MATH341 MODULE NOTES

These notes contain all of the material which will be presented on the data projector in lectures: that is, motivation, theory, definitions, and some examples. Thus you should not need to take detailed handwritten notes while the data projector is being used.

Most examples and proofs will be presented on the blackboard, and are not contained in these notes. Each time there is some blackboard material, there's a little dagger in the margin of these notes like this. †24 You should ensure that these daggers are properly cross-referenced with the relevant parts of your written notes: perhaps the simplest way to do this is to number sections of your written notes according to the number by each dagger in these notes.

The notes contain a few sections written in a smaller font like this. These contain non-examinable material "for interest only". They will not be covered in lectures.

There are also some "asides" at the end of each chapter of the notes. These cover things that most students will have met in earlier modules. They will only be covered very briefly in lectures. If you're not familiar with them, you're expected to read up on them in these notes.

Chapter 1

Metric Spaces

1.1 Introduction

The concept of distance is a familiar one. In two-dimensional space $\mathbb{R}^2$, the distance between two points is the length of the straight line joining them. If the two points have coordinates (x_1, y_1) and (x_2, y_2), we can calculate the distance between them using Pythagoras's theorem:

$$\text{distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

(see Figure 1.1).

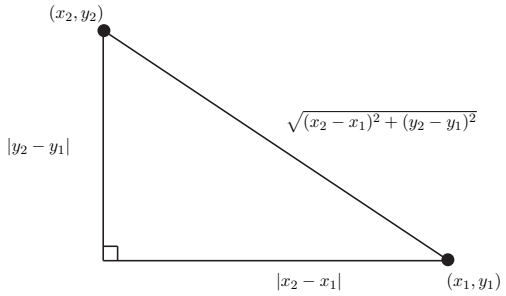

Figure 1.1: The distance between two points in $\mathbb{R}^2$

A similar formula gives the distance between two points (x_1, y_1, z_1) and (x_2, y_2, z_2) in three-dimensional space $\mathbb{R}^3$ as

$$\text{distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2},$$

and indeed we can extend this to n-dimensional space $\mathbb{R}^n$ if we wish.

The distance between two points x_1 and x_2 on the line $\mathbb{R}$ is just the size of the difference between them, i.e. $|x_1 - x_2|$: in fact this fits in with the

formulae in higher dimensions, since

$$|x_1 - x_2| = \sqrt{(x_1 - x_2)^2}$$

(try it with some values of x_1 and x_2 if you're not sure why).

Two important mathematical concepts, *limits* and *continuity*, arise from the notion of distance.

The limit of a sequence

A sequence (x_n) in a set X is just an infinite list of (some of the) elements of X (possibly with repetition): $x_0, x_1, x_2, x_3, \ldots$. Thus, for example, a sequence in $\mathbb{R}^2$ is a list of points in the plane, $(x_0, y_0), (x_1, y_1), (x_2, y_2), (x_3, y_3) \ldots$.

We say that a sequence (x_n) in X *tends to a limit* $\ell \in X$ *as* $n \to \infty$ if the x_n get closer and closer to ℓ as n gets bigger and bigger (roughly speaking – we'll see a precise definition later). In order for "closer and closer" to mean anything, we have to have a way of measuring the distance between two elements of X.

For example, the sequence $((x_n, y_n))$ in $\mathbb{R}^2$ depicted in Figure 1.2 tends (or appears to, from what we can see in the picture) to the limit (x, y).

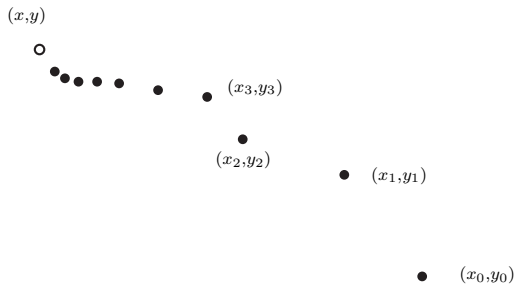

Figure 1.2: A sequence of points tending to a limit in $\mathbb{R}^2$

Continuity of a function

We have a notion that a real function $f(x)$ (that is, a function $f : \mathbb{R} \to \mathbb{R}$) is *continuous* if we can draw its graph without taking our pen off the paper. While this description is very good for understanding what continuity is all about, it has two major defects: first, it's too vague and non-mathematical – it would be very hard to prove something about all continuous functions starting from this definition. Second, it doesn't generalise to higher dimensions or other contexts: can you imagine what it would mean for it

to be possible to draw the graph of a function $f : \mathbb{R}^2 \to \mathbb{R}^2$, such as $f(x, y) = (x^2 + y^2, e^{-xy})$, without taking your pen off the paper?

(See Aside 1 on Page 48 if you're not sure about the function notation $f : X \to Y$.)

A very rough idea for a better definition of continuity is this: a function $f : X \to Y$ is continuous if $f(x)$ *is very close to* $f(y)$ *whenever* x *is very close to* y. Clearly we need a notion of distance to make sense of "very close". To see why this corresponds to our intuitive idea of continuity for functions $f : \mathbb{R} \to \mathbb{R}$, consider the graph of a discontinuous function (i.e. one which has a break in the graph), as shown in Figure 1.3. Note that although x_1 and x_2 are very close, $f(x_1)$ and $f(x_2)$ are far apart: we can find such points x_1 and x_2 precisely because of the break in the graph. We can take x_1 and x_2 to be as close to each other as we like, provided one is on each side of the break.

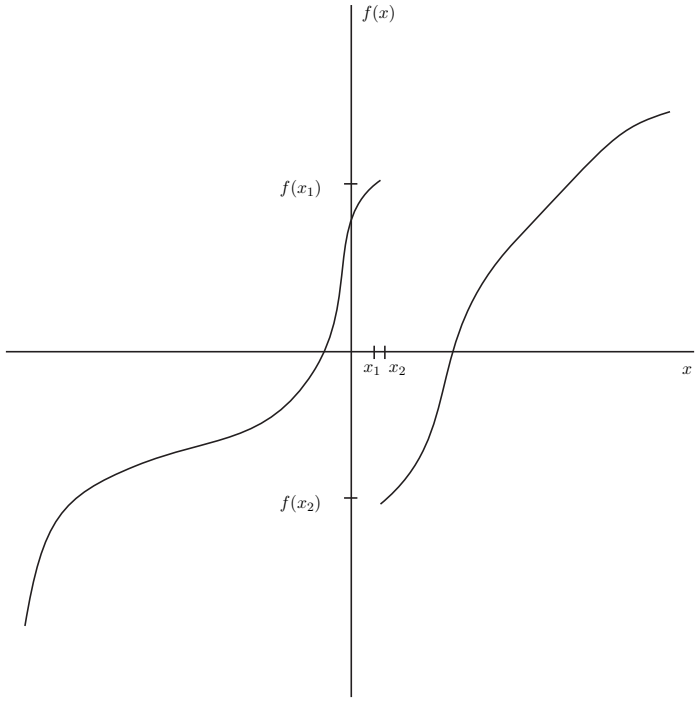

Figure 1.3: A discontinuous function $f : \mathbb{R} \to \mathbb{R}$

Thus the notion of distance makes it possible for us to talk about limits of sequences in $\mathbb{R}^n$, and continuity of functions $f : \mathbb{R}^n \to \mathbb{R}^m$ for any n and m. The start of metric space theory is when we realise that it'd be useful to be able to talk about the "distance" between two objects in other

situations than when those objects are points in n-dimensional space. Here are two examples.

The distance between shapes in the plane

Everyone would agree that the circle and hexagon on the left of Figure 1.4 are closer to each other than are the circle and the rectangle on the right.

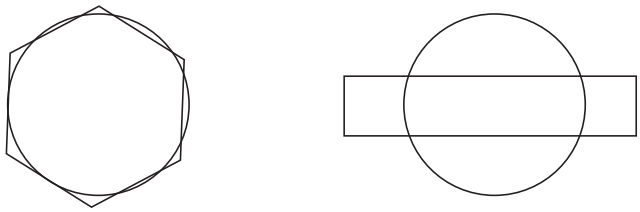

Figure 1.4: Distance between shapes in the plane

Is it possible to give a numerical value to such distances? If so, we could talk about the convergence of sequences (x_n) in the set X whose elements are "shapes in the plane". For example, we might be able to show that the sequence $x_3, x_4, x_5, x_6, \ldots$ in X depicted in Figure 1.5 tends to the circle (the elements x_3, x_4, x_5, x_6, x_{12} and x_{20} of the sequence are shown in the figure, together with the circle which they appear to tend to: x_{20} (the 20-sided polygon) is so "close" to the circle that you probably can't distinguish them).

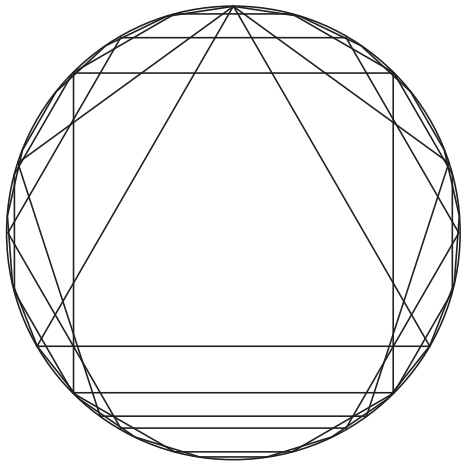

Figure 1.5: The sequence of polygons tends to the circle?

Similarly, we could talk about the continuity of functions defined on X, or taking values in X. For example, suppose we could define a function

$A : X \to \mathbb{R}$, where $A(x)$ is the area of the shape x. (Think about this for a moment. A function $X \to \mathbb{R}$ takes as input a shape in the plane (i.e. an element of X), and produces as output a real number. A good way to produce a real number from a shape in the plane is to work out its area.)

We could then ask whether or not this function is continuous: that is, if two shapes which are very close to each other always have very close areas.

In fact, in order to make sense of distances between such shapes in the plane we need to be careful about what we mean by a "shape": it will be some time before we're able to come back to this example and be more precise.

The distance between functions defined on $[0, 1]$

Everyone would agree that the two functions whose graphs are shown on the left of Figure 1.6 are closer to each other than are the two functions whose graphs are shown on the right.

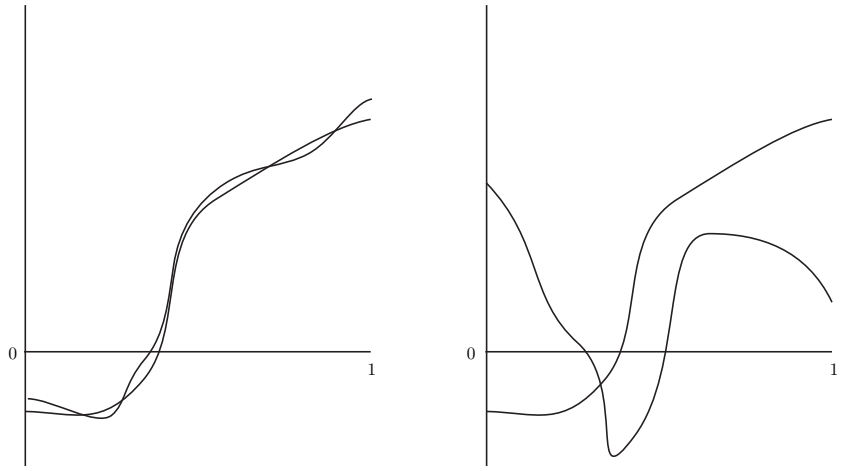

Figure 1.6: Distance between functions defined on $[0, 1]$

Is it possible to give a numerical value to such distances? If so, we could talk about the convergence of sequences (f_n) in the set X whose elements are "continuous functions $[0, 1] \to \mathbb{R}$" (note that X is an unimaginably big set).

As an example, consider the Maclaurin series expansion of $f(x) = e^x$:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots .$$

For $n \geq 0$, let $f_n \colon [0,1] \to \mathbb{R}$ be the function

$$f_n(x) = \sum_{r=0}^{n} \frac{x^r}{r!}$$

(thus $f_n(x)$ is just the first $n+1$ terms in the Maclaurin series expansion: $f_0(x) = 1$, $f_1(x) = 1 + x$, $f_2(x) = 1 + x + x^2/2$, etc.). Using our notion of distances in the set X, we might be able to show that the sequence (f_n) tends to f as n tends to ∞. (See Figure 1.7, which shows the functions f_0, f_1, f_2, f_3, f_4, and f. The function f_4 is so "close" to f that you probably can't distinguish them.)

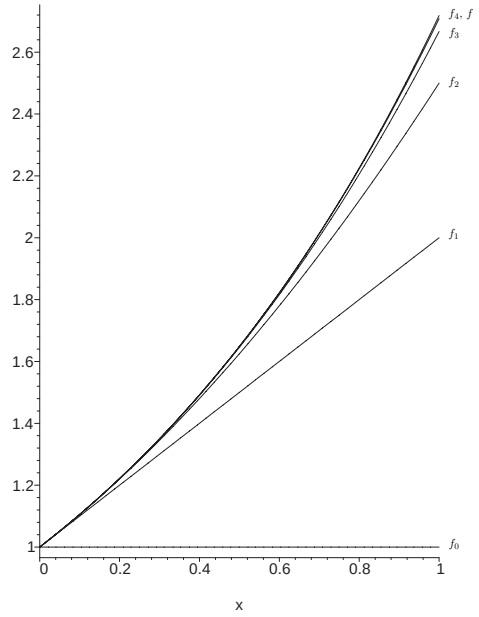

Figure 1.7: The sequence of approximations tends to the function?

Similarly, we could talk about the continuity of functions defined on X, or taking values in X. For example, there is a function $I : X \to \mathbb{R}$ defined by

$$I(f) = \int_0^1 f(x)\,\mathrm{d}x.$$

(Think about this for a moment. A function $X \to \mathbb{R}$ takes as input a continuous function defined on $[0,1]$ (i.e. an element of X), and produces as output a real number. A good way to produce a real number from a function is to integrate the function over its domain of definition.)

We could then ask whether or not this function I is continuous: that is, that if two functions $f, g\colon [0,1] \to \mathbb{R}$ are very close to each other, then their integrals $I(f) = \int_0^1 f(x)\,\mathrm{d}x$ and $I(g) = \int_0^1 g(x)\,\mathrm{d}x$ are also very close to each other (it seems reasonable that this should be true).

In contrast to the situation with shapes in the plane, we'll very soon be in a position to describe two quite different ways of defining the distance between two continuous functions $[0,1] \to \mathbb{R}$.

What's to come

In the next section we'll consider the basic properties that any sensible notion of distance ought to have, and use these to define the concept of a *metric space* which, loosely speaking, is a set where we have a means of measuring the distance between any two elements. After considering several examples of metric spaces, we'll give precise definitions of *convergence* (of a sequence) and *continuity* (of a function), and investigate these ideas in the context of different metric spaces.

The idea of isolating the notion of a metric space is a familiar one in mathematics: instead of studying specific examples (such as shapes in the plane), we study metric spaces in general. Any new concepts that we develop, or theorems that we prove, are then valid across the whole range of metric spaces. We'll see plenty of examples during the module of general results being applied across a wide range of quite different metric spaces.

1.2 Metric Spaces

Our aim is to introduce the definition of a distance, or *metric*, in any set X. We consider the conditions which any sensible notion of distance should satisfy.

We will denote the distance from a point x of X to a point y of X by $d(x,y)$. That is, d is a function

$$d : X \times X \to \mathbb{R}.$$

(See Aside 2 on Page 50 for the meaning of the *product* $X \times X$.)

We will introduce three properties which the distance function d will be required to satisfy. The first two are fairly straightforward.

1. The distance from a point to itself should be zero. The distance from a point to a different point should be greater than zero.

In terms of the distance function d, this reads:

$$\text{For all } x, y \in X, \qquad d(x, x) = 0 \quad \text{and} \quad d(x, y) > 0 \text{ if } x \neq y.$$

2. The distance from any point x to any point y should be the same as the distance from y to x (in other words, we can talk about the distance *between* two points, rather than the distance *from* one *to* the other).

   In terms of the distance function d, this reads:

$$\text{For all } x, y \in X, \qquad d(x, y) = d(y, x).$$

3. The third propery is the one which says that $d(x, y)$ is really a *distance*, rather than any old number. Intuitively, it says that going from x to y can't be further than going from x to a third point z, and then from z to y. In terms of the distance function d, this reads:

$$\text{For all } x, y, z \in X, \qquad d(x, y) \leq d(x, z) + d(z, y).$$

   See Figure 1.8, which illustrates this in the case $X = \mathbb{R}^2$. In this case, it says the length of one side of a triangle (with vertices x, y, and z) can't be greater than the combined length of the other two sides. For this reason, the condition $d(x, y) \leq d(x, z) + d(z, y)$ is known as the *triangle inequality*.

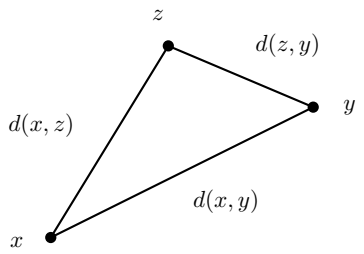

Figure 1.8: The triangle inequality: $d(x, y) \leq d(x, z) + d(z, y)$

Putting all this together, we arrive at the following definition:

Definition 1.1 (Metric Space)
Let X be a set, and $d : X \times X \to \mathbb{R}$ be a function. We say that (X, d) is a *metric space* (or, alternatively, that d *is a metric on* X) if for all $x, y, z \in X$:

1. $d(x, x) = 0$, and $d(x, y) > 0$ if $x \neq y$.

2. $d(x, y) = d(y, x)$.

3. $d(x, y) \leq d(x, z) + d(z, y)$.

Thus when we study metric space theory, what we're really studying is sets X together with functions $d : X \times X \to \mathbb{R}$ which satisfy the above three properties. Of course we have it in mind that $d(x, y)$ represents the *distance* between x and y, but this isn't part of the definition.

Note that it isn't meant to be at all obvious that this definition is the "right" one to use. I suppose that all three conditions are things that a distance should satisfy, but why shouldn't we have added some additional ones? Like many useful mathematical definitions, this one is the result of years of trial and error on the part of many different mathematicians. Finding a good definition involves getting a balance between two things:

a) If there are too many conditions, then not enough different situations fit the definition, and it isn't very useful.

b) If there are too few conditions, then too many different situations fit the definition, and it isn't possible to say much about all of those situations in general.

Definition 1.1 above, it turns out, provides an extremely good balance, and metric space theory, as a result, is very rich.

Examples 1.1 (Metric Spaces)

We're going to give a long list of examples of different metric spaces, and show that each one is indeed a metric space. In doing this, note that Definition 1.1 says that *for all* choices of x, y, z in X, *three* different conditions hold. Thus to show that (X, d) is a metric space, we should start by saying "Let x, y, z be any elements of X", and then go on to show that *each* of conditions 1, 2, and 3 holds. (Typically, some of these conditions will be absolutely obvious, so only the others will need any serious proof.)

We'll continually return to these examples in the remainder of the module to illustrate new concepts as they're introduced.

a) $X = \mathbb{R}^2$, $d_2(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$.

Important note: Here $x = (x_1, x_2)$ and $y = (y_1, y_2)$. In particular, x_2 is the "y-coordinate" of x, and y_1 is the "x-coordinate" of y. The reason for using this notation is that we conventionally refer to the *elements* of

a metric space X using the symbols x, y, z (as we did, for example, in Definition 1.1): when $X = \mathbb{R}^2$, this means that x and y refer to points in the plane, so we can't use the normal (x, y) notation to give their coordinates. This way of doing things may seem confusing at first, but hopefully you'll soon get used to it.

We won't go through a proof that this is indeed a metric in lectures, since this is just the "usual" notion of distance in the plane, which more or less motivated our definition of a metric space. In fact, it takes a surprising amount of work to show that this metric satisfies the triangle inequality.

Note that in this example and the following two, we're focussing on $\mathbb{R}^2$ in order to have something concrete to work with. We can work in $\mathbb{R}^n$ for any n in an exactly analogous manner (see page 14).

b) In fact it's possible to put other metrics on $\mathbb{R}^n$. Here's an example.

$X = \mathbb{R}^2, \quad d_1(x, y) = |x_1 - y_1| + |x_2 - y_2|.$

In other words, instead of all that tedious squaring and square-rooting, we just add the difference in the first coordinates to the difference in the second coordinates. So, for example,

$$\begin{aligned} d_1((-0.3, 1.4), (1, 1.3)) &= |-0.3 - 1| + |1.4 - 1.3| = |-1.3| + |0.1| \\ &= 1.3 + 0.1 = 1.4. \end{aligned}$$

Pictorially, the distance between (x_1, x_2) and (y_1, y_2) is the length of the L-shaped path obtained by going horizontally from (x_1, x_2) to (y_1, x_2), and then vertically from (y_1, x_2) to (y_1, y_2), as depicted in Figure 1.9. †1

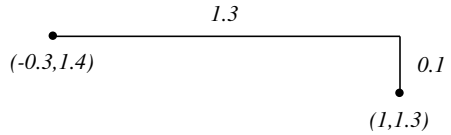

Figure 1.9: Measuring distances with the metric $d_1(x, y) = |x_1 - y_1| + |x_2 - y_2|$

To understand this metric a little better, let's look at all those points in $\mathbb{R}^2$ which are distance 1 from the origin $(0, 0)$, i.e. those points x with $d_1(x, (0, 0)) = 1$. With the usual metric d_2 on $\mathbb{R}^2$, these points would form a circle. With this new metric, we get

$$d_1(x, (0, 0)) = |x_1 - 0| + |x_2 - 0| = |x_1| + |x_2| = 1.$$

What does the set of points (x_1, x_2) satisfying $|x_1| + |x_2| = 1$ look like? If $x_1 > 0$ and $x_2 > 0$, this just says $x_1 + x_2 = 1$ (the equation of a straight line through $(0, 1)$ and $(1, 0)$). If $x_1 > 0$ and $x_2 < 0$, then $|x_2| = -x_2$, and the equation says $x_1 - x_2 = 1$ (the equation of a straight line through $(0, -1)$ and $(1, 0)$). Similar arguments in the other two quadrants ($x_1 < 0$, $x_2 > 0$; and $x_1 < 0$, $x_2 < 0$) produce the picture shown in Figure 1.10.

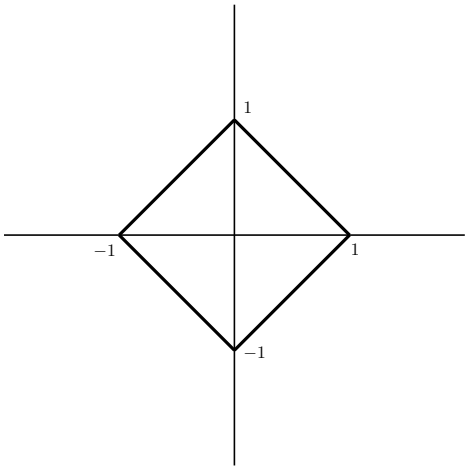

Figure 1.10: Unit circle with metric $d_1(x, y) = |x_1 - y_1| + |x_2 - y_2|$

We need a simple result before moving on to our third example:

Lemma 1.1 *Let a, b, c, and d be any real numbers. Then*

$$\max(a + b, c + d) \leq \max(a, c) + \max(b, d).$$

This result is "obvious" if you think about it... One way to see it is as follows: imagine that two students take a certain module which has both exam and continuously assessed components. Jack gets marks of a in the exam and b in CA (so his total mark is $a + b$), while Jill gets c in the exam and d in CA (so her total mark is $c + d$). Thus the LHS is the higher total mark. On the other hand, the RHS is the higher of the two exam marks plus the higher of the two CA marks, which is clearly at least as big as the higher of the two students' total marks. If that doesn't convince you, here's a proof.

Proof. It's certainly true that $a \leq \max(a, c)$ and $b \leq \max(b, d)$. Adding these gives

$$a + b \leq \max(a, c) + \max(b, d).$$

Similarly $c \leq \max(a, c)$ and $d \leq \max(b, d)$, so

$$c + d \leq \max(a, c) + \max(b, d).$$

Since both $a + b$ and $c + d$ are less than or equal to $\max(a, c) + \max(b, d)$, so is the bigger of $a + b$ and $c + d$: that is,

$$\max(a + b, c + d) \leq \max(a, c) + \max(b, d)$$

as required.

∎

c) Here's another metric we can put on $\mathbb{R}^2$.

$X = \mathbb{R}^2, \quad d_\infty(x, y) = \max(|x_1 - y_1|, |x_2 - y_2|).$

Thus we work out the difference in the x-coordinates and the difference in the y-coordinates, and say that the distance between the two points is whichever of these is bigger. So, for example,

$$\begin{aligned} d_\infty((-0.3, 1.4), (1, 1.3)) &= \max(|-0.3 - 1|, |1.4 - 1.3|) \\ &= \max(|-1.3|, |0.1|) = \max(1.3, 0.1) = 1.3. \end{aligned}$$

(In terms of the L-shaped path of Figure 1.9, the d_∞ distance between x and y is the length of the longer of the two branches of the L.) †2

The set of points in $\mathbb{R}^2$ which are distance 1 from the origin using this metric is depicted in Figure 1.11. (See exercises.)

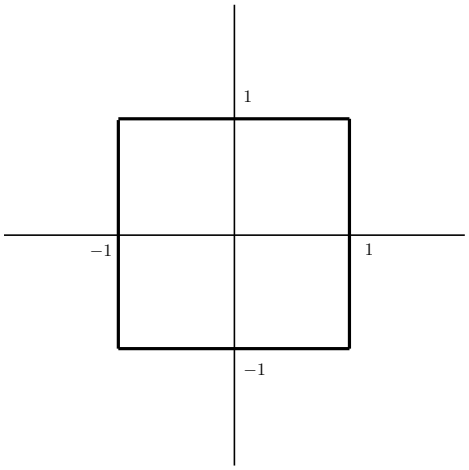

Figure 1.11: Unit circle with metric $d_\infty(x, y) = \max(|x_1 - y_1|, |x_2 - y_2|)$

We'll see shortly that for many purposes (for *topological* purposes), it's irrelevant whether we use the metric of a), b), or c) on $\mathbb{R}^n$ – we can pick whichever one suits us better. We say that the three metrics are *equivalent* (Definition 1.12 on page 43).

To extend these metrics to $\mathbb{R}^n$ we write

$$\begin{aligned} d_2(x,y) &= \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2}, \\ d_1(x,y) &= |x_1-y_1|+|x_2-y_2|+\cdots+|x_n-y_n|, \quad \text{and} \\ d_\infty(x,y) &= \max(|x_1-y_1|,|x_2-y_2|,\ldots,|x_n-y_n|). \end{aligned}$$

Note that when $n=1$ (i.e. when $X=\mathbb{R}$), they're all exactly the same as each other.

We refer to the "usual" metric d_2 on $\mathbb{R}^n$ as the *standard* metric, and often just denote it d.

Where do the symbols d_1, d_2, and d_∞ come from? More generally, for every real number $p \geq 1$, we can define a metric d_p on $\mathbb{R}^n$ by

$$d_p(x,y) = \left(|x_1-y_1|^p+|x_2-y_2|^p+\cdots+|x_n-y_n|^p\right)^{1/p}.$$

The bigger p is, the more "weight" this metric gives to co-ordinates i where $|x_i-y_i|$ is large, until in the limit as $p\to\infty$ all that matters is the maximum difference

$$d_\infty(x,y) = \max_{1\leq i\leq n}(|x_i-y_i|).$$

d) Let X be *any* set, and take

$$d(x,y) = \begin{cases} 0 & \text{if } x=y, \\ 1 & \text{if } x\neq y. \end{cases}$$

Thus any two distinct points are distance 1 apart. This is called the *discrete metric*, since each point of X is separated by a large distance from each of the other points: that is, X looks like a collection of discrete points.

We call a set X with the discrete metric *a discrete space*. †3

There's nothing special about the choice of the number 1 here – we could replace it with any positive number and get an *equivalent* metric.

We could put this metric on $\mathbb{R}^n$ if we wanted, but it's more usual to put it on sets which we think of as being discrete, such as finite sets or $\mathbb{N}$ or $\mathbb{Z}$.

e) The next three examples describe ways of making new metric spaces from old ones. First, the *subspace metric*. This is a straightforward concept.

Suppose (X, d) is a metric space, and Y is any subset of X. Then (Y, d) is also a metric space. (To be accurate, we should write something like $(Y, d|_{Y\times Y})$ here: the distance function on Y is the same as the one on X, except its domain is restricted to $Y \times Y$.)

There's very little to do to prove that (Y, d) is a metric space. For since (X, d) is a metric space, we know that for *all* $x, y, z \in X$, the three conditions in the definition of a metric space hold. So they hold for those particular $x, y, z \in X$ which happen to lie in Y.

Thus, for example, the usual metric on $\mathbb{R}$ gives us a metric d on $\mathbb{Z}$, just by restricting our attention to the world of integers rather than all real numbers. This metric is still defined by $d(m, n) = |m - n|$ (where we use the symbols m and n rather than x and y as a hint to the reader that we're talking about integers rather than any old real numbers). (In fact this metric on $\mathbb{Z}$ is *equivalent* to the discrete metric.) Similarly, there's a metric on the rational numbers $\mathbb{Q}$, and on the interval $[0, 1]$ (and indeed on the interval $[-32, 11.731)$).

f) *Bounded metrics*. We say that a metric d on X is *bounded* (or alternatively that the metric space (X, d) is bounded) if there's some number K such that $d(x, y)$ is never bigger than K. Thus there's a limit to how big the distance between two points can be. For example, the usual metric on $\mathbb{R}$ isn't bounded ($d(x, y)$ can be as big as we like), but the subspace metric on $[-1, 1]$ is bounded, since the distance between two points is never bigger than 2.

Suppose (X, d) is any metric space, and define a new function $e : X \times X \to \mathbb{R}$ by

$$e(x, y) = \min(d(x, y), 1).$$

That is, to work out $e(x, y)$ we work out $d(x, y)$, and replace it by 1 if it's bigger than 1. Then e is also a metric on X, which is bounded by 1. †4

The point is that d and e may give very different distances for points which are far apart, but for close points they are exactly the same. We'll see later the precise significance of this, but for the moment note that the ideas of convergence and continuity are expressed in terms of very

small distances, so to decide whether a sequence converges or a function is continuous we can equally well use either d or e.

There's nothing special about the number 1 in this example: we could equally well have defined $e(x,y) = \min(d(x,y), c)$ for any old number $c > 0$.

g) The *product metric*. Suppose that (X,d) and (Y,e) are both metric spaces. Then we can define a metric D on the product space $X \times Y$ by any of the following formulae:

$$\begin{aligned} D((x_1,y_1),(x_2,y_2)) &= \sqrt{d(x_1,x_2)^2 + e(y_1,y_2)^2}, \\ D((x_1,y_1),(x_2,y_2)) &= d(x_1,x_2) + e(y_1,y_2), \qquad \text{or} \\ D((x_1,y_1),(x_2,y_2)) &= \max(d(x_1,x_2), e(y_1,y_2)). \end{aligned}$$

(We've seen an example of this before: the three metrics on $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ in examples a), b), and c) are of these three forms.) †5

In fact these three metrics on $X \times Y$ are *equivalent* to each other, so for most purposes we can use whichever we find most convenient. We'll use the second metric,

$$D((x_1,y_1),(x_2,y_2)) = d(x_1,x_2) + e(y_1,y_2),$$

as the *standard* metric on a product.

The same construction holds for any finite number of spaces: suppose that $(X_1,d_1),(X_2,d_2),\ldots,(X_n,d_n)$ are all metric spaces, then we can define a metric d on the product space $X_1 \times X_2 \times \cdots \times X_n$ by setting $d((x_1,x_2,\ldots,x_n),(y_1,y_2,\ldots,y_n))$ equal to any of the following:

$$\begin{aligned} &\sqrt{d_1(x_1,y_1)^2 + d_2(x_2,y_2)^2 + \cdots + d_n(x_n,y_n)^2}, \\ &d_1(x_1,y_1) + d_2(x_2,y_2) + \cdots + d_n(x_n,y_n), \qquad \text{or} \\ &\max(d_1(x_1,y_1), d_2(x_2,y_2), \ldots, d_n(x_n,y_n)). \end{aligned}$$

Again, we use the second metric as the *standard* metric on a product of two or more spaces.

h) In this example we define a metric on a set of sequences.

Let $X = \{0,1\}^{\mathbb{N}}$ be the set of all sequences $x = (x_0, x_1, x_2, \ldots)$ of 0s and 1s. Thus an element of X might look like

$$x = (1,1,1,0,0,1,0,0,0,0,1,0,1,1,\ldots)$$

(though of course these few early entries don't tell us which element of X we're talking about): we abbreviate this to $x = 11100100001011\ldots$.

Let's explain the notation $\{0,1\}^{\mathbb{N}}$. In general, if A and B are any sets, then A^B denotes the set of all functions $B \to A$. (Note that if A and B are finite sets, with m and n elements respectively, then A^B has m^n elements, since for each of the n elements of B there's a choice of m elements of A to map it to, giving $m \times m \times \cdots \times m = m^n$ functions in all.)

Thus $\{0,1\}^{\mathbb{N}}$ denotes the set of all possible functions $f : \mathbb{N} \to \{0,1\}$. But such a function *is* really a sequence, since the function can be described exactly by a list of its values: $f(0), f(1), f(2), \ldots$, each of which is either 0 or 1.

The idea of the metric on X is that two sequences $x = (x_0, x_1, x_2, \ldots)$ and $y = (y_0, y_1, y_2, \ldots)$ will be close if they agree for a long time. Here's one way of defining a metric:

$$d(x,y) = \begin{cases} 0 & \text{if } x = y, \\ 1/2^n & \text{if } n \text{ is smallest with } x_n \neq y_n. \end{cases}$$

That is, we look for the first position where x and y differ: if this is position n then the distance between x and y is $1/2^n$. (**Caution:** the start of the sequences is position 0, not position 1.) Thus, for example

$$\begin{aligned} d(110\ldots, 010\ldots) &= 1 \\ d(001\ldots, 010\ldots) &= 1/2 \\ d(010\ldots, 011\ldots) &= 1/4 \\ d(110001010\ldots, 110001011\ldots) &= 1/2^8 = 1/256. \end{aligned}$$

†6

Another (equivalent) metric on X is given by

$$d(x,y) = \sum_{n=0}^{\infty} \frac{|x_n - y_n|}{2^n}.$$

Note that $|x_n - y_n|$ is either zero (if $x_n = y_n$) or one (if $x_n \neq y_n$). Thus this metric is similar to the previous one, but we add contributions of $1/2^n$ from each position where the sequences differ, rather than just considering the first position where they differ. The proof that this is a metric is in the exercises. We'll use the first metric as our *standard* metric on $\{0,1\}^{\mathbb{N}}$.

A similar metric can be defined on the set $Y = \{0,1\}^{\mathbb{Z}}$ of *bi-infinite* sequences of 0s and 1s, which has elements of the form

$$x = (\ldots, x_{-3}, x_{-2}, x_{-1}, x_0, x_1, x_2, x_3, \ldots)$$

(see exercises).

i) For our final example, we shall return to one of the cases considered in the introduction. Let $X = C[0,1]$, the set of all continuous functions $f : [0,1] \to \mathbb{R}$. The idea of the first metric we'll put on X is that two functions f and g should be close precisely when $f(x)$ and $g(x)$ are close to each other for all values of $x \in [0,1]$.

In order to set up the metric, we need a preliminary definition and a result which we won't be able to prove until quite a lot later (Theorem 2.7).

We say that a function $f : [0,1] \to \mathbb{R}$ (not necessarily a continuous one) is *bounded* if there is a number K with the property that $|f(x)| \leq K$ for all $x \in [0,1]$ (equivalently, this says that $-K \leq f(x) \leq K$). We write $B[0,1]$ for the set of all bounded functions $f : [0,1] \to \mathbb{R}$.

Thus, for example, the function $f(x) = x^2$ is bounded on $[0,1]$, since certainly $-1 \leq f(x) \leq 1$ for all $x \in [0,1]$, so we can take $K = 1$. Similarly the function $f(x) = 100\cos(3x) - 50\sin(2x)$ is bounded, since we certainly have $|f(x)| \leq 150$ for all values of x. Here's an example of an unbounded function $f : [0,1] \to \mathbb{R}$:

$$f(x) = \begin{cases} n & \text{if } x = \frac{1}{n} \text{ for some integer } n \geq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Thus $f(1) = 1$, $f(1/2) = 2$, $f(1/3) = 3$, $f(1/4) = 4$, $f(1/100) = 100$, and it's clear that we can make f as large as we like by taking a suitable value of x: hence f isn't bounded.

There are no very straightforward examples of unbounded functions on $[0,1]$, since any such function must be discontinuous. This is the content of the following result, which we'll prove later on:

Theorem from later (2.7) *Every continuous function $f : [0,1] \to \mathbb{R}$ is also bounded. That is, $C[0,1] \subseteq B[0,1]$.*

There's nothing special about the values 0 and 1 here: it's also true that $C[a,b] \subseteq B[a,b]$ for any $a < b$. However, it is vital that the interval is closed. It is easy to find examples of continuous functions $f : (0,1) \to \mathbb{R}$ which are *not* bounded: $f(x) = 1/x$ is one such. (This is one of a number of fundamental differences between closed intervals $[a,b]$ and open intervals (a,b). Later on (Chapter 3) we'll express the distinction by saying that $[a,b]$ is *compact* but (a,b) is not.)

We'll use another result from later on to make the metric on $C[0,1]$ a bit easier to define:

Theorem from later (2.7) *Every continuous function $f : [0,1] \to \mathbb{R}$ attains a maximum. That is, there is some $x \in [0,1]$ with $f(x) \geq f(y)$ for all $y \in [0,1]$.*

Now we're in a position to define a metric on $C[0,1]$: the *L^∞ metric* $d : C[0,1] \times C[0,1] \to \mathbb{R}$ is given by

$$d(f,g) = \max_{x \in [0,1]} |f(x) - g(x)|.$$

That is, the distance between a function f and a function g is the greatest vertical separation between their graphs (see Figure 1.12). Note that if $f(x)$ and $g(x)$ are continuous, then so is $|f(x) - g(x)|$, and so by the result just described $|f(x) - g(x)|$ does have a maximum value in $[0,1]$. †7

To look at it a different way: suppose $f : [0,1] \to \mathbb{R}$ is some given continuous function. Then the functions $g : [0,1] \to \mathbb{R}$ with $d(f,g) < \epsilon$ are precisely those whose graphs lie in an "ϵ-snake" about the graph of f, as shown in Figure 1.13.

A nearly identical formula

$$d(f,g) = \max_{x \in [a,b]} |f(x) - g(x)|.$$

could be used to define a metric on $C[a,b]$ for any $a < b$ – there's nothing special about $[0,1]$. However, it's important that the interval used is a

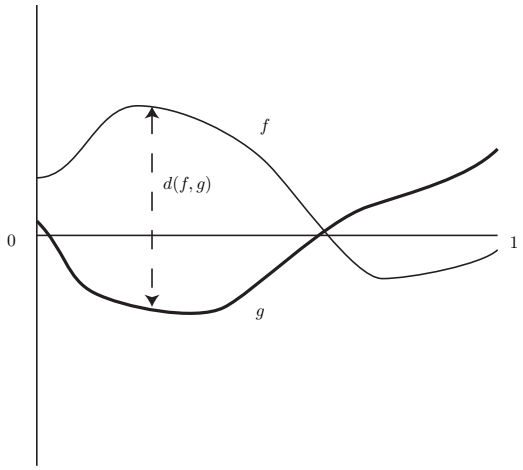

Figure 1.12: The L^∞ metric on $C[0, 1]$

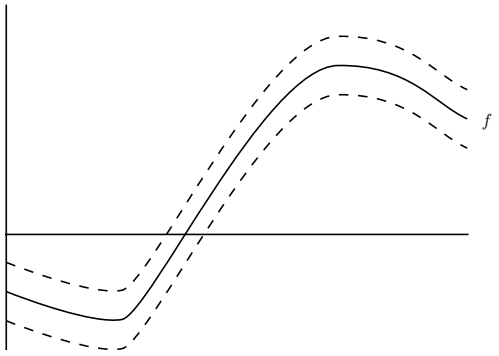

Figure 1.13: Functions distance $< \epsilon$ from f lie in the snake

closed interval. We can't define d in the same way as a function $C(0,1) \times C(0,1) \to \mathbb{R}$, since if $f, g : (0,1) \to \mathbb{R}$ are given by $f(x) = 1/x$ and $g(x) = 0$ (the zero function), then $f(x)$ and $g(x)$ get further and further apart without bound as x gets closer and closer to 0, so "$d(f,g) = \infty$", and ∞ **is not a real number**.

Here's an example of a different metric on $C[0,1]$: let the L^1 *metric* $e : C[0,1] \times C[0,1] \to \mathbb{R}$ be defined by

$$e(f,g) = \int_0^1 |f(x) - g(x)| \, \mathrm{d}x.$$

†8

This metric takes into account the difference between f and g over all of $[0,1]$, not just at the point where they differ most – it can be seen as an *average* difference between the two functions over the interval. This means that it is possible for $e(f,g)$ to be as small as we like while $d(f,g)$ is large. See, for example, the function $f : [0,1] \to \mathbb{R}$ whose graph is depicted in Figure 1.14: it is zero in most of $[0,1]$, and has a narrow tall bump around $x = 1/2$. Let $g(x) = 0$ be the zero function. Then $d(f,g) = 1$ (the functions differ by 1 at $x = 1/2$), but

$$e(f,g) = \int_0^1 |f(x) - g(x)| \, \mathrm{d}x = \int_0^1 f(x) \, \mathrm{d}x$$

is the area under the graph of $f(x)$, which can be as small as we like if we make the bump narrow enough. We'll see later that this means that the metrics d and e on $C[0,1]$ are *not equivalent*. The fact that it is possible for $e(f,g)$ to be as small as we like while $d(f,g)$ remains large will crop up several more times, in different guises, in the remainder of the module.

(We'll use the L^∞ metric on $C[0,1]$ except when we explicitly say otherwise.)

To finish, let's do an explicit calculation of the distance between two functions, $f(x) = x^2$ and $g(x) = x^3$, using each of these two metrics. Note first that $g(x) = xf(x)$, so for $0 \leq x \leq 1$ we have $g(x) \leq f(x)$, and hence $|f(x) - g(x)| = f(x) - g(x) = x^2 - x^3$.

To calculate the L^∞ distance between f and g, we have to find the maximum value of $|f(x) - g(x)| = f(x) - g(x) = x^2 - x^3$ when $x \in [0,1]$.

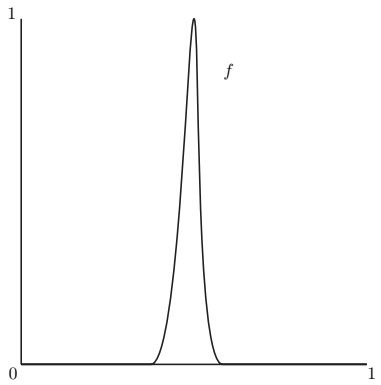

Figure 1.14: A function with a very narrow tall bump

This can be done using differentiation in the usual way. We have

$$\frac{d}{dx}(x^2 - x^3) = 2x - 3x^2,$$

which is zero when $x = 0$ or $x = 2/3$.

To find the greatest value of a function on $[0, 1]$, we have to check the turning points and the endpoints 0 and 1. Now $f(x) - g(x)$ is zero at both $x = 0$ and $x = 1$, and is $4/9 - 8/27 = 4/27$ at $x = 2/3$. So $4/27$ is the maximum value of $f(x) - g(x)$ on $[0, 1]$, and hence $d(f, g) = 4/27$.

Calculating the L^1 distance is quicker. We have

$$\begin{aligned} e(f,g) &= \int_0^1 |f(x) - g(x)| \, \mathrm{d}x \\ &= \int_0^1 (x^2 - x^3) \, \mathrm{d}x \\ &= \left[\frac{x^3}{3} - \frac{x^4}{4}\right]_0^1 \\ &= \frac{1}{3} - \frac{1}{4} = \frac{1}{12}. \end{aligned}$$

Where do the names L^∞ and L^1 come from? More generally, for every real number $p \geq 1$, we can define the L^p metric d_p on $C[0, 1]$ by

$$d_p(f,g) = \left(\int_0^1 |f(x) - g(x)|^p \, \mathrm{d}x\right)^{1/p}.$$

The bigger p is, the more "weight" this metric gives to values of x where $|f(x) - g(x)|$ is large, and in the limit as $p \to \infty$ all that matters is the maximum value of $|f(x) - g(x)|$.

1.3 Isometries

When are two metric spaces (X, d) and (Y, e) "essentially the same"?

Example 1.2 (Silly example)
Suppose we have two metric spaces (X, d) and (Y, e) given as follows:
$X = \{1, 2, 3\}$,
$d(1, 2) = 3$, $d(1, 3) = 4$, and $d(2, 3) = 6$.

$Y = \{\text{cat}, \text{dog}, \text{hen}\}$,
$e(\text{dog}, \text{hen}) = 3$, $e(\text{dog}, \text{cat}) = 4$, and $e(\text{hen}, \text{cat}) = 6$.

(When we define the spaces like this, we're taking it as read that the distance between any element and itself is zero, and that distances are symmetric (so, for example $d(2, 2) = 0$ and $d(3, 2) = 6$.) Thus to ensure that these really are metric spaces, we just have to check that the triangle inequality holds – which it does: but if we'd said $d(2, 3) = 8$ it wouldn't have done, since then we'd have had $d(2, 3) > d(2, 1) + d(1, 3)$.)

The sets X and Y are clearly very different, but when we study metric spaces we're not interested in *what* the elements of the sets are, only in *how far apart* they are from each other. From this point of view, we can see that the two metric spaces above are really the same metric space, just with different names for the elements.

To be explicit, if we make the correspondence $1 \leftrightarrow \text{dog}$, $2 \leftrightarrow \text{hen}$, $3 \leftrightarrow \text{cat}$, then we can see that the distance between any two elements of X is exactly the same as the distance between the two corresponding elements of Y. A correspondence of this sort is called an *isometry*.

In general, if there's a one-to-one correspondence (bijection) between the elements of X and the elements of Y, and the distance between corresponding pairs of elements is the same, then we can look at (Y, e) as being a version of (X, d) where we've just given different names to the elements. This gives the following definition:

Definition 1.2 (Isometry)

An *isometry* between two metric spaces (X, d) and (Y, e) is a bijection $f : X \to Y$ with the property that

$$e(f(x_1), f(x_2)) = d(x_1, x_2)$$

for all $x_1, x_2 \in X$. If such an isometry exists we say that (X, d) and (Y, e) are *isometric*.

(See Aside 3 on Page 51 for details on bijections/invertible functions.)

Examples 1.3 (Isometries)

a) $[0, 1]$ is isometric to $[2, 3]$ by $x \mapsto x + 2$ (and also by $x \mapsto 3 - x$). †9

b) Two spaces with the discrete metric are isometric by any bijection between them (if there is such a bijection). †10

c) $\{0, 1\}^{\mathbb{N}}$ is isometric to itself by a bijection which swaps 0 and 1. †11

d) If (X, d) and (Y, e) are metric spaces, then $X \times Y$ is isometric to $Y \times X$. (See exercises.)

e) $\{f \in C[0, 1] : f(1/2) = 0\}$ and $\{f \in C[0, 1] : f(1/2) = 1\}$ are isometric ($\mathcal{F}(f)(x) = f(x) + 1$). †12

While isometry expresses precisely the idea that two metric spaces are identical as metric spaces, there are times when it's too strong a notion. For example, $[0, 1]$ and $[0, 10]$ aren't isometric, but should we really regard them as being very different? One is just a "rescaled" version of the other, as though we'd chosen to measure distance in millimetres rather than centimetres, for example. Shortly we'll encounter the weaker (and more widely useful) notion of *homeomorphism* (Definition 1.14).

1.4 Convergence and Continuity

In this section we will give precise definitions of the notions of *convergence* of a sequence and *continuity* of a function. Many students find these definitions hard to come to grips with, but they will be central to the module, and so some time spent understanding them properly will be well worth it.

We start with a preliminary definition, which will be important not just here but later also.

Definitions 1.3 (Open and Closed balls)
Let (X, d) be a metric space, x be a point of X, and $r > 0$ be a real number. The *open r-ball $B_r(x)$ about x* (or the *open ball about x of radius r*) is the set of all points whose distance from x is less than r:

$$B_r(x) = \{y \in X \,:\, d(x, y) < r\}.$$

The *closed r-ball $\overline{B}_r(x)$ about x* (or the *closed ball about x of radius r*) is the set of all points whose distance from x is less than or equal to r:

$$\overline{B}_r(x) = \{y \in X \,:\, d(x, y) \leq r\}.$$

(In fact we'll only use open balls in this section, but it makes sense to define the two types of ball together.) Figure 1.15 shows an open ball in $\mathbb{R}^2$ with the standard metric: it consists of all the shaded points, the dotted boundary being intended to indicate that the boundary is not included in the set. A picture of the closed ball $\overline{B}_r(x)$ would be the same, except the boundary would be included and would be drawn with a solid line. (In fact, this isn't a bad picture of an open ball in *any* metric space. Since we can only draw pictures on paper which looks a bit like $\mathbb{R}^2$, we'll often draw pictures of general ideas applicable to any metric space schematically in this way.)

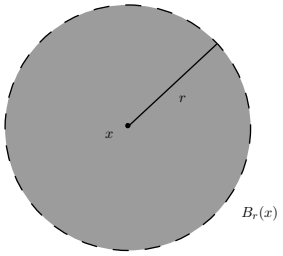

Figure 1.15: An open ball in $\mathbb{R}^2$ (standard metric)

Examples 1.4 (Open and Closed balls)

a) Balls in $\mathbb{R}$. †13

b) Balls in $\mathbb{R}^2$ with non-standard metrics (note similarity to Figures 1.10 and 1.11). †14

c) Balls $B_{1/2}(x)$ and $B_2(x)$ in discrete spaces. †15

d) Balls in $\{0,1\}^{\mathbb{N}}$. †16

e) Balls in $C[0,1]$ (L^∞ metric). †17

1.4.1 Convergence

Let (X,d) be a metric space. A *sequence* in X is an infinite list of elements of X, i.e. $x_0, x_1, x_2, x_3, \ldots$: we often write (x_n) or $(x_n)_{n\geq 0}$ to denote the sequence.

Intuitively, the sequence (x_n) tends to a limit $\ell \in X$ if the points x_n get closer and closer to ℓ as n gets larger and larger (with "closer and closer" measured using the metric d, i.e. $d(x_n,\ell)$ gets smaller and smaller as n gets larger and larger). (Recall Figure 1.2 on page 3 for a depiction of a sequence tending to a limit in $\mathbb{R}^2$.)

What do we mean by "closer and closer"?

Well, it should certainly be true that eventually all the terms of the sequence are within distance 1 of ℓ: or, in other words, in $B_1(\ell)$. By "eventually", we mean that although early terms of the sequence may be further away from ℓ, they are within distance 1 of ℓ from some point on. If that "some point" is the Nth term of the sequence, this means that $x_n \in B_1(\ell)$ *for all* $n \geq N$.

In other words,

There's some N such that $x_n \in B_1(\ell)$ for all $n \geq N$.

See Figure 1.16, which illustrates this for a sequence in $\mathbb{R}^2$. Here we would have $N = 4$, since x_n lies in $B_1(\ell)$ for all $n \geq 4$. (It's also true that x_2 is in $B_1(\ell)$, but since x_3 isn't we can't take $N = 2$.)

Now there's nothing special about the number 1. Eventually, all the terms of the sequence should be within distance $1/2$ of ℓ too. In other words,

There's some N such that $x_n \in B_{1/2}(\ell)$ for all $n \geq N$.

The N in this box will probably be bigger than the N in the previous one, since we have to go further down the sequence to ensure that all of the terms are within distance $1/2$, rather than just distance 1, of ℓ. Figure 1.17

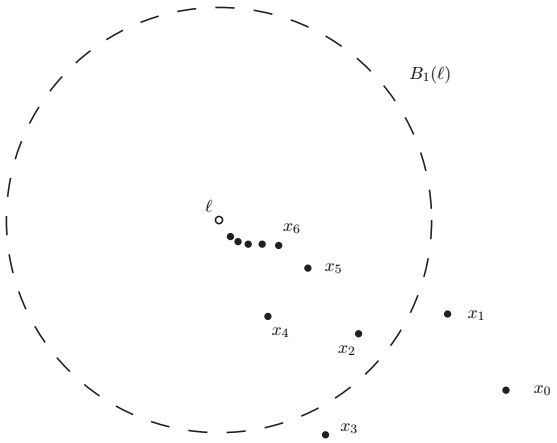

Figure 1.16: From x_4 onwards, the sequence lies in $B_1(\ell)$

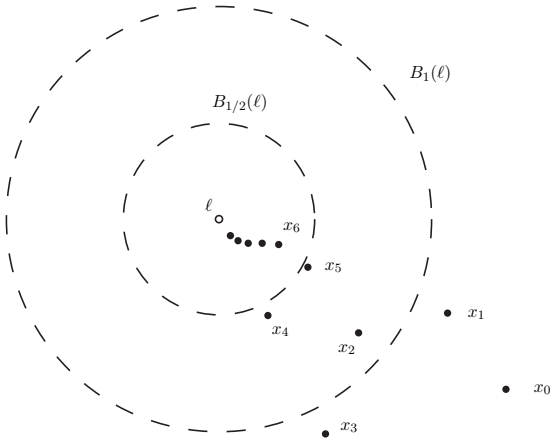

Figure 1.17: From x_6 onwards, the sequence lies in $B_{1/2}(\ell)$

shows that, for our imaginary sequence in $\mathbb{R}^2$, we have to take $N = 6$ to ensure that x_N, x_{N+1}, x_{N+2}, ... all lie in $B_{1/2}(\ell)$.

There's nothing special about 1/2 either. Taking out our magnifying glass, we can see that the sequence must lie in $B_{1/100}(\ell)$ from some x_N onwards (perhaps $N = 1357$), and that if we go even further down it will eventually lie in $B_{1/100000}(\ell)$. In fact, it must eventually lie in $B_\epsilon(\ell)$ for *any* $\epsilon > 0$.

This gives the following definition of convergence. (We describe it as "provisional" not because it's incorrect, but because it'll later be replaced by a new version (Definition 1.9) which says exactly the same, just in a better way.)

Definition 1.4 (Convergence – Provisional definition)
Let (X, d) be a metric space, (x_n) be a sequence in X, and $\ell \in X$. We say that (x_n) *tends to* ℓ *as* n *tends to* ∞ or (x_n) *converges to* ℓ, abbreviated $x_n \to \ell$ as $n \to \infty$ if

$$\boxed{\text{For all } \epsilon > 0, \quad \text{there's some } N \text{ such that } x_n \in B_\epsilon(\ell) \text{ for all } n \geq N.}$$

In your head, you should insert the words *no matter how small* after "For all $\epsilon > 0$". These words don't add anything to the mathematical meaning of the definition, but to a human reader they illustrate its purpose: however tiny ϵ is, the sequence still ends up being within ϵ of ℓ.

The important part of the discussion before the definition is that $N = N(\epsilon)$ *depends on* ϵ: the smaller the value of ϵ, the further down the sequence we have to go before we are trapped inside $B_\epsilon(\ell)$. In the example of Figures 1.16 and 1.17 we had $N(1) = 4$, $N(1/2) = 6$, and $N(1/100) = 1357$.

Since the definition says that something is true *for all* $\epsilon > 0$, the way to show that a given sequence (x_n) tends to a given ℓ is:

1. Let $\epsilon > 0$ be *any* positive number.
2. Show that there is some N such that $x_N, x_{N+1}, x_{N+2}, \ldots$ all lie in $B_\epsilon(\ell)$. This usually involves giving a formula for N in terms of ϵ.

It's worth stating exactly what it means for a sequence (x_n) *not* to tend to ℓ, too. This is exactly saying that there's an open ball about ℓ which the sequence *doesn't* eventually get trapped in. Take a look at Figure 1.18.

Here it seems clear that the sequence (x_n) doesn't converge to ℓ. If you take $\epsilon = 1/2$, then the entire sequence from x_1 onwards lies in $B_{1/2}(\ell)$; but if you take $\epsilon = 1/10$, you can see that although some points of the sequence lie in $B_{1/10}(\ell)$, it isn't true that the whole sequence eventually lies within it.

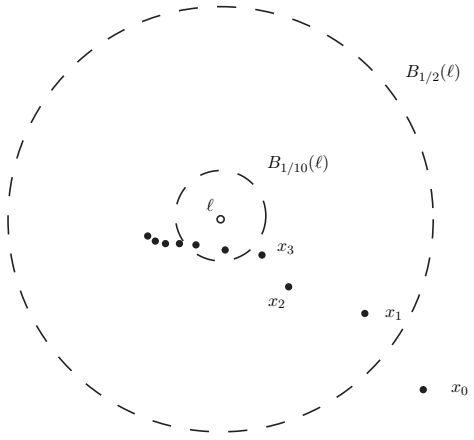

Figure 1.18: (x_n) doesn't converge to ℓ

So the way to show that a given sequence (x_n) doesn't tend to a given ℓ is:

1. Cook up (using your ingenuity) a *particular value* of ϵ (in the example of Figure 1.18 $\epsilon = 1/10$ would do but $\epsilon = 1/2$ wouldn't).

2. Show that, for this particular value of ϵ, you can find values of n as large as you like such that $x_n \notin B_\epsilon(\ell)$ (i.e. there is no N such that $x_N, x_{N+1}, \ldots$ all lie in $B_\epsilon(\ell)$).

Examples 1.5 (Convergence)

a) Convergent and non-convergent sequences in $\mathbb{R}$. †18

b) Convergent and non-convergent sequences in discrete spaces. †19

c) Convergent and non-convergent sequences in $\{0,1\}^{\mathbb{N}}$. †20

d) A sequence in $C[0,1]$ which converges in the L^1 metric but not in the L^∞ metric. †21

The following result says that sequences can have at most one limit: thus, for example, if (x_n) is a sequence in $\mathbb{R}$ which converges to 1, it's impossible

for (x_n) also to converge to 2. This may seem obvious, but if you look carefully at the proof you'll see that it uses each of the properties 1, 2, and 3 in the definition of a metric space (Definition 1.1). That is, if we'd had a weaker definition, even an "obvious" result like this one need not necessarily be true.

Lemma 1.2 (Unique limit) *Let (X, d) be a metric space, and let (x_n) be a sequence in X which converges to $\ell_1 \in X$. If $\ell_2 \in X$ and $\ell_2 \neq \ell_1$, then (x_n) does not converge to ℓ_2.*

The method of proof is by contradiction. That is, we assume that (x_n) is a sequence which *does* converge to each of two different points ℓ_1 and ℓ_2. Starting from this assumption, we argue logically until we arrive at a conclusion which is clearly absurd: a *contradiction*. This tells us that our starting assumption must have been wrong – it isn't possible for a sequence to converge to two different points. †22

If this were a module concentrating on real numbers, the following result would be very important. Since we're dealing with general metric spaces it is much less so, and we shall only prove one of the easier parts of it.

Lemma 1.3 (Operations on sequences in $\mathbb{R}$) *Suppose that (x_n) and (y_n) are sequences in $\mathbb{R}$ which converge to ℓ and m respectively, and let $c \in \mathbb{R}$. Then the sequences (cx_n), $(x_n + y_n)$, $(x_n - y_n)$, and $(x_n y_n)$ converge to $c\ell$, $\ell + m$, $\ell - m$, and ℓm respectively. Moreover, if $y_n \neq 0$ for all n and $m \neq 0$ then the sequence (x_n/y_n) converges to ℓ/m.* †23

The final lemma in this section will be useful later on – it says that convergence of a sequence in a product space is just the same as convergence of the components of the sequence in each of the spaces that the product is made of.

Lemma 1.4 (Convergence in product spaces) *Let (X, d) and (Y, e) be metric spaces, and let (z_n) be a sequence in the product space $X \times Y$. (Thus each term z_n of the sequence is of the form $z_n = (x_n, y_n)$, where $x_n \in X$ and $y_n \in Y$.) Then the following are equivalent:*

a) The sequence (z_n) converges to $z = (x, y) \in X \times Y$.

b) The sequence (x_n) converges to $x \in X$ and the sequence (y_n) converges to $y \in Y$.

This is the first of many results we'll see which state that two (or more) things are *equivalent*. This means that the two things are either both true, or are both false. There are two ways that such results are normally proved. First, we can show that if a) is true then b) is true, and that if b) is true then a) is true; second, we can show that if a) is true then b) is true, and that if a) is false then b) is false. †24

(In fact this lemma easily generalises to products $X_1 \times \cdots \times X_k$ of more than two spaces: the proof is no harder, but the notation is more complicated.)

Example 1.6 (Convergence in product spaces)
The sequence $((\frac{1}{n}, 1 - \frac{1}{n^2}))_{n\geq 1}$ in $\mathbb{R}^2$ converges to $(0, 1)$: this is precisely the same statement as saying that the real sequences $(1/n)_{n\geq 1}$ and $(1-1/n^2)_{n\geq 1}$ converge to 0 and to 1 respectively.

1.4.2 Continuity

Look again at Figure 1.3 on page 4, which shows the graph of a discontinuous function $f : \mathbb{R} \to \mathbb{R}$. We can detect that it's discontinuous because there are values x_1 and x_2, very close to each other, for which $f(x_1)$ and $f(x_2)$ are far apart. Indeed, by pushing x_1 and x_2 closer and closer to the discontinuity, we can make them *as close as we like*, while still having $f(x_1)$ and $f(x_2)$ far apart. This is the basic idea of continuity: a function f is continuous if $f(x_1)$ gets closer and closer to $f(x_2)$ as x_1 gets closer and closer to x_2. Conversely, it is discontinuous if it's possible to choose x_1 and x_2 as close to each other as we like, and still have $f(x_1)$ far from $f(x_2)$.

To turn this into a proper definition, we need to be precise about what we mean when we say "closer and closer" and "as close as we like".

Let (X, d) and (Y, e) be two metric spaces, and let $f : X \to Y$ be a function. To start with we'll just discuss the continuity of f at a particular given point $x_0 \in X$. This enables us to make a more direct parallel with the definition of convergence.

Our notion of convergence was that x_n gets closer and closer to ℓ as n gets bigger and bigger; and the idea of continuity is that $f(x)$ gets closer and closer to $f(x_0)$ as x gets closer and closer to x_0. Let's try to use this similarity to develop a definition of continuity in the same way that we developed one of convergence.

It should certainly be true that $f(x)$ is within distance 1 of $f(x_0)$ (i.e.

$f(x) \in B_1(f(x_0)))$, provided that x is close enough to x_0. "Close enough" means that there is some distance $\delta > 0$ such that any x closer than this to x_0 has $f(x) \in B_1(f(x_0))$. In other words,

There's some $\delta > 0$ such that $f(x) \in B_1(f(x_0))$ provided $x \in B_\delta(x_0)$.

See Figure 1.19, which illustrates this for a made-up function $f : \mathbb{R} \to \mathbb{R}$. Since there's no break in the graph at x_0, there must be a region around x_0 in which f lies between $f(x_0) - 1$ and $f(x_0) + 1$ (i.e. in $B_1(f(x_0))$). In the graph shown, this region is $x_0 - 3.4 < x < x_0 + 1.2$. Hence we can take $\delta = 1.2$ and have that $f(x) \in B_1(f(x_0))$ provided $x \in B_\delta(x_0)$.

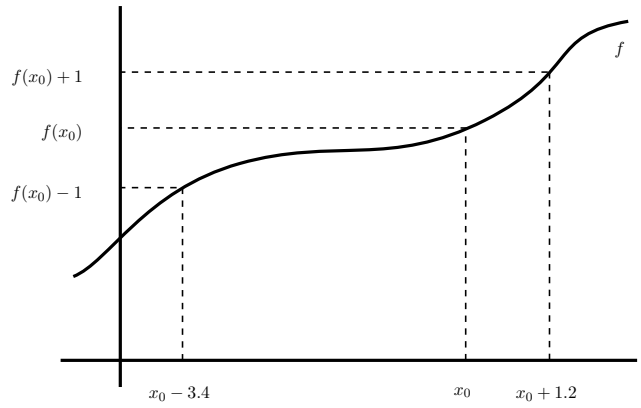

Figure 1.19: Points in $B_{1.2}(x_0)$ map into $B_1(f(x_0))$

Now there's nothing special about the number 1. $f(x)$ should also be within distance $1/2$ of $f(x_0)$ provided that x is close enough to x_0. In other words,

There's some $\delta > 0$ such that $f(x) \in B_{1/2}(f(x_0))$ provided $x \in B_\delta(x_0)$.

The δ in this box will probably be smaller than the δ in the previous one, since x has to be closer to x_0 to ensure that $f(x)$ is within distance $1/2$, rather than just distance 1, of $f(x_0)$. Continuing the example of Figure 1.19, Figure 1.20 suggests that we need to take $\delta = 0.8$ in this case.

There's nothing special about $1/2$ either. Taking out our magnifying glass, we can see that $f(x)$ should be in $B_{1/100}(f(x_0))$ if x is close enough to x_0 (perhaps $\delta = 0.002$), and that if we restrict x to be closer still to x_0, $f(x)$ will be in $B_{1/100000}(f(x_0))$. In fact, $f(x)$ must eventually lie in $B_\epsilon(f(x_0))$ for *any* $\epsilon > 0$.

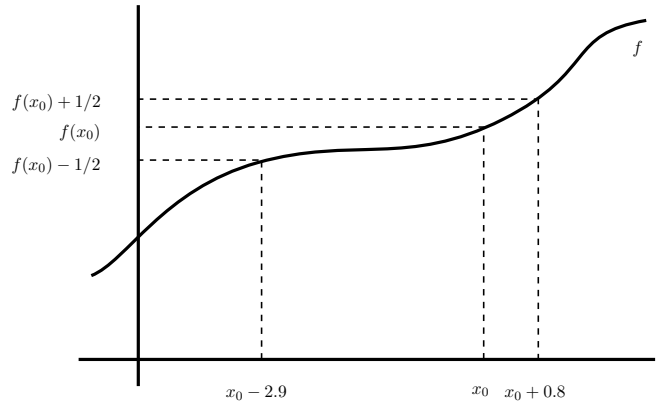

Figure 1.20: Points in $B_{0.8}(x_0)$ map into $B_{1/2}(f(x_0))$

This gives the following definition: f is continuous at x_0 if

For all $\epsilon > 0$ there's some $\delta > 0$ such that $f(x) \in B_\epsilon(f(x_0))$ provided $x \in B_\delta(x_0)$.

This can be abbreviated a bit (though the abbreviation doesn't necessarily make it any clearer). Saying "$f(x) \in A$ provided $x \in B$" is just the same as saying "$f(B) \subseteq A$": both say exactly that if we hit any point of B with f we end up in A. Thus we arrive at:

Definition 1.5 (Continuity at a point x_0)
Let (X, d) and (Y, e) be metric spaces, $f : X \to Y$ be a function, and $x_0 \in X$. Then we say that f *is continuous at* x_0 if

For all $\epsilon > 0$, there exists $\delta > 0$ such that $f(B_\delta(x_0)) \subseteq B_\epsilon(f(x_0))$.

Again, in your head you should read this as "For all $\epsilon > 0$, *no matter how small...*".

Note that the smaller the value of ϵ you choose, the smaller I'll have to choose δ in order to ensure that $f(B_\delta(x_0)) \subseteq B(f_\epsilon(x_0))$. In other words, $\delta = \delta(\epsilon)$ *depends on* ϵ.

Figure 1.21 shows this schematically. The left hand side of the figure represents the space X (where distance is measured using d), and the right hand side represents the space Y (where distance is measured using e). f takes points in X and sends them to points in Y.

Suppose we take $\epsilon = 1/2$. Then, provided f is continuous, we must be able to find some $\delta > 0$ with $f(B_\delta(x_0)) \subseteq B_{1/2}(f(x_0))$. The figure suggests

that $\delta = 0.12$ will do for this (of course these are just made up numbers). If we make ϵ smaller, say $\epsilon = 1/10$, then $\delta = 0.12$ will no longer do, since the figure shows that $f(B_{0.12}(x_0))$ doesn't fit inside $B_{1/10}(f(x_0))$. However, we can take $\delta = 0.05$, since the smaller ball $B_{0.05}(x_0)$ has $f(B_{0.05}(x_0)) \subseteq B_{1/10}(f(x_0))$. As ϵ gets smaller and smaller (i.e. the balls in Y get smaller and smaller), we need the balls in X to get smaller and smaller (i.e. δ to get smaller and smaller) in order that their images under f fit inside the balls in Y.

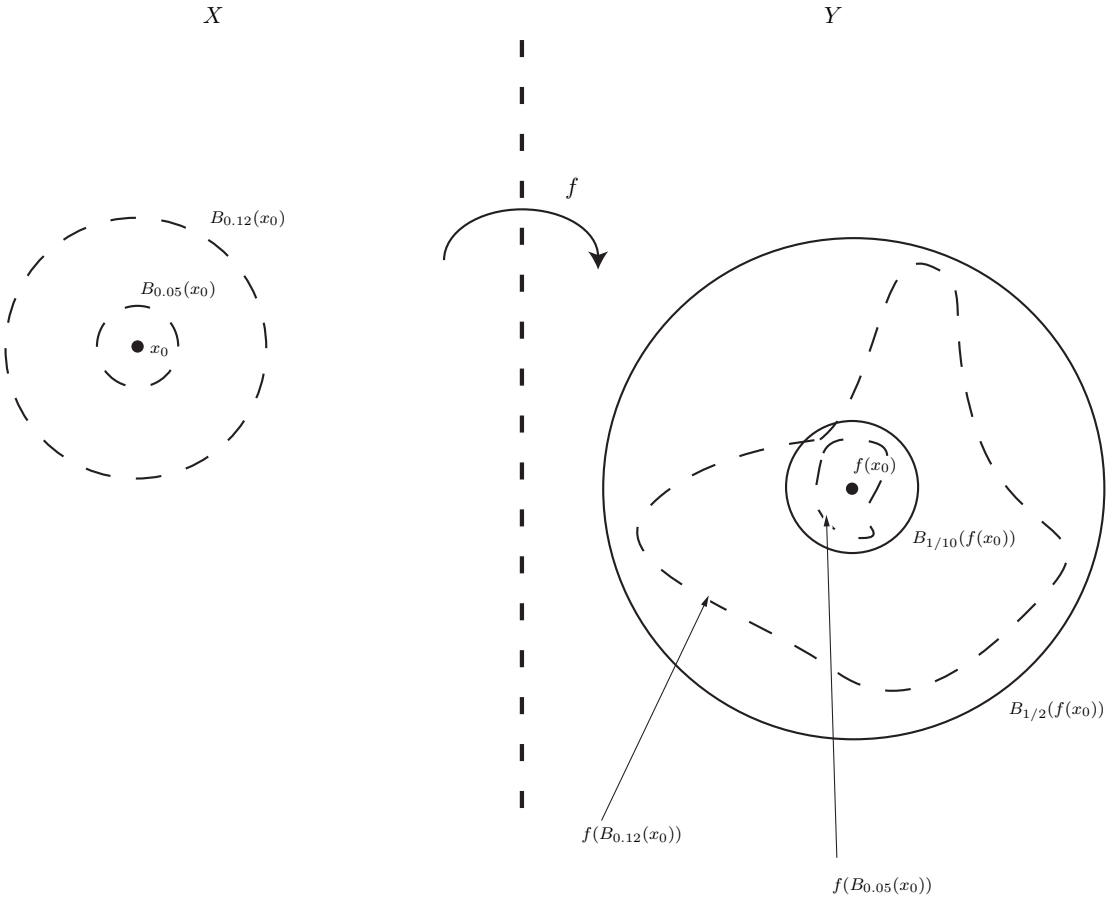

Figure 1.21: As ϵ get smaller, so does δ

A function $f : X \to Y$ is said to be *continuous* if it is continuous at every point of X (for a function $f : \mathbb{R} \to \mathbb{R}$, we only say it's continuous if there are no breaks anywhere in the graph). Once again the following definition is provisional – it will be replaced later by the equivalent Definition 1.11.

Definition 1.6 (Continuity – Provisional definition)
Let (X, d) and (Y, e) be metric spaces, and $f : X \to Y$ be a function. Then we say that f is *continuous* if it is continuous at x_0 for all $x_0 \in X$.

Note that this means that $f : X \to Y$ *isn't* continuous if there's a *single* value x_0 at which it fails to be continuous. For example, the function of Figure 1.3 is not continuous, since there's one value of x at which it fails to be continuous: the fact that it is continuous at all other values of x doesn't change this.

Since the definition of continuity says that something is true *for all* $\epsilon > 0$, the way to show that a given function $f : X \to Y$ is continuous at some $x_0 \in X$ is:

1. Let $\epsilon > 0$ be *any* positive number.
2. Show that there is some δ such that $f(B_\delta(x_0))$ is contained in $B_\epsilon(f(x_0))$. This usually involves giving a formula for δ in terms of ϵ.

It is often notationally simpler to do this without using the notation of open balls. Saying $f(B_\delta(x_0)) \subseteq B_\epsilon(f(x_0))$ is exactly the same as saying that $d(x_0, x) < \delta \implies e(f(x_0), f(x)) < \epsilon$.

It's worth stating exactly what it means for f to be *discontinuous* at x_0 too. This is exactly saying that there's an open ball about $f(x_0)$ which *doesn't* contain $f(B_\delta(x_0))$, no matter how small δ is. So the way to show that a function f *isn't* continuous at x_0 is:

1. Cook up (using your ingenuity) a *particular* value of $\epsilon > 0$.
2. Show that, for this particular value of ϵ, there is **no** value of $\delta > 0$ for which we have $f(B_\delta(x_0)) \subseteq B_\epsilon(f(x_0))$. (A typical way to show this would be to find, for each $\delta > 0$, an element x of $B_\delta(x_0)$ with $f(x) \notin B_\epsilon(f(x_0))$.)

Examples 1.7 (Continuity)

a) Continuity of $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^2$ at $x = 0$, and at general values of x. †25

b) Discontinuity of a step function $f : \mathbb{R} \to \mathbb{R}$. †26

c) Continuity of any function defined on a discrete space. †27

d) Continuity of integration $C[0, 1] \to \mathbb{R}$. †28

Recall that if $f : X \to Y$ and $g : Y \to Z$ are functions, then the *composition* $g \circ f : X \to Z$ is defined by $g \circ f(x) = g(f(x))$, i.e. first apply f and then apply g to the result. The next result says that that if we compose two continuous functions, we get a continuous result.

Lemma 1.5 (Continuity of Composition) *Let (X, d_1), (Y, d_2), and (Z, d_3) be metric spaces, and $f : X \to Y$ and $g : Y \to Z$ be continuous functions. Then $g \circ f : X \to Z$ is continuous.* †29

Another result which will be very useful to us tells us how continuity and convergence interact:

Lemma 1.6 (Continuity and Convergence) *Let (X, d) and (Y, e) be metric spaces, $f : X \to Y$ be a function, and $x^* \in X$. Then the following are equivalent:*

a) f is continuous at x^*.

b) For every sequence (x_n) in X with $x_n \to x^*$, we have $f(x_n) \to f(x^*)$.

†30

If this were a module concentrating on real numbers, the following result would be very important. Since we're dealing with general metric spaces it is much less so, and we shall only prove one of the easier parts of it.

Lemma 1.7 (Operations on continuous functions $\mathbb{R} \to \mathbb{R}$) *Suppose that $f, g : \mathbb{R} \to \mathbb{R}$ are continuous functions, and let $c \in \mathbb{R}$. Then the functions cf, $f + g$, $f - g$, and fg are also continuous. Moreover, if g has no zeros, then f/g is continuous.*

†31

1.5 Open and Closed Sets

For reasons which will soon become clear, the notion of open and closed sets will be fundamental in this module. Those of you who've done MATH241 or MATH243 will have come across this idea before.

First of all, consider those sets about which we already use the terms "open" and "closed". An open interval (a, b) is one which doesn't contain its endpoints a and b, while a closed interval $[a, b]$ is one which does contain its endpoints. Similarly, an open ball $B_r(x)$ doesn't contain any of its boundary, whereas a closed ball $\overline{B}_r(x)$ contains all of its boundary.

In general, a subset A of X will be called open in X if it contains none of its boundary, and will be called closed in X if it contains all of its boundary. This is a good intuitive way to think about open and closed sets. In fact, it's possible to define the boundary of A in such a way that this is a *definition* of open and closed. However, this approach isn't very convenient in practice, and we use alternative definitions.

To motivate the definition of *open*, suppose that a subset A of X doesn't contain any of its boundary points. That is, if we pick any point a of A, it isn't on the boundary of A. So there's room in A to squeeze in a little open ball centred on a. (The closer a is to the boundary of A, the smaller this ball will need to be.)

Definition 1.7 (Open subset)
Let (X, d) be a metric space, and A be a subset of X. We say that A is an *open* subset of X if

$$\text{For every } a \in A \text{ there is some } \epsilon > 0 \text{ with } B_\epsilon(a) \subseteq A.$$

On the other hand, if A contains all of its boundary, then if we pick any point x of X which *isn't* in A, then it isn't on the boundary of A. So there's room to squeeze a little ball around x which doesn't meet A. This is exactly saying that the complement $X \setminus A$ of A is open in X.

Definition 1.8 (Closed subset)
Let (X, d) be a metric space, and A be a subset of X. We say that A is a *closed* subset of X if $X \setminus A$ is an open subset of X. That is,

$$\text{For every } x \in X \setminus A \text{ there is some } \epsilon > 0 \text{ with } B_\epsilon(x) \subseteq X \setminus A.$$

Notice that the notions of *open* and *closed* are dual to each other. If we know what all the open subsets of X are, then we also know what all the closed subsets are (just the complements of the open subsets), and vice versa.

Important Remarks

a) Definitions 1.7 and 1.8 involve the metric space X which A is a subset of (obviously in the case of Definition 1.8, and less obviously in the case of Definition 1.7, since the set $B_\epsilon(a)$ depends on what X is). Thus **it doesn't make sense** to talk about a set A being open or closed without specifying the universal set X. See Examples 1.8 c).

b) A door is either open or closed, but a subset of X can be *neither* open nor closed; or it can be *both* open and closed. (Several of Examples 1.8 show this.)

Examples 1.8 (Open and Closed subsets)

a) Let $X = \mathbb{R}^2$. †32

i) $B_1(0)$ is open in X.

ii) $\overline{B}_1(0)$ is closed in X.

b) Let $X = \mathbb{R}$. †33

i) (a, b), (a, ∞), and $(-\infty, a)$ are all open in X.

ii) $[a, b]$, $[a, \infty)$, and $(-\infty, a]$ are all closed in X.

iii) $[a, b)$ and $(a, b]$ are neither open nor closed in X.

iv) $\mathbb{Z}$ is closed in X.

v) $\mathbb{Q}$ is neither open nor closed in X.

c) **Caution.** Whether or not A is open/closed depends not just on A, but also on the set X which A is a subset of. †34

i) Let $X = \mathbb{R}$ and $A = [0, 1)$. Then A is neither open nor closed in X.

ii) Let $X = [0, \infty)$ (with the subspace metric) and $A = [0, 1)$. Then A is open in X.

When working in $\mathbb{R}^n$, it's often easy to understand intuitively whether a subset is open or closed by thinking about its boundary. When we work in other metric spaces, it's necessary to apply the definitions more carefully.

d) Let (X, d) be any metric space. Then $\emptyset$ is both open and closed in X. Similarly X is both open and closed in X. †35

e) Let $X = \{0,1\}^{\mathbb{N}}$.

Given a *finite* sequence $s_0 s_1 \dots s_n$, write $C_{s_0 s_1 \dots s_n} \subseteq X$ for the *cylinder set*

$$C_{s_0 s_1 \dots s_n} = \{x \in X \,:\, x_0 = s_0, x_1 = s_1, \dots, x_n = s_n\}$$

(i.e. the set of all sequences which start $s_0 s_1 \dots s_n$).

Any cylinder set is both open and closed in X. †36

f) Let $X = C[0,1]$ (with the L^∞ metric), and let

$$A = \{f \in C[0,1] \,:\, f(1/2) > 0\}.$$

Then A is open in X. †37

g) Let $X = C[a,b]$, let $c < d$ be any real numbers, and let

$$A = \{f \in C[a,b] \,:\, c \leq f(x) \leq d \text{ for all } x \in [a,b]\}$$

(so A is the set of continuous functions $[a,b] \to [c,d]$.)

Then A is closed in X. †38

Examples 1.8 a) is a special case of the following more general result:

Lemma 1.8 (Open/Closed balls are open/closed) *Let (X,d) be a metric space, $x_0 \in X$, and $r > 0$. Then the open ball $B_r(x_0)$ is open in X and the closed ball $\overline{B}_r(x_0)$ is closed in X.*

†39

The following result describes some of the basic properties of open sets:

Lemma 1.9 (Properties of open subsets) *Let (X,d) be a metric space. Then*

a) Both $\emptyset$ and X are open in X.

b) Any union of open subsets of X is open in X.

c) Any finite intersection of open subsets of X is open in X.

It's important to be clear about the distinction between "any union" in b), and "any *finite* intersection" in c). b) says that if we have any collection A_j of open subsets of X (where the j runs over any index set), then their union (i.e. the set of all points of X which lie in *some* A_j) is also open in X. c) says that if $A_1, \dots, A_n$ are open subsets of X, then their intersection (i.e. the set of all points of X which lie in *every* A_j) is also open in X.

†40

Example 1.9

To illustrate the distinction, consider the *infinite* family of open subsets of $\mathbb{R}$ given by

$$A_j = \left(-\frac{1}{j}, \frac{1}{j}\right) \qquad (j = 1, 2, 3, \ldots).$$

Thus $A_1 = (-1, 1)$, $A_2 = (-1/2, 1/2)$, $A_3 = (-1/3, 1/3)$, $A_4 = (-1/4, 1/4)$, and so on. The union of all these sets is $(-1, 1)$, which is open in $\mathbb{R}$. However their intersection is $\{0\}$, which is not open in $\mathbb{R}$: this doesn't contradict Lemma 1.9 since we're intersecting an *infinite* number of sets.

The analogue for closed subsets of Lemma 1.9 is:

Lemma 1.10 (Properties of closed subsets) *Let (X, d) be a metric space. Then*

a) Both $\emptyset$ and X are closed in X.

b) Any intersection of closed subsets of X is closed in X.

c) Any finite union of closed subsets of X is closed in X.

†41

Once again, you need to appreciate the difference between "any intersection" and "any *finite* union". An example illustrating this is in the exercises.

The final result we consider in this section will be extremely useful in the remainder of the module. To understand it, suppose that A is a subset of X, and (a_n) is a sequence in X *all of whose points lie in A*. Suppose that $a_n \to \ell$ as $n \to \infty$. The fact that $a_n \to \ell$ means that ℓ is as close as we like to points of A. Thus if ℓ isn't actually in A, it must lie on its boundary. If A happens to be closed, then it contains its boundary, and hence ℓ *must* lie in A.

Lemma 1.11 *Let (X, d) be a metric space, and A be a subset of X. Then the following are equivalent:*

a) A is closed in X.

b) If (a_n) is any convergent sequence in X with $a_n \in A$ for all n then its limit lies in A.

†42

Example 1.10
As a simple example showing why the limit need not lie in A if A isn't closed, let $X = \mathbb{R}$ and $A = (0, 2)$. Consider the sequence $a_n = 1/n$. Then certainly (a_n) is a convergent sequence in $\mathbb{R}$, and $a_n \in A$ for all n: however its limit is 0, which doesn't lie in A.

1.6 Reformulation of Convergence and Continuity

In this section we give alternative (equivalent) definitions of the convergence of a sequence, and the continuity of a function: these reformulations are phrased entirely in terms of open sets, without explicit mention of the metric. We will shortly see why this is a worthwhile thing to do.

Convergence is much the easier of the two:

Theorem 1.12 *Let (X, d) be a metric space, (x_n) be a sequence in X, and $\ell \in X$. Then the following are equivalent:*

a) (x_n) converges to ℓ.

b) For every open subset U of X containing ℓ, there exists N such that for all $n \geq N$ we have $x_n \in U$.

†43

Thus we can use b) as a definition of convergence, replacing our original Definition 1.4. The new definition means exactly the same as (is equivalent to) the old one. The advantage in using it will soon become clear.

Definition 1.9 (Convergence)
Let (X, d) be a metric space, (x_n) be a sequence in X, and $\ell \in X$. We say that (x_n) *tends to ℓ as n tends to ∞* or (x_n) *converges to ℓ*, abbreviated $x_n \to \ell$ as $n \to \infty$ if

For all open subsets U of X containing ℓ,

there exists N such that $x_n \in U$ for all $n \geq N$.

Before reformulating the definition of continuity, we need to introduce some notation. You're familiar with the notation f^{-1} for the inverse of a function $f : X \to Y$, which need not necessarily exist (for example, if $f : \mathbb{R} \to \mathbb{R}$ is given by $f(x) = x^2$). We now extend the notation to a function f^{-1} taking *subsets* of Y to *subsets* of X: this function always exists (makes sense).

Definition 1.10 (The set function f^{-1})
Let X and Y be sets, and $f : X \to Y$ be a function. We write f^{-1} for the function which maps each subset U of Y to the subset

$$f^{-1}(U) = \{x \in X \,:\, f(x) \in U\}$$

of X.

That is, $f^{-1}(U)$ consists of all the points which f sends into U.

Example 1.11 (The set function f^{-1})
Let $f : \mathbb{R} \to \mathbb{R}$ be given by $f(x) = x^2$. Then

a) $f^{-1}(\{4\}) = \{-2, 2\}$.
For $x = -2$ and $x = 2$ are exactly the points with $x^2 = 4$.

b) $f^{-1}([1, 9]) = [-3, -1] \cup [1, 3]$.
For the points x with $1 \leq x^2 \leq 9$ are exactly those between 1 and 3, and those between -3 and -1.

c) $f^{-1}([-2, 1] = [-1, 1]$.
For the points x with $-2 \leq x^2 \leq 1$ are the same as those with $0 \leq x^2 \leq 1$, i.e. those between -1 and 1.

d) $f^{-1}([-2, -1] = \emptyset$.
For there are no points x with $-2 \leq x^2 \leq -1$.

Now we can reformulate the definition of continuity:

Theorem 1.13 *Let (X, d) and (Y, e) be metric spaces, and $f : X \to Y$ be a function. Then the following are equivalent:*

a) f is continuous.

b) For every open subset U of Y, $f^{-1}(U)$ is an open subset of X. †44

Thus we can use b) as a definition of continuity, replacing our original Definition 1.6. The two definitions are equivalent to each other.

Definition 1.11 (Continuity)
Let (X, d) and (Y, e) be metric spaces, and $f : X \to Y$ be a function. Then we say that f is *continuous* if

For every open subset U of Y, $f^{-1}(U)$ is an open subset of X.

Example 1.12 (A discontinuous function)
To illustrate the new definition, let's consider a function $f : \mathbb{R} \to \mathbb{R}$ which is patently discontinuous: the "step" function

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

To show that this fails to satisfy Definition 1.11, we need to find an open subset U of $\mathbb{R}$ for which $f^{-1}(U)$ is not open in $\mathbb{R}$. To do this, just take $U = (1/2, 3/2)$, which is open in $\mathbb{R}$. Now

$$f^{-1}(U) = \{x \in \mathbb{R} : 1/2 < f(x) < 3/2)\} = \{x \in \mathbb{R} : f(x) = 1\} = [0, \infty)$$

which is not open in $\mathbb{R}$.

1.7 Topology and topological concepts

The reformulations of the notions of convergence and continuity given by Theorem 1.12 b) and Theorem 1.13 b) are entirely in terms of open sets: they don't explicitly make use of the particular metrics on the spaces concerned. The fact that it is possible to write the definitions in this way means that **if two metrics d and e on X define the same open sets in X, then they are indistinguishable for the purposes of convergence and continuity**.

Definition 1.12 (Equivalent metrics)
Let X be a set, and d and e be metrics on X. We say that d and e are *equivalent* if the open subsets of X determined using d are exactly the same as the open subsets of X determined using e.

The following result gives a method of deciding whether or not two metrics are equivalent. When we need to distinguish between two metrics d and e on X, we write $B_r^d(x)$ and $B_r^e(x)$ for the open r-balls about x calculated using d and e respectively.

Theorem 1.14 (Test for equivalence of metrics) *Let X be a set, and d and e be metrics on X. Then the following are equivalent:*

a) d and e are equivalent.

b) For every $x \in X$ and every $\epsilon > 0$ there's some $\delta > 0$ such that

$$B_\delta^d(x) \subseteq B_\epsilon^e(x) \quad \text{and} \quad B_\delta^e(x) \subseteq B_\epsilon^d(x).$$

That is: there's no open e-ball so small you can't fit a little d-ball inside it, and no open d-ball so small you can't fit a little e-ball inside it. †45

Examples 1.13 (Equivalent metrics)

a) The three metrics on $\mathbb{R}^2$ given in Examples 1.1 a), b), and c) are equivalent to each other. †46

b) Let (X, d) be any metric space, and let e be the bounded metric on X given by

$$e(x, y) = \min(d(x, y), 1)$$

(see Examples 1.1 f)). Then d and e are equivalent. (So we can replace any metric with an equivalent bounded metric.) †47

c) The L^∞ and L^1 metrics on $C[0, 1]$ are *not* equivalent. †48

Wherever possible, we'll define concepts exclusively in terms of open sets. This has the advantage that we know that the concepts don't change their meaning when we replace one metric with another equivalent one (for example, with a bounded metric).

We can develop this idea further by introducing the notion of topology. The *topology* of a metric space is precisely the collection of its open sets: thus *equivalent metrics are ones which define the same topology.* We can generalise the notion of metric spaces to *topological spaces* where we simply specify the open sets, without giving a metric from which they're derived, or even assuming that such a metric exists. Here's the definition:

Definition 1.13 (Topological Space)
A *topological space* is a set X together with a collection of subsets of X (which we call "open sets"), satisfying the following properties:

a) $\emptyset$ and X are open.

b) Any union of open sets is open.

c) Any finite intersection of open subsets is open.

We also say that a collection of subsets of X satisfying these properties *defines a topology* on X.

Notice that Lemma 1.9 says precisely that the open sets in a metric space (X, d) define a topology on X. However, there are many (and important) examples of topological spaces where the open sets aren't given by any metric. That is, topological spaces are genuinely *more general* than metric spaces.

Example 1.14 (Indiscrete topology)
This example is *not* an important one, but is a straightforward one which shows that there are topological spaces where the open sets aren't given by any metric.

Let X be any set with at least two elements, and define a topology on X by saying that the only open sets in X are $\emptyset$ and X. This is a topology, since a) is clearly satisfied, and b) and c) follow from the fact that if we take unions and intersections of $\emptyset$ and X, the only results we can get are again $\emptyset$ and X.

There is no metric on X which generates this topology. †49

Any concept (such as convergence or continuity) which can be defined entirely in terms of open sets makes sense for any topological space, and is called a *topological concept*. (We have to be a bit careful about what we mean by "defined entirely in terms of open sets". For example, we can make use of closed sets in our definitions (since closed sets are just the complements of open sets), and of any topological notion we've already defined (e.g. continuity and convergence). What we can't use is the metric $d(x, y)$ itself, or concepts like $B_r(x)$ which can change their meaning when we replace our metric with an equivalent one.)

A topological space is a set together with a collection of subsets designated as open. Suppose X and Y are both topological spaces, and that

there's a bijection (invertible map) $f : X \to Y$ which carries the open sets of X precisely onto the open sets of Y. Then X and Y are essentially the same topological space: we've just renamed each point x of X as $f(x)$ in Y. In this case we say that X and Y are *homeomorphic*, and the map f is called a *homeomorphism*. (So homeomorphisms preserve all the topological structure: they play the same role as isomorphisms do in group theory, for example.)

There's another way to say that f carries the open subsets of X precisely onto the open subsets of Y. It can be unpacked into the following two statements:

a) For each open subset U of X, $f(U)$ is an open subset of Y.

b) Each open subset V of Y is f of an open subset of X: that is, $f^{-1}(V)$ is an open subset of X.

But (referring to Theorem 1.13) these say precisely that: a) $f^{-1} : Y \to X$ is continuous; and b) $f : X \to Y$ is continuous.

Definition 1.14 (Homeomorphism)
Let (X, d) and (Y, e) be metric spaces. A bijection $f : X \to Y$ is a *homeomorphism* if both f and f^{-1} are continuous. If such a homeomorphism exists, we say that X and Y are *homeomorphic*.

Examples 1.15 (Homeomorphisms)

a) $[0, 1]$ and $[-1, 1]$ are homeomorphic. †50

b) $(0, 1)$ and $\mathbb{R}$ are homeomorphic. †51

c) $\{f \in C[0, 1] : 0 \leq f(x) \leq 1\}$ and $\{f \in C[0, 1] : 0 \leq f(x) \leq 2\}$ are homeomorphic. †52

Homeomorphisms are our promised generalisation of isometries. Note that two homeomorphic metric spaces need not be isometric (e.g. $[0, 1]$ and $[-1, 1]$).

Let's finish this chapter with an example of a non-topological concept *(this final part of the chapter will probably be omitted)*.

Definition 1.15 (Totally bounded)
We say that a metric space (X, d) is *totally bounded* if for all $\epsilon > 0$, there are a finite number of points $x_1, x_2, \ldots, x_n$ of X such that every point x of X has $d(x, x_i) < \epsilon$ for some i.

That is, for any tiny ϵ you propose, I can find *finitely many* points of X which come within distance ϵ of every point of X. Of course the smaller you choose ϵ to be, the more points of X I'm likely to need.

Examples 1.16 (Totally bounded)

a) $(0, 1)$ is totally bounded. †53

b) $\mathbb{R}$ is not totally bounded. †54

c) A discrete space is totally bounded if and only if it is finite. †55

(Example c) shows that total boundedness is not the same as boundedness – any discrete space is bounded.)

To see that being totally bounded is not a topological notion, note that $(0, 1)$ and $\mathbb{R}$ are homeomorphic (topologically identical) to each other (Examples 1.15 b)), but that $(0, 1)$ is totally bounded and $\mathbb{R}$ is not.

Aside 1 (Function notation)

The function notation

$$f : X \to Y$$

will be used extensively in this module. If you're not quite sure about it, now's the time to get to grips with it.

When we write $f : X \to Y$, we mean that f is a function *from* the set X *to* the set Y. That is, for every element x of X, there is an associated element of Y which is denoted $f(x)$. It may be helpful to regard f as some sort of machine which is given as *input* an element x of X, and produces as *output* an element $f(x)$ of Y.

We can describe $f(x)$ in any way we like, but the function **must** give some output for **every** value of $x \in X$, and this output **must** be a **single** element of Y.

The set X is called the *domain* of the function f, and the set Y is called its *range*.

Examples 1.17 (Function Notation)

a) $f : \mathbb{R} \to \mathbb{R}$ denotes a "normal" real-valued function, which takes a real number x as input and produces a real number $y = f(x)$ as output. We can describe such a function by a formula such as

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

or by some other means. For example, we could define a function $g : \mathbb{R} \to \mathbb{R}$ by

$$g(x) \text{ is the smallest integer greater than or equal to } x.$$

(In this example, we'd have $g(1.3) = 2$, $g(2) = 2$, $g(2.71) = 3$, $g(\pi) = 4$. Note that for every possible input $x \in \mathbb{R}$, there is a single output $g(x)$ which we have specified exactly.)

b) $h : \{0, 1, 2\} \to \mathbb{Z}$ denotes a function which associates an integer $h(x)$ to each of $x = 0$, $x = 1$, and $x = 2$. We could describe the function by a formula such as

$$h(x) = x^3 - 4x + 3,$$

or by listing the values which it takes explicitly:

$$h(0) = 3, \quad h(1) = 0, \quad h(2) = 3.$$

(This is the same function h as the one given by the formula, but we could have defined a function by choosing *any* three integers for $h(0)$, $h(1)$, and $h(2)$.)

c) Note that there is no requirement for $f : X \to Y$ to take every possible value in Y. For example, the function $g : \mathbb{R} \to \mathbb{R}$ above only takes integer values: if y is a non-integer, then there is no $x \in \mathbb{R}$ with $g(x) = y$. Similarly the function $h : \{0, 1, 2\} \to \mathbb{Z}$ only takes the values 0 and 3.

If it happens that $f : X \to Y$ *does* take every possible value in Y, then we say that f is *surjective* (see Aside 3 below).

d) Nor is there any requirement that different inputs give different outputs. For example, the function $g : \mathbb{R} \to \mathbb{R}$ above has $g(1.3) = g(2) = 2$. Similarly, the function $h : \{0, 1, 2\} \to \mathbb{Z}$ has $h(0) = h(2) = 3$.

If it happens that $f : X \to Y$ *does* always give different outputs for different inputs, then we say that f is *injective* (see Aside 3 below).

Two functions f and g are equal if they have the same domain X, the same range Y, and $f(x) = g(x)$ for every possible input $x \in X$. Note in particular that this means that the function $g : \mathbb{R} \to \mathbb{R}$ defined above is *not* equal to the function $k : \mathbb{R} \to \mathbb{Z}$ given by

$$k(x) \text{ is the smallest integer greater than or equal to } x.$$

Although $g(x) = k(x)$ for every value $x \in \mathbb{R}$, the functions have different ranges and so are not equal.

Aside 2 (Cartesian Products)

If X and Y are sets, then $X \times Y$ denotes the set consisting of all pairs (x, y), where x is an element of X and y is an element of Y. It is called the *Cartesian product* (or just the *product*) of X and Y.

Examples 1.18 (Cartesian Products)

a) $\mathbb{R} \times \mathbb{R}$ is the set consisting of all pairs (x, y), where both x and y are real numbers. Thus it is the set which we are used to denoting $\mathbb{R}^2$.

Notice that when we take the product of a set with itself like this, the *order* of the elements in the pair matters. That is, $(1, 1.5)$ is *not* the same element of $\mathbb{R} \times \mathbb{R}$ as $(1.5, 1)$.

b) $\{1, 2\} \times \{2, 3, 4\}$ has 6 elements:

$$(1, 2), \quad (1, 3), \quad (1, 4), \quad (2, 2), \quad (2, 3), \quad \text{and} \quad (2, 4).$$

In general, if X and Y are finite sets with m and n elements respectively, then $X \times Y$ has mn elements, since there is a choice of m first entries in the pair, and n second entries.

c) We can extend the notation to more than two sets: for example, $X \times Y \times Z$ denotes the set of all triples (x, y, z), where $x \in X$, $y \in Y$, and $z \in Z$. Thus $\mathbb{R} \times \mathbb{R} \times \mathbb{R}$ is the set which we are accustomed to denoting $\mathbb{R}^3$: it consists of all triples (x, y, z) where $x, y, z \in \mathbb{R}$.

Aside 3 (Bijections (Invertible functions))

Let $f : X \to Y$ be a function. In general, f need not take every possible value in Y. If it does, then we say that it is *surjective* or *a surjection*.

Examples 1.19 (Surjections)

a) The function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$ is not surjective. For $-1 \in \mathbb{R}$, and there is no $x \in \mathbb{R}$ with $x^2 = -1$.

b) The function $g : \mathbb{R} \to [0, \infty)$ defined by $g(x) = x^2$ is surjective. For given any $y \in [0, \infty)$, we have $g(\sqrt{y}) = y$. (Note that f and g are *not* the same function: see Aside 1 above.)

c) The function $h : \mathbb{Z} \to \mathbb{Z}$ defined by $h(n) = 2n$ is not surjective. For $1 \in \mathbb{Z}$, and there is no $n \in \mathbb{Z}$ with $2n = 1$.

d) The function $k : \mathbb{Z} \to \mathbb{Z}$ defined by $k(n) = n + 3$ is surjective. For given any $m \in \mathbb{Z}$ we have $k(m - 3) = m$.

In general, f need not give different outputs for different inputs. If it does (that is, if $f(x_1) \neq f(x_2)$ whenever $x_1 \neq x_2$), then we say that it is *injective* or *an injection*.

Examples 1.20 (Injections)

a) The function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$ is not injective. For $f(-1) = f(1) = 1$.

b) The function $\ell : [0, \infty) \to [0, \infty)$ defined by $\ell(x) = x^2$ is injective. For if $0 \leq x_1 < x_2$, then $\ell(x_1) < \ell(x_2)$.

c) The function $h : \mathbb{Z} \to \mathbb{Z}$ defined by $h(n) = 2n$ is injective. For if $n_1 \neq n_2$ then $2n_1 \neq 2n_2$.

d) The function $k : \mathbb{Z} \to \mathbb{Z}$ defined by $k(n) = n + 3$ is injective. For if $n_1 \neq n_2$ then $n_1 + 3 \neq n_2 + 3$.

If $f : X \to Y$ is both surjective and injective, then we say that it is *bijective* or *a bijection*. Thus of the functions considered in the examples above, only k and ℓ are bijections (they are both surjective and injective).

Putting together the definitions of surjective and injective, a function $f : X \to Y$ is bijective if

every $y \in Y$ is equal to $f(x)$ for **exactly** *one $x \in X$.*

(The fact that it is surjective means that $y = f(x)$ for *at least one* $x \in X$, and the fact that it is injective means that $y = f(x)$ for *at most one* $x \in X$.)

Bijections are precisely those function which have inverses: that is, $f : X \to Y$ is a bijection if and only if there is a function $f^{-1} : Y \to X$ with the property that $f^{-1}(f(x)) = x$ for all $x \in X$, and $f(f^{-1}(y)) = y$ for all $y \in Y$ (i.e. f^{-1} is f "in reverse"). In fact, if $f : X \to Y$ is a bijection, then we can define $f^{-1} : Y \to X$ by

$$f^{-1}(y) = \text{ the unique } x \in X \text{ with } f(x) = y.$$