb{Q}$, i.e. every integer is a rational number.

Since rational numbers (by definition) can be written as quotients (or fractions) involving integers, the sets $\mathbb{Q}$ and $\mathbb{Z}$ are closely related. However, on the number line these sets do not resemble each other at all. As mentioned above, the integers are "spaced out" on the number line and there are gaps between them. By contrast, the rational numbers are "densely packed" in the number line. The diagram below is intended to show that the stretch of the number line between 1 and 2 contains infinitely many rational numbers - we can't draw infinitely many of them in a picture, but hopefully this picture indicates how we can keep adding more and more of them indefinitely. By contrast with the situation for $\mathbb{Z}$, there are no stretches of the number line that are without rational numbers.



Note: Does this picture persuade you that there are infinitely many rational numbers between 1 and 2? If not, do you believe this statement? It is up to you to consider its plausibility and reach a conclusion.

Related to these remarks is the observation that the set of rational numbers is not *well-ordered*. Given a particular rational number, there is no *next* rational number after it. For example 0 is a rational number, but there is no *next rational number* after 0 on the number line. This is the same as saying that there is no *smallest* positive rational number. If we had a candidate for this title, we could divide by 2 and we would have a smaller but still positive rational number.

Exercise 1.1.1. *Choose a stretch of the number line - for example the stretch from $-\frac{7}{4}$ to $-\frac{11}{8}$ (but pick your own example). Persuade yourself that your stretch contains infinitely many rational numbers.*

So the rational numbers are not sparse like the integers. They come close to covering the whole number line in the sense that any stretch (however short) of the number line contains infinitely many rational numbers. The idea of visualizing the points corresponding to rational numbers as a "mist" on the number line has been suggested.

Now imagine an infinite straight line, on which the integers are marked (in order) by an infinite set of evenly spaced dots. Imagine that the rational numbers have also been marked by dots, so that the dot representing $\frac{3}{2}$ is halfway between the dot representing 1 and the dot representing 2, and so on. (Of course it is not physically possible to do all this marking, but it's possible to imagine what the picture would look like). At this stage a lot of dots have been marked - every stretch of the line, no matter how short, contains an infinite number of marked dots.

However, many points on the line remain unmarked. For example, somewhere between the dot representing the rational number 1.4142 and the dot representing the rational number 1.4143 is an unmarked point that represents the real number $\sqrt{2}$. This number is not rational - it cannot be expressed in the form $\frac{a}{b}$ for integers a and b. Somewhere between the dot that marks 3.1415 and the dot that marks 3.1416 is an unmarked point representing the irrational number π. The set $\mathbb{R}$ of *real numbers* is the set of numbers corresponding to *all* points on the line, marked or not. It includes both the rational and irrational numbers.

Note: Because the examples of irrational numbers that are usually cited are things like $\sqrt{2}, \pi$ and e, you could get the impression that irrational numbers are special and rare. This is far from being true. In a very precise way that we will see later, the irrational numbers are more numerous that the rational numbers. If you think of the points representing irrational numbers as a "mist" on the number line, it would be a denser mist than the one for rational numbers. If all the rational numbers were coloured blue on the number line and all the irrational numbers were coloured red, the whole number line would be a jumble of blue and red points, but there would be more red than blue. If you zoomed in, as far as you like, on any section of the number line, however short, both blue and red would still appear, and there would still be more red than blue.

Exercise 1.1.2. *Write down five irrational numbers between 4 and 5.*

Hint: If you don't know what to do, start with $\sqrt{2}$ or some other number that you know is irrational. The number $\sqrt{2}$ is not between 4 and 5 obviously. What adjustments can be made?

Exercise 1.1.3. *Suppose that a is a rational number and b is an irrational number.*

- *Might $a + b$ be irrational?*
- *Must $a + b$ be irrational?*
- *Might ab be irrational?*
- *Must ab be irrational?*
- *Might the product of two irrational numbers be irrational?*
- *Must the product of two irrational numbers be irrational?*

Hint : If in doubt, give yourself some examples and investigate.

To conclude this section we propose two different ways of thinking about the set of real numbers.

1. (*Arithmetic description*) The set $\mathbb{R}$ of real numbers consists of *all* numbers that can be written as (possibly non-terminating and possibly non-repeating) decimals.

   This description is accurate and conceptually it is valuable, but it is not of much practical use because it is not possible to write out a non-terminating non-repeating decimal or do calculations with it. In reality, when we do calculations with decimals (either by hand or by machine), we truncate them at some point and work with approximations which are rational numbers. The set of numbers that a calculator works with is not the set of real numbers or even the set of rational numbers - it is some subset of $\mathbb{Q}$ that depends on the precision of the instrument.

   This arithmetic description of the real numbers highlights the following point. *All* numbers that can be expressed as decimals means *all* numbers that can be written as sequences of the digits 0,1,…,9 (with a decimal point somewhere) with no pattern of repetition necessary in the digits. In the universe of all such things, the ones that terminate (i.e. end in an infinite string of zeroes) or have a repeating pattern from some point onwards are special and rare. These are the rational numbers. The ones that have all zeroes after the decimal point are even more special - these are the integers.

Later we will look at the following questions, which might seem at first glance not to even make sense, but which have interesting and surprising answers.

- Are there more rational numbers than integers?
- Are there more real numbers than rational numbers?

2. (*Geometric description*) The set $\mathbb{R}$ of real numbers is the set of *all* points on the number line. This is a *continuum* - there are no gaps in the real numbers and no point on the line that doesn't correspond to a real number.

Note: As this course proceeds you will need to know what the symbols $\mathbb{Z}$, $\mathbb{Q}$ and $\mathbb{R}$ mean and be able to recall this information easily. You'll need to be familiar with all the notation involving sets etc. that is used in this section and to be able to use it in an accurate way.

1.2 Subsets of $\mathbb{R}$

In this section we consider the notions of *finite* and *infinite* sets, and the *cardinality* of a set. Reasonable goals for this section are to become familiar with these ideas and to practice interpreting descriptions of sets that are presented in terse mathematical notation (this means, amongst other things, distinguishing between different kinds of brackets : $\{\}$, $[\,]$, $(\,)$, etc.).

Definition 1.2.1. *A set is* finite *if it is possible to list its distinct elements one by one, and this list comes to an end.*
A set is infinite *if any attempt at listing its distinct elements continues indefinitely.*

Example 1.2.2. *The point of this example is not only to show some finite and infinite sets but also to consider the notation that is used to describe them.*

1. $\{1, 2, 3, 4, 5\}$ is a *finite* set - its only elements are the integers 1,2,3,4,5, there are five of them.

2. The interval $[1, 3]$ is an *infinite* set - it consists of *all* the real numbers that are at least equal to 1 and at most equal to 3.
$$[1, 3] := \{x \in \mathbb{R} : 1 \leqslant x \leqslant 3\}.$$
Note: The symbol ":=" here means this is a statement of the *definition* of $[1, 3]$.

3. $\mathbb{Z}$ and $\mathbb{Q}$ are *infinite* sets.

4. The set of real solutions of the equation
$$x^5 + 2x^4 - x^2 + x + 17 = 0$$
is a *finite* set. We don't know how many elements it has, but it has at most five, since each one corresponds to a factor of degree 1 of this polynomial of degree 5.

5. The set of prime numbers is *infinite*.
A pair of *twin primes* is a pair of primes that differ by 2 : e.g. 3 and 5, 11 and 13, 59 and 61. It is not known whether the set of pairs of twin primes is finite or infinite.

Definition 1.2.3. *The* cardinality *of a finite set* S*, denoted* $|S|$*, is the number of elements in* S*.*

Example 1.2.4.

1. If $S = \{5, 7, 8\}$ then $|S| = 3$.

2. $|\{4, 10, \pi\}| = 3$

3. $|\{x \in \mathbb{Z} : \pi < x < 3\pi\}| = 6$.
Note: $\{x \in \mathbb{Z} : \pi < x < 3\pi\} = \{4, 5, 6, 7, 8, 9\}$.

4. The cardinality of $\mathbb{Q}$ is infinite.

Remarks

1. The notation "$|\cdot|$" is severely overused in mathematics. This can be a bit annoying since mathematical text is supposed to be entirely unambiguous. If x is a real number, $|x|$ means the absolute value of x. If S is a set, $|S|$ means the cardinality of S. If A is a matrix $|A|$ means the determinant of A. There are other usages as well. It is supposed to be clear from the context what is meant.

2. Defining the concept of cardinality for infinite sets is trickier, since you can't say how many elements they have. We will be able to say though what it means for two infinite sets to have the same (or different) cardinalities.

Example 1.2.5. *(A silly example) In a hotel, keys for all the guest rooms are kept on hooks behind the reception desk. If a room is occupied, the key is missing from its hook because the guests have it. If the receptionist wants to know how many rooms are occupied, s/he doesn't have to visit all the rooms to check - s/he can just count the number of hooks whose keys are missing.*

There is nothing deep about this example, but it illustrates a point that is important in mathematics. In the example, the occupied rooms are in *one-to-one correspondence* with the empty hooks. This means that each occupied room corresponds to *one and only one* empty hook, and each empty hook corresponds to *one and only one* occupied room. So the number of empty hooks is the same as the number of occupied rooms and we can count one by counting the other.

Definition 1.2.6. *Suppose that* A *and* B *are sets. Then a* one-to-one correspondence *or a* bijective correspondence *between* A *and* B *is a pairing of each element of* A *with an element of* B*, in such a way that every element of* B *is matched to exactly one element of* A.

Definition 1.2.7. *Suppose that* A *and* B *are sets. A function* $f : A \longrightarrow B$ *is called a* bijection *if*

- *Whenever* a_1 *and* a_2 *are* different *elements of* A, $f(a_1)$ *and* $f(a_2)$ *are* different *elements of* B.
- Every *element* b *of* B *is the image of some element* a *of* A.

Note: Definitions 1.2.6 and 1.2.7 are not really much different from each other, but the second one has a bit more technical machinery of a sort that is sometimes useful in trying to describe correspondences. The example about the keys and rooms is a case of both. The sets A and B here are the respectively the set of empty hooks and occupied rooms. The bijective correspondence is the matching of each empty hook to the room opened by its key, and the "function" f associates to each hook the corresponding room. The fact that different hooks have different images under f says that each key opens only one room, and the fact that every element of B is the image of *some* element of A says that every occupied room is opened by *some* key belonging to an empty hook.

The translation between the concrete context of Example 1.2.5 and the formal definition 1.2.7 is completely unnecessary in terms of understanding the example, but the point is that sometimes the objects we are dealing with don't have a concrete context like this and the formal language is actually necessary to describe the situation. The point of the example is just to show that Definition 1.2.7 is not as obscure or as complicated as it might seem at first glance.

Quite often, in order to determine the cardinality of a set, it is easiest to determine the cardinality of another set with which we know it is in bijective correpsondence.

Example 1.2.8. *How many integers between 1 and 1000 are perfect squares?*

Solution: The list of perfect squares in our range begins as follows

$$1,\ 4,\ 9,\ 16,\dots$$

One way of solving the problem would be to keep writing out successive terms of this sequence until we hit one that exceeds 1000, and then delete that one and count the terms that we have. This is actually more work than we are asked to do, since we are not asked for the list of squares but just the number of them.

Alternatively, we could notice that $(31)^2 = 961$ and $(32)^2 = 1024$.
So the numbers $1^2, 2^2, \dots, (31)^2$ are all in the range 1 to 1000 and these are the only perfect squares in that range, the answer to our question is 31.

What we are using here, technically, is the fact that the set of perfect squares in the range of interest is in *bijective correspondence* with the set $\{1, 2, 3, \dots, 31\}$ - the numbers in question are the squares of the first 31 natural numbers. To get the answer 31, it's not really the squares in the range 1 to 1000 that we are counting but the integers in the range 1 to 31.

The following example shows that it could be possible to know that there is a bijective correspondence between two finite sets, without knowing the cardinality of either of them. While this example is a bit contrived, the point of it is to see that sometimes we can show that two sets must be in bijective correspondence even if we don't know much about their elements. This can be a useful device.

Example 1.2.9. *Show that the equations*

$$x^3 + 2x + 4 = 0 \textit{ and } x^3 + 3x^2 + 5x + 7 = 0$$

have the same number of real solutions.

Solution: One way of doing this without having to solve the equations is to demonstrate a bijective correspondence their sets of real equations. We can write

$$\begin{aligned} x^3 + 3x^2 + 5x + 7 &= (x^3 + 3x^2 + 3x + 1) + 2x + 6 \\ &= (x+1)^3 + (2x+2) + 4 \\ &= (x+1)^3 + 2(x+1) + 4. \end{aligned}$$

This means that a real number a is a solution of the second equation if and only if

$$(a+1)^3 + 2(a+1) + 4 = 0$$

i.e. if and only if $a+1$ is a solution of the first equation.

The correspondence $a \longleftrightarrow a+1$ is a bijective correspondence between the solution sets of the two equations. So they have the same number of real solutions.

Note: This number is either 1 or 3. Why?

1.3 Infinite sets and cardinality

Recall from the last section that

- The cardinality of a finite set is defined as the number of elements in it.
- Two sets A and B have *the same cardinality* if (and only if) it is possible to match each element of A to an element of B in such a way that every element of each set has exactly one "partner" in the other set. Such a matching is called a *bijective correpondence* or *one-to-one correspondence*. A bijective correspondence between A and B may be expressed as a *function* from A to B that assigns different elements of B to all the elements of A and "uses" all the elements of B. A function that has these properties is called a *bijection*.

In the case of finite sets, the second point above might seem to be overcomplicating the issue, since we can tell if two finite sets have the same cardinality by just counting their elements and noting that they have the same number. The notion of bijective correspondence is emphasized for two reasons. First, as we saw in Example 1.2.9, it is occasionally possible to establish that two finite sets are in bijective correspondence without knowing the cardinality of either of them. More importantly, we would like to develop some notion of cardinality for infinite sets aswell. We can't count the number of elements in an infinite set. However, for a given pair of infinite sets, we could possibly show that it is or isn't possible to construct a bijective correspondence between them.

Definition 1.3.1. *Suppose that* A *and* B *are sets (finite or infinite). We say that* A *and* B *have the same cardinality (written* $|A| = |B|$*) if a bijective correspondence exists between* A *and* B.

In other words, A and B have the *same cardinality* if it's possible to match each element of A to a different element of B in such a way that every element of both sets is matched exactly once. In order to say that A and B have *different* cardinalities we need to establish that it's impossible to match up their elements with a bijective correspondence. If A and B are infinite sets, showing that such a thing is *impossible* is a formidable challenge.

The remainder of this section consists of a collection of examples of pairs of sets that have the same cardinality. Recall the following definition.

Definition 1.3.2. *The set* $\mathbb{N}$ *of natural numbers ("counting numbers") consists of all the positive integers.*

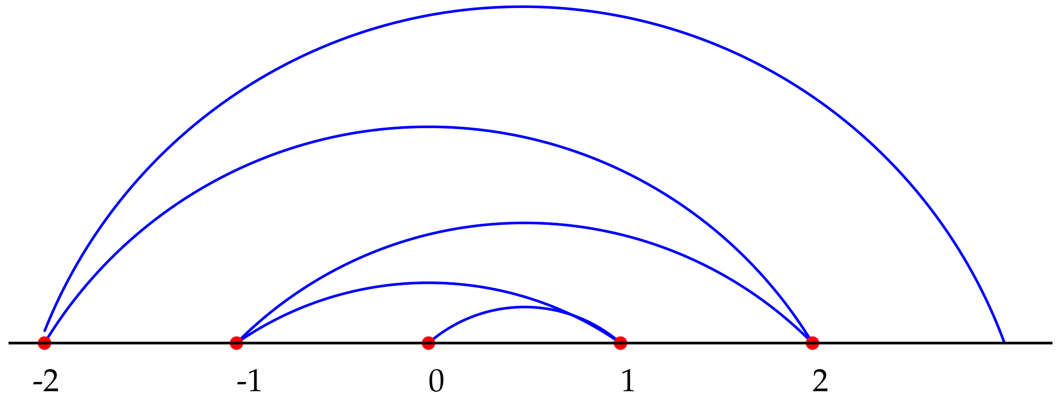
$$\mathbb{N} = \{1, 2, 3, \dots\}.$$

Example 1.3.3. *Show that* $\mathbb{N}$ *and* $\mathbb{Z}$ *have the same cardinality.*

The sets $\mathbb{N}$ and $\mathbb{Z}$ are both infinite obviously. In order to show that $\mathbb{Z}$ has the same cardinality of $\mathbb{N}$, we need to show that the right-hand column of the table below can be filled in with the integers *in some order*, in such a way that each integer appears there exactly once.

$\mathbb{N}$		$\mathbb{Z}$
1	$\longleftrightarrow$	?
2	$\longleftrightarrow$	?
3	$\longleftrightarrow$	?
4	$\longleftrightarrow$	?
$\vdots$	$\longleftrightarrow$	$\vdots$

So we need to list all the integers on the right hand side, in such a way that every integer appears once. Just following the natural order on the integers won't work, because then there is no first entry for our list. Starting at a particular integer like 0 and then following the natural order won't work, because then we will never get (for example) any negative integers in our list. Something that *will* work is suggested by following the arcs, starting from 0, in the picture below.



We can start with 0, then list 1 and then -1, then 2 and then -2, then 3 and then -3 and so on. This is a systematic way of writing out the integers, in the sense that given any integer, we can identify the one position where it will appear in our list. For example the integer 10 will be Item 20 in our list, the integer -11 will be Item 22.

Our table becomes

$\mathbb{N}$		$\mathbb{Z}$
1	$\longleftrightarrow$	0
2	$\longleftrightarrow$	1
3	$\longleftrightarrow$	-1
4	$\longleftrightarrow$	2
5	$\longleftrightarrow$	-2
6	$\longleftrightarrow$	3
$\vdots$	$\longleftrightarrow$	$\vdots$

If we want to be fully explicit about how this bijective correspondence works, we can even give a formula for the integer that is matched to each natural number. The correspondence above describes a bijective function $f : \mathbb{N} \longrightarrow \mathbb{Z}$ given for $n \in \mathbb{N}$ by

$$f(n) = \begin{cases} \frac{n}{2} & \text{if} \quad n \text{ is even} \\ -\left(\frac{n-1}{2}\right) & \text{if} \quad n \text{ is odd} \end{cases}$$

Exercise 1.3.4. *What integer occurs in position 50 in our list? In what position does the integer -65 appear?*
As well as understanding this example at the informal/intuitive level suggested by the picture above, think about the formula above, and satisfy yourself that it does indeed descibe a bijection between $\mathbb{N}$ and $\mathbb{Z}$. If you are convinced that the two questions above (and all others like them) have unique answers that can be worked out, this basically says that our correspondence between $\mathbb{N}$ and $\mathbb{Z}$ is bijective.

Example 1.3.3 demonstrates a curious thing that can happen when considering cardinalities of infinite sets. The set $\mathbb{N}$ of natural numbers is a *proper subset* of the the set $\mathbb{Z}$ of integers (this means that every natural number is an integer, but the natural numbers do not account for all the integers). Yet we have just shown that $\mathbb{N}$ and $\mathbb{Z}$ are in bijective correspondence. So it is possible for an infinite set to be in bijective correspondence with a proper subset of itself, and hence to have the same cardinality as a proper subset of itself. This can't happen for finite sets (why?).

Putting an infinite set in bijective correspondence with $\mathbb{N}$ amounts to providing a robust and unambiguous scheme or instruction for listing all its elements starting with a first, then a second, third, etc., in such a way that it can be seen that every element of the set will appear exactly once in the list.

Definition 1.3.5. *A set is called* countably infinite *(or* denumerable*) if it can be put in bijective correspondence with the set of natural numbers. A set is called* countable *if it is either finite or countably infinite.*

Basically, an infinite set is countable if its elements can be listed in an inclusive and organised way. “Listable” might be a better word, but it is not really used. Example 1.3.3 shows that the set $\mathbb{Z}$ of integers is countable. To fully appreciate the notion of countability, it is helpful to look at an example of an infinite set that is not countable. This is coming up in Section 1.4.

Thus according to Definition 1.3.1, the sets $\mathbb{N}$ and $\mathbb{Z}$ have the same cardinality. Maybe this is not so surprising, because $\mathbb{N}$ and $\mathbb{Z}$ have a strong geometric resemblance as sets of points on the number line. What is more surprising is that $\mathbb{N}$ (and hence $\mathbb{Z}$) has the same cardinality as the set $\mathbb{Q}$ of all rational numbers. These sets do not resemble each other much in a geometric sense. The natural numbers are sparse and evenly spaced, whereas the rational numbers are densely packed into the number line. Nevertheless, as the following construction shows, $\mathbb{Q}$ is a countable set.

Example 1.3.6. *Show that $\mathbb{Q}$ is countable.*

We need to show that the rational numbers can be organized into a numbered list in a systematic way that includes all of them. Such a list is a one-to-correspondence with the set $\mathbb{N}$ of natural numbers. To construct such a list, start with the following array of fractions.

$$\begin{array}{ccccccccc}
\dots & -3/1 & -2/1 & -1/1 & 0/1 & 1/1 & 2/1 & 3/1 & \dots \\
\dots & -3/2 & -2/2 & -1/2 & 0/2 & 1/2 & 2/2 & 3/2 & \dots \\
\dots & -3/3 & -2/3 & -1/3 & 0/3 & 1/3 & 2/3 & 3/3 & \dots \\
\dots & -3/4 & -2/4 & -1/4 & 0/4 & 1/4 & 2/4 & 3/4 & \dots \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots &
\end{array}$$

In these fractions, the numerators increase through all the integers as we travel along the rows, and the denominators increase through all the natural numbers as we travel downwards through the columns. Every rational number occurs somewhere in the array. In order to construct a bijective correspondence between $\mathbb{N}$ and the set of fractions in our array, we construct a systematic path that will visit every fraction in the array exactly once. One way of doing this (not the only way) is suggested by the arrows in the following diagram.

$$\begin{array}{ccccccccccccccc}
\dots & -3/1 & & -2/1 & \leftarrow & -1/1 & & 0/1 & \rightarrow & 1/1 & & 2/1 & \rightarrow & 3/1 & \dots \\
 & \uparrow & & \downarrow & & \uparrow & & & & \downarrow & & \uparrow & & \downarrow & \\
\dots & -3/2 & & -2/2 & & -1/2 & \leftarrow & 0/2 & \leftarrow & 1/2 & & 2/2 & & 3/2 & \dots \\
 & \uparrow & & \downarrow & & & & & & & & \uparrow & & \downarrow & \\
\dots & -3/3 & & -2/3 & \rightarrow & -1/3 & \rightarrow & 0/3 & \rightarrow & 1/3 & \rightarrow & 2/3 & & 3/3 & \dots \\
 & \uparrow & & & & & & & & & & & & \downarrow & \\
\dots & -3/4 & \leftarrow & -2/4 & \leftarrow & -1/4 & \leftarrow & 0/4 & \leftarrow & 1/4 & \leftarrow & 2/4 & \leftarrow & 3/4 & \dots \\
 & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots &
\end{array}$$

This path determines a listing of all the fractions in the array, that starts as follows

$$0/1,\ 1/1,\ 1/2,\ 0/2,\ -1/2,\ -1/1,\ -2/1,\ -2/2,\ -2/3,\ -1/3,\ 0/3,\ 1/3,\ 2/3,\ 2/2,\ 2/1,\ 3/1,\ 3/2,\ 3/3,\ 3/4,\ \dots$$

What this example demonstrates is a bijective correspondence between the set $\mathbb{N}$ of natural numbers and the set of all fractions in our array. A bijective correspondence between some infinite set and $\mathbb{N}$ is really just an ordered listing of the elements of that set (“ordered” here just means for the purpose of the list, and the order in which elements are listed doesn’t need to relate to any natural order on the set). This is not (exactly) a bijective correspondence between $\mathbb{N}$ and $\mathbb{Q}$.

Exercise 1.3.7. *Why not? (Think about this before reading on.)*

The reason why not is that every rational number appears many times in our array. Already in the section of the list above we have $1/1$, $2/2$ and $3/3$ appearing - these represent different entries in our array but they all represent the same rational number. Equally, the fraction $3/4$ appears in the segment of the list above, but $6/8$, $9/12$ and $90/120$ will all appear later.

In order to get a bijective correspondence between $\mathbb{N}$ and $\mathbb{Q}$, construct a list of all the rational numbers from the array as above, but whenever a rational number is encountered that has already appeared, leave it out. Our list will begin

$$0/1,\ 1/1,\ 1/2,\ -1/2,\ -1/1,\ -2/1,\ -2/3,\ -1/3,\ 1/3,\ 2/3,\ 2/1,\ 3/1,\ 3/2,\ 3/4,\ \dots$$

We conclude that the rational numbers are countable.

Note : Unlike our one-to-one correspondence between $\mathbb{N}$ and $\mathbb{Z}$ in Example 1.3.3, in this case we cannot write down a simple formula to tell us what rational number will be Item 34 on our list (i.e. corresponds to the natural number 34) or where in our list the rational number $292/53$ will appear.

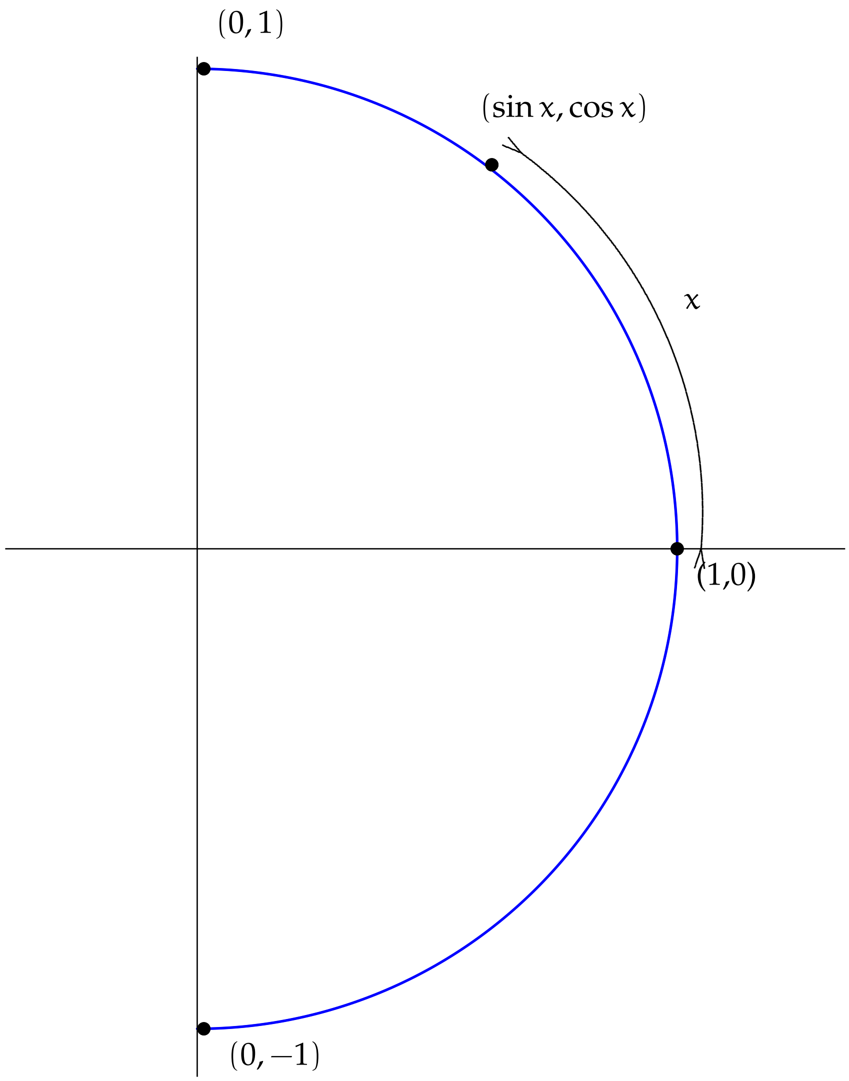
In our next example we show that the set of all the real numbers has the same cardinality as an open interval on the real line.

Example 1.3.8. *Show that $\mathbb{R}$ has the same cardinality as the open interval $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$*

Note: $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right) = \{x \in \mathbb{R} : -\frac{\pi}{2} < x < \frac{\pi}{2}$. The open interval $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ consists of all those real numbers that are strictly between $-\frac{\pi}{2}$ and $\frac{\pi}{2}$, not including $-\frac{\pi}{2}$ and $\frac{\pi}{2}$ themselves.

In order to do this we have to establish a bijective correspondence between the interval $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ and the full set of real numbers.

An example of a function that provides us with such a bijective correspondence is familiar from calculus/trigonometry. Recall that for a number x in the interval $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ $\tan x$ is defined as follows : travel from $(1, 0)$ a distance $|x|$ along the circumference of the unit circle, anti-clockwise if x is positive and clockwise if x is negative. We arrive at a point which is in the right-hand side of the unit circle, because we have travelled a distance of less than $\frac{\pi}{2}$ which would be one-quarter of a full lap. Now $\tan x$ is the ratio of the y-coordinate of this point to the x-coordinate (which are $\sin x$ and $\cos x$ respectively.



Now $\tan 0 = \dfrac{\sin 0}{\cos 0} = \dfrac{0}{1} = 0$, and as x increases from 0 towards $\frac{\pi}{2}$, $\tan x$ is increasing, since $\sin x$ (the y-coordinate of a point on the circle) is increasing and $\cos x$ (the x-coordinate) is decreasing. Moreover, since $\sin x$ is approaching 1 and $\cos x$ is approaching 0 as x approaches $\frac{\pi}{2}$, $\tan x$ is increasing without limit as x approaches $\frac{\pi}{2}$. Thus the values of $\tan x$ run through *all* the positive real numbers as x increases from 0 to $\frac{\pi}{2}$.

For the same reason, the values of $\tan x$ include *every* negative real number exactly once as x runs between 0 and $-\frac{\pi}{2}$.

Thus for $x \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ the correspondence

$$x \longleftrightarrow \tan x$$

establishes a bijection between the open interval $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ and the full set of real numbers. We conclude that the interval $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ has the same cardinality as $\mathbb{R}$.

Notes:

1. We don't know yet if $\mathbb{R}$ (or $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$) has the same cardinality as $\mathbb{N}$ - we don't know if $\mathbb{R}$ is countable.

2. The interval $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ may seem like an odd choice for an example like this. A reason for using it in this example is that is convenient for using the tan function to exhibit a bijection

with $\mathbb{R}$. However, note that the interval $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ is in bijective correspondence with the interval $(-1,1)$, via the function that just multiplies everything by $\frac{2}{\pi}$.

$$\begin{array}{ccc} \left(-\dfrac{\pi}{2}, \dfrac{\pi}{2}\right) & \longleftrightarrow & (-1,1) \\ x & \longleftrightarrow & \frac{2x}{\pi}. \end{array}$$

Exercise 1.3.9. *Show that the the open interval* $(0,1)$ *has the same cardinality as*

1. *The open interval* $(-1,1)$
2. *The open interval* $(1,2)$
3. *The open interval* $(2,6)$.

We finish this section now with a digression about bounded and unbounded subsets of $\mathbb{R}$.
Basically, a subset X of $\mathbb{R}$ is *bounded* if, on the number line, its elements are all enclosed within some region. In other words there exist real numbers a and b with $a < b$, for which all the points of X are in the interval (a, b).

Definition 1.3.10. *Let* X *be a subset of* $\mathbb{R}$. *Then* X *is* bounded below *if there exists a real number* a *with* $a < x$ *for* all *elements* x *of* X. *(Note that* a *need not belong to* X *here).*
The set X *is* bounded above *if there exists a real number* b *with* $x < b$ *for* all *elements* x *of* X. *(Note that* b *need not belong to* X *here).*
The set X *is* bounded *if it is bounded above and bounded below (otherwise it's* unbounded*).*

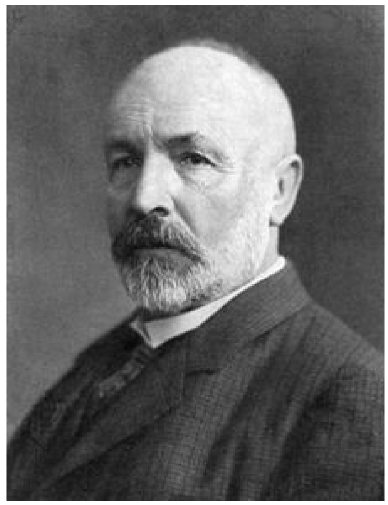
Example 1.3.11.

1. $\mathbb{Q}$ *is unbounded.*
2. $\mathbb{N}$ *is bounded below but not above.*
3. $(0,1)$, $[0,1]$, $[2,100]$ *and all open and closed intervals are bounded.*
4. $\{\cos x : x \in \mathbb{R}\}$ *is bounded, since* $\cos x$ *can only have values between* -1 *and* 1.

Remark: Example 1.3.8 shows that it is possible for a bounded subset of $\mathbb{R}$ to have the same cardinality as the full set $\mathbb{R}$ of real numbers.

1.4 $\mathbb{R}$ is uncountable

Our goal in this section is to show that the set $\mathbb{R}$ of real numbers is *uncountable* or *non-denumerable*; this means that its elements cannot be *listed*, or cannot be put in one-to-one correspondence with the natural numbers. We saw at the end of Section 1.3 that $\mathbb{R}$ has the same cardinality as the interval $(-\frac{\pi}{2}, \frac{\pi}{2})$, or the interval $(-1, 1)$, or the interval $(0, 1)$. We will show that the open interval $(0, 1)$ is uncountable. This assertion and its proof date back to the 1890's and to Georg Cantor. The proof is often referred to as "Cantor's diagonal argument" and applies in more general contexts than we will see in these notes.



Georg Cantor : born in St Petersburg (1845), died in Halle (1918)

Theorem 1.4.1. *The open interval $(0, 1)$ is not a countable set.*

Before embarking on a proof, we recall precisely what this set is. It consists of all real numbers that are greater than zero and less than 1, or equivalently of all the points on the number line that are to the right of 0 and to the left of 1. It consists of all numbers whose decimal representation have only 0 before the decimal point (except $0.000\ldots$ which is equal to 0, and $0.99999\ldots$ which is equal to 1). Note that the digits after the decimal point may terminate in an infinite string of zeros, or may have a repeating pattern to their digits, or may not have either of these properties. The interval $(0, 1)$ includes all these possibilities.

Our goal is to show that the interval $(0, 1)$ cannot be put in bijective correspondence with the set $\mathbb{N}$ of natural numbers. Our strategy is to show that no attempt at constructing a bijective correspondence between these two sets can ever be complete; it can never involve *all* the real numbers in the interval $(0, 1)$ no matter how it is devised. In order to achieve this, we will imagine that we had a listing of the elements of the interval $(0, 1)$; i.e. a bijective correspondence between this interval and $\mathbb{N}$. Such a correspondence would have to look something like the following.

$\mathbb{N}$		$(0,1)$
1	$\longleftrightarrow$	$0.13567324\ldots$
2	$\longleftrightarrow$	$0.10000000\ldots$
3	$\longleftrightarrow$	$0.32323232\ldots$
4	$\longleftrightarrow$	$0.56834662\ldots$
5	$\longleftrightarrow$	$0.79993444\ldots$
$\vdots$		$\vdots$

Note: The exact numbers that appear in the right-hand column above are not important, the point is that a bijective correspondence between $\mathbb{N}$ and $(0, 1)$ would have this general form. We don't know whether any particular decimal number in the right hand side terminates in zeros (or repeats) or not, but we know that some do and some don't.

So the entries in the right hand column above are basically infinite sequences of digits from 0 to 9. The right hand column must then consist somehow of a list of *all* such sequences. Our problem is to show that this is impossible : that no matter how the right hand column is constructed, it can't contain *every* sequence of digits from 1 to 9. We can do this by exhibiting an example of a sequence that can't possibly be there.

Suppose our list starts as follows.

$$\begin{array}{ccl} \mathbb{N} & & (0,1) \\ \hline \\ 1 & \longleftrightarrow & 0.\mathbf{1}3567324\ldots \\ 2 & \longleftrightarrow & 0.1\mathbf{0}000000\ldots \\ 3 & \longleftrightarrow & 0.32\mathbf{3}23232\ldots \\ 4 & \longleftrightarrow & 0.568\mathbf{3}4662\ldots \\ 5 & \longleftrightarrow & 0.7999\mathbf{3}444\ldots \\ \vdots & & \vdots \end{array}$$

We will construct an element x of $(0,1)$ that is not in the list. To do so :

1. Look at the *first* digit after the decimal point in Item 1 in the list. If this is 1, write 2 as the first digit after the decimal point in x. Otherwise, write 1 as the first digit after the decimal point in x. *So x differs in its first digit from Item 1 in the list*.

2. Look at the *second* digit after the decimal point in Item 2 in the list. If this is 1, write 2 as the second digit after the decimal point in x. Otherwise, write 1 as the second digit after the decimal point in x. *So x differs in its second digit from Item 2 in the list*.

3. Look at the *third* digit after the decimal point in Item 3 in the list. If this is 1, write 2 as the third digit after the decimal point in x. Otherwise, write 1 as the third digit after the decimal point in x. *So x differs in its third digit from Item 3 in the list*.

4. Continue to construct x digit by digit in this manner. At the nth stage, look at the nth digit after the decimal point in Item n in the list. If this is 1, write 2 as the nth digit after the decimal point in x. Otherwise, write 1 as the nth digit after the decimal point in x. *So x differs in its nth digit from Item n in the list*.

What this process constructs is an element x of the interval $(0,1)$ that does not appear in the proposed list. The number x is not Item 1 in the list, because it differs from Item 1 in its 1st digit, it is not Item 2 in the list because it differs from Item 2 in its 2nd digit, it is not Item n in the list because it differs from Item n in its nth digit.

We conclude that the set of real numbers $\mathbb{R}$ is *not countable* (or *uncountable*).

Note:

1. In our example, the number x would start $0.21111\ldots$.

2. According to our construction, our x will always have all its digits equal to 1 or 2. So not only have we shown that the interval $(0,1)$ is uncountable, we have even shown that the set of all numbers in this interval whose digits are all either 1 or 2 is uncountable.

3. A challenging exercise : why would the same proof not succeed in showing that the set of rational numbers in the interval $(0,1)$ is uncountable?

Informally, Cantor's diagonal argument tells us that the "infinity" that is the cardinality of the real numbers is "bigger" than the "infinity" that is the cardinality of the natural numbers, or integers, or rational numbers. He was able to use the same argument to construct examples of infinite sets of different (and bigger and bigger) cardinalities. So he actually established the notion of infinities of different magnitudes.

The work of Cantor was not an immediate hit within his own lifetime. It met some opposition from the finitist school which held that only mathematical objects that can be constructed in a finite number of steps from the natural numbers could be considered to exist. Foremost among the proponents of this viewpoint was Leopold Kronecker. From the book "The Honors Class" by Ben Yandell :

> Many mathematicians, Leopold Kronecker in Berlin, in particular, were bothered by this headlong leap into the infinite, accessible only by inference, not finite construction. Georg Cantor (1845-1918), teaching at Halle in 1888, had invented set theory in the 1870s and was writing about infinities of different sizes and even doing arithmetic with them. But Kronecker would admit only numbers or other mathematical objects that were finitely 'constructible'.

Leopold Kronecker (1823-1891)

> God made the integers, all else is the work of man.
>
> What good your beautiful proof on π? Why investigate such problems, given that irrational numbers do not even exist?

The work of Cantor had influential admirers too, among them David Hilbert, who set the course of much of 20th Century mathematics in his address to the International Congress of Mathematicians in Paris in 1900.

David Hilbert (1862-1943)

> No one shall expel us from the paradise that Cantor has created for us.
>
> What new methods and new facts in the wide and rich field of mathematical thought will the new centuries disclose?

Hilbert's address to the Paris Congress is one of the most famous mathematical lectures ever. In it he posed 23 unsolved problems, the first of which was Cantor's *Continuum Hypothesis*. The Continuum Hypothesis proposes that every subset of $\mathbb{R}$ is either countable (i.e. has the same cardinality as $\mathbb{N}$ or $\mathbb{Z}$ or $\mathbb{Q}$) or has the same cardinality as $\mathbb{R}$. This seems like a question to which the answer is either a straightforward yes or no, but it took the work of Kurt Gödel in the 1930s and Paul Cohen in the 1960s to reach the remarkable conclusion that the answer to the question is *undecidable*. This means essentially that the standard axioms of set theory do not provide enough structure to determine the answer to the question. Both the Continuum Hypothesis and its negation are consistent with the working rules of mathematics. People who work in set theory can legitimately assume that either the Continuum Hypothesis is satisfied or not. Fortunately most of us can get on with our mathematical work without having to worry about this very often.

1.5 The Completeness Axiom in $\mathbb{R}$

The rational numbers and real numbers are closely related, even though the set $\mathbb{Q}$ of rational numbers is countable and the set $\mathbb{R}$ of real numbers is not, and in this sense there are many more real numbers than rational numbers. However, $\mathbb{Q}$ is "dense" in $\mathbb{R}$. This means that every interval of the real number line, no matter how short, contains infinitely many rational numbers. This statement has a practical impact as well, which we use all the time whether consciously or not.

Lemma 1.5.1. *Every real number (whether rational or not) can be approximated by a rational number with a level of accuracy as high as we like.*

Justification for this claim : 3 is a rational approximation for π.
3.1 is a closer one.
3.14 is closer again.
3.14159 is closer still.
3.1415926535 is even closer than that, and we can keep improving on this by truncating the decimal expansion of π at later and later stages. For example if we want a rational approximation that differs from the true value of π by less that 10^{-20} we can truncate the decimnal approximation of π at the 21st digit after the decimal point. This is what is meant by "a level of accuracy as high as we like" in the statement of the lemma.

Notes:

1. The fact that all real numbers can be approximated with arbitary closeness by rational numbers is used all the time in everyday life. Computers basically don't deal with all the real numbers or even with all the rational numbers, but with some specified level of precision. They really work with a subset of the rational numbers.

2. The *sequence*
$$3,\ 3.1,\ 3.14,\ 3.141,\ 3.1415,\ 3.14159,\ 3.141592,\ \dots$$
is a list of numbers that are steadily appraoching π. All of these numbers are less than π; they are increasing and they are approaching π. We say that this sequence *converges* to π and we will investigate the concept of convergent sequences in Chapter 2.

3. We haven't looked yet at the question of how the numbers in the above sequence can be calculated, i.e. how we can get our hands on better and better approximations to the value of the irrational number π. That's another thing that we will look at in Chapter 2.

The goal of this last section of Chapter 1 is to pinpoint one essential property of subsets of $\mathbb{R}$ that is not shared by subsets of $\mathbb{Z}$ or of $\mathbb{Q}$. We need a few definitions and some terminology in order to describe this.

Definition 1.5.2. *Let S be a subset of $\mathbb{R}$. An element b of $\mathbb{R}$ is an* upper bound *for S if $x \leqslant a$ for all $x \in S$. An element a of $\mathbb{R}$ is a* lower bound *for S if $a \leqslant x$ for all $x \in S$.*

So an upper bound for S is a number that is to the right of all elements of S on the real line, and a lower bound for S is a number that is to the left of all points of S on the real line. Note that if b is an upper bound for S, then so is every number b' with $b < b'$. If a is a lower bound for S then so is every number a' with $a' < a$. So if S has an upper bound at all it has infinitely many upper bounds, and if S has a lower bound at all it has infinitely many lower bounds. Recall that

- S is *bounded above* if it has an upper bound,
- S is *bounded below* if it has a lower bound,
- S is *bounded* if it is bounded both above and below.

In this section we are mostly interested in sets that are bounded on at least one side.

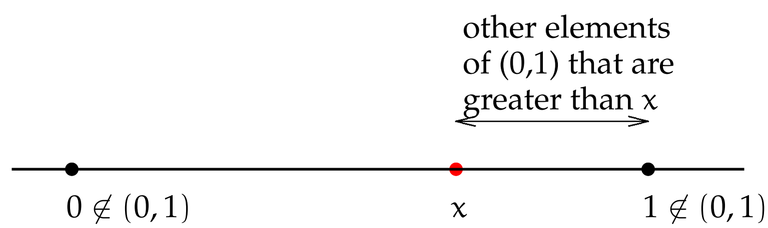
Definition 1.5.3. *Let S be a subset of $\mathbb{R}$. If there is a number m that is both an element of S and an upper bound for S, then m is called the* maximum element *of S and denoted* $\max(S)$.
If there is a number l that is both an element of S and a lower bound for S, then l is called the minimum element *of S and denoted by* $\min(S)$.

Notes

1. A set can have at most one maximum (or minimum) element. For suppose that both m and m' are maximum elements of S according to the definition. Then $m' \leqslant m$ because m is a maximum element of S, and $m \leqslant m'$ because m' is a maximum element of S. The only way that both of these statements can be true is if $m = m'$.

2. Pictorially, on the number line, the maximum element of S is the rightmost point that belongs to S, if such a point exists. The minimum element of S is the leftmost point on the number line that belongs to S, if such a point exists.

3. There are basically two reasons why a subset S of $\mathbb{R}$ might fail to have a maximum element. First, S might not be bounded above - then it certainly won't have a maximum element. Secondly S might be bounded above, but might not contain an element that is an upper bound for itself. Probably the easiest example of this to think about is an open interval like $(0, 1)$. This set is certainly bounded above - for example by 1. However, take any element x of $(0, 1)$. Then x is a real number that is strictly greater than 0 and strictly less than 1. Between s and 1 there are more real numbers all of which belong to $(0, 1)$ and are greater than x. So x is not an upper bound for the interval $(0, 1)$.



An open interval like $(0, 1)$, although it is bounded, has no maximum element and no minimum element.

An example of a subset of $\mathbb{R}$ that *does* have a maximum and a minimum element is a *closed* interval like $[2, 3]$. The minimum element of $[2, 3]$ is 2 and the maximum element is 3.

Remark : Every *finite* subset of $\mathbb{R}$ has a maximum element and a minimum element (these may be the same if the set has only one element).

For bounded subsets of $\mathbb{R}$, there are notions called the *supremum* and *infimum* that are closely related to maximum and minimum. Every subset of $\mathbb{R}$ that is bounded above has a supremum and every subset of $\mathbb{R}$ that is bounded below has an infimum. This is the *Axiom of Completeness* for $\mathbb{R}$.

Definition 1.5.4. *Let S be a subset of $\mathbb{R}$ that is bounded above. Then the set of all upper bounds for S has a minimum element. This number is called the* supremum *of S and denoted* $\sup(S)$.

Let S be a subset of $\mathbb{R}$ that is bounded below. Then the set of all lower bounds for S has a maximum element. This number is called the infimum *of S and denoted* $\inf(S)$.

Notes

1. The *supremum* of S is also called the *least upper bound (lub)* of S. It is the least of all the numbers that are upper bounds for S.

2. The *infimum* of S is also called the *greatest lower bound (glb)* of S. It is the greatest of all the numbers that are lower bounds for S.

3. Definition 1.5.4 is simultaneously a definition of the terms *supremum* and *infimum* and a statement of the *Axiom of Completeness* for the real numbers.

To see why this statement says something special about the real numbers, temporarily imagine that the only number system available to us is $\mathbb{Q}$, the set of rational numbers. Look at the set

$$S := \{x \in \mathbb{Q} : x^2 < 2\}.$$

So S consists of all those rational numbers whose square is less than 2. It is bounded below, for example by -2, and it is bounded above, for example by 2. This is saying that every rational number whose square is less than 2 is itself between -2 and 2 (of course we could narrow this interval with a bit more care). The positive elements of S are all those positive rational numbers that are less than the real number $\sqrt{2}$.

Claim 1: S does not have a maximum element.

To see this, suppose that x is a candidate for being the maximum element of S. Then x is a rational number and $x^2 \leqslant 2$. For any (very large) integer n, $x + \frac{1}{n}$ is a rational number and

$$\left(x + \frac{1}{n}\right)^2 = x^2 + 2\frac{x}{n} + \frac{1}{n^2}.$$

We can choose n large enough that $2\frac{x}{n} + \frac{1}{n^2}$ is so small that $\left(x + \frac{1}{n}\right)^2$ is still less than 2. Then the number $x + \frac{1}{n}$ belongs to S, and it is bigger than x, contrary to the hypothesis that x could be a maximum element of S.

Claim 2 : S has no least upper bound in $\mathbb{Q}$

To see this, suppose that x is a candidate for being a least upper bound for S in $\mathbb{Q}$. Then $x^2 > 2$. Note x^2 cannot be equal to 2 because $x \in \mathbb{Q}$.

For a (large) integer n

$$\left(x - \frac{1}{n}\right)^2 = x^2 - 2\frac{x}{n} + \frac{1}{n^2} = x^2 - \frac{1}{n}\left(2x - \frac{1}{n}\right).$$

Choose n large enough that $x^2 - \frac{1}{n}\left(2x - \frac{1}{n}\right)$ is still greater than 2. Then $x - \frac{1}{n}$ is still an upper bound for S, and it is less than x.
So S has no least upper bound in $\mathbb{Q}$.

If we consider the same set S as a subset of $\mathbb{R}$, we can see that $\sqrt{2}$ is the supremum of S in $\mathbb{R}$ (and $-\sqrt{2}$ is the infimum of S in $\mathbb{R}$).

This example demonstrates that the *Axiom of Completeness* does not hold for $\mathbb{Q}$, i.e. a bounded subset of $\mathbb{Q}$ need not have a supremum in $\mathbb{Q}$ or an infimum in $\mathbb{Q}$.

Example 1.5.5 (Summer Examinations 2011)**.** *Determine with proof the supremum and infimum of the set*

$$T = \left\{5 - \frac{5}{n} : n \in \mathbb{N}\right\}.$$

Solution: (Supremum) First, look at the numbers in the set. All of them are less than 5. They can be very close to 5 if n is large.
Guess: $\sup(T) = 5$.
We need to show :

1. 5 is an upper bound for T.
   To see this, suppose that $x \in T$. Then $x - 5 - \frac{5}{k}$ for some $k \in \mathbb{N}$. Since k is positive, $\frac{5}{k}$ is positive and $x < 5$. Hence 5 is an upper bound for T.

2. If b is an upper bound for T, then b cannot be less than 5.
   To see this suppose that $b < 5$, so $5 - b$ is a positive real number. We can choose a natural number m so large that $\frac{5}{m} < 5 - b$. Then $5 - \frac{5}{m} > b$, which means that b is not an upper bound for T, as $5 - \frac{5}{m} \in T$.

Thus 5 is the *least upper bound (supremum)* of T.

Solution: (Infimum) Look for the least elements of T. These occur when n is small : when $n = 1$ we get that $5 - \frac{5}{1} = 0$ belongs to T.
Guess: $\inf(T) = 0$.
We need to show :

1. 0 is a lower bound for T.
   To see this, suppose that $x \in T$. Then $x = 5 - \frac{5}{k}$ for some $k \in \mathbb{N}$. Since $k \in \mathbb{N}$, $k \geqslant 1$ and $\frac{5}{k} \leqslant 5$. Thus $5 - \frac{5}{k} \geqslant 5 - 5$ which means $x \geqslant 0$ and 0 is a lower bound for T.

2. If a is a lower bound for T, then a cannot be greater than 0.
   No number greater than 0 can be a lower bound for T, since $0 \in T$. Thus 0 is the *minimum element* (and therefore the infimum) of T.

Chapter 2

Sequences, Series and Convergence

2.1 Introduction to sequences and series

Example 2.1.1. *Does it make sense to talk about the "number"*

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = 1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \frac{1}{25} + \ldots ?$$

What does the question "does it make sense" mean? What we are talking about is the sum of infinitely many specified positive numbers. We can't actually do the addition and calculate what this "number" is based on the definition above. But we can add up any finite collection of the given terms and get an answer for that. Does this sum "settle down" to some identifiable value if we keep adding more terms (whatever that means)? Does it keep growing and growing without bound? Are there ways of finding out? Why would we want to know?

The following experiment might give a slightly vague but hopefully convincing answer to some of these questions. Partially evaluating the sum above for various "initial segments" gives the following results.

- $1 + \frac{1}{4} = 1.25$
- $1 + \frac{1}{4} + \frac{1}{9} \approx 1.361111$
- $1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} \approx 1.423611$
- $1 + \frac{1}{4} + \frac{1}{9} + \cdots + \frac{1}{(10)^2} \approx 1.549767$
- $1 + \frac{1}{4} + \frac{1}{9} + \cdots + \frac{1}{(200)^2} \approx 1.639947$
- $1 + \frac{1}{4} + \frac{1}{9} + \cdots + \frac{1}{(10000)^2} \approx 1.644834$

This experiment goes as far as computing the first 100000 terms of the sum, and it appears that the values are not increasing without limit but "settling down" at around 1.6449. What is the significance of this?

$$\frac{\pi^2}{6} \approx 1.644934$$

The *series* $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n^2}$ *converges* to the number $\dfrac{\pi^2}{6}$ (we will have precise definitions for the italicized terms a bit later). This fact is remarkable - there is no obvious connection between π and squares of the form $\frac{1}{n^2}$; moreover all the terms in the series are rational but $\dfrac{\pi^2}{6}$ is certainly not. This example gives us in principle a way of calculating the digits of π or at least of π^2. (In practice there are similar but better ways, as the convergence in this example is very slow).

Example 2.1.2. *What about*

$$\sum_{i=1}^{\infty} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots ?$$

Try experimenting with initial segments again :

-

There's no sign of this "settling down" or converging to anything that we can identify from this information. This doesn't tell us anything of course - maybe there is convergence but it can't be seen until we take many more millions of terms into our calculation? How could we know that this *doesn't* converge to anything?

Example 2.1.3. *What about*

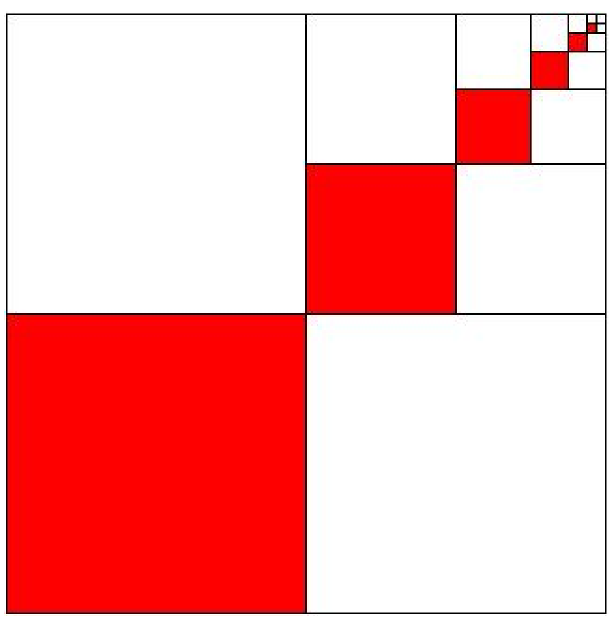
$$\sum_{i=1}^{\infty} \frac{1}{2^{2n}} = \frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \ldots ?$$

Experimenting reveals

- $\frac{1}{4} + \frac{1}{16} = \frac{5}{16}$
- $\frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \frac{1}{256} + \frac{1}{1024} = \frac{341}{1024} \approx 0.33301$

These calculations can be verified directly using properties of sums of geometric progressions. It appears that this series is converging to $\frac{1}{3}$.

The following picture gives some graphical evidence for this hypothesis. The large square has area 1, and the red squares have areas $\frac{1}{4}$, $\frac{1}{16}$, etc. The picture is intended to indicate that the red squares occupy one-third of the total area, since every red square is "accompanied" by two white squares of the same area, and all these squares together make up the total area 1. This picture is not really a proof, as it is not possible to actually draw squares representing all the terms of the series, but it is a visual way of understanding the statement that the series $\sum_{n=1}^{\infty} \frac{1}{2^{2n}}$ converges to $\frac{1}{3}$.



2.2 Sequences

Note: Chapter 11 of Stewart's Calculus is a good reference for this chapter of our lecture notes.

Definition 2.2.1. *A* sequence *is basically an infinite ordered list*

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

- The items in list a_1, a_2 etc. are called *terms* (1st term, 2nd term, and so on).
- In our context the terms will generally be real numbers - but they don't have to be.
- The sequence $a_1, a_2, \ldots$ can be denoted by $\{a_n\}$ or by $\{a_n\}_{n=1}^{\infty}$.
- There may be an overall formula for the terms of the sequence, or a "rule" for getting from one to the next, but there doesn't have to be.

Example 2.2.2. *1.* $\{(-1)^n + 1\}_{n=1}^{\infty} : a_n = (-1)^n + 1$
$a_1 = -1 + 1 = 0,\ a_2 = (-1)^2 + 1 = 2, a_3 = (-1)^3 + 1 = 0, \ldots$

$$0, 2, 0, 2, 0, 2, \ldots$$

2. $\{\sin(\frac{n\pi}{2})\}_{n=1}^{\infty} : a_n = \sin(\frac{n\pi}{2})$
$a_1 = \sin(\frac{\pi}{2}) = 1,\ a_2 = \sin(\pi) = 0,\ a_3 = \sin(\frac{3\pi}{2}) = -1,\ a_4 = \sin(2\pi) = 0, \ldots.$

$$1, 0, -1, 0, 1, 0, -1, 0, \ldots$$

3. $\{\frac{1}{n}\sin(\frac{n\pi}{2})\}_{n=1}^{\infty} : a_n = \sin(\frac{n\pi}{2})$
$a_1 = \sin(\frac{\pi}{2}) = 1,\ a_2 = \frac{1}{2}\sin(\pi) = 0,\ a_3 = \frac{1}{3}\sin(\frac{3\pi}{2}) = -\frac{1}{3},\ a_4 = \frac{1}{4}\sin(2\pi) = 0, \ldots.$

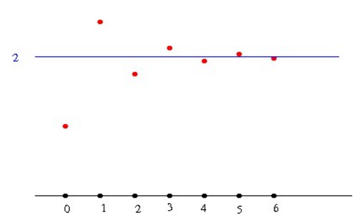
$$1, 0, -\frac{1}{3}, 0, \frac{1}{5}, 0, -\frac{1}{7}, 0, \ldots$$

One way of visualizing a sequence is to consider is as a function whose domain is the set of natural numbers and think of its graph, which will be a collection of isolated points, one for each natural number.

Example 2.2.3. $\{2 + (-1)^n 2^{1-n}\}_{n=1}^{\infty}$. *Write* $a_n = 2 + (-1)^n 2^{1-n}$. *Then*

$$a_1 = 2 - 2^0 = 1,\ a_2 = 2 + 2^{-1} = \frac{5}{2},\ a_3 = 2 - 2^{-2} = \frac{7}{4}, a_4 = 2 + 2^{-3} = \frac{17}{8}.$$

Graphical representation of $\{a_n\}$*:*



As n gets very large the positive number $\frac{1}{2^{n-1}}$ gets very small. By taking n as large as we like, we can make $\frac{1}{2^{n-1}}$ as small as we like.

Hence for very large values of n, the number $2 + (-1)^n \frac{1}{2^{n-1}}$ is very close to 2. By taking n as large as we like, we can make this number as close to 2 as we like.

We say that the sequence *converges* to 2, or that 2 is the *limit* of the sequence, and write

$$\lim_{n\to\infty}\left(2 + (-1)^n \frac{1}{2^{n-1}}\right) = 2.$$

Note: Because $(-1)^n$ is alternately positive and negative as n runs through the natural numbers, the terms of this sequence are alternately greater than and less than 2.

We now state the formal definition of convergence of a sequence. This is reminiscent of the definition of a limit for a function. A sequence converges to the number L if no matter how restrictive your notion of "near L" is, there is a point in the sequence beyond which every term is near L.

Definition 2.2.4. *The sequence* $\{a_n\}$ converges *to the number* L *(or has* limit L*) if for every positive real number* ε *(no matter how small) there exists a natural number* N *with the property that the term* a_n *of the sequence is within* ε *of* L *for all terms* a_n *beyond the* N*th term. In more compact language :*

$$\forall\varepsilon > 0,\ \exists N \in \mathbb{N} \text{ for which } |a_n - L| < \varepsilon\ \forall n > N.$$

Notes

- If a sequence has a limit we say that it *converges* or *is convergent*. If not we say that it *diverges* or *is divergent*.
- If a sequence converges to L, it means that no matter how small a radius around L we choose, there is a point in the sequence beyond which all terms are within that radius of L. This means (at least) that beyond a certain point all terms of the sequence are *very close together* (and very close to L). Where that point is depends on how you interpret "very close together".

Being convergent is a very strong property for a sequence to have, and there are lots of different ways for a sequence to be divergent.

Example 2.2.5. *1.* $\{\max(\{(-1)^n, 0\})\}_{n=1}^{\infty}$: 0, 1, 0, 1, 0, 1, ...
This sequence alternates between 0 and 1 and does not approach any limit.

2. *A sequence can be divergent by having terms that increase (or decrease) without limit.*
$\{2^n\}_{n=1}^{\infty}$: 2, 4, 8, 16, 32, 64, ...

3. *A sequence can have haphazard terms that follow no overall pattern, such as the sequence whose* n*th term is the* n*th digit after the decimal point in the decimal representation of* π*.*

Remark: The notion of a convergent is sometimes described informally with words like "the terms get closer and closer to L as n gets larger". It is *not true* however that the terms in a sequence that converges to a limit L *must* get *progressively* closer to L as n increases, as the following example shows.

Example 2.2.6. *The sequence* a_n *is defined by*

$$a_n = 0 \textit{ if } n \textit{ is even, } a_n = \tfrac{1}{n} \textit{ if } n \textit{ is odd.}$$

This sequence begins :

$$1,\ 0,\ \frac{1}{3},\ 0,\ \frac{1}{5},\ 0,\ \frac{1}{7},\ 0,\ \frac{1}{9},\ 0, \ldots$$

It converges to 0 *although it is not true that every step takes us closer to zero.*

The following is an example of a convergent sequence.

Example 2.2.7. *Find* $\lim_{n\to\infty} \dfrac{n}{2n-1}$.
Solution*: As if calculating a limit as* $x \to \infty$ *of an expression involving a continuous variable* x*, divide above and below by* n*.*

$$\lim_{n\to\infty} \frac{n}{2n-1} = \lim_{n\to\infty} \frac{n/n}{2n/n - 1/n} = \lim_{n\to\infty} \frac{1}{2 - \frac{1}{n}} = \frac{1}{2}.$$

So the sequence $\left\{\dfrac{n}{2n-1}\right\}$ *converges to* $\frac{1}{2}$.

As for subsets of $\mathbb{R}$, there is a concept of *boundedness* for sequences. Basically a sequence is bounded (or bounded above or bounded below) if the set of its terms, considered as a subset of $\mathbb{R}$, is bounded (or bounded above or bounded below). More precisely :

Definition 2.2.8. *The sequence* $\{a_n\}$ *is* bounded above *if there exists a real number* M *for which* $a_n \leqslant M$ *for all* $n \in \mathbb{N}$.
The sequence $\{a_n\}$ *is* bounded below *if there exists a real number* m *for which* $m \leqslant a_n$ *for all* $n \in \mathbb{N}$.
The sequence $\{a_n\}$ *is* bounded *if it is bounded both above and below.*

Example 2.2.9. *The sequence* $\{n\}$ *is bounded below (for example by* 0 *or* 1*) but not above. The sequence* $\{\sin n\}$ *is bounded below (for example by* -1*) and above (for example by* 1*).*

Theorem 2.2.10. *If a sequence is convergent it must be bounded.*

Proof

Note : what we have to do here is use the definitions of convergent and bounded to reason that every sequence that is convergent must be bounded.

Suppose that $\{a_n\}_{n=1}^{\infty}$ is a convergent sequence with limit L.

Then (by definition of convergence) there exists a natural number N such that every term of the sequence after a_N is between $L-1$ and $L+1$.

(Note: there is nothing special here about $L-1$ *and* $L+1$ *- you could choose* $L-\frac{1}{2}$ *and* $L+\frac{1}{2}$ *or anything like that - the point is that when you choose a certain "window" around* L*, there is a point (*N*) beyond which all the terms of the sequence are in this "window".)*

The set consisting of the first N terms of the sequence is a finite set : it has a maximum element M_1 and a minimum element m_1.
Let $M = \max\{M_1, L+1\}$ and let $m = \min\{m_1, L-1\}$.

(So M *is defined to be either* M_1 *or* $L+1$*, whichever is the greater, and* m *is defined to be either* m_1 *or* $L-1$*, whichever is the lesser.)*

Then $\{a_n\}$ is bounded above by M and bounded below by m.
So our sequence is bounded.

Increasing and decreasing sequences

Definition 2.2.11. *A sequence* $\{a_n\}$ *is called* increasing *if* $a_n \leqslant a_{n+1}$ *for all* $n \geqslant 1$.
A sequence $\{a_n\}$ *is called* strictly increasing *if* $a_n < a_{n+1}$ *for all* $n \geqslant 1$.
A sequence $\{a_n\}$ *is called* decreasing *if* $a_n \geqslant a_{n+1}$ *for all* $n \geqslant 1$.
A sequence $\{a_n\}$ *is called* strictly decreasing *if* $a_n > a_{n+1}$ *for all* $n \geqslant 1$.

Definition 2.2.12. *A sequence is called* monotonic *if it is* either increasing or decreasing.
Similar terms : monotonic increasing, monotonic decreasing, monotonically increasing/decreasing.

Note: These definitions are not *entirely* standard. Some authors use the term *increasing* for what we have called *strictly increasing* and/or use the term *nondecreasing* for what we have called *increasing*.
Examples

1. An increasing sequence is bounded below but need not be bounded above. For example

$$\{n\}_{n=1}^{\infty} : 1,\ 2,\ 3,\ \ldots$$

2. A bounded sequence need not be monotonic. For example

$$\{(-1)^n\} : -1,\ 1,\ -1,\ 1,\ -1,\ \ldots$$

3. A convergent sequence need not be monotonic. For example
$$\left\{\frac{(-1)^{n+1}}{n}\right\}_{n=1}^{\infty} : 1, -\frac{1}{2}, \frac{1}{3}, -\frac{1}{4}, \dots$$
This sequence converges to 0 but is neither increasing nor decreasing.

4. A montonic sequence need not be convergent, as Example 1 above shows.

However, if a sequence is *bounded* and *monotonic*, it is *convergent*. This is the *Monotone Convergence Theorem*, which is the major theorem of this section.

Theorem 2.2.13 (The Monotone Convergence Theorem)**.** *If a sequence* $\{a_n\}_{n=1}^{\infty}$ *is montonic and bounded, then it is convergent.*

Proof: *(We can start by giving ourselves a monotonic bounded sequence - we can take it to be increasing; the argument for a decreasing sequence is similar.)*
Suppose that $\{a_n\}$ is *increasing* and bounded. Then the set $\{a_n : n \in \mathbb{N}\}$ is a bounded subset of $\mathbb{R}$ and by the *Axiom of Completeness* it has a *least upper bound* (or supremum) L.

(We are just giving the name L *here to the supremum of the set of values of the sequence. We are supposed to be showing that the sequence is convergent, i.e. has a limit :* L *is our candidate for that limit)*

We will show that the sequence $\{a_n\}$ converges to L.

Choose a (very small) $\varepsilon > 0$. Then $L - \varepsilon$ is *not an upper bound* for $\{a_n : n \in \mathbb{N}\}$, becasue L is the *least* upper bound for this set.

This means there is some $N \in \mathbb{N}$ for which $L - \varepsilon < a_N$. Since L is an upper bound for $\{a_n : n \in \mathbb{N}\}$, this means
$$L - \varepsilon < a_N \leqslant L$$
(i.e. a_N *is between* $L - \varepsilon$ *and* L*).*

Since the sequence $\{a_n\}$ is increasing and its terms are bounded above by L, *every* term after a_N is between a_N and L, and therefore between $L - \varepsilon$ and L. These terms are all within ε of L.

Using the fact that our sequence is increasing and bounded, we have

- Identified L as the least upper bound for the set of terms in our sequence
- Showed that no matter how small an ε we take, there is a point in our sequence beyond which *all* terms are within ε of L.

This is exactly what it means for the sequence to converge to L. This concludes the proof.

Example 2.2.14 (from 2011 Summer Exam)**.** *A sequence* $\{a_n\}$ *is defined by*
$$a_1 = 0, \quad a_{n+1} = \sqrt{a_n + 6} \text{ for all } n \geqslant 1.$$
Show that this sequence is bounded above by 3 and that it is increasing.
Deduce that the sequence is convergent and find its limit.

Note: This is an example of a sequence that is defined *recursively*. This means that the first term is given and subsequent terms are defined (one by one) in terms of previous ones. We are not given a general formula for the nth term although one may exist.

Solution:

1. 3 *is an upper bound.*
Suppose that $a_k < 3$ for some k. Then
$$a_{k+1} = \sqrt{a_k + 6} < \sqrt{3 + 6} = 3.$$
This says that if $a_k \leqslant 3$, then $a_{k+1} \leqslant 3$ also.
Then, since $a_1 < 3$, we have $a_2 < 3$, then $a_3 < 3$, etc.

2. *The sequence is increasing*
   Let $k \in \mathbb{N}$. We need to show that $a_k \leqslant a_{k+1}$.
   We know that $0 \leqslant a_k < 3$: note this implies that

$$a_k = \sqrt{a_k^2} < \sqrt{3a_k} = \sqrt{a_k + 2a_k} < \sqrt{a_k + 6} = a_{k+1}.$$

   Then $a_k < a_{k+1}$ for each k, which means the sequence is increasing.

3. *The sequence converges*
   Since the sequence is increasing and bounded, it converges by the Monotone Convergence Theorem.

   Let L be the limit. Then, taking limits as $n \to \infty$ on both sides of the equation

$$a_{n+1} = \sqrt{a_n + 6}$$

   we find that

$$L = \sqrt{L+6} \Longrightarrow L^2 = L + 6 \Longrightarrow (L-3)(L+2) = 0.$$

   Thus $L = 3$ or $L = -2$, and since all the terms of our sequence are between 0 and 3 it must be that $L = 3$.

2.3 Introduction to Infinite Series

Definition 2.3.1. *A* series *or* infinite *series is the sum of all the terms in a sequence.*

Example 2.3.2 (Examples of infinite series)**.**

1. $\sum_{n=1}^{\infty} n = 1 + 2 + 3 + \dots$

2. *A* geometric *series*

$$\sum_{n=1}^{\infty} \frac{1/2^n}{=} 1 + \frac{1}{2} + \frac{1}{2^2} + \dots$$

Every term in this series is obtained from the previous one by multiplying by the common ratio $\frac{1}{2}$. *This is what* geometric *means.*

3. *The* harmonic series

$$\sum_{n=1}^{\infty} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \dots$$

4. *An* alternating *series*

$$\sum_{n=0}^{\infty} (-1)^n = 1 + (-1) + 1 + (-1) + \dots$$

Notes

1. For now these infinite sums are just formal expressions or arrangements of symbols. Whether it is meaningful to think of them as numbers or not is something that can be investigated.

2. A *series* is not the same thing as a *sequence* and it is important not to confuse these terms. A *sequence* is just a list of numbers. A *series* is an infinite sum.

3. The "sigma" notation for sums : *sigma* (lower case σ, upper case Σ) is a letter from the Greek alphabet, the upper case $\sum$ is used to denote sums. The notation

$$\sum_{n=i}^{j} a_n$$

means : i and j are integers and $i \leqslant j$. For each n from i to j the number a_n is defined; the expression above means the sum of the numbers a_n where n runs through all the values from i to j, i.e.

$$\sum_{n=i}^{j} a_n = a_i + a_{i+1} + a_{i+2} + \dots + a_{j-1} + a_j.$$

For example

$$\sum_{n=2}^{5} n^2 = 2^2 + 3^2 + 4^2 + 5^2 = 54.$$

For infinite sums it is possible to have $-\infty$ and/or ∞ (instead of fixed integers i and j) as subscripts and superscripts for the summation.

What does it mean to talk about the sum of *infinitely many* numbers? We cannot add infinitely many numbers together in practice, although we can (in principle) at least, add up any finite collection of numbers. In the examples above we can start from the beginning, adding terms at the start of the series. Adding term by term we get the following lists.

1. $\sum_{n=1}^{\infty} n = 1 + 2 + 3 + \dots$

$$1, 1+2, 1+2+3, 1+2+3+4, 1+2+3+4+5, \dots : 1, 3, 6, 10, 15, \dots$$

   Since the terms being added on at each stage are getting bigger, the numbers in the list above will keep growing (faster and faster as n increases) - we can't associate a numberical value with this infinite sum.

2. A *geometric* series

$$\sum_{n=1}^{\infty} \frac{1/2^n}{=} 1 + \frac{1}{2} + \frac{1}{2^2} + \dots$$

$$1, 1+\frac{1}{2}, 1+\frac{1}{2}+\frac{1}{2^2}, 1+\frac{1}{2}+\frac{1}{2^2}+\frac{1}{2^3} \dots \ : \ 1, \ \frac{3}{2}, \ \mathtt{frac74}, \ \mathtt{frac158}, \ \frac{31}{16}, \ \frac{63}{32} \dots$$

   In this example the terms that are being added on at each step ($\frac{1}{2^n}$) are getting smaller and smaller as n increases, and the numbers in the list appear to be converging to 2.

3. The *harmonic series*

$$\sum_{n=1}^{\infty} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \dots$$

$$1, 1+\frac{1}{2}, 1+\frac{1}{2}+\frac{1}{3}, 1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4} \dots \ : \ 1, \ \frac{3}{2}, \ \frac{11}{6}, \ \frac{25}{12}, \ \frac{137}{60}, \dots$$

   It is harder to see what is going on here.

4. An *alternating* series

$$\sum_{n=0}^{\infty} (-1)^n = 1 + (-1) + 1 + (-1) + \dots$$

$$1, \ 1-1, \ 1-1+1, \ 1-1+1-1, \ 1-1+1-1+1 \dots \ : \ 1, 0, 1, 0, 1, \dots$$

   The terms being "added on" at each step are alternating between 1 and -1, and as we proceed with the summation the "running total" alternates between 0 and 1. So there is no numerical value that we can associate with the infinite sum $\sum_{n=0}^{\infty} (-1)^n$.

Note: The series in 2. above *converges* to 2, the series in 1. and 4. are both *divergent* and it is not obvious yet but the series in 3. is *divergent* as well. Our next task is to give precise meanings to these terms for series. In order to do this we need some terminology. Bear in mind that we know what it means for a sequence to converge, but we don't yet have a definition of convergence for series.

Definition 2.3.3. *For a series* $\sum_{n=1}^{\infty} a_n$, *and for* $k \geqslant 1$, *let*

$$s_k = \sum_{n=1}^{k} a_n = a_1 + a_2 + a_3 + \dots + a_k.$$

Thus $s_1 = a_1, \ s_2 = a_1 + a_2, \ s_3 = a_1 + a_2 + a_3$ *etc.*
Then s_k *is called the* kth partial sum *of the series, and the sequence* $\{s_k\}_{k=1}^{\infty}$ *is called the* sequence of partial sums *of the series.*
If the sequence of partial sums converges to a limit s, *the series is said to* converge *and* s *is called its sum. In this situation we can write*

$$\sum_{n=1}^{\infty} a_n = s.$$

If the sequence of partial sums diverges, *the series is said to* diverge.

Example 2.3.4 (Convergence of a geometric series)**.** *Recall the second example above :*

$$\sum_{n=0}^{\infty} \frac{1}{2^n} = 1 + \frac{1}{2} + \frac{1}{2^2} + \dots$$

In this example, for $k \geqslant 0$,

$$\begin{aligned} s_k = \sum_{n=0}^{k} \frac{1}{2^n} &= 1 \quad + \quad \frac{1}{2} + \frac{1}{4} + \dots \frac{1}{2^k} \\ \frac{1}{2}s_k = \sum_{n=1}^{k} \frac{1}{2^{n+1}} &= \qquad\quad \frac{1}{2} + \frac{1}{4} + \dots \frac{1}{2^k} + \frac{1}{2^{k+1}} \end{aligned}$$

Then

$$s_k - \frac{1}{2}s_k = \frac{1}{2}s_k = 1 - \frac{1}{2^{k+1}} \Longrightarrow s_k = 2 - \frac{1}{2^k}.$$

So the sequence of partial sums has kth term $2 - \frac{1}{2^k}$. This sequence converges to 2 so the series converges to 2; we can write

$$\sum_n = 0^\infty \frac{1}{2^k} = 2.$$

General geometric series : Consider the sequence of partial sums for the geometric series

$$\sum_{n=0}^{\infty} ar^n = a + ar + ar^2 + \dots$$

(This is a *geometric series* with initial term a and *common ratio* r.) The kth partial sum s_k is given by

$$\begin{aligned} s_k = \sum_{n=0}^{k} ar^n &= a \quad + \quad ar + \dots ar^k \\ rs_k = \sum_{n=0}^{k} ar^{n+1} &= \qquad\quad ar + ar^2 + \dots ar^{k+1} \end{aligned}$$

Then $(1-r)s_k = a - ar^{k+1} \Longrightarrow s_k = \dfrac{a(1-r^{k+1})}{1-r}$. If $|r| < 1$, then $r^{k+1} \to 0$ as $k \to \infty$, and the sequence of partial sums (hence the series) converges to $\dfrac{a}{1-r}$. If $|r| \geqslant 1$ the series is divergent.

Next we show that the harmonic series is divergent.

Theorem 2.3.5. *The harmonic series* $\sum_{n=1}^{\infty} \frac{1}{n}$ *is divergent.*

Proof: We show that the sequence of partial sums of the harmonic series is not bounded above.

- The first term is 1.
- The second term is $\frac{1}{2}$.
- The sum of the 3rd and 4th terms exceeds $\frac{1}{2}$:

$$\frac{1}{3} + \frac{1}{4} > \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

- The sum of the 5th, 6th, 7th and 8th terms exceeds $\frac{1}{2}$:

$$\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}>\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}=\frac{1}{2}.$$

- For the same reason, the sum of the next 8 terms (terms 9 through 16) also exceeds $\frac{1}{2}$.

- In general the sum of the 2^{n-1} terms $\dfrac{1}{2^{n-1}+1}$ through $\dfrac{1}{2^n}$ exceeds $\frac{1}{2}$.

So, as we list terms in the sequence of partial sums of the harmonic series, we keep accumulating non-overlapping stretches of terms that add up to more than $\frac{1}{2}$. Thus the entire series has infinitely many non-overlapping stretches all individually summing to more than $\frac{1}{2}$. Then the sum of this series is not finite and the series diverges.

Note: A necessary condition for the series $\sum_{n=1}^{\infty} a_n$ to *converge* is that the *sequence* $\{a_n\}_{n=1}^{\infty}$ converges to 0; i.e. that $a_n \to 0$ as $n \to \infty$. If this does *not* happen, then the sequence of partial sums has no possibility of converging.

The example of the harmonic series shows that the condition $a_n \to 0$ as $n \to \infty$ is not *sufficient* to guarantee that the series $\sum_{n=1}^{\infty}$ will converge.

2.4 Introduction to power series

Definition 2.4.1. *A* polynomial *in the* variable x *is an expression of the form*

$$\sum_{i=0}^{n} a_i x^i = a_0 + a_1 x + a_2 x^2 + \cdots + a_{n-1} x^{n-1} + a_n x^n$$
$$\text{or} \quad a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0,$$

where the coefficients a_i *are real numbers and* n *is a natural number. The* degree *of the polynomial is the highest* k *for which* x^k *appears with non-zero coefficient.*

Examples

1. $x^3 + 5x^2$ is a polynomial of degree 3 (also called cubic).
2. $2 + 2x + 2x^2 + 2x^7$ is a polynomial of degree 7.
3. $x^4 + x^3 + \frac{3}{x^2}$ is *not* a polynomial, because it involves a negative power of x.

The point is that a polynomial can has a constant term (which may be zero) and a *finite number* of terms involving particular positive powers of x that have numbers as coefficients. A polynomial may be regarded as a function of x, and polynomials are functions of a special type.

Definition 2.4.2. *A* power series *in the variable* x *resembles a polynomial, except that it may contain infinitely many positive powers of* x*. It is an expression of the type*

$$\sum_{i=0}^{\infty} a_i x^i = a_0 + a_1 x + a_2 x^2 + \ldots,$$

where each a_i *is a number.*

Example 2.4.3.

$$\sum_{n=0}^{\infty} x^n = 1 + x + x^2 + x^3 + \ldots$$

is a power series.

Question: Does it make sense to think of a power series as a *function* of x? We investigate this question for the example above.

So define a "function" by

$$f(x) = \sum_{n=0}^{\infty} x^n = 1 + x + x^2 + \ldots$$

- If we try to evaluate this function at $x = 2$, we get a *series* of real numbers.

  $$f(2) = \sum_{n=0}^{\infty} 2^n = 1 + 2 + 2^2 + \ldots$$

  This series is divergent, so our power series does not define a function that can be evaluated at 2.

- If we try evaluating at 0 (and allow that the first term x^0 of the power series is interpreted as 1 for *all* values of x), we get

  $$f(0) = 1 + 0 + 0^2 + \cdots = 1.$$

  So it does make sense to "evaluate" this function at $x = 0$.

- If we try evaluating at $x = \frac{1}{2}$, we get

$$f(\frac{1}{2}) = \sum_{n=0}^{\infty} \left(\frac{1}{2}\right)^n = 1 + \frac{1}{2} + \left(\frac{1}{2}\right)^2 + \dots$$

  From our work on geometric series in Section 2.3 we know that this is a geometric series with first term $a = 1$ and common ratio $r = \frac{1}{2}$. We know that if $|r| < 1$, such a series converges to the number $\dfrac{a}{1-r}$. In this case

$$\frac{a}{1-r} = \frac{1}{1-\frac{1}{2}} = 2,$$

  and we have $f(\frac{1}{2}) = 2$.

In general we know that a geometric series of this sort converges provided that the absolute value of its common ratio is less than 1. So for example if we put $x = \frac{1}{3}$ we find that $f(\frac{1}{3})$ is the sum of a geomtric series with first term 1 and common ratio $\frac{1}{3}$; this is

$$\frac{1}{1-\frac{1}{3}} = \frac{3}{2}.$$

In general for any value of x whose absolute value is less than 1 (i.e. any x in the interval $(-1, 1)$), we find that $f(x)$ is a convergent geometric series, converging to $\dfrac{1}{1-x}$.

Conclusion: For values of x in the interval $(-1, 1)$ (i.e. $|x| < 1$), the function $f(x) = \frac{1}{1-x}$ coincides with the power series $\sum_{n=0}^{\infty} x^n$.

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n, \text{ for } |x| < 1.$$

The interval $(-1, 1)$ is called the *interval of convergence* of the power series, and 1 is the *radius of convergence*. We say that the *power series representation* of the function $f(x) = \frac{1}{1-x}$ is $\sum_{n=0}^{\infty} x^n$, for values of x in the interval $(-1, 1)$.
Note: The expression $\frac{1}{1-x}$ makes sense of course for all values of x except $x = 1$. We are not saying that the domain of the function $f(x) = \frac{1}{1-x}$ only consists of the interval $(-1, 1)$, but just that it is only on this interval that our power series represents this function.

Remark: The fact that for certain values of x we can represent $\frac{1}{1-x}$ with a power series might be interesting (at least to some people!), but it is not of particular use if you want to calculate $\frac{1}{1-x}$ for some particular value of x, because this is easily done directly. However, if we could obtain a power series representation for a function like $\sin x$ and use it to evaluate (or approximate) $\sin(1)$ or $\sin(9)$ or $\sin(20)$, that might be of real practical use. These numbers are *not* easy to obtain directly because the definition of $\sin x$ doesn't tell us how to calculate $\sin x$ for a particular x - you can use a calculator of course but how does the calculator do it? If we had a power series representation for $\sin x$ and we knew it converged for the value of x we had in mind, we couldn't necessarily write down the limit but we could calculate partial sums to get an estimation as accurate as we like.

Questions: What functions can be represented by power series, and on what sorts of interval or subsets of $\mathbb{R}$? If a function could be represented by a power series, how would we calculate the coefficients in this series?

We are not going to give a full answer to these questions, but a partial one involving *Maclaurin* or *Taylor* series.

Suppose that $f(x)$ is an infinitely differentiable function (this means that all the deriviatives of f are themselves differentiable), and suppose that f is represented by the power series

$$f(x) = \sum_{n=0}^{\infty} c_n x^n.$$

We can work out appropriate values for the coefficients c_n as follows.

- Put $x = 0$. Then $f(0) = c_0 + \sum_{n=1}^{\infty} c_n (0)^n \implies f(0) = c_0$.
  The constant term in the power series is the value of f at 0.
- To calculate c_1, look at the value of the first derivative of f at 0, and differentiate the power series term by term. We expect

  $$f'(x) = c_1 + 2c_2 x + 3c_3 x^2 + \cdots = \sum_{n=1}^{\infty} n c_n x^{n-1}.$$

  Then we should have

  $$f'(0) = c_1 + 2c_2 \times 0 + 3c_3 \times 0 + \cdots = c_1.$$

  Thus $c_1 = f'(0)$.
- For c_2, look at the second derivative of f. We expect

  $$f''(x) = 2(1)c_2 + 3(2)c_3 x + 4(3)c_4 x^2 + 5(4)c_5 x^3 + \cdots = \sum_{n=2}^{\infty} n(n-1)c_n x^{n-2}.$$

  Putting $x = 0$ gives $f''(0) = 2(1)c_2$ or $c_2 = \dfrac{f''(0)}{2(1)}$.
- For c_3, look at the third derivative $f^{(3)}(x)$. We have

  $$f^{(3)}(x) = 3(2)(1)c_3 + 4(3)(2)c_4 x + 5(4)(3)c_5 x^2 + \cdots = \sum_{n=3}^{\infty} n(n-1)(n-2)c_n x^{n-3}.$$

  Setting $x = 0$ gives $f^{(3)}(0) = 3(2)(1)c_3$ or $c_3 = \frac{f^{(3)}(0)}{3(2)(1)}$

Continuing this process, we obtain the following general formula for c_n:

$$c_n = \frac{1}{n!} f^{(n)}(0).$$

Definition 2.4.4. *For a positive integer* n*, the number* n factorial*, denoted* $n!$ *is defined by*

$$n! = n \times (n-1) \times (n-2) \times \ldots 3 \times 2 \times 1.$$

The number $0!$ *(zero factorial) is defined to be 1.*

Example 2.4.5 (Power series representation of e^x)**.**

The coefficient of x^n in the Maclaurin series expansion of e^x is

$$c_n = \frac{1}{n!} \frac{d^n}{dx^n} (e^x)|_{x=0} = \frac{1}{n!} e^0 = \frac{1}{n!}.$$

Thus the Maclaurin series for e^x is given by

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

Note that if we differentiate this series term by term we get exactly the same series back, which is what we would expect for a power series that represents e^x, since $\dfrac{d}{dx}(e^x) = e^x$.

Theorem 2.4.6. *For* every *real number* x*, the above power series converges to* e^x*. The interval of convergence of this power series is all of* $\mathbb{R}$ *and the radius of convergence is infinite.*

Example 2.4.7 (Power series representation of $\sin x$).

Write $f(x) = \sin x$, and write $\sum_{n=0}^{\infty} c_n x^n$ for the Maclaurin series of $\sin x$. Then

- $f(0) = \sin 0 = 0 \Longrightarrow c_0 = 0$
- $f'(0) = \cos 0 = 1 \Longrightarrow c_1 = 1$
- $f''(0) = -\sin 0 = 0 \Longrightarrow c_2 = \frac{0}{2!} = 0$
- $f^{(3)}(0) = -\cos 0 = -1 \Longrightarrow c_3 = \frac{-1}{3!} = -\frac{1}{6}$
- $f^{(4)}(0) = \sin 0 = 0 \Longrightarrow c_4 = \frac{0}{4!} = 0$

This pattern continues : if k is even then $f^{(k)}(0) = \pm sin 0 = 0$, so $c_k = 0$.
If k is odd and $k \equiv 1 \mod 4$ then $f^{(k)}(0) = \cos 0 = 1$ and $c_k = \frac{1}{k!}$.
If k is odd and $k \equiv 3 \mod 4$ then $f^{(k)}(0) = -\cos 0 = -1$ and $c_k = -\frac{1}{k!}$.
Thus the Maclaurin series for $\sin x$ is given by

$$\sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} x^{2k+1} = x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \frac{1}{7!}x^7 + \dots$$

Note that this series only involves odd powers of x - this is not surprising because sin is an *odd function*; it satisfies $\sin(-x) = -\sin x$.

Theorem 2.4.8. *For every real number* x*, the above series converges to* $\sin x$*.*

Thus computing partial sums of this series gives us an effective way of approximating $\sin x$ for any real number x.

Exercise 2.4.9. *Show that the Maclaurin series for* $\cos x$ *is given by*

$$\sum_{k=0}^{\infty} \frac{(-1)^k}{(2k!)} x^{2k}.$$

(Note that this can be obtained by differentiating term-by-term the series for $\sin x$, as we would expect since $\frac{d}{dx}(\sin x) = \cos x$.)

Chapter 3

Integral Calculus

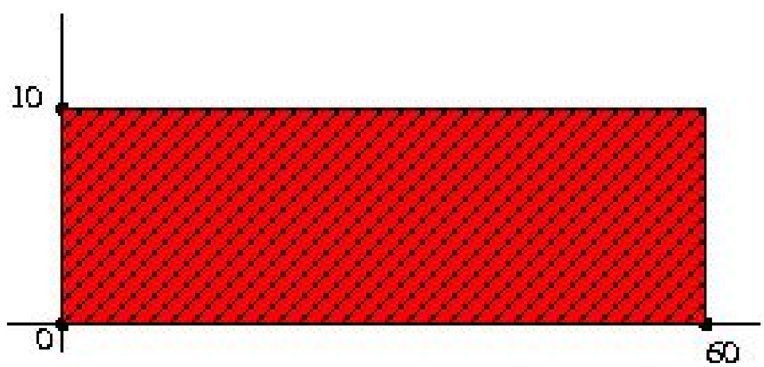
3.1 Areas under curves - introduction and examples

Problem 3.1.1. *A car travels in a straight line for one minute, at a constant speed of* $10\mathrm{m/s}$*. How far has the car travelled in this minute?*

SOLUTION: The car is travelling for 60 seconds, and covering 10 metres in each second, so in total it covers $60 \times 10 = 600$ metres.
That wasn't very hard.

An easy example like this one can be a starting point for studying more complicated problems. What makes this example easy is that the car's speed is not changing so all we have to do is multiply the distance covered in one second by the number of seconds. Note that we can interpret the answer graphically as follows.

Suppose we draw a graph of the car's speed against time, where the x-axis is labelled in seconds and the y-axis in $\mathrm{m/s}$. The graph is just the horizontal line $y = 10$ of course.



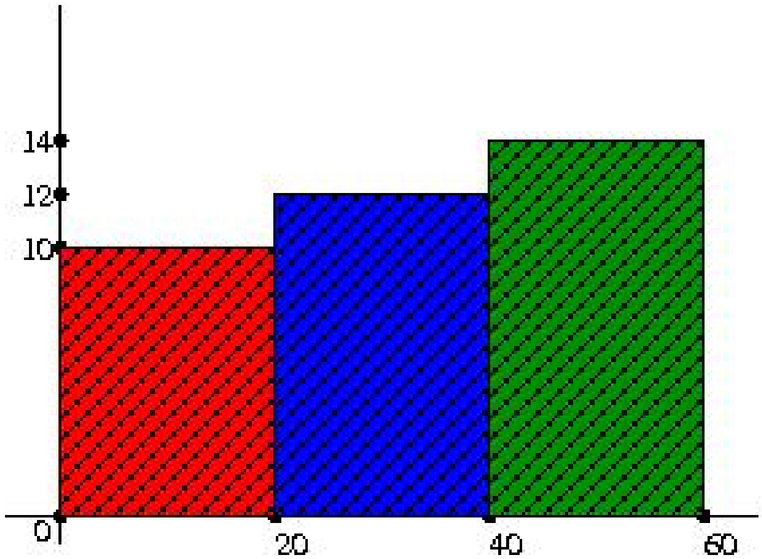
We can label the time when we start observing the car's motion as $t = 0$ and the time when we stop as $t = 60$. Note then that the total distance travelled – 600m – is the area enclosed under the graph, between the x-axis, the horizontal line $y = 10$, and the vertical lines $x = 0$ (or time $t = 0$) and $x = 60$ marking the beginning and end of the period of observation. This is no coincidence; if we divide this rectangular region into vertical strips of width 1, one for each second, what we get are 60 vertical strips of width 1 and height 10, each accounting for 10 units of area, and each accounting for 10 metres of travel.
The next problem is a slightly harder example of the same type.

Problem 3.1.2. *Again the car travels in one direction for one minute. This time it travels at* $10\mathrm{m/s}$ *for the first 20 seconds, at* $12\mathrm{m/s}$ *for the next 20 seconds, and at* $14\mathrm{m/s}$ *for the last 20 seconds.*
What is the total distance travelled?

SOLUTION: This is not much harder really (although it may be a physically unrealistic problem - why?). This time, the car covers

- $20 \times 10 = 200$ metres in the first 20 seconds
- $20 \times 12 = 240$ metres in the next 20 seconds and
- $20 \times 14 = 280$ metres in the last 20 seconds,

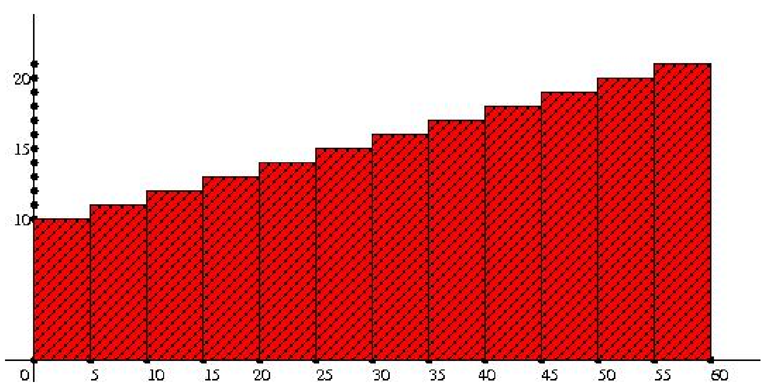
so the total distance is 720 metres.



Once again the total distance travelled is the area of the region enclosed between the lines $x = 0$, $x = 60$, the y-axis and the graph showing speed against time. The region whose area represents the distance travelled is the union of three rectangles, all of width 20, and of heights 10, 12 and 14.

Problem 3.1.3. *Same set up, but this time the car's speed is* $10\mathrm{m/s}$ *for the first 5 seconds,* $11\mathrm{m/s}$ *for the next 5, and so on, increasing by* $1\mathrm{m/s}$ *every five seconds so that the speed is 21* $\mathrm{m/s}$ *for the last five seconds. Again the problem is to calculate the total distance travelled in metres.*

The answer is left as an exercise, but this time the distance is the area indicated below.

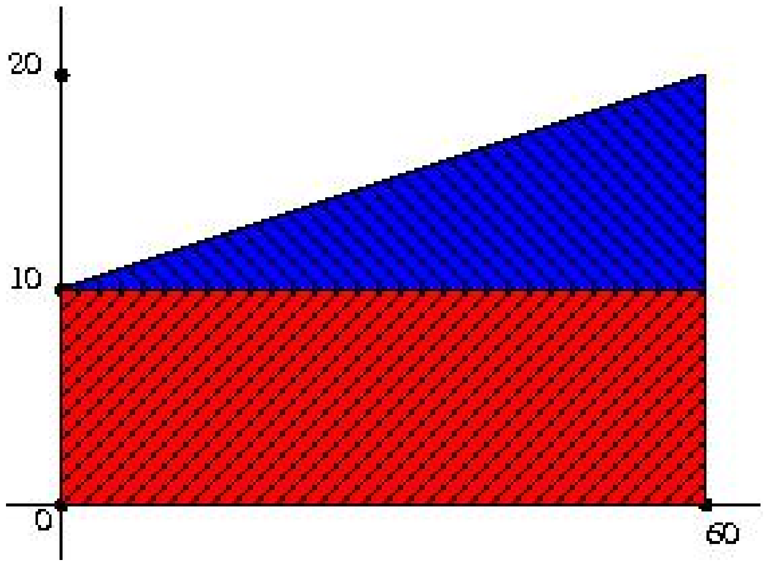


Problem 3.1.4. *Again our car is travelling in one direction for one minute, but this time its speed increases at a constant rate from* $10\mathrm{m/s}$ *at the start of the minute, to* $20\mathrm{m/s}$ *at the end. What is the distance travelled?*

NOTE: This is a more realistic problem, in which the speed is increasing at a constant rate. This constant acceleration would apply for example in the case of an object falling freely under gravity.

SOLUTION: This is a different problem from the others. Because the speed is varying *all the time* this problem cannot be solved by just multiplying the speed by the time or by a combination of such steps as in Problems 1.1.1 and 1.1.2.

The following picture shows the graph of the speed against time.



If the total distance travelled is represented in this example, as in the others, by the area under the speed graph between $t = 0$ and $t = 60$, then we can observe that it's the area of a region consisting of a rectangle of width 60 and height 10, and a triangle of width 60 and perpendicular height 10. Thus the total distance travelled is given by

$$(60 \times 10) + \frac{1}{2}(60 \times 10) = 900\mathrm{m}.$$

QUESTION: Should we believe this answer? Just because the distance is given by the area under the graph when the speed is constant, how do we know the same applies in cases where the speed is varying continuously? Here is an argument that might justify this claim.

In Problem 1.1.4, the speed increases steadily from $10\mathrm{m/s}$ to $20\mathrm{m/S}$ over the 60 seconds. We want to calculate the distance travelled.

We can approximate this distance as follows.

- Suppose we divide the one minute into 30 two-second intervals.
- At the start of the first two-second interval, the car is travelling at $10\mathrm{m/s}$. We make the *simplifying assumption* that the car travels at $10\mathrm{m/s}$ throughout the first two seconds, thereby covering $20\mathrm{m}$ in the first two seconds. Note that this actually underestimates the true distance travelled in the first second, because in fact the speed is increasing from $20\mathrm{m/s}$ during these two seconds.
- At the start of the second two-second interval, the car has completed one-thirtieth of its acceleration from $10\mathrm{m/s}$ to $20\mathrm{m/s}$, so its speed is

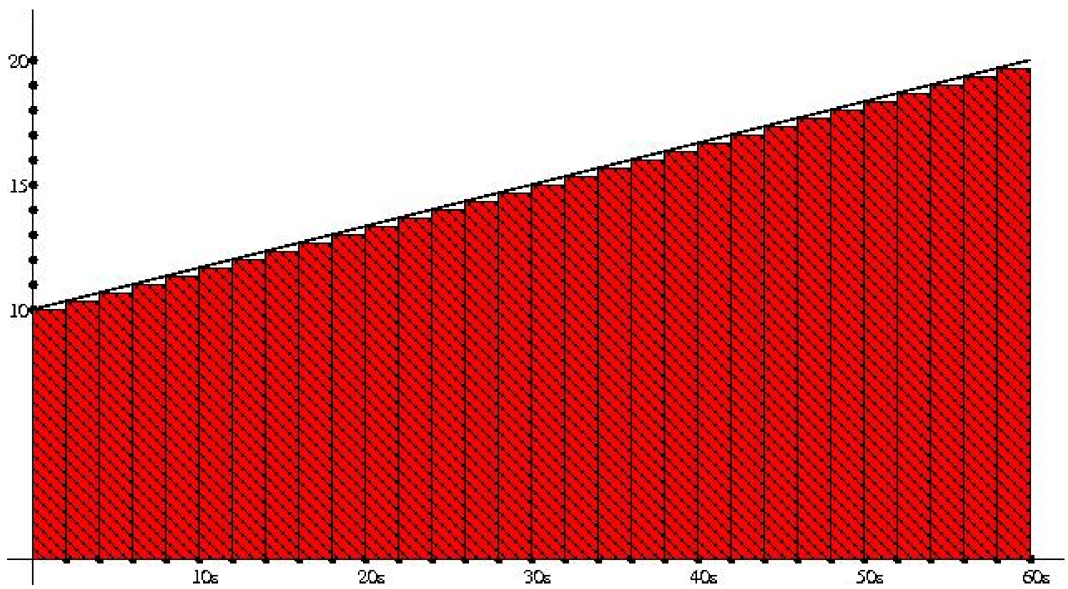
$$10 + \frac{10}{30} = 10\frac{1}{3}\mathrm{m/s}.$$

If we make the *simplifying assumption* that the speed remains constant at $10\frac{1}{3}\mathrm{m/s}$ throughout the second two-second interval, we estimate that the car travels $20\frac{2}{3}\mathrm{m}$ during the second two-second interval. This underestimates the true distance beacuse the car is actually accelerating *from* $10\frac{1}{3}\mathrm{m/s}$ during these two seconds.

- If we proceed in this manner we would estimate that the car travels
  - $20\mathrm{m}$ in the first two seconds;
  - $20\frac{2}{3}\mathrm{m}$ in the next two seconds;
  - $21\frac{1}{3}\mathrm{m}$ in the next two seconds, and so on;
  - $\ldots 39\frac{1}{3}\mathrm{m}$ in the 30th two-second interval.

This would give us a total of $890\mathrm{m}$ as the estimate for distance travelled, but that's not really the point of this discussion.

The distance that we estimate using the assumption that the speed remains constant for each of the 30 two-second intervals, is indicated by the area in red in the diagram below, where the black line is the true speed graph. Note that the red area includes all the area under the speed graph, except for 30 small triangles of base length 2 and height $\frac{1}{3}$.



Suppose now that we refine the estimate by dividing our minute of time into 60 one-second intervals and assuming the the speed remains constant for each of these, instead of into 30 two-second intervals.

If do this we will estimate that the car travels

- $10\mathrm{m}$ in the first seconds;
- $10\frac{1}{6}\mathrm{m}$ in the next seconds;
- $10\frac{2}{6}\mathrm{m}$ in the next second, and so on;
- $\ldots 19\frac{5}{6}\mathrm{m}$ in the 60th one-second interval.

This would give us a total of $895\mathrm{m}$ as the estimate for distance travelled. What is the corresponding picture? Draw it, or at least part of it, as an exercise.

Note that this *still underestimates* the distance travelled in each second, because it assumes that the speed remains constant at its starting point for the duration of each second, whereas in reality it increases. But this estimate is closer to the true answer than the last one, because this estimate takes into account speed increases every second, instead of every two seconds.

The corresponding "area" picture has sixty rectangles of width 1 instead of thirty of width 2, and it includes all the area under the speed graph, except for *sixty triangles of base length 1 and height* $\frac{1}{6}$.

If we used the same strategy but dividing our minute into 120 half-second intervals, we would expect to get a better estimate again. As the number of intervals increases and their width decreases, the red rectangles in the picture come closer and closer to filling *all* the area under the speed graph. The true distance travelled is the limit of these improving estimates, as the length of the subintervals approaches zero. This is *exactly* the area under the speed graph, between $x = 0$ and $x = 60$.

So we can now assert more confidently that the answer to Problem 1.1.4 is $900\mathrm{m}$.

Problem 3.1.5. *Again our car is travelling in one direction for one minute, but this time its speed* v *increases from* $10\mathrm{m/s}$ *to* $20\mathrm{m/s}$ *over the minute, according to the formula*

$$v(t) = 20 - \frac{1}{360}(60 - t)^2,$$

where t *is measured in seconds, and* $t = 0$ *at the start of the minute.*
What is the distance travelled?

NOTE: The formula means that after t seconds have passed, the speed of the car in $\mathrm{m/s}$ is $20 - \frac{1}{360}(60 - t)^2$. So for example after 30 seconds the car is travelling at a speed of

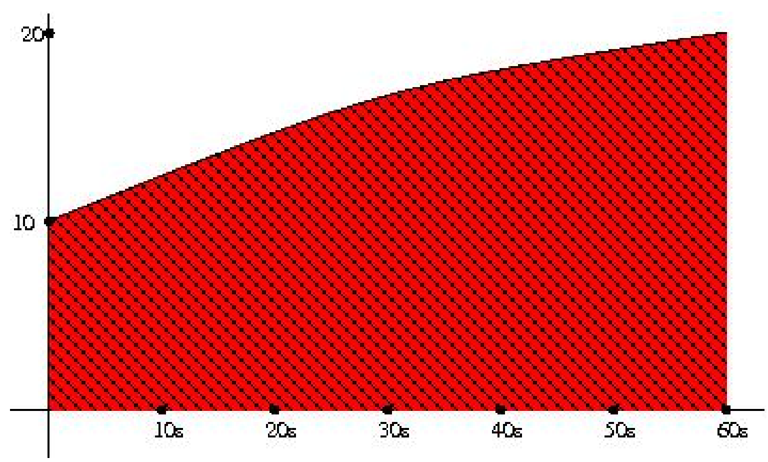
$$20 - \frac{1}{360}(60 - 30)^2 = 20 - \frac{1}{360}900 = 17.5\mathrm{m/s}.$$

Note that $60 - t$ is decreasing as t increases from 0 to 60, so $(60 - t)^2$ is decreasing also. Thus the expression

$$20 - \frac{1}{360}(60 - t)^2$$

is *increasing* as t goes from 0 to 60. So the car is accelerating throughout the minute.

Below is the graph of the speed (in $\mathrm{m/s}$) against time (in s), with the area below it (between $t = 0$ and $t = 60$) coloured red.



The argument above works in exactly the same way for this example, to persuade us that the distance travelled should be given by the area under the speed graph, between $t = 0$ and $t = 60$. This is the area that is coloured red in the picture above.

PROBLEM! The upper boundary of this area is a part of a parabola not a line segment. The region is not a combination of rectangles and triangles as in Problem 1.1.4. We can't calculate its area using elementary techniques.

So : what we need is a theory or a method that will allow us to calculate the area bounded by a section of the graph of a function and the x-axis, over a specified interval.

IMPORTANT NOTE: The problem of calculating the distance travelled by an object from knowledge of how its speed is changing is just one example of a scientific problem that can be solved by calculating the area of a region enclosed between a graph and the x-axis. Here are just a few more examples.

1. The fuel consumption of an aircraft is a function of its speed. The total amount of fuel consumed on a journey can be calculated as the area under the graph showing speed against time.

2. The energy stored by a solar panel is a function of the light intensity, which is itself a function of time. The total energy stored in one day can be modelled as the area under a graph of the light intensity against time for that day.

3. The volume of (for example) a square pyramid can be interpreted as the area of a graph of its horizontal cross-section area against height above the base.

4. In medicine, if a drug is administered intravenously, the quantity of the drug that is in the person's bloodstream can be calculated as the area under the graph of a function that depends both on the rate at which the drug is administered and on the rate at which it breaks down.

5. The quantity of a pollutant in a lake can be estimated by calculating areas under graphs of functions describing the rate at which the pollutant is being introduced and the the rate it which is is dispersing or being eliminated.

6. The concept of area under a graph is widely used in probability and statistics, where for example the probability that a randomly chosen person is aged between 20 and 30 years is the area under the graph of the appropriate *probability density function*, over the relevant interval.

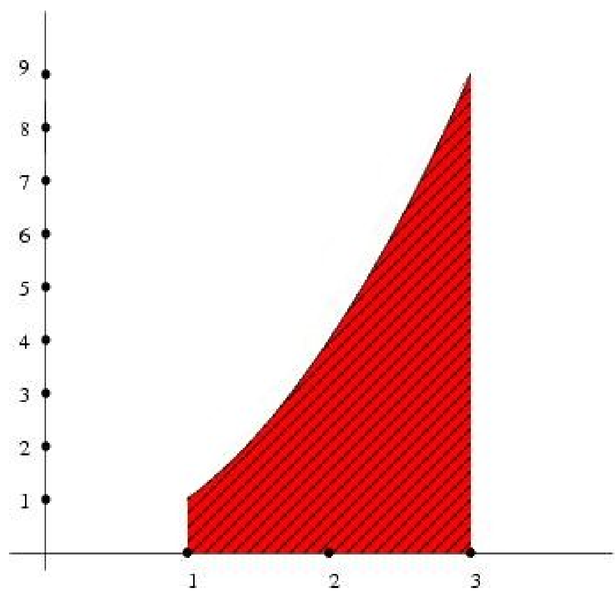
3.2 The Definite Integral

In the last section we concluded that a theory for discussing (and hopefully calculating) areas enclosed between the graphs of known functions and the x-axis, within specified intervals, would be useful. Such a theory does exist and it forms a large part of what is called integral calculus. In order to develop and use this theory we need a technical language and notation for talking about areas under curves. The goal of this section is to understand this notation and be able to use it - it is a bit cumbersome and not the most intuitively appealing, but with a bit of practice it is quite manageable.

Example 3.2.1. *Suppose that f is the function defined by $f(x) = x^2$. Note that $f(x)$ is positive when $1 \leqslant x \leqslant 3$. This means that in the region between the vertical lines $x = 1$ and $x = 5$, the graph $y = f(x)$ lies completely above the $x - \mathrm{axis}$. The area that is enclosed between the graph $y = f(x)$, the x-axis, and the vertical lines $x = 1$ and $x = 3$ is called the* definite integral of x^2 from $x = 1$ to $x = 3$, *and denoted by*

$$\boxed{\int_1^3 x^2\,dx.}$$

This diagram shows the region whose area is the definite integral $\int_1^3 x^2\,dx$.



NOTE: At the moment we are not trying to actually *calculate* this red area, we are just thinking about how the integral notation is used and what it means.

Example 3.2.2. *Suppose the function f is defined by*

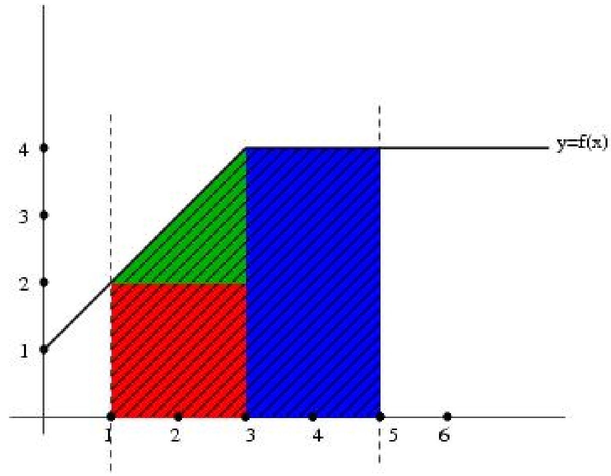
$$f(x) = \begin{cases} x+1 & \text{if } 0 \leqslant x \leqslant 3 \\ 4 & \text{if } x \geqslant 3 \end{cases}$$

Then the graph of f consists of the section of the line $y = x + 1$ between $x = 0$ and $x = 3$ (this is the line segment joining the points $(0, 1)$ and $(3, 4)$, and the constant line $y = 4$ from $x = 3$ onwards.

Now $\displaystyle\int_1^5 f(x)\,dx$ represents the area enclosed by the graph $y = f(x)$, the x-axis, and the vertical lines $x = 1$ and $x = 5$. From the diagram below we can see that this area consists of

- A (green) triangle of base length 2 and height 2, area 2;

- A (red) rectangle of base length 2 and height 2, area 4;
- A (blue) rectangle of base length 2 and height 4, area 8.



Adding these three areas, we can conclude that

$$\int_1^5 f(x)\,dx = 2 + 4 + 8 = 14.$$

In this example we are able to calculate the actual value of the definite integral because the region whose area is involved is just an arrangement of rectangles and triangles. Note from this example that in general for a function f and numbers a and b, the definite integral $\int_a^b f(x)\,dx$ is *a number*.

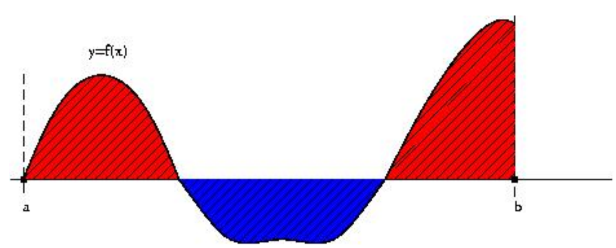
We now move on to the general definition of a definite integral.

Definition 3.2.3. *Let a and b be fixed real numbers, with $a < b$ (so a is to the left of b on the number line). Let f be a function for which it makes sense to talk about the* area *enclosed between the graph of f and the x-axis, over the interval from a to b. Then the* definite integral from a to b of f, *denoted*

$$\int_a^b f(x)\,dx$$

is defined to be the number obtained by subtracting the area enclosed below the x-axis by the graph $y = f(x)$ and the vertical lines $x = a$ and $x = b$ from the area enclosed above the x-axis by the graph $y = f(x)$ and the vertical lines $x = a$ and $x = b$.

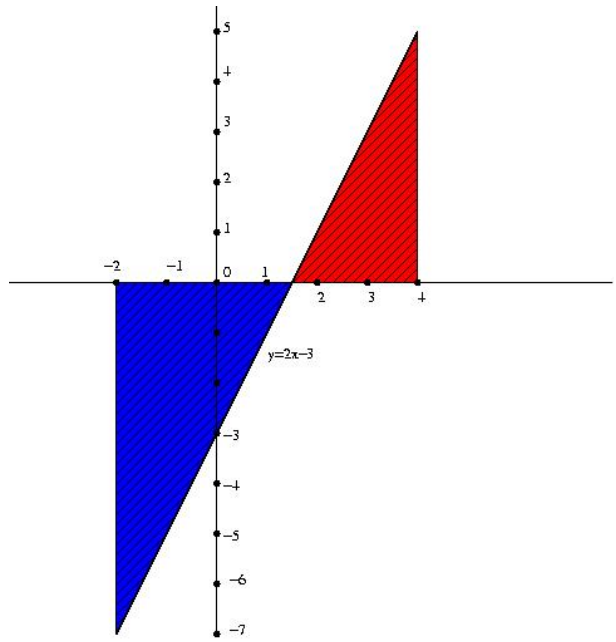
Example 3.2.4. *If the graph $y = f(x)$ is as shown in the diagram below, then $\displaystyle\int_a^b f(x)\,dx$ is the number obtained by subtracting the total area that is coloured blue from the total area that is coloured red.*



Example 3.2.5. *(a) Calculate $\displaystyle\int_{-2}^4 2x - 3\,dx$.*
(b) Calculate the total area enclosed between the x-axis and the line $y = 2x-3$, between $x = -2$ and $x = 4$.

SOLUTION: (a) We need to describe the areas enclosed by the curve $y = 2x - 3$, above and below the x-axis, between $x = -2$ and $x = 4$.
The curve $y = 2x - 3$ is a line; it passes through the points $(-2, -7)$ and $(4, 5)$ and it intercepts the x-axis at $x = \frac{3}{2}$.

The diagram below describes the problem :



The area of the red triangle is

$$\frac{1}{2} \times \frac{5}{2} \times 5 = \frac{25}{4},$$

and the area of the blue triangle is

$$\frac{1}{2} \times \frac{7}{2} \times 7 = \frac{49}{4}.$$

Thus $\int_{-2}^{4} 2x - 3\,dx = \frac{25}{4} - \frac{49}{4} = -\frac{24}{4} = -6.$

(b) The *total area* enclosed between the x-axis and the line $y = 2x - 5$, between $x = -2$ and $x = 4$, is the sum of the areas of the red and blue triangles, which is $\frac{25}{4} + \frac{49}{4} = \frac{37}{2}$.

Note the difference between the two parts of this question, and be careful about this distinction.

NOTES

1. In Definition 3.2.3, What is meant by the phrase "for which it makes sense to talk about the *area* enclosed between the graph of f and the x-axis" is (more or less) that the graph $y = f(x)$ is not just a scattering of points, but consists of a curve or perhaps more than one curve. There is a formal theory about "integrable functions" that makes this notion precise.

2. *Note on Notation*
   The notation surrounding definite integrals is a bit unusual. This note explains the various components involved in the expression

$$\int_{a}^{b} f(x)\,dx.$$

- "$\int$" is the *integral sign*.
- The "dx" indicates that f is a function of the variable x, and that we are talking about area between the graph of $f(x)$ against x and the x-axis.
- The "$f(x)$" in $\int_a^b f(x)\,dx$ is called the *integrand*. It is the function whose graph is the upper (or lower) boundary of the region whose area is being described.
- The numbers a and b are respectively called the *lower* and *upper* (or left and right) *limits of integration*. They determine the left and right boundaries of the region whose area is being described.

  In the expression $\int_a^b f(x)\,dx$, the limits of integration a and b are taken to be values of the variable x - this is included in what is to be interpreted from "dx". If there is any danger of ambiguity about this, you can write

  $$\int_{x=a}^{x=b} f(x)\,dx \text{ instead of } \int_a^b f(x)\,dx.$$

  Please do not confuse this use of the word "limit" with its other uses in calculus.

- Important note about this notation : neither the symbol "$\int$" nor the symbol "dx" in this setup is meaningful by itself : they must always accompany each other. You could think of them as being a bit like left and right parentheses or left and right quotation marks - a phrase that is opened with a left parenthesis "(" must be closed by a right parenthesis ")" - neither of these parentheses is meaningful by itself. In the language of definite integrals, an expression that is opened with the integral sign "$\int$" must be closed with the symbol "dx" (or "dt" or "du" as appropriate) indicating the variable involved. The symbol "dx" doesn't really have a meaning by itself - it is a companion to the integral sign.

Some Historical Remarks

The notation that is currently in use for the definite integral was introduced by Gottfried Leibniz around 1675. The rationale for it is as follows :
Areas were estimated as we did in Section 1.1. The interval from a to b would be divided into narrow subintervals, each of width Δx. The name x_i would be given to the left endpoint of the ith subinterval, and the height of the graph above the point x_i would be given by $f(x_i)$. So the area under the graph on this ith subinterval would be approximated by that of a rectangle of width Δx and height $f(x_i)$. The total area would be approximated by the sum of the areas of all of these narrow rectangles, which was written as

$$\sum f(x_i)\Delta x.$$

The accuracy of this estimate improves as the width of the subintervals gets smaller and the number of them gets larger; the true area is the limit of this process as $\Delta x \to 0$. The notation "dx" was introduced as an expression to replace Δx in this limit, and the integral sign $\int$ is a "limit version" of the summation sign $\sum$. The integral symbol itself is based on the "long s" character which was in use in English typography until about 1800.

The idea of calculating areas of regions by taking finer and finer subdivisions in this manner dates back to the ancient Greeks; early examples of what is now called "integration" can be found in the work of Archimedes (circa 225 BC). The idea of computing areas under graphs by taking narrower and narrower vertical columns was put on a firm theoretical basis by Bernhard Riemann in the 1850s.

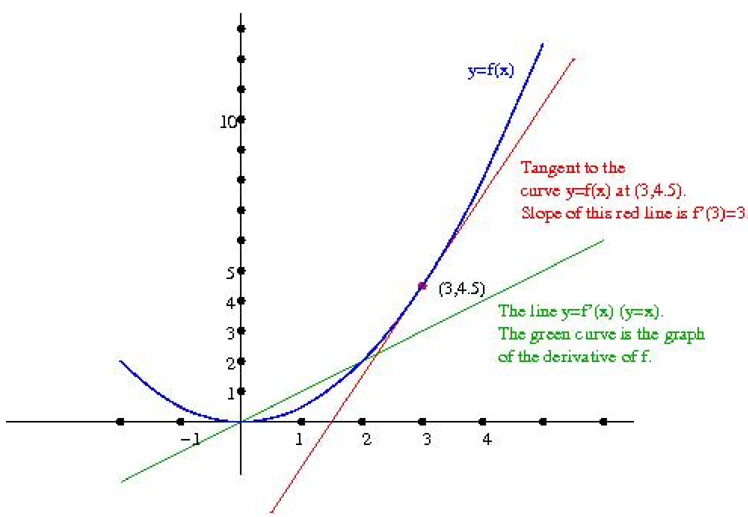
For more information on the history of calculus and of mathematics generally, see `http://www-history.mcs.st-and.ac.uk/index.html`.

3.3 The Fundamental Theorem of Calculus

In this section, we discuss the *Fundamental Theorem of Calculus* which establishes a crucial link between differential calculus and the problem of calculating definite integrals, or areas under curves. At the end of this section, you should be able to explain this connection and demonstrate with some examples how the techniques of differential calculus can be used to calculate definite integrals.

Differential calculus is about *how functions are changing*. Suppose for example, that you are thinking of temperature (in °C) as a function of time (in hours). You might write temperature as $T(t)$ to indicate that the temperature T varies with time t. The *derivative* of the function $T(t)$, denoted $T'(t)$, tells us how the temperature is *changing over time*. If you know that at 10.00am yesterday the derivative of T was 0.5 (°C/hr), then you know that the temperature was *increasing* by half a degree per hour at that time. However this does not tell you anything about what the temperature actually was at this time. If you know that by 10.00pm last night the derivative of the temperature was -2°C/hr you still don't know anything about what the temperature was at the time, but you know that it was cooling at a rate of 2 degrees per hour. The derivative T' is itself a function of time, as the rate of increase or decrease of temperature will not remain constant throughout the day. Knowing about $T'(t)$ doesn't tell us anything about how warm or cold it was at any given time, but it gives us such information as when it was getting warmer, when it was getting colder, when it stopped getting warmer and started to cool, and so on.

RECALL : Suppose that f is a function of a variable x. Then $f'(x)$ is the *derivative* of f, also a function of x. The value of f' at a particular point a is the *slope* of the tangent line to the graph $y = f(x)$ at the point $(a, f(a))$. The diagram below shows the graph of the function defined by $f(x) = \frac{1}{2}x^2$ and the tangent line to this graph at the point (3,4.5). The *slope* of this tangent line (which happens to be 3) is the *derivative* of f when $x = 3$, i.e. it is $f'(3)$. As x varies – as we move along the graph from left to right – the slope of the tangent line varies too, so f' is a function of x; as we know it is given in this example by the formula $f'(x) = \frac{d}{dx}\left(\frac{1}{2}x^2\right) = x$.



Now we are going to define a new function related to definite integrals and consider its derivative - we start with an example.

Example 3.3.1. *At time $t = 0$ an object is travelling at 5 metres per second. After t seconds its speed in m/s is given by*

$$v(t) = 5 + 2t.$$

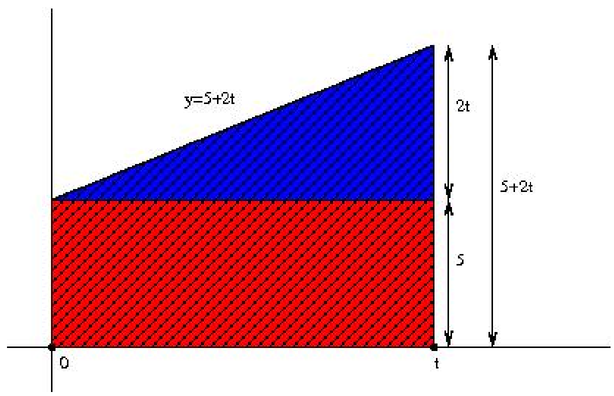
Let $s(t)$ denote the distance travelled by the object after t seconds. So $s(t)$ depends on t obviously since the object is moving over time. From our work in Section 3.1 we know that

$s(t)$ is the area under the graph of $v(t)$ against t, between the vertical lines through 0 and t. We can calculate this in terms of t, by drawing a picture of the graph.

Look at the shape of the region between the graph and the x-axis, between the vertical lines through 0 and t. It is a trapezoid with

- bottom edge formed by a segment of the x-axis of length t;
- left and right edges formed by segments of the vertical lines through 0 and t, of lengths 5 and $5 + 2t$ respectively;
- Top edge formed by part of the graph $y = 5 + 2t$.

The area of this region is $s(t)$. As shown in the diagram, it is the sum of the areas of a rectangle of width t and height 5 (area 5t) and a triangle of width t and height 2t (area t^2). This means : *for any $t \geqslant 0$, the distance covered by this object in the first t seconds of its movement is given by $s(t) = 5t+t^2$.*



IMPORTANT NOTE: The function $s(t)$ associates to t the area under the graph $y = v(t)$ from time 0 to time t. As t increases (i.e. as time passes), this area increases (it represents the distance travelled which is obviously increasing). Note that the derivative of $s(t)$ is exactly $v(t)$.

$$s(t) = 5t + t^2;\ s'(t) = 5 + 2t = v(t).$$

We shouldn't really be surprised by this given the physical context of the problem : $s(t)$ is the total distance travelled at time t, and $s'(t)$ at time t is $v(t)$, the speed at time t. So this is saying that the *instantaneous rate of change* of the distance travelled at a particular moment is the *speed* at which the object is travelling at that moment - which makes sense.

However, there is another way to interpret this statement, which makes sense for definite integrals generally :

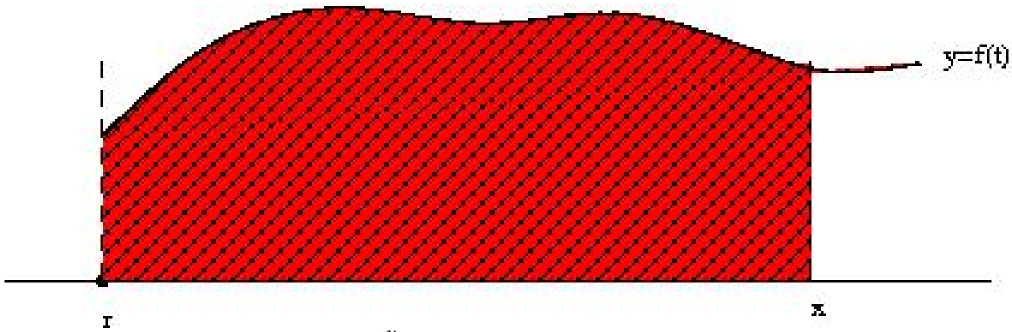
- v is a function whose graph we are looking at.
- For a positive number t, $s(t)$ is the area under the graph of v, to the right of 0 and to the left of t.
- Then the derivative of s is just v, the function under whose graph the area is being measured, i.e $s'(t) = v(t)$.

The more general version of this statement is the *Fundamental Theorem of Calculus*, stated below.

Theorem 3.3.2. *(Fundamental Theorem of Calculus (FToC))*
Let f be a (suitable) function, and let r be a fixed number. Define a function A by

$$A(x) = \int_r^x f(t)\,dt.$$

This means : for a number x, $A(x)$ is the area enclosed by the graph of f and the $x-\mathrm{axis}$, between the vertical lines through r and x. The picture below shows what the function A does.



$$A(x) = \int_r^x f(t)\,dt \text{ is the area shown in red.}$$

The function A *depends on the variable* x*, via the right limit in the definite integral. The Fundamental Theorem of Calculus tells us that the function* f *is exactly the derivative of this area accumulation function* A*. Thus*

$$A'(x) = f(x).$$

Example 3.3.3. *Define a function* F *for* $x \geqslant -6$ *by*

$$F(x) = \int_{-6}^{x} \cos(\pi e^{t^2-4})\,dt.$$

Find $F'(-2)$.

Solution: By the FToC,

$$F'(x) = \cos(\pi e^{x^2-4}), \text{ for } x > -6.$$

Then $F'(-2) = \cos(\pi e^{(-2)^2-4}) = \cos(\pi) = -1$.

NOTES

1. We won't prove the Fundamental Theorem of Calculus, but to get a feeling for what it says, look again at the picture above, and think about how $A(x)$ changes when x moves a little to the right. If $f(x) = 0$, $A(x)$ doesn't change at all as no area is accumulating under the graph of f. If $f(x)$ is positive and large, $A(x)$ increases quickly as x moves to the right. If $f(x)$ is positive but smaller, $A(x)$ increases more slowly with x, because area accumulates more slowly under the "lower" curve. If $f(x)$ is negative, then $A(x)$ will decrease as x increases, because we will be accumulating "negative" area.

2. The Fundamental Theorem of Calculus is *interesting* because it connects differential calculus to the problem of calculating definite integrals, or areas under curves.

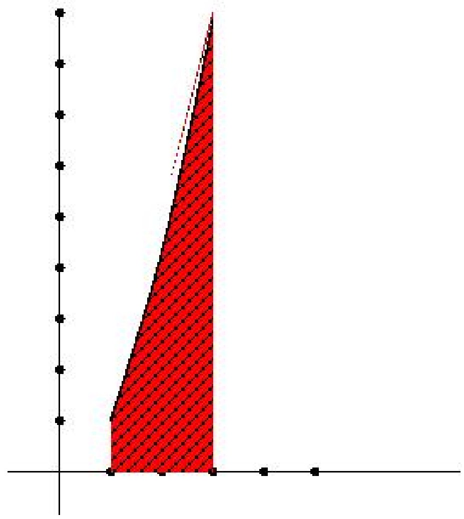
3. The Fundamental Theorem of Calculus is *useful* because we know a lot about differential calculus. Using the machinery of differentiation (the product rule, chain rule etc), we can calculate the derivative of just about anything that can be written in terms of elementary functions (like polynomials, trigonometric functions, exponentials and so on). So we have a lot of theory about differentiation that is all of a sudden relevant to calculating definite integrals as well.

4. The Fundamental Theorem of Calculus can be traced back to work of *Isaac Barrow* and *Isaac Newton* in the mid 17th Century.

Finally we show how to use the Fundamental Theorem of Calculus to calculate definite integrals.

Example 3.3.4. *Calculate* $\int_1^3 x^2\,dx$*.*

Solution: The area that we want to calculate is shown in the picture below.



Imagine that r is some point to the left of 1, and that the function A is defined for $x \geqslant r$ by

$$A(x) = \int_r^x x^2 \, dx,$$

i.e. $A(x)$ is the area under the graph of x^2 between r and x. Then

$$\int_1^3 x^2 \, dx = A(3) - A(1);$$

this is the area under the graph that is to the left of 3 but to the right of 1. So - if we could calculate $A(x)$, we could evaluate this function at $x = 3$ and at $x = 0$.

What we know about the function $A(x)$, from the Fundamental Theorem of Calculus, is that its derivative is given by $A'(x) = x^2$. What function A has derivative x^2?

The derivative of x^3 is $3x^2$, so the derivative of $\frac{1}{3}x^3$ is x^2.

Note : $\frac{1}{3}x^3$ is *not the only* expression whose derivative is x^2. For example $\frac{1}{3}x^3 + 1$, $\frac{1}{3}x^3 - 5$ and any expression of the form $\frac{1}{3}x^3 + C$ for any constant C, also have derivative x^2. All of these are candidates for $A(x)$: basically they just correspond to different choices for the point r. All of these choices for $A(x)$ give the same outcome when we use them to evaluate $\int_1^3 x^2 \, dx$ as suggested above.

So : take $A(x) = \frac{1}{3}x^3$. Then

$$\int_1^3 x^2 \, dx = A(3) - A(1) = \frac{1}{3}(3^3) - \frac{1}{3}(1^3) = 9 - \frac{1}{3} = \frac{26}{3}.$$

This technique is described in general terms in the following version of the Fundamental Theorem of Calculus :

Theorem 3.3.5. *(Fundamental Theorem of Calculus, Part 2) Let* f *be a function. To calculate the definite integral*

$$\int_a^b f(x) \, dx,$$

first find a function F *whose derivative is* f*, i.e. for which* $F'(x) = f(x)$*. (This might be hard). Then*

$$\int_a^b f(x) \, dx = F(b) - F(a).$$

3.4 Techniques of Integration

Recall the following strategy for evaluating definite integrals, which arose from the Fundamental Theorem of Calculus (see Section 3.3). To calculate

$$\int_a^b f(x)\,dx$$

1. Find a function F for which $F'(x) = f(x)$, i.e. find a function F whose derivative is f.
2. Evaluate F at the limits of integration a and b; i.e. calcuate $F(a)$ and $F(b)$. This means replacing x separately with a and b in the formula that defines $F(x)$.
3. Calculate the number $F(b) - F(a)$. This is the definite integral $\displaystyle\int_a^b f(x)\,dx$.

Of the three steps above, the first one is the hard one. There are many examples of (very reasonable looking) functions f for which it is not possible to write down a function F whose derivative is f in a manageable way. But there are many also for which it is, and they will be the focus of our attention in this chapter.

Suppose for example we look at the function g defined by $g(x) = \sin(x^2 + x)$. From the chain rule for differentiation we know that $g'(x) = (2x + 1)\cos(x^2 + x)$. But suppose that we started with

$$(2x + 1)\cos(x^2 + x)$$

and we wanted to find something whose derivative with respect to x was equal to this expression. *How would we get back to* $\sin(x^2 + x)$? In this section we will develop answers to this question, but it doesn't have a neat answer. The answer consists of a collection of strategies, techniques and observations that have to be employed judiciously and adapted for each example. It takes some careful practice to become adept at reversing the differentiation process which is basically what we have to do.

Recall the following notation: if F is a function that satisfies $F'(x) = f(x)$, then

$$F(x)|_a^b \text{ or } F(x)|_{x=a}^{x=b} \text{ means } F(b) - F(a).$$

We also need the following definition:

Definition 3.4.1. *Let f be a function. Another function F is called an* antiderivative *of f if the derivative of F is f, i.e. if $F'(x) = f(x)$, for all (relevant) values of the variable x.*

Thus for example x^2 is an antiderivative of $2x$. Note that $x^2 + 1$, $x^2 + 5$ and $x^2 - 20e$ are also antiderivatives of $2x$. So we talk about *an* antiderviative of a function or expression rather that *the* antiderivative. So: a function may have more than one antiderivative, but different antiderivatives of a particular function will always differ from each other by a constant.
Note : Two functions will have the same derivative if their graphs differ from each other only by a vertical shift; in this case the tangent lines to these graphs for particular values of x will always have the same slope.

Definition 3.4.2. *Let f be a function. The* indefinite integral *of f, written*

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

is the "general antiderivative" of f. If $F(x)$ is a particular antiderivative of f, then we would write

$$\int f(x)\,dx = F(x) + C,$$

to indicate that the differnt antiderivatives of f look like $F(x) + C$, where C maybe any constant. (In this context C is often referred to as a constant of integration).

Example 3.4.3. *We would write*

$$\int 2x\,dx = x^2 + C$$

to indicate that every antiderivative of $2x$ *has the form* x^2+C *for some constant* C*, and that every expression of the form* $x^2 + C$ *(for a constant* C*) has derivative equal to* $2x$*.*

In this section we will consider examples where antiderivatives can be determined without recourse to any sophisticated techniques (which doesn't necessarily mean easily).

The following table reminding us of the derivatives of some elementary functions may be helpful.

$f(x)$	$f'(x)$
x	1
x^2	$2x$
x^3	$3x^2$
$\frac{1}{x^2}$	$-\frac{2}{x^3}$
x^n	nx^{n-1}

$f(x)$	$f'(x)$
$\sin x$	$\cos x$
$\cos x$	$-\sin x$
$\sin 2x$	$2\cos 2x$
e^x	e^x
e^{3x}	$3e^{3x}$

Basically our goal is to figure out how to get from the right to the left column in a table like this.

Example 3.4.4. *Find (i)* $\int x^2\,dx$*, (ii)* $\int_4^6 x^2\,dx$

SOLUTION: (i) $\dfrac{d}{dx}(x^3) = 3x^2$ - so x^3 is not an antiderivative of x^2, it is "too big" by a factor of 3. Thus $\frac{1}{3}x^3$ should be an antiderivative of x^2; indeed

$$\frac{d}{dx}\left(\frac{1}{3}x^3\right) = \frac{1}{3}3x^2 = x^2.$$

We conclude

$$\int x^2\,dx = \frac{1}{3}x^3 + C.$$

This means that every antiderivative of x^2 has the form $\frac{1}{3}x^3 + C$ for some constant C.

(ii) By FTC (Part 2) we have

$$\int_4^6 x^2\,dx = \left.\frac{x^3}{3}\right|_4^6 = \frac{6^3}{3} - \frac{4^3}{3} = \frac{153}{3}.$$

Example 3.4.5. *Determine* $\int \cos 2x\,dx$*.*

SOLUTION: The question is : what do we need to differentiate to get $\cos 2x$? Well, what do we need to differentiate to get something involving cos?
(If you can't answer this question fairly quickly, you are advised to brush up on your knowledge of derivatives of trigonometric functions - don't forget that the SUMS centre can help in this situation).
We know that the derivative of $\sin x$ is $\cos x$.
So a reasonable guess would say that the derivative of $\sin 2x$ might be "something like" $\cos 2x$.
By the chain rule, the derivative of $\sin 2x$ is in fact $2\cos 2x$.
So, in our search for an antiderivative of $\cos 2x$, $\sin 2x$ is pretty close but it gives us twice what we want - we are out by a factor of 2.
So we should compensate for this by taking $\dfrac{1}{2}\sin 2x$; its derivative is

$$\frac{1}{2}(2\cos 2x) = \cos 2x.$$

CONCLUSION: $\displaystyle\int \cos 2x\, dx = \frac{1}{2}\sin 2x + C.$
NOTE: The reason for the commentary on this example is to give you an idea of the sorts of thought processes a person might go through while figuring out an antiderivative of $\cos 2x$. You would not be expected to provide this sort of commentary if you were answering a question like this in an assessment - it would be enough to just write the line labelled "CONCLUSION" above.
The following examples are similar, with less commentary as we continue.

Example 3.4.6. *Determine* $\displaystyle\int e^{\frac{1}{2}x}\, dx$

SOLUTION: We are looking for something whose derivative is $e^{\frac{1}{2}x}$. We know that the derivative of e^x is e^x, so the answer should be something like $e^{\frac{1}{2}x}$. But this is not exactly right because the derivative of $e^{\frac{1}{2}x}$ is

$$\frac{1}{2}e^{\frac{1}{2}x},$$

which is only half of what we want - we are out by a factor of $\frac{1}{2}$ - what we want is twice what we have. We can compensate for this by multiplying what we have by 2 (or dividing it by $\frac{1}{2}$ which is the same). So what we want is $2e^{\frac{1}{2}x}$ - use the chain rule to confirm that the derivative of this expression is $e^{\frac{1}{2}x}$ as required.
CONCLUSION: $\displaystyle\int e^{\frac{1}{2}x}\, dx = 2e^{\frac{1}{2}x} + C$

Example 3.4.7. *Determine* $\displaystyle\int x^5\, dx$

SOLUTION: The derivative of x^6 is $6x^5$. So the derivative of $\dfrac{1}{6}x^6$ is x^5. Hence

$$\int x^5\, dx = \frac{1}{6}x^6 + C.$$

IMPORTANT NOTE: We know that in order to calculate the derivative of an expression like x^n, we reduce the index by 1 to $n-1$, and we multiply by the constant n. So

$$\frac{d}{dx}x^n = nx^{n-1}$$

in general. To find an *antiderivative* of x^n we have to reverse this process. This means that the index *increases* by 1 to $n+1$ and we multiply by the constant $\dfrac{1}{n+1}$. So

$$\int x^n\, dx = \frac{1}{n+1}x^{n+1} + C.$$

This makes sense as long as the number n is not equal to -1 (in which case the fraction $\frac{1}{n+1}$ wouldn't be defined). We will see later how to manage $\int x^{-1}\,dx$ or $\int \frac{1}{x}\,dx$.
Note: included in the general description of $\int x^n\,dx$ above is the statement that

$$\int 1\,dx = x + C.$$

This makes sense when we ask ourselves what we need to differentiate in order to get 1. The answer is x.

Example 3.4.8. *Determine* $\int 3x^2 + 2x + 4\,dx$.

SOLUTION: $\int 3x^2 + 2x + 4\,dx = 3(x^3/3) + 2(x^2/2) + 4x + C = x^3 + 2x^2 + 4x + C.$
Remark: Here we are separately applying our ability to integrate expressions of the form x^n to the x^3 term, the x^2 term, and the constant term. We are also making use of the following fact that indefinite integration behaves *linearly*. This means : if $f(x)$ and $g(x)$ are expressions involving x and a and b are real numbers, we have

$$\int af(x) + bg(x)\,dx = a\int f(x)\,dx + b\int g(x)\,dx.$$

Example 3.4.9. *Determine* $\int_0^\pi \sin x + \cos x\,dx$.

SOLUTION: We need to write down *any* antiderivative of $\sin x + \cos x$ and evaluate it at the limits of integration :

$$\begin{aligned}\int_0^\pi \sin x + \cos x\,dx &= -\cos x + \sin x|_0^\pi \\ &= (-\cos\pi + \sin\pi) - (-\cos 0 + \sin 0) \\ &= -(-1) + 0 - (-1 + 0) = 2.\end{aligned}$$

NOTE: In case you don't find it easy to remember things like cosine and sine of π, $\frac{\pi}{2}$ etc, it is easy enough if you think about it in terms of the definitions of the trigonometric functions. To determine $\cos\pi$, start at the point $(1, 0)$ and travel counter–clockwise around the unit circle through an angle of π radians (180 degrees), arriving at the point $(-1, 0)$. The x-coordinate of the point you are at now is $\cos\pi$, and the y-coordinate is $\sin\pi$.

Example 3.4.10. *Determine* $\int x^{1/3}\,dx$.

SOLUTION: $\int x^{1/3}\,dx = \dfrac{1}{4/3}x^{4/3} + C = \dfrac{3}{4}x^{4/3} + C.$

3.4.1 Substitution - Reversing the Chain Rule

The Chain Rule of Differentation tells us that in order to differentiate the expression $\sin x^2$, we should regard this expression as sin("something") whose derivative (with respect to "something" is cos("something"), then multiply this expression by the derivative of the "something" with respect to x. Thus

$$\frac{d}{dx}(\sin x^2) = \cos x^2 \frac{d}{dx}(x^2) = 2x\cos x^2.$$

Equivalently

$$\int 2x\cos x^2\,dx = \sin x^2 + C.$$

In this section, through a series of examples, we consider how one might go about reversing the differentiation process to get from $2x\cos x^2$ back to $\sin x^2$.

Example 3.4.11. *Determine* $\int 2x\sqrt{x^2+1}\,dx$.

SOLUTION: Notice that the integrand (i.e. the expression to be integrated) involves both the expressions x^2+1 and $2x$. Note also that $2x$ is the derivative of x^2+1.
Introduce the notation u and set $u = x^2+1$. Note $\dfrac{du}{dx} = 2x$.
Then $2x\sqrt{x^2+1} = \frac{du}{dx}\sqrt{u} = u^{\frac{1}{2}}\dfrac{du}{dx}$.

Suppose we were able to find a function F of u for which $\dfrac{d}{du}(F(u)) = u^{\frac{1}{2}}$. Then by the chain rule we would have

$$\frac{d}{dx}(F(u)) = \frac{d}{du}(F(u))\frac{du}{dx} = u^{\frac{1}{2}}(2x) = 2x\sqrt{x^2+1}.$$

So $F(u)$ would be an antiderivative (with respect to x) of $2x\sqrt{x^2+1}$.
Thus we want

$$\frac{d}{du}(F(u)) = u^{\frac{1}{2}}$$

So take

$$F(u) = \int u^{\frac{1}{2}}\,du = \frac{2}{3}u^{\frac{3}{2}} + C.$$

(At this stage we are just using the note after Example 3.4.7, with $n = \frac{1}{2}$).
Thus

$$\int 2x\sqrt{x^2+1}\,dx = \frac{2}{3}(x^2+1)^{\frac{3}{2}} + C.$$

We usually formulate this procedure of "integration by substitution" in the following more concise way.
To find $\int 2x\sqrt{x^2+1}\,dx.$:
Let $u = x^2+1$.
Then $\dfrac{du}{dx} = 2x \Longrightarrow du = 2x\,dx$. Then

$$\int 2x\sqrt{x^2+1}\,dx = \int \sqrt{x^2+1}(2x\,dx) = \int u^{\frac{1}{2}}\,du = \frac{2}{3}u^{\frac{3}{2}} + C.$$

So

$$\int 2x\sqrt{x^2+1}\,dx = \frac{2}{3}(x^2+1)^{\frac{3}{2}} + C.$$

Example 3.4.12. *Determine* $\displaystyle\int x\sin(2x^2)\,dx$

SOLUTION: Let $u = 2x^2$.
Then $\dfrac{du}{dx} = 4x\,dx;\ x\,dx = \dfrac{1}{4}du$. So

$$\int x\sin(2x^2)\,dx = \frac{1}{4}\int \sin u\,du = -\frac{1}{4}\cos u + C = \frac{1}{4}\cos(2x^2) + C.$$

REMARK: It is good practice to check your answer to a problem like this, either mentally or on paper. Check that the derivative of $-\frac{1}{4}\cos(2x^2)$ is indeed equal to $x\sin(2x^2)$.

Example 3.4.13. *Determine* $\displaystyle\int (1-\cos t)^2 \sin t\,dt$

SOLUTION: Write $u = 1 - \cos t$.
Then $\dfrac{du}{dt} = sint;\ du = \sin t\, dt$.
So
$$\int (1-\cos t)^2 \sin t\, dt = \int u^2 du = \frac{1}{3}u^3 + C = \frac{1}{3}(1-\cos t)^3 + C.$$
QUESTION: How do we know what expression to extract and refer to as u?
Really what we are doing in this process is changing the integration problem in the variable t to a (hopefully easier) integration problem in a new variable u - there is a change of variables taking place.

There is no easy answer to the question of how to decide what to rename as "u", but with practice we can develop a sense of what might work. In this example the integrand involves the expression $1 - \cos t$ and also its derivative $\sin t$. This is what makes the substitution $u = 1 - \cos t$ effective for this problem. The "$\sin t$" part of the integrand gets "absorbed" into the "du" in the change of variables, and the "$1 - \cos t$" part is obviously easily written in terms of u. We could try the alternative $u = \sin t$, but this is not likely to be helpful, since it is not so easy to see how to express $1 - \cos t$ in terms of this u, or what would happen with du which would be effectively $\cos t\, dt$.

Example 3.4.14. *To determine* $\displaystyle\int \frac{(1+\sqrt{x})^3}{\sqrt{x}}\, dx$

SOLUTION: How are we to choose u? Well, what are the candidates?
The integrand involves the expressions $1 + \sqrt{x}$ and $\frac{1}{\sqrt{x}}$. The derivative of $1 + \sqrt{x}$ is "something like" $\frac{1}{\sqrt{x}}$, so setting $u = 1 + \sqrt{x}$ might be worth a try.

Let $u = 1 + \sqrt{x}$.
Then $\dfrac{du}{dx} = \dfrac{1}{2}u^{-\frac{1}{2}} = \dfrac{1}{2}\dfrac{1}{\sqrt{x}};\ \dfrac{1}{\sqrt{x}}dx = 2\, du$.
So
$$\int \frac{(1+\sqrt{x})^3}{\sqrt{x}}\, dx = 2\int u^3\, du = \frac{2}{4}u^4 + C = \frac{1}{2}(1+\sqrt{x})^4 + C.$$

Example 3.4.15. *Determine* $\displaystyle\int \frac{16x}{\sqrt{8x^2+1}}\, dx$

SOLUTION: Let $u = \sqrt{8x^2+1}$.
Then $\dfrac{du}{dx} = \dfrac{1}{2}(8x^2+1)^{-\frac{1}{2}}(16x) = \dfrac{8x}{\sqrt{8x^2+1}}$.
Thus $\dfrac{16x}{\sqrt{8x^2+1}}\, dx = 2du$, and
$$\int \frac{16x}{\sqrt{8x^2+1}}\, dx = 2\int du = 2u + C = 2\sqrt{x^2+1} + C.$$
NOTE: An alternative here would have been to set $u = 8x^2 + 1$. That this would also be successful is left for you to check as an exercise.

DIGRESSION - IMPORTANT NOTE: *The Integral* $\displaystyle\int \frac{1}{x}\, dx$

Suppose that $x > 0$ and $y = \ln x$. Recall this means (by definition) that $e^y = x$. Differentiating both sides of this equation (with respect to x) gives
$$e^y \frac{dy}{dx} = 1 \Longrightarrow \frac{dy}{dx} = \frac{1}{e^y} = \frac{1}{x}.$$
Thus the derivative of $\ln x$ is $\dfrac{1}{x}$, and
$$\int \frac{1}{x}\, dx = \ln x + C,\ \text{for } x > 0.$$

If $x < 0$, then

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

This latter formula applies for all $x \neq 0$.

Example 3.4.16. *To determine* $\int \frac{\sec^2 x}{\tan x}\,dx$

Note : the derivative of $\tan x$ is $\sec^2 x$, suggesting the substitution $u = \tan x$. You are not necessarily expected to know the derivative of $\tan x$ (or of any of the trigonometric functions) of the top of your head, but you should know where to find them in the "Formulae and Tables" booklet.
Let $u = \tan x$.
Then $\frac{du}{dx} = \sec^2 x;\ du = \sec^2 x\,dx$. Thus $\frac{\sec^2 x}{\tan x}\,dx = \frac{1}{u}du$, and

$$\int \frac{\sec^2 x}{\tan x}\,dx = \int \frac{1}{u}\,du = \log|u| + C = \log|\tan x| + C.$$

Substitution and Definite Integrals

Example 3.4.17. *Evaluate* $\int_0^1 \frac{5r}{(4+r^2)^2}\,dr$.

SOLUTION: To find an antiderivative, let $u = 4 + r^2$.
Then $\frac{du}{dr} = 2r,\ du = 2r\,dr;\ 5r\,dr = \frac{5}{2}du$.
So

$$\int \frac{5r}{(4+r^2)^2}\,dr = \frac{5}{2}\int \frac{1}{u^2}\,du = \frac{5}{2}\int u^{-2}\,du.$$

Thus $\int \frac{5r}{(4+r^2)^2}\,dr = -\frac{5}{2}\times\frac{1}{u} + C$, and we need to evaluate $-\frac{5}{2}\times\frac{1}{u}$ at $r = 0$ and at $r = 1$. We have two choices :

1. Write $u = 4 + r^2$ to obtain

$$\begin{aligned}\int_0^1 \frac{5r}{(4+r^2)^2}\,dr &= -\frac{5}{2}\frac{1}{4+r^2}\Big|_{r=0}^{r=1}\\ &= -\frac{5}{2}\frac{1}{4+1^2} - \left(-\frac{5}{2}\times\frac{1}{4+0^2}\right)\\ &= -\frac{5}{2}\times\frac{1}{5} + \frac{5}{2}\times\frac{1}{4}\\ &= \frac{1}{8}.\end{aligned}$$

2. Alternatively, write the antiderivative as $-\frac{5}{2}\frac{1}{u}$ and replace the limits of integration with the corresponding values of u.
   When $r = 0$ we have $u = 4 + 0^2 = 4$.
   When $r = 1$ we have $u = 4 + 1^2 = 5$.
   Thus

$$\begin{aligned}\int_0^1 \frac{5r}{(4+r^2)^2}\,dr &= -\frac{5}{2}\times\frac{1}{u}\Big|_{u=4}^{u=5}\\ &= -\frac{5}{2}\times\frac{1}{5} - \left(-\frac{5}{2}\times\frac{1}{4}\right)\\ &= \frac{1}{8}.\end{aligned}$$

3.4.2 Integration by parts - reversing the product rule

In this section we discuss the technique of "integration by parts", which is essentially a reversal of the product rule of differentiation.

Example 3.4.18. *Find* $\int x \cos x \, dx$.

There is no obvious substitution that will help here.
How could $x \cos x$ arise as a derivative?
Well, $\cos x$ is the derivative of $\sin x$. So, if you were differentiating $x \sin x$, you would get $x \cos x$ but according to the product rule you would also get another term, namely $\sin x$. Thus

$$\begin{aligned} \frac{d}{dx}(x \sin x) &= x \cos x + \sin x \\ \Longrightarrow \frac{d}{dx}(x \sin x) - \sin x &= x \cos x. \end{aligned}$$

Note that $\sin x = \frac{d}{dx}(-\cos x)$. So

$$\frac{d}{dx}(x \sin x) - \frac{d}{dx}(-\cos x) = x \cos x \Longrightarrow \frac{d}{dx}(x \sin x + \cos x) = x \cos x.$$

CONCLUSION: $\int x \cos x \, dx = x \sin x + \cos x + C$.

What happened in this example was basically that the product rule was reversed. This process can be managed in general as follows. Recall from differential calculus that if u and v are expressions involving x, then

$$(uv)' = u'v + uv'.$$

Suppose we integrate both sides here with respect to x. We obtain

$$\int (uv)' \, dx = \int u'v \, dx + \int uv' \, dx \Longrightarrow uv = \int u'v \, dx + \int uv' \, dx.$$

This can be rearranged to give the *Integration by Parts Formula* :

$$\int uv' \, dx = uv - \int u'v \, dx.$$

Strategy : when trying to integrate a product, assign the name u to one factor and v' to the other. Write down the corresponding u' (the derivative of u) and v (an antiderivative of v').

The integration by parts formula basically allows us to exchange the problem of integrating uv' for the problem of integrating $u'v$ - which might be easier, if we have chosen our u and v' in a sensible way.

Here is the first example again, handled according to this scheme.

Example 3.4.19. *Use the integration by parts technique to determine* $\int x \cos x \, dx$.

SOLUTION: Write

$$\begin{array}{ll} u = x & v' = \cos x \\ u' = 1 & v = \sin x \end{array}$$

Then

$$\begin{aligned} \int x \cos x \, dx &= \int uv' \, dx = uv - \int u'v \, dx \\ &= x \sin x - \int 1 \sin x \, dx \\ &= x \sin x + \cos x + C. \end{aligned}$$

NOTE: We could alternatively have written $u = \cos x$ and $v' = x$. This would be less successful because we would then have $u' = -\sin x$ and $v = \frac{x^2}{2}$, which looks worse than v'. The integration by parts formula would have allowed us to replace

$$\int x \cos x \, dx \quad \text{with} \quad \int \frac{x^2}{2} \sin x \, dx,$$

which is not an improvement.

So it matters which component is called u and which is called v'.

Example 3.4.20. *To determine* $\int \ln x \, dx$.

SOLUTION: Let $u = \ln x$, $v' = 1$. Then $u' = \frac{1}{x}$, $v = x$.

$$\begin{aligned}\int \ln x \, dx &= \int uv' \, dx = uv - \int u'v \, dx \\ &= x \ln x - \int \frac{1}{x} x \, dx \\ &= x \ln x - x + C.\end{aligned}$$

NOTE: Example 3.4.20 shows that sometimes problems which are not obvious candidates for integration by parts can be attacked using this technique.

Sometimes two applications of the integration by parts formula are needed, as in the following example.

Example 3.4.21. *To evaluate* $\int x^2 e^x \, dx$.

SOLUTION: Let $u = x^2$, $v' = e^x$. Then $u' = 2x$, $v = e^x$.

$$\begin{aligned}\int x^2 e^x \, dx &= \int uv' \, dx = uv - \int u'v \, dx \\ &= x^2 e^x - \int 2x e^x \, dx \\ &= x^2 e^x - 2 \int x e^x \, dx.\end{aligned}$$

Let $I = \int x e^x \, dx$.

To evaluate I apply the integration by parts formula a second time.

$$\begin{array}{ll} u = x & v' = e^x \\ u' = 1 & v = e^x. \end{array}$$

Then $I = \int x e^x \, dx = x e^x - \int e^x \, dx = x e^x - e^x + C$. Finally

$$\int x^2 e^x \, dx = x^2 e^x - 2x e^x + 2e^x + C.$$

The next example shows another mechanism by which a second application of the integration by parts formula can succeed where the first is not enough.

Example 3.4.22. *Determine* $\int e^x \cos x \, dx$.

Let

$$\begin{array}{ll} u = e^x & v' = \cos x \\ u' = e^x & v = \sin x. \end{array}$$

Then

$$\int e^x \cos x \, dx = e^x \sin x - \int e^x \sin x \, dx.$$

For $\int e^x \sin x \, dx$: Let

$$\begin{array}{ll} u = e^x & v' = \sin x \\ u' = e^x & v = -\cos x. \end{array}$$

Then

$$\int e^x \sin x \, dx = -e^x \cos x + \int e^x \cos x \, dx,$$

and

$$\int e^x \cos x \, dx = e^x \sin x - \left(-e^x \cos x + \int e^x \cos x \, dx \right)$$

$$\implies 2 \int e^x \cos x \, dx \quad = \quad e^x \sin x + e^x \cos x + C$$

$$\implies \int e^x \cos x \, dx = \frac{1}{2}(e^x \sin x + e^x \cos x) + C$$

Finally, an example of a definite integral evaluated using the integration by parts technique.

Example 3.4.23. *Evaluate* $\int_0^1 (x+3)e^{2x} \, dx$.

SOLUTION: Write

$$\begin{array}{ll} u = x + 3 & v' = e^{2x} \\ u' = 1 & v = \frac{1}{2}e^{2x} \end{array}$$

Then

$$\begin{aligned} \int_0^1 (x+3)e^{2x} \, dx &= \int uv' \, dx = (uv)|_0^1 - \int_0^1 u'v \, dx \\ &= \frac{x+3}{2}e^{2x}\Big|_0^1 - \frac{1}{2}\int_0^1 e^{2x} \, dx \\ &= \frac{x+3}{2}e^{2x}\Big|_0^1 - \frac{1}{2} \times \frac{1}{2}e^{2x}\Big|_0^1 \\ &= \frac{4}{2}e^2 - \frac{3}{2}e^0 - \frac{1}{4}e^2 + \frac{1}{4}e^0 \\ &= \frac{7}{4}e^2 - \frac{5}{4}. \end{aligned}$$

3.4.3 Partial Fraction Expansions - Integrating Rational Functions

We know how to integrate polynomial functions; for example

$$\int 2x^2 + 3x - 4 \, dx = \frac{2}{3}x^3 + \frac{3}{2}x^2 - 4x + C.$$

We also know that

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

This section is about integrating *rational functions*; i.e. quotients in which the numerator and denominator are both polynomials.

REMARK: If we were presented with the task of adding the expressions $\frac{2}{x+3}$ and $\frac{1}{x+4}$, we would take $(x+3)(x+4)$ as a *common denominator* and write

$$\frac{2}{x+3}+\frac{1}{x+4}=\frac{2(x+4)}{(x+3)(x+4)}+\frac{1(x+3)}{(x+3)(x+4)}=\frac{2(x+4)+1(x+3)}{(x+3)(x+4)}=\frac{3x+11}{(x+3)(x+4)}.$$

Question: Suppose we were presented with the expression $\frac{3x+11}{(x+3)(x+4)}$ and asked to rewrite it in the form $\frac{A}{x+3}+\frac{B}{x+4}$, for *numbers* A and B. How would we do it?

Another Question Why would we want to do such a thing?

Answer to the second question: Maybe if we want to integrate the expression : we know how to integrate things like $\frac{1}{x+3}$, but not things like $\frac{3x+11}{(x+3)(x+4)}$.

Answer to the first question: Write

$$\frac{3x+11}{(x+3)(x+4)}=\frac{A}{x+3}+\frac{B}{x+4}.$$

Then

$$\frac{3x+11}{(x+3)(x+4)}=\frac{A(x+4)}{(x+3)(x+4)}+\frac{B(x+3)}{(x+3)(x+4)}=\frac{(A+B)x+4A+3B}{(x+3)(x+4)}.$$

This means $3x+11=(A+B)x+4A+3B$ for all x, which means

$$A+B=3, \text{ and } 4A+3B=11.$$

Thus $-4A-4B=-12,\ -B=-1,\ B=1$ and $A=2$. So

$$\frac{3x+11}{(x+3)(x+4)}=\frac{2}{x+3}+\frac{1}{x+4}.$$

Alternative Method: We want

$$3x+11=A(x+4)+B(x+3),$$

for *all* real numbers x. If this statement is true for all x, then in particular it is true when $x=-4$. Setting $x=-4$ gives

$$-12+11=A(0)+B(-1)\Longrightarrow B=1.$$

Setting $x=-3$ gives

$$-9+11=A(1)+B(0)\Longrightarrow A=2.$$

Thus

$$\frac{3x+11}{(x+3)(x+4)}=\frac{2}{x+3}+\frac{1}{x+4}.$$

Expansions of rational functions of this sort are called *partial fraction expansions*.

Example 3.4.24. *Determine* $\int \frac{3x+11}{(x+3)(x+4)}\,dx$.

SOLUTION : Write

$$\int \frac{3x+11}{(x+3)(x+4)}\,dx=\int \frac{2}{x+3}\,dx+\int \frac{1}{x+4}\,dx$$

Then

$$\int \frac{3x+11}{(x+3)(x+4)}\,dx=2\ln|x+3|+\ln|x+4|+C=\ln(x+3)^2+\ln|x+4|+C.$$

Example 3.4.25. *Determine* $\int \frac{1}{x^2+5x+6}\,dx$.

SOLUTION: Write $\frac{1}{x^2+5x+6} = \frac{1}{(x+2)(x+3)}$ in the form

$$\frac{A}{x+2} + \frac{B}{x+3},$$

for constants A and B. This means

$$\frac{1}{(x+2)(x+3)} = \frac{A(x+3)+B(x+2)}{(x+2)(x+3)},$$

i.e. $1 = A(x+3) + B(x+2)$ for all x.
Thus

$$0x + 1 = (A+B)x + (3A+2B),$$

which means $A + B = 0$ and $3A + 2B = 1$. This pair of equations has the unique solution $A = 1,\ B = -1$. Thus

$$\begin{aligned} \frac{1}{(x+2)(x+3)} &= \frac{1}{x+2} = \frac{1}{x+3} \\ \Longrightarrow \int \frac{1}{(x+2)(x+3)} &= \int \frac{1}{x+2} - \frac{1}{x+3}\,dx \\ &= \ln|x+3| - \ln|x+2| + C. \end{aligned}$$

NOTE: Any expression of the form $\frac{f(x)}{g(x)}$ where

1. $f(x)$ and $g(x)$ are polynomials and $g(x)$ has higher degree than $f(x)$, and
2. $g(x)$ can be factorized as the product of distinct linear factors

$$g(x) = (x-a_1)(x-a_2)\dots(x-a_k)$$

has a *partial fraction expansion* of the form

$$\frac{f(x)}{g(x)} = \frac{A_1}{x-a_1} + \frac{A_2}{x-a_2} + \dots + \frac{A_k}{x-a_k},$$

where $A_1,\ A_2,\dots,A_k$ are numbers.

Example 3.4.26. *Determine* $\int \frac{x^3+3x+2}{x+1}\,dx$.

In this example the degree of the numerator exceeds the degree of the denominator, so first apply long division to find the quotient and remainder upon dividing x^3+3x+2 by $x = 1$.

We find that the quotient is $x^2 - x + 4$ and the remainder is -2. Hence

$$\frac{x^3+3x+2}{x+1} = x^2 - x + 4 + \frac{-2}{x+1}.$$

Thus

$$\int \frac{x^3+3x+2}{x+1}\,dx = \int x^2 - x + 4\,dx - 2\int \frac{1}{x+1}\,dx = \frac{1}{3}x^3 - \frac{1}{2}x^2 + 4x - 2\ln|x+1| + C.$$

NOTE: In the above example we had $\dfrac{f(x)}{g(x)}$ with $f(x)$ of greater degree than $g(x)$. In such cases we can always write

$$\frac{f(x)}{g(x)} = q(x) + \frac{r(x)}{g(x)},$$

where the polynomials $q(x)$ and $r(x)$ are the quotient and remainder respectively on dividing $f(x)$ by $g(x)$, and the degree of $r(x)$ is less than that of $g(x)$.

Example 3.4.27. *Determine* $\displaystyle\int \frac{x+1}{(2x+1)^2(x-2)}\,dx.$

In this case the denominator has a repeated linear factor $2x+1$. It is necessary to include both $\dfrac{A}{2x+1}$ and $\dfrac{B}{(2x+1)^2}$ in the partial fraction expansion. We have

$$\frac{x+1}{(2x+1)^2(x-2)} = \frac{A}{2x+1} + \frac{B}{(2x+1)^2} + \frac{C}{x-2}.$$

Then

$$\frac{x+1}{(2x+1)^2(x-2)} = \frac{A(2x+1)(x-2) + B(x-2) + C(2x+1)^2}{(2x+1)^2(x-2)}.$$

This means that the polynomials $x+1$ and $A(2x+1)(x-2)+B(x-2)+C(2x+1)^2$ are equal, and therefore have the same value when x is replaced by any real number.

$$\begin{array}{lll} x = 2: & 3 = C(5)^2 & C = \frac{3}{25} \\ x = -\frac{1}{2}: & \frac{1}{2} = B\left(-\frac{5}{2}\right) & B = -\frac{1}{5} \\ x = 0: & 1 = A(1)(-2) + B(-2) + C(1)^2 & A = -\frac{6}{25} \end{array}$$

Thus

$$\frac{x+1}{(2x+1)^2(x-2)} = \frac{-6/25}{2x+1} + \frac{-1/5}{(2x+1)^2} + \frac{3/25}{x-2}$$

and

$$\int \frac{x+1}{(2x+1)^2(x-2)}\,dx = -\frac{6}{25}\int \frac{1}{2x+1}\,dx - \frac{1}{5}\int \frac{1}{(2x+1)^2}\,dx + \frac{3}{25}\int \frac{1}{x-2}\,dx.$$

Call the three integrals on the right above $I_1,\ I_2,\ I_3$ respectively.

- I_1 : Let $u = 2x+1,\ du = 2dx,\ dx = \frac{1}{2}du.$
  $\int \frac{1}{2x+1}\,dx = \frac{1}{2}\int \frac{du}{u} = \frac{1}{2}\ln|u|(+C_1) = \frac{1}{2}\ln|2x+1|(+C_1).$
- I_2 : Let $u = 2x+1,\ du = 2dx,\ dx = \frac{1}{2}du.$

  $\int \frac{1}{(2x+1)^2}\,dx = \frac{1}{2}\int u^{-2}\,du = -\frac{1}{2}u^{-1}(+C_2) = -\frac{1}{2(2x+1)}(+C_2).$
- I_3 : $\int \frac{1}{x-2}\,dx = \ln|x-2|(+C_3).$

Thus

$$\int \frac{x+1}{(2x+1)^2(x-2)}\,dx = -\frac{3}{25}\ln|2x+1| + \frac{1}{10(2x+1)} + \frac{3}{25}\ln|x-2| + C.$$

3.5 Improper Integrals

Suppose that $f(x)$ is a continuous function that satifies

$$\lim_{x\to\infty} f(x) = 0;$$

for example $f(x) = e^{-x}$ has this property. Then we can consider the total area between the graph $y = f(x)$ and the x-axis, to the right of (for example) $x = 1$. This area is denoted by

$$\int_1^\infty f(x)\,dx$$

and referred to as an *improper integral*. For a given function, it is not clear whether the area involved is finite or infinite (if it is infinite, the improper integral is said to *diverge* or to be *divergent*). One question that arises is how we can determine if the relevant area is finite or infinite, another is how to calculate it if it is finite.

Definition 3.5.1. *If the function* f *is continuous on the interval* $[a, \infty)$*, then the* improper integral $\displaystyle\int_a^\infty f(x)\,dx$ *is defined by*

$$\int_a^\infty f(x)\,dx = \lim_{b\to\infty}\int_a^b f(x)\,dx$$

provided this limit exists. In this case the improper integral is called convergent *(otherwise it's divergent). Similarly, if* f *is continuous on* $(-\infty, a)$*, then*

$$\int_{-\infty}^a f(x)\,dx := \lim_{b\to-\infty}\int_b^a f(x)\,dx$$

Remarks:

1. So to calculate an improper integral of the form $\int_1^\infty f(x)\,dx$ (for example), we first calculate

   $$\int_1^b f(x)\,dx$$

   for a general b. This will typically be an expression involving b. We then take the limit as $b \to \infty$.

2. The condition that $f(x)$ is continuous in the definition above is a bit stronger than we really need. In order to make the definition

   $$\int_a^\infty f(x)\,dx = \lim_{b\to\infty}\int_a^b f(x)\,dx$$

   what we really need is that $\int_a^b f(x)\,dx$ exists for all b with $b \geqslant a$.

3. If both $\int_{-\infty}^a f(x)\,dx$ and $\int_a^\infty f(x)\,dx$ exist for some a, then the improper integral $\displaystyle\int_{-\infty}^\infty f(x)\,dx$ is defined by

   $$\int_\infty^\infty f(x)\,dx = \int_{-\infty}^a f(x)\,dx + \int_a^\infty f(x)\,dx.$$

Example 3.5.2. *Show that the improper integral* $\displaystyle\int_1^\infty \frac{1}{x}\,dx$ *is divergent.*

Solution:

$$\int_1^b \frac{1}{x}\,dx = \ln x|_1^b = \ln b - \ln 1 = \ln b.$$

Since $\ln b \to \infty$ as $b \to \infty$, $\lim_{b\to\infty} \ln b$ does not exist and the integral diverges.
EXERCISE: Think about how this is related to the divergence of the harmonic series $\sum \frac{1}{n}$.

Example 3.5.3. *Evaluate* $\displaystyle\int_{-\infty}^{-2} \frac{1}{x^2}\,dx.$

Solution:

$$\int_b^{-2} \frac{1}{x^2}\,dx = -\frac{1}{x}\bigg|_b^{-2} = \frac{1}{2} + \frac{1}{b}$$

Then $\lim_{b\to\infty}\left(\frac{1}{2}+\frac{1}{b}\right) = \frac{1}{2}$, and

$$\int_{-\infty}^{-2} \frac{1}{x^2}\,dx = \frac{1}{2}.$$

Example 3.5.4. *Evaluate* $\displaystyle\int_2^{\infty} xe^{-x}\,dx.$

Solution: Integrating by parts gives

$$\begin{aligned}\int_2^b xe^{-x}\,dx &= -xe^{-x}\big|_2^b + \int_2^b e^{-x}\,dx \\ &= -be^{-b} + 2e^{-2} - e^{-b} + e^2\end{aligned}$$

Taking the limit as $b \to \infty$, we obtain

$$\int_2^{\infty} xe^{-x}\,dx = \frac{2}{e^2} + e^2.$$

ANOTHER TYPE OF IMPROPER INTEGRAL

If the graph $y = f(x)$ has a vertical asymptote for a value of x in the interval $[c, d]$, these needs to be considered when computing the integral $\int_c^d f(x)\,dx$, since this integal describes the area of a region that is infinite in the vertical direction at the asymptote.

- If the vertical asymptote is at the left endpoint c, then we define

$$\int_c^d f(x)\,dx = \lim_{b\to a^+} \int_b^d f(x)\,dx.$$

- If the vertical asymptote is at the right endpoint d, then we define

$$\int_c^d f(x)\,dx = \lim_{b\to d^-} \int_c^b f(x)\,dx.$$

- If the vertical asymptote is at an interior point m of the interval $[c, d]$, then we define

$$\int_c^d f(x)\,dx = \int_c^m f(x)\,dx + \int_m^c f(x)\,dx,$$

  and the two improper integrals involving m are handled as above.

As in the case of improper integrals of the first type, these improper integrals are said to *converge* if the limits in question can be evaluated and to diverge if these limits do not exist. Divergence means that the area involved is infinite.

Example 3.5.5. *Determine whether the improper integral* $\int_{-2}^{4} \frac{1}{x^2}\,dx$ *is convergent or divergent.*

SOLUTION: What makes this integral improper is the fact that the graph $y = \frac{1}{x^2}$ has a vertical asymptote at $x = 0$. Thus

$$\int_{-2}^{4} \frac{1}{x^2}\,dx = \int_{-2}^{0} \frac{1}{x^2}\,dx + \int_{0}^{4} \frac{1}{x^2}\,dx$$

For the first of these two integrals we have

$$\begin{aligned} \int_{-2}^{0} \frac{1}{x^2}\,dx &= \lim_{b\to 0^-} \int_{-2}^{b} \frac{1}{x^2}\,dx \\ &= \lim_{b\to 0^-} \left(-\frac{1}{x}\right)\Big|_{-2}^{b} \\ &= \lim_{b\to 0^-} \left(-\frac{1}{b} + \frac{1}{2}\right) \end{aligned}$$

Since $\lim_{b\to 0^-}(-\frac{1}{b})$ does not exist, the improper integral $\int_{-2}^{0} \frac{1}{x^2}\,dx$ *diverges*. This means that the area enclosed between the graph $y = \frac{1}{x^2}$ and the x-axis over the interval $[-2, 0]$ is *infinite*.

Now that we know that the first of the two improper integrals in our problem diverges, we don't need to bother with the second. The improper integral $\int_{-2}^{4} \frac{1}{x^2}\,dx$ is divergent.