University of Sheffield

School of Mathematics and Statistics

Metric Spaces

MAS331/6352

2018–19

1 Metric spaces

This course is entitled "Metric spaces". You may be wondering what such a thing should be. We'll see a formal definition a little later in the course, but here's a rough definition: it is a space in which we can discuss convergence and do analysis, just as was done for the real numbers $\mathbb{R}$ in MAS221 *Analysis*[1]. Sequences can be of many types: we can consider sequences of numbers, sequences of points in a plane, sequences of functions and so on. Whenever we talk about a sequence approaching a limit, then implicitly we are assuming that we have some notion of *distance*, so that we can talk about two things being *close* or about the terms in a sequence getting *close* to a limit. Roughly speaking, a metric space is a set (of numbers, points, functions etc.) on which we have a way to measure distance.

To define notions such as convergence, we need some idea of what it means for points to be 'close' to one another. It is therefore necessary to consider spaces with some sort of 'distance' on them. This leads quickly to the definition of a metric space.

We will isolate three fundamental properties of 'distance', and base all our deductions on these three properties alone. With so few properties, this makes our proofs rather simple, and also more general as they will apply to any situation where we have a notion of distance.

The three fundamental properties are

- The distance between any two points is non-negative and the distance between two points is zero precisely when the two points are the same.
- The distance from one point to a second is the same as the distance back from the second point to the first.
- The distance from one point to a second is at most the distance from one point to a third and from the third to the second, i.e., the distance from the first to the second *via* the third.

Hopefully you agree that these properties are reasonably sensible ones to adopt for a notion of distance.

[1] The equivalent module for Nanjing students is MAS207 *Continuity and Integration*.

1.1 Metrics.

We will need to handle pairs of elements of a set X. The *Cartesian product* $X \times X$ consists of all ordered pairs (x, y) where $x \in X$ and $y \in X$. We measure distance using a function $d\colon X \times X \to \mathbb{R}$. Given $x, y \in X$, such a function determines a real number $d(x, y)$.

Definition 1.1. A *metric space* consists of a non-empty set X, together with a *distance function*, or *metric*, $d\colon X \times X \to \mathbb{R}$ satisfying the following axioms:

M1: For all $x, y \in X$, $d(x, y) \geqslant 0$ and $d(x, y) = 0$ if and only if $x = y$.

M2: For all $x, y \in X$, $d(x, y) = d(y, x)$.

M3: For all $x, y, z \in X$, $d(x, z) \leqslant d(x, y) + d(y, z)$ [the *triangle inequality*].

Example 1.2. There is an obvious way to measure distance in $\mathbb{R}$. For $x, y \in \mathbb{R}$, let $d(x, y) = |x - y|$. We next check that this is indeed a metric.

Lemma 1.3. *The function $d(x, y) = |x - y|$ is a metric on $\mathbb{R}$.*

Proof. We must carefully check that the three axioms M1, M2 and M3 hold. Let's begin with M1. Let $x, y \in \mathbb{R}$. Remember that $|x - y|$ is defined to be $x - y$ or $y - x$, whichever is non-negative. Thus $d(x, y) \geqslant 0$. If $x = y$ then $d(x, x) = |x - x| = |0| = 0$, and if $d(x, y) = 0$ then either $x - y = 0$ or $y - x = 0$, and so $x = y$. This proves axiom M1.

Axiom M2 is easy: for all $x, y \in X$,

$$d(x, y) = |x - y| = |y - x| = d(y, x).$$

Now we shall prove axiom M3. Let $x, y, z \in \mathbb{R}$. We want to show that $|x - z| \leqslant |x - y| + |y - z|$. By the definition of absolute value, we have

$$\begin{aligned} -|x - y| \leqslant x - y \leqslant |x - y|, \\ -|y - z| \leqslant y - z \leqslant |y - z|. \end{aligned}$$

Adding these, we get

$$-(|x-y|+|y-z|) \leqslant x-z \leqslant |x-y|+|y-z|,$$

from which it follows that $|x-z| \leqslant |x-y|+|y-z|$. This proves M3. $\square$

Now consider k-dimensional Euclidean space $\mathbb{R}^k$ for $k \geq 1$. We shall introduce three metrics on $\mathbb{R}^k$. The first one should be familiar, at least for $k=2,3$. We first define the metrics and later check that they are indeed metrics.

Example 1.4. The *Euclidean metric* d_2 on $\mathbb{R}^k$ is given by the rule

$$d_2(x,y) = d_2((a_1,\ldots,a_k),(b_1,\ldots,b_k)) = \sqrt{(a_1-b_1)^2+\cdots+(a_k-b_k)^2}.$$

Usually, when we think of $\mathbb{R}^k$ as a metric space, we will use this Euclidean metric, unless otherwise stated.

When $k=2$, the distance from (a_1,a_2) to (b_1,b_2) is

$$d_2((a_1,a_2),(b_1,b_2)) = \sqrt{(a_1-b_1)^2+(a_2-b_2)^2}.$$

Thus $(\mathbb{R}^2, d_2)$ is just the plane with its usual notion of distance.

If we identify each point $(a,b) \in \mathbb{R}^2$ with the complex number $z = a+ib$ then we get a metric on $\mathbb{C}$ with

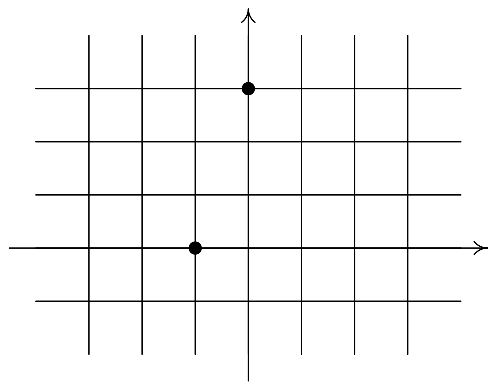
$$d(w,z) = |w-z| = \sqrt{(a_1-a_2)^2+(b_1-b_2)^2}$$

where $w = a_1+ib_1$, $z = a_2+ib_2$.

Example 1.5. The *taxicab metric* d_1 on $\mathbb{R}^k$ is given by

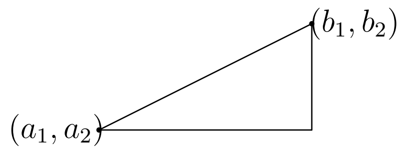
$$d_1(x,y) = d_1((a_1,\ldots,a_k),(b_1,\ldots,b_k)) = |a_1-b_1|+\cdots+|a_k-b_k|.$$

When $k=2$ this represents the distance between two points if we can only move along the lines of a square grid (think of a city with a grid street plan).



Example 1.6. The *maximum metric* or *supremum metric* d_∞ on $\mathbb{R}^k$ is given by the rule

$$d_\infty(x, y) = d_\infty((a_1, \ldots, a_k), (b_1, \ldots, b_k)) = \max\{|a_1 - b_1|, \ldots, |a_k - b_k|\}.$$



In this diagram, $d_1((a_1, a_2), (b_1, b_2))$ is the sum of the lengths of the vertical and horizontal sides of the triangle and $d_\infty((a_1, a_2), (b_1, b_2))$ is the maximum of these two lengths whereas $d_2((a_1, a_2), (b_1, b_2))$ is the length of the hypotenuse.

In general, (see Problem 1.3(a))

$$d_\infty((a_1, a_2), (b_1, b_2)) \leq d_2((a_1, a_2), (b_1, b_2)) \leq d_1((a_1, a_2), (b_1, b_2)).$$

For example

$$d_1((-1, 0), (0, 3)) = 4, \quad d_2((-1, 0), (0, 3)) = \sqrt{10}, \quad d_\infty((-1, 0), (0, 3)) = 3.$$

Remark 1.7. When $k = 1$, $d_1(x, y) = d_2(x, y) = d_\infty(x, y) = |x - y|$ for all $x, y \in \mathbb{R}$.

Although we've said "Euclidean metric", "taxicab metric", and "maximum metric", we haven't actually *proved* that any of these are metrics (except when $k = 1$). We are now going to prove that d_∞ and d_2

are metrics. The proof that d_1 is a metric is on the Problem Sheet (Problem 1.4).

Proposition 1.8. *d_∞ is a metric on $\mathbb{R}^k$.*

Proof. We need to check the three axioms for $(\mathbb{R}^k, d_\infty)$ to be a metric space. Let

$$x = (a_1, \ldots, a_k),\ y = (b_1, \ldots, b_k),\ z = (c_1, \ldots, c_k) \in \mathbb{R}^k.$$

For axiom M1, $d_\infty(x, y) = \max\{|a_1 - b_1|, \ldots, |a_k - b_k|\} \geq 0$. Also $d_\infty(x, y) = 0$ if and only if $|a_i - b_i| = 0$ for all i if and only if $x = y$. Thus M1 holds.

Axiom M2 is clear because $|a_i - b_i| = |b_i - a_i|$ for each i, so that

$$\begin{aligned} d_\infty(x, y) &= \max\{|a_1 - b_1|, \ldots, |a_k - b_k|\} \\ &= \max\{|b_1 - a_1|, \ldots, |b_k - a_k|\} \\ &= d_\infty(y, x). \end{aligned}$$

For axiom M3, note that, by the proof of Lemma 1.3, $|a_i - c_i| \leqslant |a_i - b_i| + |b_i - c_i|$ for each i, so

$$\begin{aligned} d_\infty(x, z) &= \max\{|a_i - c_i|\} \\ &\leqslant \max\{|a_i - b_i| + |b_i - c_i|\} \\ &\leqslant \max\{|a_i - b_i|\} + \max\{|b_i - c_i|\} \\ &= d_\infty(x, y) + d_\infty(y, z), \end{aligned}$$

which proves M3. □

Before showing the Euclidean metric d_2 is indeed a metric we need the following lemma.

Lemma 1.9 (Cauchy–Schwarz inequality)**.** *For $e_1, \ldots, e_k, f_1, \ldots, f_k \in \mathbb{R}$ we have*

$$|e_1 f_1 + e_2 f_2 + \cdots + e_k f_k| \leqslant \sqrt{e_1^2 + \cdots + e_k^2}\sqrt{f_1^2 + \cdots + f_k^2}.$$

Proof. If $e_1 = \cdots = e_n = 0$ the result is obvious. So let's assume that this is not the case, and consider the quadratic polynomial

$$P(x) = (e_1 x + f_1)^2 + (e_2 x + f_2)^2 + \cdots + (e_k x + f_k)^2 = Ax^2 + 2Bx + C,$$

where

$$\begin{aligned} A &= e_1^2 + \cdots + e_k^2, \\ B &= e_1 f_1 + \cdots + e_k f_k, \\ C &= f_1^2 + \cdots + f_k^2. \end{aligned}$$

Since $P(x) \geqslant 0$ for any $x \in \mathbb{R}$ we must have $4B^2 - 4AC \leqslant 0$, which rearranges to $|B| \leqslant \sqrt{AC}$, which is exactly the inequality we want.

$\square$

Theorem 1.10. *The Euclidean metric d_2 is a metric on $\mathbb{R}^k$.*

Proof. We must check that d_2 satisfies the three axioms for $(\mathbb{R}^k, d_2)$ to be a metric space. Let $x = (a_1, \ldots, a_k)$, $y = (b_1, \ldots, b_k)$, $z = (c_1, \ldots, c_k) \in \mathbb{R}^k$. For M1, $d_2(x,y) = (\sum (a_i - b_i)^2)^{\frac{1}{2}} \geqslant 0$. Also $d_2(x,y) = 0 \Leftrightarrow (a_i - b_i)^2 = 0$ for each $i \Leftrightarrow a_i = b_i$ for each $i \Leftrightarrow x = y$. This proves axiom M1.

Axiom M2 is clear because $(a_i - b_i)^2 = (b_i - a_i)^2$ for all i, so that

$$\begin{aligned} d_2(x,y) &= \left(\sum (a_i - b_i)^2\right)^{\frac{1}{2}} \\ &= \left(\sum (b_i - a_i)^2\right)^{\frac{1}{2}} \\ &= d_2(y,x). \end{aligned}$$

Finally let's prove axiom M3:

$$\begin{aligned} &(d_2(x,y) + d_2(y,z))^2 \\ =&d_2(x,y)^2 + d_2(y,z)^2 + 2d_2(x,y)d_2(y,z) \\ =&\sum (a_i - b_i)^2 + \sum (b_i - c_i)^2 + 2\left(\sum (a_i - b_i)^2\right)^{\frac{1}{2}} \left(\sum (b_i - c_i)^2\right)^{\frac{1}{2}} \\ \geqslant&\sum (a_i - b_i)^2 + \sum (b_i - c_i)^2 + 2\sum (a_i - b_i)(b_i - c_i) \\ &\quad \text{(by Cauchy-Schwarz with } e_i = a_i - b_i \text{ and } f_i = b_i - c_i) \\ =&\sum ((a_i - b_i) + (b_i - c_i))^2 \\ =&\sum (a_i - c_i)^2 \\ =&d_2(x,z)^2, \end{aligned}$$

so that, taking square roots, $d_2(x,z) \leqslant d_2(x,y)+d_2(y,z)$ as required. $\square$

Remark 1.11. The three metrics d_1, d_2 and d_∞ on $\mathbb{R}^k$ are actually part of an infinite family of metrics, but these are the only ones you are ever likely to use in practice. The general definition is the following: If $p \in (1,\infty)$, there is a metric d_p on $\mathbb{R}^n$ defined by

$$d_p(x,y) = d_p((a_1,\ldots,a_k),(b_1,\ldots,b_k)) = \left(\sum_{i=1}^{k} |a_i - b_i|^p\right)^{\frac{1}{p}}.$$

- When $n = 1$, all these metrics are equal to the usual metric on $\mathbb{R}$.
- When $p = 1$, d_p is just the taxicab metric d_1.
- When $p = 2$, d_p is the Euclidean metric d_2.
- When $p \to \infty$, $d_p(x,y) \to d_\infty(x,y)$.

Try to convince yourself about the last point. You might like to use your calculator to compute $1+2$, $\sqrt{1^2+2^2}$, $\sqrt[3]{1^3+2^3}$, $\ldots$, and see what happens.

Example 1.12 (The discrete metric)**.** Let X be a non-empty set, and define the *discrete metric* d_0 on X by:

$$d_0(x,y) = \begin{cases} 0 & \text{if } x = y, \\ 1 & \text{if } x \neq y. \end{cases}$$

See Problem 8 for M1,2,3. This may seem silly but it is the first case of a useful metric in communications and is significant if all we need to know about two objects is whether they are the same. Fix a positive integer n and let X be the set of all finite sequences in which each term is 0 or 1. E.g. with $n = 7$ we might have 0111001 and 0101010. For two sequences x, y let $d(x,y)$ be the number of positions in which x and y differ. E.g. $d(0111001, 0101010) = 3$. This is a metric called the *Hamming distance*. When $n = 1$ it is the same as the discrete metric. We could do it with $a, b, c, \ldots, z$ in place of $0, 1$, for example, with four-letter words, $d(\text{love}, \text{hate}) = 3$.

Historical Background. The concept of a metric space seems to have first been introduced by the French mathematician Maurice Fréchet (1878-1973) in his PhD thesis that was submitted in 1906. The German mathematician Felix Hausdorff (1868-1942) first used the "metric space" terminology. You can use Google and Wikipedia to learn more about these founders of the subject.

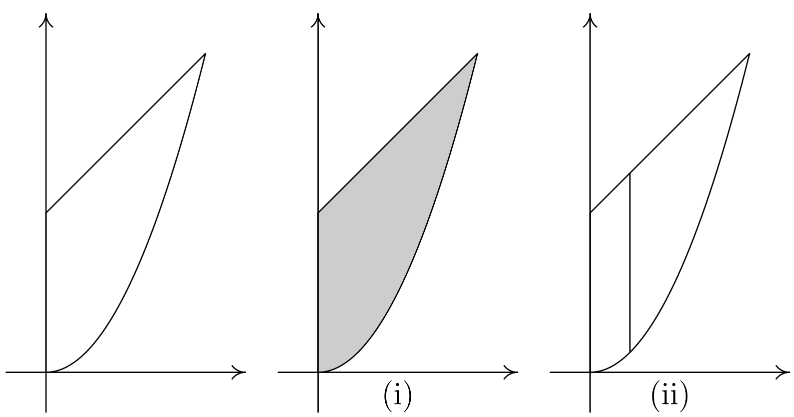
1.2 Function spaces.

Just as a real number can arise as the limit of a sequence or solution of an equation and an element of $\mathbb{R}^n$ can arise as the limit of a sequence or solution of a (simultaneous) equation, a function can arise as the limit of a sequence of functions or as a solution of a differential or integral equation. The theory of metric spaces covers these three situations. To apply it to functions, we need to be able to define a meaningful *distance* between two functions. The theory works best with continuous functions on closed bounded intervals.

Definition 1.13. Let $I = [a, b]$ be a closed bounded interval in $\mathbb{R}$, such as $[0, 1]$. The *space of continuous functions on* I, denoted $C(I)$, is

$$C(I) = \{f\colon I \to \mathbb{R} : f \text{ is continuous}\}.$$

How do we measure the distance between two functions such as $f(x) = x^2$ and $g(x) = x + 2$ on $[0, 2]$?



(i) (ii)

Comparing two points $(a_1, \ldots, a_n)$ and $(b_1, \ldots, b_n)$ in $\mathbb{R}^n$ using d_1, we take account of n non-negative numbers $|a_i - b_i|$ and take their sum. For functions, there are infinitely many points $|f(x) - g(x)|$ to take into account and we replace summation by integration. In other words we measure how far apart the functions are by taking the area between their graphs. (Diagram (i))

For d_∞ on $\mathbb{R}^n$, we took the maximum of the numbers $|a_i - b_i|$ and we can do the same for functions, taking the maximum value of $|f(x) - g(x)|$ as x ranges through the interval I. In other words we measure how far apart the functions are by taking the maximum distance between their graphs. (Diagram (ii))

Definition 1.14. Let $I = [a, b]$ be a closed and bounded interval.

(i) The metric d_1 on $C[a, b]$ is defined by

$$d_1(f, g) = \int_a^b |f(x) - g(x)| dx.$$

This exists because

- as f and g are continuous, so are $f(x) - g(x)$ and $|f(x) - g(x)|$;
- from MAS221, every continuous function on a closed and bounded interval has a Riemann integral.

(ii) The *supremum metric* d_∞ on $C[a, b]$ is defined by

$$d_\infty(f, g) = \sup\{|f(x) - g(x)| = \max\{|f(x) - g(x)| : x \in I\}.$$

This works because

- as above;
- from the boundedness theorem (MAS221), every continuous function on a closed and bounded interval is bounded and attains a maximum value.

Example 1.15. Let $I = [0, 2]$. If $f(x) = x^2$, $g(x) = x + 2$, find (i) $d_1(f, g)$ and (ii) $d_\infty(f, g)$.

Solution Note that throughout the interval $[0, 2]$, $g(x) \geq f(x)$ so $|f(x) - g(x)| = g(x) - f(x) = x + 2 - x^2$.

(i)

$$d_1(f,g) = \int_0^2 (x+2-x^2)\,dx = \left[\frac{x^2}{2} + 2x - \frac{x^3}{3}\right]_0^2 = 3\frac{1}{3}.$$

(ii) When $f(x)$ and $g(x)$ are differentiable, as in this case, $|f(x) - g(x)|$ must take its maximum value on $[a,b]$ at a or at b or at a stationary point of $f-g$. $|f(x)-g(x)| = x+2-x^2$ has derivative $1-2x$ so it has a stationary point at $x=\frac{1}{2}$. The values of $|f(x) - g(x)|$ at 0, $\frac{1}{2}$ and 2 are 2, $2\frac{1}{4}$ and 0 so the maximum value $2\frac{1}{4}$, is at the stationary point $\frac{1}{2}$. Thus $d_\infty(f,g) = 2\frac{1}{4}$.

Remark 1.16. The fact that I was closed and bounded in Definition 1.14 is very important. To see this, try to define d_∞ and d_1 on $C((0,1])$. Let $f(x) = 1/x$ and $g(x) = 0$ and try to work out $d_\infty(f,g)$ and $d_1(f,g)$:

$$\begin{aligned} d_\infty(f,g) &= \sup\{|\tfrac{1}{x} - 0| \,:\, x \in (0,1]\} = \sup\{\tfrac{1}{x} \,:\, x \in (0,1]\}, \\ d_1(f,g) &= \int_0^1 \frac{1}{x} dx = [\ln(x)]_0^1 . \end{aligned}$$

Neither of these makes any sense!

Also you might think that, in d_∞, an alternative distance to use would be where the graphs are closest together : $d(f,g) = \min(|f(x) - g(x)|$. But if the graphs intersect, as they do at 0 and 2 in the example, this would give $d(f,g) = 0$, contradicting axiom M1 unless $f = g$.

Proposition 1.17. *Let $I = [a,b]$ be a closed and bounded interval. Then d_∞ is a metric on $C(I)$.*

Proof. We must verify the three axioms for $(C(I), d_\infty)$ to be a metric space.

Let $f,g,h \in C(I)$, so that f,g and h are continuous functions from I to $\mathbb{R}$. For M1, $d_\infty(f,g) = \sup\{|f(x) - g(x)| \,:\, x \in I\} \geqslant 0$. Also, $d_\infty(f,g) = 0 \Leftrightarrow |f(x) - g(x)| = 0$ for all $x \Leftrightarrow f(x) = g(x)$ for all $x \Leftrightarrow f = g$. This proves M1.

M2 is immediate because $|f(x) - g(x)| = |g(x) - f(x)|$ for all x, so that

$$\begin{aligned} d_\infty(f,g) &= \sup\{|f(x)-g(x)| : x \in I\} \\ &= \sup\{|g(x)-f(x)| : x \in I\} \\ &= d_\infty(g,f). \end{aligned}$$

For M3, note that, by M3 in $\mathbb{R}$,

$$|f(x)-h(x)| \leqslant |f(x)-g(x)| + |g(x)-h(x)|$$

for all $x \in I$. Now we have

$$\begin{aligned} d_\infty(f,h) &= \sup\{|f(x)-h(x)| : x \in I\} \\ &\leqslant \sup\{|f(x)-g(x)| + |g(x)-h(x)| : x \in I\} \\ &\leqslant \sup\{|f(x)-g(x)| : x \in I\} + \sup\{|g(x)-h(x)| : x \in I\} \\ &= d_\infty(f,g) + d_\infty(g,h), \end{aligned}$$

as required. □

Theorem 1.18. *Let $I = [a,b]$ be a closed and bounded interval. Then d_1 is a metric on $C(I)$.*

The odd thing about this theorem, in comparison with all of the similar results we've seen so far, is that M1 is the hardest part to prove while M2 and M3 are by now very routine. In fact, we're only going to prove M1; the rest is an exercise.

Exercise 1.19. Prove axioms M2 and M3 for d_1 on $C(I)$.

Proof that axiom M1 holds for d_1. Let $f, g \in C(I)$, so that $f, g\colon I \to \mathbb{R}$ are continuous functions. Then

$$d_1(f,g) = \int_a^b |f(x)-g(x)|dx \geqslant 0$$

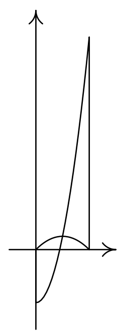
and $d_1(f,g) = 0$ when $f = g$. Now suppose that $f \neq g$. We have to show that $d_1(f,g) > 0$.

There exists $c \in (a,b)$ for which $f(c) \neq g(c)$, otherwise $f(x)-g(x) = 0$ for $a < c < b$ and, by continuity of $f-g$, $f(a)-g(a) = 0 = f(b)-g(b)$. Let $A = |f(c)-g(c)|$. Then, since $|f(x)-g(x)|$ is continuous, we can find $\delta > 0$ such that $|f(x)-g(x)| > A/2$ whenever $|x-c| < \delta$. (Diagram on board in lecture.) Then

$$d_1(f,g) = \int_a^b |f(x)-g(x)|dx \geqslant \int_{c-\delta}^{c+\delta} |f(x)-g(x)|dx \geqslant \int_{c-\delta}^{c+\delta} A/2dx = \delta A > 0.$$

□

Example 1.20. In $C[0,1]$, let $f(x) = 5x^2 - 1$ and $g(x) = x - x^2$. Work out their distance apart using each of d_1 and d_∞.



If $h(x) = f(x) - g(x) = 6x^2 - x - 1$ you can check that

- the only stationary point of h is at $x = \frac{1}{12}$, where $h(x) = -25/24$ and $|h(x)| = 25/24$.
- $h(x) = 0$ when $x = \frac{1}{2}$, $h(x) \leq 0$ when $0 \leq x \leq \frac{1}{2}$ and $h(x) \geq 0$ when $\frac{1}{2} \leq x \leq 1$.
- at the endpoints, $|h(0)| = 1$ and $|h(1)| = 4$.

$$\begin{aligned} d_1(f,g) &= \int_0^1 |f(x) - g(x)|\, dx \\ &= \int_0^{\frac{1}{2}} (1 + x - 6x^2)\, dx + \int_{\frac{1}{2}}^1 (6x^2 - x - 1)\, dx \\ &= \frac{5}{4}. \end{aligned}$$

- As the maximum must be at an endpoint of the interval or at a stationary point, $d_\infty(f,g) = 4$.

1.3 Useful Inequalities

Suppose that (X, d) is an arbitrary metric space. Here are two useful inequalities that are worth knowing about. They are both consequences of the triangle inequality:

$$|d(x,z) - d(y,z)| \leq d(x,y) \tag{1}$$

for all $x, y, z \in X$.

$$|d(x,y) - d(a,b)| \leq d(x,a) + d(y,b) \tag{2}$$

for all $x, y, a, b \in X$.

I won't give proofs of these here - they are an exercise for you to do in Assessment 1.

1.4 Balls and subspaces.

Definition 1.21. Let (X, d) be a metric space, let $x \in X$, and let $r \geqslant 0$.

(i) The *closed ball centred at x with radius r* is

$$B[x,r] = \{y \in X : d(x,y) \leqslant r\},$$

which is all the points in X that are at most a distance r from x.

(ii) The *open ball centred at x with radius r* is

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

which is all the points in X that are a distance less than r from x.

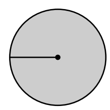
If we want to keep track of which space is involved, we may write $B_X[x,r]$ or $B_X(a,r)$, but most of the time we will drop the subscript if the metric space is clear from the context.

Example 1.22. Consider $\mathbb{R}$ with its usual metric. Set $x = 1$, $r = \frac{1}{2}$. Then

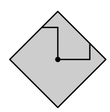
$$B[1, \tfrac{1}{2}] = \{y \in \mathbb{R} : d(1, y) \leqslant \tfrac{1}{2}\} = \{y \in \mathbb{R} : |y - 1| \leqslant \tfrac{1}{2}\} = [\tfrac{1}{2}, \tfrac{3}{2}],$$
$$B(1, \tfrac{1}{2}) = \{y \in \mathbb{R} : d(1, y) < \tfrac{1}{2}\} = \{y \in \mathbb{R} : |y - 1| < \tfrac{1}{2}\} = (\tfrac{1}{2}, \tfrac{3}{2}).$$

Example 1.23. The diagrams below show the closed balls in $\mathbb{R}^2$, centred at (a, b), for the three metrics d_2, d_1, d_∞:

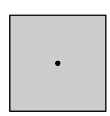
Using d_2 we get the usual closed disc in $\mathbb{R}^2$ that we are familiar with:



Using d_1 the analogue of the disc is the square rotated through $\pi/4$ radians:



Using d_∞ the analogue of the disc is the square with vertices (from bottom left, clockwise) $(a-r, b-r), (a+r, b-r), (a+r, b+r), (a-r, b+r)$:



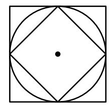
Since by Problem 1.3(a)

$$d_\infty((a_1, a_2), (b_1, b_2)) \le d_2((a_1, a_2), (b_1, b_2)) \le d_1((a_1, a_2), (b_1, b_2)),$$

we have

$$B_1((a, b), r) \subseteq B_2((a, b), r) \subseteq B_\infty((a, b), r),$$

where the subscript indicates which metric we are using. You can prove this in Problem 1.5.



Definition 1.24. Let (X, d) be a metric space, and let Y be a subset of X. Then there is a new metric space (Y, d_Y), with d_Y defined by

$$d_Y(y_1, y_2) = d(y_1, y_2),$$

so we are just restricting the original metric to Y. With this metric, Y is a (metric) *subspace* of X, and d_Y is sometimes called the *restriction* of d from X to Y (or just the "restricted metric").

Example 1.25. Let $(X, d) = (\mathbb{R}^2, d_\infty)$, and let Y be the line $\{(y, y) : y \in \mathbb{R}\}$. Then, for $(y_1, y_1), (y_2, y_2) \in Y$,

$$\begin{aligned} d_Y((y_1, y_1), (y_2, y_2)) &= d_\infty((y_1, y_1), (y_2, y_2)) \\ &= \max(|y_1 - y_2|, |y_1 - y_2|) \\ &= |y_1 - y_2|. \end{aligned}$$

1.5 Appendix - Normed Spaces

In the study of metric spaces we generalise the concept of "distance". You might also think it makes sense to generalise the related concept of "length". To do this you need to combine the notions of metric space and vector space together to get a *normed space*. Here's a formal definition:

Let V be a vector space defined over the real numbers $\mathbb{R}$ (in fact you can replace these with the complex numbers or by an arbitrary field). A mapping $||\cdot||$ from V to $\mathbb{R}$ is called a *norm* if it satisfies the following axioms

(N1) $||x|| \geq 0$ for all $x \in V$ and $||x|| = 0$ if and only if $x = 0$.

(N2) $||\alpha x|| = |\alpha|.||x||$ for all $\alpha \in \mathbb{R}$ and for all $x \in V$.

(N3) $||x + y|| \leq ||x|| + ||y||$ for all $x, y \in V$.

The pair $(V, ||\cdot||)$ is called a *normed space*. Every normed space is a metric space where the metric d is defined by

$$d(x, y) = ||x - y||,$$

for $x, y \in V$. The metrics d_1, d_2 and d_∞ on $\mathbb{R}^n$ and also d_1 and d_∞ on $C[a, b]$ arise in this way. The study of normed spaces is part of the subject of *functional analysis*, which MMath students can study in MAS436.

2 Convergence and Sequences.

We can now discuss convergence of sequences in the context of metric spaces. Intuitively, a sequence converges when it terms are getting closer and closer to some *limit*. How close the terms are to the limit can be measured using a metric so metric spaces provide the right context for studying convergence.

2.1 The Definition of Convergence.

A *sequence* in a set X means an ordered list $x_1, x_2, x_3, \ldots$ of elements of X, and we will often write such a sequence as (x_n). We want to have some notion of limits of sequences, and convergence.

The idea here is just the same as it is in $\mathbb{R}$ (from MAS221). Let's quickly remind ourselves of how it works in $\mathbb{R}$. I give you a sequence, say

$$1, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{4}}, \frac{1}{\sqrt{5}}, \ldots,$$

and I claim that this sequence converges to 0. Then you might say:

> 'Convince me that the terms here eventually get within 0.01 of 0 and stay that close,'

then I can say:

> 'Look at the terms from the 10001st onwards: they are
>
> $$\frac{1}{\sqrt{10001}}, \frac{1}{\sqrt{10002}}, \frac{1}{\sqrt{10003}}, \ldots,$$
>
> all of which are clearly within 0.01 of 0.'

You might then try to make my task harder by saying:

> 'Now convince me that the terms eventually get within 0.001 of 0 and stay that close,'

but I can say:

'Look at the terms from the 1000001st onwards: they are

$$\frac{1}{\sqrt{1000001}}, \frac{1}{\sqrt{1000002}}, \frac{1}{\sqrt{1000003}}, \ldots,$$

all of which are clearly within 0.001 of 0.'

We could continue like this for a long time. You challenge me with a very small number $\epsilon > 0$, and I have to find some term, the Nth say, such that all the subsequent terms, the $(N+1)$st, $(N+2)$nd, ... are within ϵ of the limit 0.

Informally, for a sequence to tend to a limit, then for any 'margin of error', all the terms are eventually within this margin of error from the limit. We'll write ϵ for our error margin, which of course will be some positive real number. The word 'eventually' will mean 'for all terms beyond a certain point', or in other words, all terms beyond the Nth for some N, and then we get the following definition:

Definition 2.1. Let (x_n) be a sequence in the metric space $X = (X, d)$. Let $a \in X$. We say that $x_n \to a$, or that (x_n) *has limit* $a \in X$, or that (x_n) *converges to* a, if, given $\epsilon > 0$, there exists $N \in \mathbb{N}$ so that, for all $n > N$, $x_n \in B(a, \epsilon)$, that is, $d(x_n, a) < \epsilon$.

Example 2.2. Show that $\frac{1}{\sqrt{n}} \to 0$ in $\mathbb{R}$.

Solution Let $\epsilon > 0$. By the Archimedean property of $\mathbb{R}$, there exists $N \in \mathbb{N}$ so that $N > \frac{1}{\epsilon^2}$. For $n > N$,

$$d\left(\frac{1}{\sqrt{n}}, 0\right) = \frac{1}{\sqrt{n}} < \frac{1}{\sqrt{N}} < \epsilon,$$

so $\frac{1}{\sqrt{n}} \to 0$.

Look carefully at the proof and make sure that you understand that it really does show that $\frac{1}{\sqrt{n}} \to 0$.

Now we should proceed to sequences in other metric spaces. The good news is that, to decide whether a sequence (x_n) tends to a in a metric space X, we only need to be able to handle the sequence $(d(x_n, a))$ of non-negative real numbers in $\mathbb{R}$ with its usual metric.

Proposition 2.3. *Let (x_n) be a sequence in a metric space X and let $a \in X$. Then $x_n \to a$ in (X, d) if and only if $d(x_n, a) \to 0$ in $\mathbb{R}$.*

Proof. As $d(x_n, a) \geq 0$,

$$d(x_n, a) \in B_{\mathbb{R}}(0, \epsilon) \Leftrightarrow d(x_n, a) < \epsilon \Leftrightarrow x_n \in B_X(a, \epsilon).$$

By the definition of convergence, $x_n \to a$ in X if and only if $d(x_n, a) \to 0$ in $\mathbb{R}$. $\square$

Remark 2.4. This can be useful even when working in $\mathbb{R}$ with its usual metric. In that case it says that, for a real sequence (x_n), $x_n \to a$ if and only if $|x_n - a| \to 0$.

Summary 2.5. Proposition 2.3 reduces issues of convergence in metric spaces to issues of convergence of sequences of non-negative real numbers. From MAS221, we know that, in $\mathbb{R}$,

(i) (Algebra of limits)

  (a) if $y_n \to y$ and $c \in \mathbb{R}$ then $cy_n \to cy$. In particular, if $y_n \to 0$ and $c \in \mathbb{R}$ then $y_n \to 0$

  (b) if $y_n \to y$ and $z_n \to z$ then $y_n + z_n \to y + z$. In particular, if $y_n \to 0$ and $z_n \to 0$ then $y_n + z_n \to 0$.

  (c) if $y_n \to y$ and $z_n \to z$ then $y_n z_n \to yz$. In particular, if $y_n \to 0$ and $z_n \to z$ then $y_n z_n \to 0$.

  (d) if $y_n \to y$, $z_n \to z$, each $z_n \neq 0$ and $z \neq 0$ then $y_n/z_n \to y/z$.

(ii) (Sandwich Rule) If $w_n \leq y_n \leq z_n$ for all n and $w_n \to \ell$ and $z_n \to \ell$ then $y_n \to \ell$. In particular, taking $w_n = 0 = \ell$, if $0 \leq y_n \leq z_n$ and $z_n \to 0$ then $y_n \to 0$.

(iii) The following sequences tend to 0:

  (a) $1/n^p (= n^{-p})$ for any $p > 0$. This includes $1/n$, $1/n^2$ and $1/\sqrt{n}$.

  (b) a^n for any a such that $-1 < a < 1$ (only $0 \leq a < 1$ will be relevant here).

Example 2.6. Show that $\left(\frac{1}{\sqrt{n}}, \frac{n-1}{n}\right) \to (0, 1)$ in $\mathbb{R}^2$ with the Euclidean metric.

Solution Note that

$$\begin{aligned} d\left(\left(\frac{1}{\sqrt{n}}, \frac{n-1}{n}\right), (0,1)\right) &= \sqrt{\left(\frac{1}{\sqrt{n}} - 0\right)^2 + \left(\frac{n-1}{n} - 1\right)^2} \\ &= \sqrt{\frac{1}{n} + \frac{1}{n^2}} \\ &\leqslant \sqrt{\frac{2}{n}} = \sqrt{2}\frac{1}{\sqrt{n}}. \end{aligned}$$

By Summary 2.5(iii)(a) and (i)(a), $\sqrt{2}\frac{1}{\sqrt{n}} \to 0$. By the Sandwich Rule,

$$d\left(\left(\frac{1}{\sqrt{n}}, \frac{n-1}{n}\right), (0,1)\right) \to 0$$

so, by Proposition 2.3, $\left(\frac{1}{\sqrt{n}}, \frac{n-1}{n}\right) \to (0,1)$.

Example 2.7. Show that the sequence (f_n) in $C[0,1]$ defined by

$$f_n \colon [0,1] \to \mathbb{R}, \qquad x \mapsto x^n/n,$$

converges to the constant function $f(x) = 0$ in $C[0,1]$ under both d_∞ and d_1.

Solution For d_∞,

$$d_\infty(f_n, f) = \sup\{|f_n(x) - f(x)| \,:\, x \in [0,1]\} = \sup\{x^n/n \,:\, x \in [0,1]\} = \frac{1}{n},$$

as f_n is increasing on $[0,1]$. We know that $\frac{1}{n} \to 0$ so, by Proposition 2.3, $f_n \to 0$.

For d_1,

$$d_1(f_n, f) = \int_0^1 \left|\frac{x^n}{n} - 0\right| dx = \int_0^1 \frac{x^n}{n} dx = \frac{1}{n(n+1)} < \frac{1}{n^2}.$$

We know that $\frac{1}{n^2} \to 0$ so, by the Sandwich Rule, $d_1(f_n, f) \to 0$ in $\mathbb{R}$ and hence $f_n \to f$ in $(C[0,1], d_1)$. Note that the details are different for the two metrics.

We will need to know that a convergent sequence in a metric space has a unique limit.

Proposition 2.8. *A sequence (x_n) in a metric space (X, d) has at most one limit.*

Proof. Suppose that $x_n \to a$ and $x_n \to b$. By axiom M3,

$$0 \leqslant d(a, b) \leqslant d(a, x_n) + d(x_n, b)$$

for all n. Now since $x_n \to a$ and $x_n \to b$, $d(a, x_n) \to 0$ and $d(x_n, b) \to 0$. By the algebra of limits (2.5), $d(a, x_n) + d(x_n, b) \to 0$ and, by the Sandwich Rule, the constant sequence $d(a, b) \to 0$. Hence $d(a, b) = 0$, and, by Axiom M1, $a = b$. □

In the next part, I will use the notation $\mathbf{x} = (x^1, x^2, \ldots, x^m)$ for points in $\mathbb{R}^m$, rather than the familiar $(x_1, x_2, \ldots, x_m)$. This is so that we can use the subscript to label terms in a sequence. You can think of a sequence $(\mathbf{x_n}) = (x_n^1, \ldots, x_n^m)$ in $\mathbb{R}^m$ as being made up of m sequences (x_n^1), ..., (x_n^m) in $\mathbb{R}$ in the co-ordinate positions. We shall see that $(\mathbf{x_n})$ converges to $\mathbf{x}$ as $n \to \infty$ if and only if each x_n^j converges to x^j. We establish this first for d_1, and then it will follow for d_2 and d_∞.

Proposition 2.9. *Let $(\mathbf{x_n}) = (x_n^1, \ldots, x_n^m)$ be a sequence in $\mathbb{R}^m$ and let $\mathbf{x} = (x^1, \ldots, x^m) \in \mathbb{R}^m$. Under the metric d_1, $\mathbf{x_n} \to \mathbf{x}$, as $n \to \infty$ if and only if $x_n^j \to x^j \in \mathbb{R}$, as $n \to \infty$ for all $1 \leq j \leq m$.*

Proof. First suppose that $\mathbf{x_n} \to \mathbf{x}$ in $\mathbb{R}^m$ under d_1. Let $1 \leq j \leq m$.

$$0 \leqslant d(x_n^j, x^j) = |x_n^j - x^j| \leqslant \sum_{i=1}^{m} |x_n^i - x^i| = d_1(\mathbf{x_n}, \mathbf{x}).$$

By Proposition 2.3, $d_1(\mathbf{x_n}, \mathbf{x}) \to 0$ so, by the Sandwich Rule, $d(x_n^j, x^j) \to 0$, and so $x_n^j \to x^j$, again by Proposition 2.3.

Conversely, suppose that $x_n^j \to x^j$ for each j. Then $d(x_n^j, x^j) \to 0$ for each j. But

$$d_1(\mathbf{x_n}, \mathbf{x}) = \sum_{j=1}^{m} |x_n^j - x^j| \to 0$$

and so $\mathbf{x_n} \to \mathbf{x}$ under d_1. □

Remark 2.10. Proposition 2.9 also holds for d_2 and d_∞ because a sequence in $\mathbb{R}^m$ converges under d_1 if and only if it converges under d_2 if and only if it converges under d_∞. See Chapter 2, Problem 3. Note that Example 2.6 could now be done by showing that $\frac{1}{\sqrt{n}} \to 0$ and $\frac{n-1}{n} \to 1$.

Definition 2.11. There's one more elementary notion we need to discuss, namely *subsequences*. Let $n_1 < n_2 < n_3 < \dots$ be an increasing sequence of natural numbers and let $(x_n) = x_1, x_2, x_3, \dots$ be a sequence in a metric space (X, d). Taking $x_{n_1}, x_{n_2}, x_{n_3}, \dots$, we get a new sequence (x_{n_k}), which is a *subsequence* of (x_n). For example we might have $x_1, x_3, x_5, \dots$ or $x_1, x_3, x_6, x_7, x_{12}, \dots$. Note that $n_k \geqslant k$.

Proposition 2.12. *Let (x_n) be a sequence in a metric space (X, d) converging to a limit $a \in X$. Then any subsequence (x_{n_k}) also converges to a.*

Proof. Let $x_{n_1}, x_{n_2}, x_{n_3}, \dots$ be any subsequence of (x_n). For any $\epsilon > 0$, there is a natural number N such that if $n > N$, $d(x_n, a) < \epsilon$. Now choose the natural number K so that $n_K > N$. If $k > K$, then $n_k > n_K > N$, and we conclude that $d(x_{n_k}, a) < \epsilon$. Thus $(x_{n_k}) \to a$ as $n \to \infty$. □

2.2 Convergence in Function Spaces.

In MAS221, you will have started to get used to using ϵ's when studying convergence. However, it may have seemed slightly pointless to you there, because your intuition of what convergence ought to mean is so strong, and the results you got by using ϵ's was always in agreement with your intuition. However, your intuition is probably not so strong in the less standard cases of metric spaces, particularly function spaces. It is in this situation that the full power of working with abstract formulations becomes clear. In this section we'll look at the function space $C[a, b]$ with its two metrics d_∞ and d_1. We'll see how convergence is a subtle business that can behave in ways we don't expect.

Let (f_n) be a sequence of continuous functions $f_n\colon [a, b] \to \mathbb{R}$, so that (f_n) is a sequence in $C[a,b]$. We already have two ideas of what it could mean for f_n to converge to a function $f \in C[a,b]$:

- $f_n \to f$ in the metric space $(C[a,b], d_1)$.
- $f_n \to f$ in the metric space $(C[a,b], d_\infty)$.

We saw an example of both sorts of convergence in Example 2.7. However there is another, more intuitive, form of convergence:

Definition 2.13. We say that (f_n) *converges pointwise* to f if, for each $x \in [a,b]$, the sequence $(f_n(x))$ converges to $f(x)$ in $\mathbb{R}$.

Think carefully about the definition. It says that pointwise convergence is what happens when you forget that each f_n is a continuous function, and just remember the sequences $(f_n(x))$ for each x.

Example 2.14. Let (f_n) be the sequence of functions defined by

$$f_n(x) = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!}.$$

You know that, for any x, $e^x = \sum_{i=0}^{\infty} x^i/i!$. What this means is that for each x, $\sum_{i=0}^{n} x^i/i! \to e^x$ as $n \to \infty$, and this exactly means that (f_n) converges pointwise to e^x.

Show that, for $b > 0$, (f_n) converges to e^x in $C([0,b])$ under d_∞.

Solution Here $|f_n(x) - e^x| = e^x - f_n(x)$. This has derivative $e^x - f_{n-1}(x) = \frac{x^n}{n!} + \frac{x^{n+1}}{(n+1)!} + \cdots \geq 0$ for all $x \in [0,b]$ so it is increasing and must take its maximum value at b. So $d_\infty(f_n(x), e^x) = |f_n(b) - e^b| \to 0$ as $f_n \to f$ pointwise. Therefore (f_n) converges to e^x in $C([0,b])$. What about d_1? See the next result.

We now have *three* ideas of convergence of functions in $C[a,b]$: convergence in the d_∞ metric, convergence in the d_1 metric, and convergence pointwise.[2] It's natural to ask whether they are related. For example, "If $f_n \to f$ in $(C[a,b], d_1)$, does this mean that $f_n \to f$ pointwise?" In general the answer to such questions is 'no', as we'll see in an example. However, we do have:

[2]From MAS221, you can see that convergence in the metric d_∞ is the same as *uniform convergence*. This is also explored in the Appendix to this section.

Proposition 2.15. *If f_n converges to f in $(C[a,b], d_\infty)$, then f_n converges to f in $(C[a,b], d_1)$ and also f_n converges to f pointwise.*

Proof. Suppose that f_n converges to f in $(C[a,b], d_\infty)$. For each n, let $k_n = d_\infty(f_n, f)$. Then $k_n \to 0$ and $|f_n(x) - f(x)| \leq k_n$ whenever $a \leqslant x \leqslant b$.

First let's prove that $f_n \to f$ in $(C[a,b], d_1)$. Then

$$d_1(f_n, f) = \int_a^b |f_n(x) - f(x)| dx \leq \int_a^b k_n dx = (b-a)k_n.$$

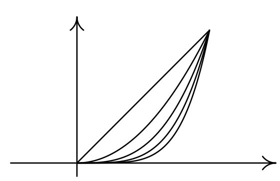
By the algebra of limits and the Sandwich Rule, $d_1(f_n, f) \to 0$ so, by Proposition 2.3, $f_n \to f$ in $(C[a,b], d_1)$.

For pointwise convergence, let $x \in [a,b]$. We must prove that $f_n(x) \to f(x)$. In $\mathbb{R}$,

$$|f_n(x) - f(x)| \leq k_n.$$

As $k_n \to 0$, $|f_n(x) - f(x)| \to 0$ by the Sandwich Rule. By Proposition 2.3, $f_n(x) \to f(x)$. That is, (f_n) converges pointwise to f. $\square$

Example 2.16. Here's an example of how strange convergence in function spaces can be. In the space $C([0,1])$, let $f_n(x) = x^n$. Does the sequence (f_n) converge in any of the senses we've described? Sketch the first few functions in the sequence.



On the one hand, it looks as if the graphs are getting closer and closer together, and approaching some limit graph. On the other hand, this limit graph seems to be

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

and this isn't continuous. It turns out that whether this sequence converges or not depends on our definition of convergence. Show that

(i) (f_n) does not converge pointwise to any continuous function $f \in C[0,1]$.

(ii) (f_n) does not converge in $(C[0,1], d_\infty)$.

(iii) (f_n) converges to the zero function $g(x) = 0$ in $(C[0,1], d_1)$

Solution

(i) For $x \in [0,1]$, $f_n(x) = x^n$, which converges to 0 if $x < 1$, and converges to 1 if $x = 1$. Therefore if f_n converges pointwise to a function $f \in C[0,1]$, then

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

But this isn't continuous, $f \notin C[0,1]$.

(ii) Remember from Proposition 2.15 that if (f_n) converged in $(C[0,1], d_\infty)$, then it would converge pointwise. But we know it doesn't converge pointwise.

(iii) For all n,

$$d_1(f_n, g) = \int_0^1 |f_n(x) - g(x)| dx = \int_0^1 x^n dx = \frac{1}{n+1} < \frac{1}{n}.$$

We know that $\frac{1}{n} \to 0$ so, by the Sandwich Rule, $d_1(f_n, g) \to 0$ and, by Proposition 2.3, $f_n \to g$.

The last example is a very good indication that the two natural metrics on function spaces like $C[0,1]$ have very different properties. When we deal with function spaces, we will try to be very careful always to specify the metric. There are other examples of this kind on the question sheet.

Appendix: More on Uniform Convergence

Uniform convergence is our name for convergence in $C[a,b]$ with the d_∞ metric. There is also a neat $\epsilon - N$ way of describing this convergence and it is as follows:

A sequence of functions (f_n) converges uniformly to f if given any $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that if $n > N$ then $|f_n(x) - f(x)| < \epsilon$ for all $x \in [a, b]$ (*).

Compare this with pointwise convergence - given any $\epsilon > 0$, for each $x \in [a, b]$ there exists $N(x) \in \mathbb{N}$ such that if $n > N(x)$ then $|f_n(x) - f(x)| < \epsilon$. For uniform convergence once we've chosen ϵ, N is the same for all values of x but for pointwise convergence, each value of x needs it own N.

Why is (*) equivalent to the definition using the d_∞ metric? Well that tells us that given $\epsilon > 0$ there exists N such that if $n > N$ then

$$\begin{aligned} d_\infty(f_n, f) < \epsilon \quad &\Leftrightarrow \quad \sup_{x \in [a,b]} |f_n(x) - f(x)| < \epsilon \\ &\Leftrightarrow \quad |f_n(x) - f(x)| < \epsilon \text{ for all } x \in [a, b], \end{aligned}$$

and that's what we need.

3 Closed and Open Sets.

We now know what it means for a sequence in a metric space to converge. Later on, in Chapter 6, we will want to formulate general conditions under which we can guarantee that sequences generated by iterating a function x, $f(x)$, $f(f(x))$, $f(f(f(x)))$ will converge.

It turns out that we'll need extra conditions on the space to guarantee this convergence, and the next couple of chapters will study these conditions. After all, we've seen examples of sequences that converge and examples which don't. We will need to know when sequences do converge.

3.1 Closed Sets

If we have a sequence lying in some given subset of a metric space, it would be good to be able to guarantee that the limit, if it exists, should lie within the set also.

For an example where this doesn't happen, think about the sequence $(\frac{1}{n})$. This is a sequence in the set $(0, 1]$, whose limit in $\mathbb{R}$ does exist (0, of course!), but which is not in the set $(0, 1]$. So the sequence lies in $(0, 1]$ but the limit does not. On the other hand, we'll see that every sequence in $[0, 1]$ that converges to some limit, the limit must be a member of $[0, 1]$. So the subset $[0, 1]$ of $\mathbb{R}$ has a property which is not shared by $(0, 1]$.

On the other hand, and we'll check this in a moment, every sequence in $[0, 1]$ that converges to some real number, will again converge to a member of $[0, 1]$. So the subset $[0, 1]$ of $\mathbb{R}$ has a property which is not shared by $(0, 1]$ or $\mathbb{Q}$. The next definition describes this desirable property.

Definition 3.1. Let X be a metric space, and let A be a subset of X. We say that A is a *closed* subset of X if whenever we have a sequence $x_1, x_2, \ldots$ in A which converges to a limit $x \in X$, then the limit x also lies in A.

Proposition 3.2. *Let (x_n) be a convergent sequence in $\mathbb{R}$ with limit x. If $x_n \geqslant 0$ for all x then $x \geqslant 0$. Therefore $[0, \infty)$ is closed in $\mathbb{R}$.*

Proof. Suppose that $x < 0$ and let $\epsilon = \frac{-x}{2} > 0$. There exists $N \in \mathbb{N}$ such that $|x_n - x| < \epsilon$ for all $n > N$. For $n > N$,

$$x_n - x \leqslant |x_n - x| < \epsilon = \frac{-x}{2}$$

and so

$$x_n < x + \frac{-x}{2} = \frac{x}{2} < 0,$$

a contradiction. Therefore $x \geqslant 0$. In other words, if each $x_n \in [0, \infty)$ then $x \in [0, \infty)$. So $[0, \infty)$ is closed. □

Lemma 3.3. *A closed ball $B[a, r]$ in a metric space (X, d) is a closed subset.*

Proof. Let $x_n \to x$ be a convergent sequence in X with each $x_n \in B[a, r]$. Then

$$d(a, x) \leqslant d(a, x_n) + d(x_n, x) \leq r + d(x_n, x),$$

so

$$d(x_n, x) + r - d(a, x) \geq 0.$$

But $d(x_n, x) \to 0$ so, by the algebra of limits,

$$d(x_n, x) + r - d(a, x) \to r - d(a, x).$$

By Proposition 3.2, $r - d(a, x) \geq 0$ so $d(a, x) \leqslant r$, that is $x \in B[a, r]$. Thus $B[a, r]$ is closed. □

Example 3.4. For $a < b$, the closed interval $[a, b] = B[\frac{a+b}{2}, \frac{b-a}{2}]$ in $\mathbb{R}$ is a closed ball, so it is closed in the sense of Def. 3.1

To show that a subset F of a metric space is not **closed**, we just need to produce one sequence within F with a limit outside F.

Example 3.5. In $\mathbb{R}$, let $F = (0, \infty)$. Let $x_n = \frac{1}{n}$. Then $x_n \in F$ for all n but $x_n \to 0 \notin F$. Therefore F is not closed.

Example 3.6. In $(\mathbb{R}_2, d_2)$, the subset $F = \{(x, y) \in \mathbb{R}^2 : x > 0\}$ of $\mathbb{R}^2$ is not closed. Let $x_n = (\frac{1}{n}, 0)$. Then $x_n \in F$ for all n but $x_n \to (0, 0) \notin F$.

Example 3.7. $\mathbb{Q}$ is not closed in $\mathbb{R}$ because the sequence $1, 1.4, 1.41, 1.414, \cdots$ has limit $\sqrt{2} \notin \mathbb{Q}$ but each term, having a finite decimal expansion, is rational.

Example 3.8. Let $b > 0$ and, in $C[0, b]$, let F be the set of polynomial functions from $[0, b]$ to $\mathbb{R}$. In Example 2.14, we saw a sequence of elements of F converging to e^x under d_∞. As $e^x \notin F$, F is not closed in $(C[0, b], d_\infty)$.

Example 3.9. Let

$$F = \{f \in C[0, 1] : f(1) = 1\}.$$

Show that F is closed when the metric is d_∞ but not when the metric is d_1.

Solution Let (f_n) be a sequence of elements of F that converges to some $f \in C[0, 1]$ under d_∞. Then $f_n(1) = 1$ for all n. To show that F is closed we must prove that $f \in F$, that is $f(1) = 1$. By Proposition 2.15, $f_n \to f$ pointwise. In particular, $f_n(1) \to f(1)$, and since $f_n(1) = 1$ for all n this means that $f(1) = 1$. Thus $f \in F$, as required.

By Example 2.16, the sequence $f_n(x) = x^n$ converges to the *zero* function $f(x) = 0$ under d_1. Here each $f_n \in F$, because $f_n(1) = 1$, whereas $f \notin F$ because $f(1) = 0$. Thus F is not closed.

3.2 Open Sets

There is a complementary notion to a closed subset, which, as you might expect, is the notion of an open subset. A set is *open* if every point of the set can be surrounded by an open ball that is also contained in the set.

Definition 3.10. A subset A of a metric space is *open* if for each $a \in A$ there is $r > 0$ such that $B(a, r) \subseteq A$.

Lemma 3.11. *An open ball $B(x, t)$ in a metric space (X, d) is an open subset in the sense of 3.10.*

Proof. Let $a \in B(x, t)$. Thus $d(a, x) < t$. Let $r = t - d(a, x) > 0$

and let $y \in B(a, r)$. Thus $d(a, y) < r$. Then

$$\begin{aligned} d(x, y) &\leq d(x, a) + d(a, y) \\ &< d(x, a) + r = t. \end{aligned}$$

Thus $y \in B(x, t)$. Therefore $B(a, r) \subseteq B(x, t)$, and so $B(x, t)$ is open. $\square$

Example 3.12. For $a < b$, the open interval $(a, b) = B(\frac{a+b}{2}, \frac{b-a}{2})$ in $\mathbb{R}$ is an open subset of $\mathbb{R}$.

Example 3.13. In $(\mathbb{R}^2, d_2)$, let

$$F = \{(x, y) : x > 0\}.$$

Show that F is open. (We saw in Ex. 3.6 that F is not closed.)

Solution Let $(x, y) \in F$, and set $r = x > 0$. Let $(a, b) \in B((x, y), r)$. Then

$$|a - x| \leqslant \sqrt{(a - x)^2 + (b - y)^2} < x$$

so $-x < a - x < x$. Adding x, $0 < a (< 2x)$. Thus $(a, b) \in F$. Therefore $B((x, y), r) \subseteq F$ and hence F is open.

Example 3.14. $(0, 1] \subset \mathbb{R}$ is not an open subset. Suppose it is open. Then there is an open ball $B(1, r) \subseteq (0, 1]$. But $B(1, r) = (1 - r, 1 + r)$ always contains $1 + \frac{r}{2} \notin (0, 1]$. This is a contradiction so $(0, 1]$ is not open. The method will work for any interval of the form $[a, b)$, $[a, b]$ or $(a, b]$.

We've now seen enough examples. Although the words *open* and *closed* suggest a connection, this may not be clear from the definitions. The connection involves set complements: the *complement* of A in X is the set $X \backslash A = \{x \in X : x \notin A\}$.

Theorem 3.15. *A subset A of a metric space X is open if and only if the complement $X \setminus A$ is closed. (Applying this to the complement $X \backslash A$, which has complement A, we get that X is closed if and only if the complement $X \setminus A$ is open.)*

Proof. Let A be open. We need to show that $X \setminus A$ is closed. So let (x_n) be a sequence in $X \setminus A$ converging to x. Suppose that $x \in A$. Then we can find $r > 0$ such that $B(x, r) \subset A$. As $x_n \to x$, there exists x_n with $x_n \in B(x, r)$. But then $x_n \in A$, contradicting our assumption that $x_n \in X \setminus A$. Therefore $x \notin A$ i.e. $x \in X \setminus A$.

Conversely let $X \setminus A$ be closed. Suppose that A is not open. Then there exists $a \in A$ such that no open ball $B(a, r)$ is completely inside A, i.e. $B(a, r) \cap (X \setminus A) \neq \emptyset$ for all $r > 0$. Hence, for each $n \in \mathbb{N}$ there exists $x_n \in B(a, \frac{1}{n})$ with $x_n \notin A$. Since $d(a, x_n) < 1/n$ the sequence $x_n \to a$. Since $x_n \in X \setminus A$ and $X \setminus A$ is closed, the limit $a \in X \setminus A$, a contradiction as $a \in A$. Therefore A is open. $\square$

Remark 3.16. Note that subsets may be open but not closed, closed but not open, neither open nor closed, or both open and closed. All these can be seen inside the real line: $(0, 1)$ is open but not closed, $[0, 1]$ is closed but not open, $[0, 1)$ is neither closed nor open, and the whole real line is both open and closed as a subset of itself.

It is particularly important to remember that sets can be neither closed nor open. A common error is to say that a set is open because it is not closed. Sets are not doors!

3.3 Unions and intersections

Let X be a set and let $A_i, i \in I$ be subsets of X, indexed by some set I. We can define unions and intersections as for pairs of sets:

$$\begin{aligned}
\bigcup_{i \in I} A_i &= \{x : x \in A_i \text{ for some } i \in I\}, \\
\bigcap_{i \in I} A_i &= \{x : x \in A_i \text{ for all } i \in I\}.
\end{aligned}$$

Recall also the *complement* of A in X or *set difference*: $X \setminus A = \{x \in X : x \notin A\}$. There are important relationships (the de Morgan Laws) between these operations that we'll use in the next section:

$$X \setminus \cup A_i = \cap (X \setminus A_i),$$
$$X \setminus \cap A_i = \cup (X \setminus A_i).$$

So taking complement turns unions into intersections and vice versa.

Proposition 3.17. *Let (X, d) be a metric space. Then:*

(i) X and $\emptyset$ are open subsets of X.

(ii) The union of any number of open subsets of X is again open.

(iii) Let $A_1, A_2, \ldots, A_n$ be a finite collection of open sets in a metric space X. Then $A_1 \cap A_2 \cap \cdots \cap A_n$ is also open.

Proof. (i) From the definition, it is clear that X is open. For $\emptyset$ to fail to be open, there would have to exist $a \in \emptyset$ such that $\emptyset$ contains no open ball $B(a, r)$. There can be no such a so $\emptyset$ is open.

(ii) Let $A_i \subset X$ be open sets, and let $a \in A = \cup A_i$. Then $a \in A_i$ for some i so there exists $\varepsilon > 0$ such that $B(a, \varepsilon) \subseteq A_i \subseteq A$, Hence A is open.

(iii) Let $a \in A_1 \cap A_2 \cap \cdots \cap A_n$. Then $a \in A_i$ for each i so, for each i, there exists $\varepsilon_i > 0$ such that $B(a, \varepsilon_i) \subseteq A_i$. Let $\varepsilon = \min_{1 \le i \le n} \varepsilon_i$. For each i, $0 < \varepsilon \le \varepsilon_i$ so $B(a, \varepsilon) \subseteq B(a, \varepsilon_i) \subseteq A_i$ and therefore $B(a, \varepsilon) \subseteq A_1 \cap A_2 \cap \cdots \cap A_n$. Hence $A_1 \cap A_2 \cap \cdots \cap A_n$ is open. □

Taking complements swaps unions and intersections and, by Theorem 3.15, it swaps open and closed. It follows that there is a companion result:

Proposition 3.18. *Let (X, d) be a metric space. Then:*

(i) X and $\emptyset$ are closed subsets of X.

(ii) The intersection of any number of closed subsets of X is again closed.

(iii) Let $A_1, A_2, \ldots, A_n$ be a finite collection of closed sets in a metric space X. Then $A_1 \cup A_2 \cup \cdots \cup A_n$ is also closed.

Remark 3.19. Note the subtle difference between (ii) and (iii) in Propositions 3.17 and 3.18. In both finiteness is needed in (iii). For example, in $\mathbb{R}$, the intersection of the open sets $(-\frac{1}{n}, \frac{1}{n})$ is $\{0\}$ which is not open and the union of the closed sets $[\frac{1}{n}, 1]$ is $(0, 1]$, which is not closed.

A metric is a useful thing to have. However we can talk about convergence and continuity whenever we have a notion of *nearness*. And for that open sets suffice! (Simply imagine small open balls around a point to measure nearness without mentioning distances.) This leads to the idea of a topological space where, instead of having a metric, we are only told what the open sets are (and consequently what the closed sets are by complementation). Here is the full definition.

> A *topology* on a set X is a collection τ of subsets of X subject to the following conditions:
>
> (i) $\varnothing, X \in \tau$;
>
> (ii) An arbitrary union of elements of τ is in τ; and,
>
> (iii) A finite intersection of subsets of X in τ is again in τ.
>
> The pair (X, τ) is referred to as a *topological space*; the subsets of X in τ are the open sets (for the topology τ).

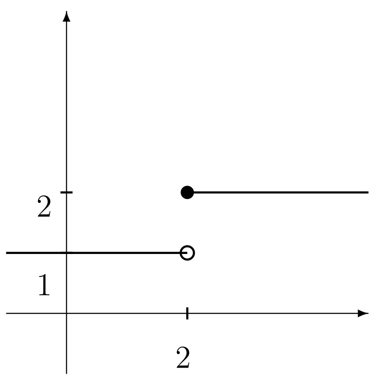
Clearly a metric space is a topological space with topology given by the open sets coming from the metric.

4 Continuity.

Here we establish the important notion of continuity of functions between metric spaces.

4.1 Continuity in $\mathbb{R}$.

Let's start by reminding ourselves how to define continuity for functions $f: \mathbb{R} \to \mathbb{R}$, which was mentioned in MAS170 (and treated properly in MAS221). We will give two definitions, both of which are designed to reflect the condition that the graph of f should not have any 'jumps'. (Rather informally, one should be able to draw the function without taking the pen off the paper.) Here's an example of a *discontinuity*:



Our first definition uses *sequences*. Let's take the sequence $1, 1\frac{1}{2}, 1\frac{2}{3}, 1\frac{3}{4}, \ldots$, which converges to 2. The values $f(1), f(1\frac{1}{2}), f(1\frac{2}{3}), f(1\frac{3}{4}), \ldots$ taken by the function at these points are all 1, whereas the value $f(2)$ at the limit is 2. The fact that the values taken by the function on the sequence do not tend to the value of the function at the limit of the sequence is reflecting the fact that there is a jump in the function. This leads to the sequential definition of continuity:

> A function $f: \mathbb{R} \to \mathbb{R}$ is continuous at $x \in \mathbb{R}$ if, given any sequence $x_n \to x$, then the sequence $f(x_n) \to f(x)$.

A second notion of continuity arises as follows. A naive definition of a continuous function $f\colon \mathbb{R} \to \mathbb{R}$ is that small changes in x should lead to only small changes in $f(x)$. Look back at the picture of the discontinuity we saw earlier.

The problem in the picture is that a very small change in x (from just below 2 to exactly 2) gives a very big jump in the value of $f(x)$ (from 1 to 2). We want to ensure that small changes in x don't give big changes in $f(x)$. We want our definition of continuity to prevent the 'jump', where a non-zero change in the value of $f(x)$ occurs over arbitrarily small changes in the value of x. We can reformulate this in a similar way to the ϵ-definition of convergence:

> A function $f\colon \mathbb{R} \to \mathbb{R}$ is continuous at $x \in \mathbb{R}$ if for all $\epsilon > 0$, there is a $\delta > 0$ such that if $|x - y| < \delta$, then $|f(x) - f(y)| < \epsilon$.

Here is how the definition works. I claim that f is continuous. Then you can challenge me by giving some maximum jump along the y-axis (this is ϵ), and then I have to show you some positive distance (this is δ) along the x-axis such that if two x-values differ by less than my number, then the corresponding y-values differ by less than your challenge. So, in the example above, you can challenge me by saying

> 'Convince me that a jump of 10 in the y-values cannot occur over an arbitrarily small interval in the x-values,'

and I can easily win this challenge by saying

> 'Well, if I choose any interval of length 2, say (any positive number would do!), the y-values in this interval are only going to vary by at most 1 (since all the y-values are 1 or 2).'

Clearly you lost because 10 was too big. You can do better by choosing a smaller jump:

> 'Convince me that a jump of 0.1 in the y-values cannot occur over an arbitrarily small interval in the x-values,'

If I try to respond in the same way:

> 'In any interval of length ℓ, the y-values in this interval are going to vary by at most 0.1',

you can point out that I am wrong – in any interval around $x = 2$, the y-value jumps from 1 to 2. So you would win this game. Clearly this game is exactly reflecting the property of whether there are discontinuities in the function – if there are not, I will win, and otherwise you will win.

4.2 The Definition of Continuity.

We'll give two definitions of continuity, just like in the last passage, and then we'll show that they are equivalent.

Definition 4.1. A function $f\colon X \to Y$ between two metric spaces (X, d_X) and (Y, d_Y) is *continuous at* $x \in X$ if:

> Given an $\epsilon > 0$ we can find a $\delta > 0$ such that, whenever $d_X(x, y) < \delta$, we have $d_Y(f(x), f(y)) < \epsilon$.

We can rewrite this in terms of open balls:

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

To see this recall that if $A \subseteq X$ then $f(A) \subseteq Y$ is defined by

$$f(A) = \{f(x) \in Y, x \in A\}.$$

Now $B(x, \delta) = \{a \in X; d_X(x, a) < \delta\}$. But by definition of continuity if $d_X(x, a) < \delta$ then $d_Y(f(x), f(a)) < \epsilon$; equivalently, if $a \in B(x, \delta)$ then $f(a) \in B(f(x), \epsilon)$, i.e.

$$\{f(a); a \in B(x, \delta)\} \subseteq B(f(x), \epsilon),$$

i.e. $f(B(x, \delta)) \subseteq B(f(x), \epsilon)$.

Definition 4.2. A function $f\colon X \to Y$ between two metric spaces (X, d_X) and (Y, d_Y) is *continuous at* $x \in X$ if:

> Whenever we have a sequence $x_1, x_2, \ldots$ of elements of X converging to x, then the sequence $f(x_1), f(x_2), \ldots$ in Y converges to the limit $f(x) \in Y$.

Definition 4.3. We say that a function $f \colon X \to Y$ between metric spaces X and Y is *continuous* if f is continuous at every $x \in X$.

Theorem 4.4. *Let $f \colon (X, d_X) \to (Y, d_Y)$ be a map between metric spaces. Take $x \in X$. Then the following are equivalent:*

(i) f is continuous at x in the sense of Definition 4.1.

(ii) f is continuous at x in the sense of Definition 4.2.

Proof. (1) $\Rightarrow$ (2). Suppose that f is continuous at $x \in X$ in the sense of Definition 4.1, and let $x_n \to x$. We need to show that $f(x_n) \to f(x)$. Choose $\epsilon > 0$. Then there exists $\delta > 0$ such that $d_Y(f(x_n), f(x)) < \epsilon$ whenever $d_X(x_n, x) < \delta$. But, since $x_n \to x$, there is an N such that $d_X(x_n, x) < \delta$ for all $n > N$. Combining these, we see that for all $n > N$, $d_Y(f(x_n), f(x)) < \epsilon$, and so $(f(x_n))$ tends to $f(x)$.

(2) $\Rightarrow$ (1). We suppose that f is continuous at x in the sense of Definition 4.2, and suppose for a contradiction that Definition 4.1 fails. Then there is an $\epsilon > 0$ such that:

> For all possible $\delta > 0$, there is a $y \in X$ such that $d_X(x, y) < \delta$, but $d_Y(f(x), f(y)) \geqslant \epsilon$.

We'll now use this property to find a sequence that makes Definition 4.2 fail, giving us a contradiction. Let $\delta = \frac{1}{n}$. Then by the above there is some y_n such that $d_X(x, y_n) < \frac{1}{n}$ while $d_Y(f(x), f(y_n)) \geqslant \epsilon$. Now consider the sequence (y_n); we have $d_X(x, y_n) < \frac{1}{n}$, and so $y_n \to x$. However, $d_Y(f(x), f(y_n)) \geqslant \epsilon$ for all n, so $f(y_n) \not\to f(x)$. This gives the required contradiction to (2). $\square$

Example 4.5. The function $f \colon \mathbb{R} \to \mathbb{R}$, defined by

$$f(x) = \begin{cases} 1 & \text{if } x < 2, \\ 2 & \text{if } 2 \leqslant x, \end{cases}$$

is not continuous. We can use either definition; (4.1): there does not exist $\delta > 0$ such that $f(B(2,\delta)) \subseteq B(2,\frac{1}{2})$, (4.2): the sequence $x_n = 2 - \frac{1}{n} \to 2$ but $f(x_n) = 1 \to 1 \neq f(2)$. (This is the example from the discussion at the start of the section.)

Example 4.6. Recall from Example 2.16 that if $f_n(x) = x^n$ then, in $C[0,1]$ under d_1, (f_n) converges to f where $f(x) = 0$ for all x. Define $\theta : C[0,1] \to \mathbb{R}$ by $\theta(g) = g(1)$. Then $\theta(f_n) = f_n(1) = 1 \to 1$ but $\theta(f) = f(1) = 0$. Hence θ is not continuous (at f) when $C[0,1]$ has the metric d_1 and $\mathbb{R}$ has its usual metric.

Now consider the same map θ when $C[0,1]$ has the metric d_∞ and $\mathbb{R}$ has its usual metric. Let (f_n) be any convergent sequence in $(C[0,1], d_\infty)$ with limit f. Then by Proposition 2.15, (f_n) converges pointwise to f. In particular, $f_n(1) \to f(1)$, that is $\theta(f_n) \to \theta(f)$. So here θ is continuous.

Example 4.7. Let $f\colon (\mathbb{R}^2, d_2) \to \mathbb{R}$ (usual metric on $\mathbb{R}$) be defined by $f(x,y) = x^2 - y^3$. Show that f is continuous.

Solution Let $((x_n, y_n)) \to (x,y)$ be a convergent sequence in $\mathbb{R}^2$. Then $(x_n) \to x$ and $y_n \to y$ by the d_2 version of Proposition 2.9. By the algebra of limits, $x_n^2 - y_n^3 \to x^2 - y^3$, that is, $f((x_n, y_n)) \to f((x,y))$. Thus f is continuous.

Example 4.8. Let A be any 2×2 matrix of real numbers, and regard it as a function $A\colon \mathbb{R}^2 \to \mathbb{R}^2$, where $\mathbb{R}^2$ has the d_∞ metric. This is a continuous function, as we'll now prove. Let's write

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

and let $\mathbf{x} = (x_1, x_2)$, $\mathbf{y} = (y_1, y_2)$ be elements of $\mathbb{R}^2$. Then a computation that we'll omit (see Appendix to Chapter 4) shows that $d_\infty(A\mathbf{x}, A\mathbf{y}) \leqslant 2M d_\infty(\mathbf{x}, \mathbf{y})$, where $M = \max\{|a|, |b|, |c|, |d|\}$. Now we are ready to prove continuity.

> Let $\mathbf{x} \in \mathbb{R}^2$ and let $\epsilon > 0$. Then set $\delta = \frac{\epsilon}{2M}$, so that when $d_\infty(\mathbf{x}, \mathbf{y}) < \delta$, we have
>
> $$d_\infty(A\mathbf{x}, A\mathbf{y}) \leqslant 2M d_\infty(\mathbf{x}, \mathbf{y}) < 2M\delta = 2M\frac{\epsilon}{2M} = \epsilon.$$

This proves that A is continuous at $\mathbf{x}$, and it works for all choices of $\mathbf{x}$, so that A is continuous.

The proof fails when A is the sero matrix, as we can't divide by $M = 0$, but it is obvious that A is continuous in that case. (Why?)

Example 4.9. Here's a slightly unusual, but important, example of continuity. We shall show that the integration map is a continuous map from $C[a, b]$ to $\mathbb{R}$. This means that we will make the definition of continuity apply in a situation where we have no intuitive notion of what continuity should mean.

Consider $C[a, b]$ with the supremum metric $d_\infty(f, g) = \sup\{|f(t) - g(t)| : t \in [a, b]\}$, and define $I \colon C[a, b] \to \mathbb{R}$ to be the function that sends $f \in C[a, b]$ to

$$I(f) = \int_a^b f(t)\, dt \ \in \mathbb{R}.$$

Thus, given a function $f \in C[a, b]$, the value $I(f)$ is the 'area under the graph' of f. We will prove that I is continuous.

Now fix a function $f \in C[a, b]$. We show that I is continuous at f. Let $\epsilon > 0$, and take any δ such that $0 < \delta < \frac{\epsilon}{b-a}$. If $g \in C[a, b]$ is such that $d_\infty(f, g) < \delta$, then $|f(t) - g(t)| < \delta$ for all $t \in [a, b]$, and so

$$\begin{aligned}
d(I(f), I(g)) &= |I(f) - I(g)| \\
&= \left| \int_a^b f(t)\, dt - \int_a^b g(t)\, dt \right| \\
&= \left| \int_a^b (f(t) - g(t))\, dt \right| \\
&\leqslant \int_a^b |f(t) - g(t)|\, dt \leqslant \int_a^b \delta\, dt = \delta(b - a) < \epsilon.
\end{aligned}$$

This completes the proof of continuity at f.

Almost all of the familiar operations one can carry out with continuous functions carry over. We give a small sample here.

Proposition 4.10. *Let (X, d_X), (Y, d_Y), (Z, d_Z) be metric spaces.*

(i) *(Constant function) If $f\colon X \to Y$ maps all of X to a single point y of Y then f is continuous.*

(ii) *(Composition) If $f\colon X \to Y$, $g\colon Y \to Z$ are continuous then the composite $g \circ f\colon X \to Z$ is continuous.*

(iii) *(Inclusion) Let $A \subset X$. Then the inclusion $j\colon A \to X$ (defined by $j(a) = a$, for all $a \in A$) is a continuous function (with A considered as a subspace, as in Definition 1.24).*

Proof. (i) If $x_n \to x$ in X then $f(x_n) = f(x) = y$, and so $f(x_n) \to f(x)$.

(ii) Let $x_n \to x$ in X. As f is continuous, we must have $f(x_n) \to f(x)$ in Y. As g is continuous we must have $g(f(x_n)) \to g(f(x))$ as required.

(iii) If $a_n \to a$ then $j(a_n) = a_n$ tends to $a = j(a)$.

□

The next result shows that the distance function on an arbitrary metric space is itself a continuous function.

Theorem 4.11. *Let (X, d) be a metric space and fix $x \in X$. Then the mapping $a \to d(a, x)$ is continuous from (X, d) to $\mathbb{R}$ with its usual metric.*

Proof. Let (a_n) be a sequence converging to a in X. Then $d(a_n, a) \to 0$ as $n \to \infty$. Using the inequality (1) of section 1.3, we have

$$|d(a_n, x) - d(a, x)| \leq d(a_n, a),$$

so by the sandwich rule, $|d(a_n, x) - d(a, x)| \to 0$ as $n \to \infty$, i.e. $d(a_n, x) \to d(a, x)$ in $\mathbb{R}$ as $n \to \infty$ and the result follows by Definition 2 of continuity. □

4.3 Continuity and Closed and Open Sets.

Next we'll see how continuous maps interact with open and closed sets. Here's a reminder of some notation:

Definition 4.12. Let $f\colon X \to Y$ be a function between sets X and Y, and let U be a subset of Y. Then the *preimage* or *inverse image* of U under f, written $f^{-1}(U)$, is defined to be

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

Thus $f^{-1}(U)$ consists of all the points in X that f sends into U. Note the following useful properties:

- $f^{-1}(Y) = X$.
- $f^{-1}(Y \setminus A) = X \setminus f^{-1}(A)$.
  This holds because

$$\begin{aligned} & x \in f^{-1}(Y \setminus A) \\ \Leftrightarrow\ & f(x) \in Y \setminus A \\ \Leftrightarrow\ & f(x) \notin A \\ \Leftrightarrow\ & x \in X \setminus f^{-1}(A) \end{aligned}$$

Theorem 4.13. *Let $f\colon X \to Y$ be a function between metric spaces. The following are equivalent:*

(i) f is continuous.

(ii) For every closed subset $A \subseteq Y$, $f^{-1}(A)$ is closed.

(iii) For every open subset $A \subseteq Y$, $f^{-1}(A)$ is open.

Proof. (i)$\Rightarrow$(ii) Suppose that f is continuous and let $A \subseteq Y$ be closed. Let (x_n) be a sequence in $f^{-1}(A)$ such that $x_n \to x$ in X. Then, by Definition 4.2, $f(x_n) \to f(x)$. But each $x_n \in f^{-1}(A)$ so each $f(x_n) \in A$ which is closed. So $f(x) \in A$ and $x \in f^{-1}(A)$, which is therefore closed.

(ii)$\Rightarrow$(iii) Suppose that (ii) holds and let $A \subseteq Y$ be open. Then $Y \backslash A$ is closed so $f^{-1}(Y \backslash A)$ is closed. But $f^{-1}(Y \backslash A) = X \backslash f^{-1}(A)$ therefore $X \backslash f^{-1}(A)$ is closed and $f^{-1}(A)$ is open.

(iii)$\Rightarrow$(i). Finally, suppose that $f^{-1}(A)$ is open for every open set $A \subseteq Y$. We want to prove that f is continuous. Let $x \in X$ and let $\epsilon > 0$. The open ball $B = B(f(x), \epsilon)$ in Y is open and contains $f(x)$ so $f^{-1}(B)$ is open and contains x. Therefore $B(x, \delta) \subseteq f^{-1}(B)$

for some $\delta > 0$. But then $f(B(x, \delta)) \subseteq B = B(f(x), \epsilon)$. Hence f is continuous by Definition 4.1. $\square$

Example 4.14. In $(\mathbb{R}^2, d_2)$, let $A = \{(x, y) \in \mathbb{R}^2 : 3 < x^2 - y^3 < 21\}$. Show that A is open in $\mathbb{R}^2$ under d_2.

Solution $A = f^{-1}((3, 21))$, where $f : \mathbb{R}^2 \to \mathbb{R}$ is the continuous function from Example 4.7. As $(3, 21)$ is open in $\mathbb{R}$, A is open in $\mathbb{R}^2$ by Theorem 4.13.

Example 4.15. It's worth noting that if A is closed (resp. open) in X, and if $f\colon X \to Y$ is continuous, then $f(A)$ need not be closed (resp. open) in Y. For example, consider the maps

$$\begin{aligned} f\colon \mathbb{R} \to \mathbb{R}, &\quad x \longmapsto e^{-x}, \\ g : \mathbb{R} \to \mathbb{R}, &\quad x \longmapsto x^2. \end{aligned}$$

These are continuous, $\mathbb{R}$ is closed and open but $f(\mathbb{R}) = (0, \infty)$, which is not closed in $\mathbb{R}$ and $g(\mathbb{R}) = [0, \infty)$ which is not open.

Appendix: Example 4.8 continued

To prove that $d_\infty(A\mathbf{x}, A\mathbf{y}) \leq 2M d_\infty(\mathbf{x}, \mathbf{y})$, where $M = \max\{|a|, |b|, |c|, |d|\}$.

First note that

$$A\mathbf{x} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} ax_1 + bx_2 \\ cx_1 + dx_2 \end{pmatrix}.$$

and similarly $A\mathbf{y} = \begin{pmatrix} ay_1 + by_2 \\ cy_1 + dy_2 \end{pmatrix}$. So

$$\begin{aligned} d_\infty(A\mathbf{x}, A\mathbf{y}) &= \max\{|(ax_1 + bx_2) - a(y_1 + by_2)|, |(cx_1 + dx_2) - (cy_1 + dy_2)|\} \\ &\leq \max\{|a||x_1 - y_1| + |b||x_2 - y_2|, |c||x_1 - y_1| + |d||x_2 - y_2|\} \\ &\leq M(|x_1 - y_1| + |x_2 - y_2|) \\ &\leq 2M \max\{|x_1 - y_1|, |x_2 - y_2|\} \\ &= 2M d_\infty(\mathbf{x}, \mathbf{y}), \end{aligned}$$

where on the last line but one we used the inequality

$$a + b \leq 2 \max\{a, b\},$$

for $a, b \geq 0$.

5 Cauchy Sequences and Completeness.

There is a difficulty with the definition of convergence that we met earlier. We don't know any way to test for convergence without first guessing a limit a, and then verifying Definition 2.1. If we don't know what a should be we cannot start to handle $d(x_n, a)$. It may not be possible to guess exactly what the limit must be. It may only be possible to guess an approximation. But if we know that a limit exists then there may be an algorithm that will find out (see the next chapter). To say that a sequence converges to a limit means, roughly, that the terms of the sequence get closer and closer to the limit. One consequence of this is that the terms will get closer and closer *to each other*. We might enquire if this statement — that the terms get closer together — is enough to guarantee that the sequence converges to a limit?

5.1 Cauchy Sequences

To say that a sequence converges to a limit means, roughly, that the terms of the sequence get closer and closer to the limit. One consequence of this is that the terms will get closer and closer *to each other*. Maybe this statement — that the terms get closer together — is enough to guarantee that the sequence converges to a limit. This is the idea behind a Cauchy sequence.

Definition 5.1. We say that a sequence $x_1, x_2, \ldots$ in a metric space X is a *Cauchy sequence* if for all $\epsilon > 0$, there exists N such that $d(x_m, x_n) < \epsilon$ whenever $m, n > N$.

To see that this is a good internal test for convergence, we'd need to check that sequences converge precisely when they satisfy this Cauchy condition. Our first indication that this is a good test comes with the following result, which shows that convergent sequences are Cauchy.

Lemma 5.2. *In a metric space (X, d), every convergent sequence is Cauchy.*

Proof. Suppose that the sequence $(x_n) \to x$. Let $\epsilon > 0$. Then there exists N such that $d(x_n, x) < \frac{\epsilon}{2}$ for all $n > N$. For $n, m > N$, we then have

$$d(x_n, x_m) \leqslant d(x_n, x) + d(x, x_m) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

so that (x_n) is Cauchy. □

It is important to appreciate that Cauchy sequences do not always converge. Indeed, in $\mathbb{Q}$, the sequence $\frac{3}{1}, \frac{31}{10}, \frac{314}{100}, \ldots$, where the nth term is π to $n-1$ decimal places, is Cauchy (since it is a convergent sequence in $\mathbb{R}$) but its limit is π, which is not in $\mathbb{Q}$. So the sequence is not convergent in the metric space $\mathbb{Q}$, even though it is a Cauchy sequence. The problem here is that the rational numbers are not enough to do analysis!

Example 5.3. Let (X, d) be a metric space, let (x_n) be a sequence in X, and suppose that

$$d(x_{n+1}, x_{n+2}) \leqslant \tfrac{1}{2} d(x_n, x_{n+1})$$

for all $n \geqslant 1$. This says that the distance between consecutive terms x_m, x_{m+1} halves every time m increases.

Then (x_n) is Cauchy. To show this, first note that by the assumption we have $d(x_n, x_{n-1}) \leqslant \frac{1}{2^{n-2}} d(x_2, x_1)$, so that if $n > m$ we have

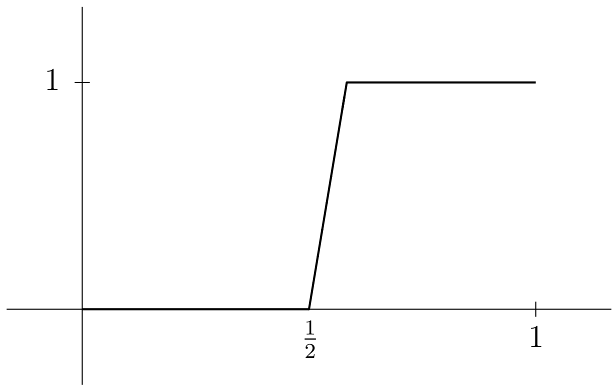
$$\begin{aligned} d(x_m, x_n) &\leqslant d(x_m, x_{m+1}) + \cdots + d(x_{n-1}, x_n) \\ &\leqslant d(x_1, x_2) \times \left[\frac{1}{2^{m-1}} + \cdots + \frac{1}{2^{n-2}}\right] \\ &\leqslant d(x_1, x_2) \times \left[\frac{1}{2^{m-1}} + \frac{1}{2^m} + \cdots\right] \\ &= \frac{d(x_2, x_1)}{2^{m-2}}, \end{aligned}$$

where we used the formula for geometric series on the last line. This last quantity converges to 0 as $m \to \infty$. So for $\epsilon > 0$, take N to be sufficiently large so that $\frac{d(x_2, x_1)}{2^{m-2}} < \epsilon$ for all $m > N$. Then $d(x_n, x_m) < \epsilon$ whenever $n, m > N$. This proves that (x_n) is Cauchy.

Example 5.4. Consider the space $C[0, 1]$ of continuous functions on the closed interval $[0, 1]$, with the metric d_1. Define the sequence

(f_n) by

$$f_n(x) = \begin{cases} 0 & \text{if } x \leqslant \frac{1}{2}, \\ n(x - \frac{1}{2}) & \text{if } \frac{1}{2} \leqslant x \leqslant \frac{1}{2} + \frac{1}{n}, \\ 1 & \text{if } \frac{1}{2} + \frac{1}{n} \leqslant x. \end{cases}$$



Then (f_n) is a Cauchy sequence. To prove this, first verify that

$$d_1(f_n, f_m) = \frac{1}{2}\left(\frac{1}{m} - \frac{1}{n}\right) < \frac{1}{2m} \tag{3}$$

for $m < n$. Then, given $\epsilon > 0$, take N to be the smallest natural number that exceeds $\frac{1}{2\epsilon}$. If $m, n > N$, then assume that $n > m$, so that

$$d_1(f_n, f_m) < \frac{1}{2m} < \frac{1}{2N} < \epsilon,$$

and therefore (f_n) is Cauchy.

You should try to do the calculation in (3). Details will be posted on the course website in a week or two.

5.2 Completeness

In this section we'll single out and study the following important property of a metric space.

Definition 5.5. The metric space X is *complete* if every Cauchy sequence (x_n) in X converges to a limit $x \in X$. More generally, a subset $A \subseteq X$ is *complete* if every Cauchy sequence in A converges to a limit in A.

Example 5.6. $\mathbb{Q}$ is not complete. In $\mathbb{Q}$, the sequence $\frac{1}{1}, \frac{14}{10}, \frac{141}{100}, \ldots$, where the nth term is $\sqrt{2}$ to $n-1$ decimal places, is Cauchy (since it is a convergent sequence in $\mathbb{R}$) but its limit is $\sqrt{2}$, which is not in $\mathbb{Q}$. So the sequence is not convergent in the metric space $\mathbb{Q}$, even though it is a Cauchy sequence. The problem here is that the rational numbers are not enough to do analysis!

We need to show that $\mathbb{R}$ is complete and that is the next aim. To prove that $\mathbb{R}$ is complete is slightly tricky, and to prove it we'll need two preliminary results.

Theorem 5.7. *Let (x_n) be a Cauchy sequence in a metric space (X, d) and let $a \in X$. If (x_n) has a subsequence (x_{n_k}) converging to a, then (x_n) converges to a.*

Proof. Let $\epsilon > 0$. As (x_n) is Cauchy, there exists N such that $d(x_n, x_m) < \frac{\epsilon}{2}$ for all $n, m > N$. As $x_{n_k} \to a$, there exists K such that $d(x_{n_k}, a) < \frac{\epsilon}{2}$ for all $k > K$. Choose an integer M such that $M > N$ and $M > K$ and let $m = n_M$. Then $m \geqslant M > N$. Let $n > m$. Then $n, m > N$ so $d(x_n, x_m) < \frac{\epsilon}{2}$. But $x_m = x_{n_M}$ is in the subsequence and $M > K$ so $d(x_m, a) < \frac{\epsilon}{2}$. Then

$$d(x_n, a) \leqslant d(x_n, x_m) + d(x_m, a) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Thus $x_n \to a$. □

Reminder 5.8 (Bolzano-Weierstrass Theorem, MAS221)**.** Let (x_n) be a sequence of real numbers which is *bounded*, i.e. there are some α, β such that $\alpha \leqslant x_n \leqslant \beta$ for all n. Then the sequence (x_n) has a convergent subsequence.

Theorem 5.9. $\mathbb{R}$ *is complete.*

Proof. Let $x_1, x_2, \ldots$ be a Cauchy sequence in $\mathbb{R}$. There exists $N \in \mathbb{N}$ such that $|x_m - x_n| < 1$ whenever $m, n > N$. Then

$$|x_m| \leqslant |x_{N+1}| + |x_m - x_{N+1}| \leqslant |x_{N+1}| + 1$$

whenever $m > N$, so that

$$|x_m| \leqslant \max\{|x_1|, \dots, |x_N|, |x_{N+1}| + 1\},$$

and so (x_m) is bounded. By the Bolzano-Weierstrass Theorem there is a convergent subsequence (x_{n_k}) tending to ℓ, say. By Theorem 5.7, $x_n \to \ell$. Hence $\mathbb{R}$ is complete. $\square$

Proposition 5.10. *Let X be a complete metric space, and let A be a closed subset of X. Then A is complete.*

Proof. Let $a_1, a_2, \dots$ be a Cauchy sequence in A. Then it is also a Cauchy sequence in X, and therefore converges to some $x \in X$. Since A is a closed subset of X, $x \in A$, so A is complete. $\square$

Proposition 5.11. *Let A be a complete subset of a metric space X. Then A is closed.*

Proof. Let (x_n) be a sequence in A that converges to $x \in X$. We need to show that $x \in A$. By Lemma 5.2, (x_n) is a Cauchy sequence and so, since A is complete, it converges to some $a \in A$. Since limits are unique, $a = x$ and $x \in A$ as required. $\square$

Theorem 5.12. *For $k \geq 1$, $\mathbb{R}^k$, with the Euclidean metric d_2, is complete.*

Proof. Let $(\mathbf{x_n}) = ((x_n^1, \dots, x_n^k))$ be a Cauchy sequence in $\mathbb{R}^k$. Let $\epsilon > 0$, and let $N \in \mathbb{N}$ be such that $d_2(\mathbf{x_n}, \mathbf{x_m}) < \epsilon$ for all $n, m > N$. Then for each $1 \leqslant i \leqslant k$ we have

$$|x_n^i - x_m^i| \leqslant \sqrt{(x_n^1 - x_m^1)^2 + \cdots + (x_n^k - x_m^k)^2} = d_2(\mathbf{x_n}, \mathbf{x_m}) < \epsilon.$$

Therefore the sequence (x_n^i) is Cauchy in $\mathbb{R}$, and so it converges to some limit $x^i \in \mathbb{R}$. It follows from the d_2 version of Proposition 2.9 that $(\mathbf{x_n})$ converges to $\mathbf{x} = (x^1, \dots, x^k)$. Thus $\mathbb{R}^k$ is complete. $\square$

Within the next example, we will use the following important property of the Riemann integral: if g is a non-negative continuous function on $[a, b]$ and $\int_a^b g(x)dx = 0$, then $g(x) = 0$ for all $a \leq x \leq b$.

Example 5.13. $C[a,b]$, with the metric d_1, is *not* complete.

We'll demonstrate this in the case of $C[0,1]$. Remember from Example 5.4 that we had a Cauchy sequence of functions (f_n) defined by

$$f_n(x) = \begin{cases} 0 & \text{if } x \leqslant \frac{1}{2}, \\ n(x - \frac{1}{2}) & \text{if } \frac{1}{2} \leqslant x \leqslant \frac{1}{2} + \frac{1}{n}, \\ 1 & \text{if } \frac{1}{2} + \frac{1}{n} \leqslant x. \end{cases}$$

Suppose for a contradiction that this Cauchy sequence does converge, i.e. that $f_n \to f$ for some *continuous* function f. Then $d_1(f_n, f) \to 0$ as $n \to \infty$. But for all $n \in \mathbb{N}$

$$d_1(f_n, f) = \int_0^1 |f_n(t) - f(t)|dt \geqslant \int_0^{\frac{1}{2}} |f_n(t) - f(t)|dt = \int_0^{\frac{1}{2}} |f(t)|dt \geqslant 0,$$

so that we must have $\int_0^{\frac{1}{2}} |f(t)|dt = 0$. Since f is continuous, this implies that $f(t) = 0$ for $0 \leqslant t \leqslant \frac{1}{2}$. The same argument shows that

$$d_1(f_n, f) \geqslant \int_{\frac{1}{2}+1/n}^1 |f(t) - f_n(t)|dt \to \int_{\frac{1}{2}}^1 |f(t) - 1|dt \geqslant 0,$$

and since f is continuous, then $f(t) = 1$ for all $\frac{1}{2} \leqslant t \leqslant 1$. No such f can exist, since we have $f(\frac{1}{2}) = 0$ and $f(\frac{1}{2}) = 1$. Thus $C[0,1]$ is not complete with respect to the metric d_1.

The last example showed that $(C[a,b], d_1)$ is not complete. However, we know that the two metrics on $C[a,b]$ behave very differently; indeed, they behave differently enough that with one metric the space is not complete but with the other it *is*:

Theorem 5.14. *$(C[a,b], d_\infty)$ is complete.*

Proof. Let (f_n) be a Cauchy sequence in $C[a,b]$. We will prove the theorem in three steps.

Step 1 (f_n) converges pointwise to a function f defined on $[a,b]$.

To prove step 1, suppose that $\epsilon > 0$. Then there exists $N \in \mathbb{N}$ so that $d_\infty(f_n, f_m) < \epsilon$ for all $m, n > N$. But then for all $x \in [a, b]$,

$$\begin{aligned} |f_n(x) - f_m(x)| &\leq \sup_{y\in[a,b]} |f_n(y) - f_m(y)| \\ &= d_\infty(f_n, f_m) < \epsilon, \end{aligned}$$

so the sequence $(f_n(x))$ is Cauchy in $\mathbb{R}$; hence it converges by completeness of $\mathbb{R}$. Our desired function f is then defined by

$$f(x) := \lim_{n\to\infty} f_n(x).$$

Step 2 $\lim_{n\to\infty} d_\infty(f_n, f) = 0$.

To prove step 2, this time we choose $N \in \mathbb{N}$ so that $d_\infty(f_n, f_m) < \epsilon/2$ for all $m, n > N$. Then for all $x \in [a, b]$

$$\begin{aligned} |f_n(x) - f(x)| &\leq |f_n(x) - f_m(x)| + |f_m(x) - f(x)| \\ &\leq d_\infty(f_n, f_m) + |f_m(x) - f(x)| \\ &< \epsilon/2 + |f_m(x) - f(x)|. \end{aligned}$$

By the result of step 1, we know that $\lim_{m\to\infty} |f_m(x) - f(x)| = 0$. Hence, by taking limits as $m \to \infty$ on both sides of the last inequality we deduce that for all $x \in [a, b]$

$$|f_n(x) - f(x)| \leq \epsilon/2 < \epsilon,$$

and so the sequence (f_n) converges uniformly to f, and this is equivalent to the claim of Step 2.

Step 3 f is continuous.

Let $x \in [a, b]$ and choose $\epsilon > 0$. By the result established in Step 2, we can choose n large enough that $|f_n(x) - f(x)| < \frac{\epsilon}{3}$ for all $x \in [a, b]$, and since f_n is continuous at x we can find $\delta > 0$ so that $|f_n(x) - f_n(y)| < \frac{\epsilon}{3}$ whenever $|x - y| < \delta$. Then for $|x - y| < \delta$ we have

$$\begin{aligned} |f(x) - f(y)| &\leqslant |f(x) - f_n(x)| + |f_n(x) - f_n(y)| + |f_n(y) - f(y)| \\ &< \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon. \end{aligned}$$

Hence f is continuous.

The result now follows by combining the conclusions in Steps 2 and 3. $\square$

As you see, most of our main examples of metric spaces, $\mathbb{R}$, $\mathbb{R}^n$, $C[0,1]$, are complete (as long as we use the right metric!). This will be very useful indeed, when we study the main application of the general theory to convergence of iterations, in the next chapter.

6 The Contraction Mapping Principle.

In this chapter we apply the theory developed so far to solve equations of the form $f(x) = x$, that is to find fixed points of f, by picking an initial value x_0 and iterating

$$x_{n+1} = f(x_n)$$

to obtain a sequence (x_n). If this sequence has a limit x and f is continuous, then x is indeed a fixed point of f, because

$$f(x) = f(\lim_{n\to\infty} x_n) = \lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} x_{n+1} = x.$$

The question is whether we can find conditions to tell us when this idea will actually work! In this chapter we'll prove a powerful theorem that tells us some circumstances where this happens in general, and we'll apply this widely.

6.1 Contractions

We would like a condition on a function $f : X \to X$ which guarantees that the iterative method will lead to a unique solution of $x = f(x)$. Here is one possibility that doesn't quite work:

$$d(f(x), f(y)) < d(x, y) \qquad (*)$$

for all x, y distinct in X (i.e. for all $x \neq y$). This tells us that if we apply f to a fixed point x and some other point y then $f(y)$ is closer to x than y is.

Example 6.1. The functions $\cos : \mathbb{R} \to \mathbb{R}$ and $\sin : \mathbb{R} \to \mathbb{R}$ satisfy property $(*)$. We use the standard fact that $|\sin\theta| < |\theta|$ whenever $\theta \neq 0$. If $x \neq y$, then:

$$\begin{aligned} |\cos x - \cos y| &= \left|2\sin\left(\tfrac{x-y}{2}\right)\sin\left(\tfrac{x+y}{2}\right)\right| \\ &\leqslant 2\left|\sin\left(\tfrac{x-y}{2}\right)\right| \\ &< 2\tfrac{|x-y|}{2} \\ &= |x-y|. \end{aligned}$$

Also

$$\begin{aligned} |\sin x - \sin y| &= |\cos(\frac{\pi}{2} - x) - \cos(\frac{\pi}{2} - y)| \\ &< |(\frac{\pi}{2} - x) - (\frac{\pi}{2} - y)| \\ &= |x - y|. \end{aligned}$$

If we iterate $x_{n+1} = \cos(x_n)$, starting with $x_0 = 1$, then we find that from x_{20} onwards, all the x_n's have value 0.739 to three decimal places. It seems that in this case the iteration does converge to a solution of $x = \cos x$.

Example 6.2. There are functions $f(x)$ which do satisfy $(*)$ for all distinct x, y, but which do not have a fixed point. Let $f\colon [1, \infty) \to [1, \infty)$ be such that $f(x) = x + \frac{1}{x}$ for all $x, y \in [1, \infty)$. For $x, y \in [1, \infty)$ with $x \neq y$,

$$\begin{aligned} |f(x) - f(y)| &= \left| x + \frac{1}{x} - \left(y + \frac{1}{y} \right) \right| \\ &= \left| (x - y) + \left(\frac{1}{x} - \frac{1}{y} \right) \right| \\ &= \left| (x - y) + \left(\frac{y - x}{xy} \right) \right| \\ &= |x - y| \left| 1 - \frac{1}{xy} \right|. \end{aligned}$$

Since $x, y \geqslant 1$ and $x \neq y$, we have $0 < 1 - \frac{1}{xy} < 1$. Therefore $|f(x) - f(y)| < |x - y|$ if $x \neq y$.

A fixed point x for f would satisfy $x = x + \frac{1}{x}$ which is impossible as $\frac{1}{x}$ is never 0.

Definition 6.3. Let $f\colon X \to X$ be a function on the metric space (X, d). Then f is a *contraction* if there exists a constant $0 \leqslant k < 1$ such that

$$d(f(x), f(y)) \leqslant k d(x, y)$$

for all $x, y \in X$.

Examples 6.4. (i) If $X \subseteq \mathbb{R}$, $f : X \to X$ satisfies $(*)$, $0 < k < 1$ and $kf(x) \in X$ for all $x \in X$ then $kf : X \to X$ is clearly a contraction. For example, for $0 < k < 1$, the function $g : \mathbb{R} \to \mathbb{R}$ given by $g(x) = k \cos x$ is a contraction.

(ii) Fix k with $\frac{1}{2} \leq k < 1$ and let $f(x) = (x + \frac{1}{x})$. Then $g := kf$ defines a function $g : [1, \infty) \to [1, \infty)$ (see Chapter 6 Problem 1), and it is a contraction as f satisfies $(*)$.

(iii) Let $f : \mathbb{R} \to \mathbb{R}$ be given by $f(x) = \frac{2}{3} \cos x + \frac{1}{5} \sin x$. Then

$$\begin{aligned} & |f(x) - f(y)| \\ = \ & \left|\tfrac{2}{3}(\cos x - \cos y) + \tfrac{1}{5}(\sin x - \sin y)\right| \\ \leqslant \ & \left|\tfrac{2}{3}(\cos x - \cos y)\right| + \left|\tfrac{1}{5}(\sin x - \sin y)\right| \\ \leqslant \ & \tfrac{2}{3}|x - y| + \tfrac{1}{5}|x - y| \ \text{ (as cos, sin satisfy (*))} \\ = \ & \tfrac{13}{15}|x - y|. \end{aligned}$$

So f is a contraction with $k = \frac{13}{15}$.

Proposition 6.5. *Let X be a metric space and let $f : X \to X$ be a contraction. Then f is continuous.*

Proof. Let $x \in X$ and let (x_n) be a sequence that converges to x in X. Then $d(x_n, x) \to 0$. For all n,

$$d(f(x_n), f(x)) \leq kd(x_n, x), \text{ where } k < 1,$$

so, by the algebra of limits and the Sandwich Rule, $d(f(x_n), f(x)) \to 0$, that is, $f(x_n) \to f(x)$. Hence f is continuous. □

6.2 The Contraction Mapping Principle

Theorem 6.6 (Contraction Mapping Principle)**.** *Let $f : X \to X$ be a contraction of the complete metric space (X, d). Then f has a unique fixed point.*

Proof. Let $0 \leq k < 1$ be such that

$$d(f(x), f(y)) \leqslant kd(x, y)$$

for all $x, y \in X$.

If $k = 0$, then $d(f(x), f(y)) = 0$ for all $x, y \in X$, and so $f(x) = f(y)$. Hence there exists $x' \in X$ such that $f(x) = x'$ for all $x \in X$. But then $f(x') = x'$ and x' is the required fixed point.

If $0 < k < 1$, we choose $x_0 \in X$, and define (x_n) recursively by iterating

$$x_{n+1} = f(x_n).$$

We will prove that (x_n) is a Cauchy sequence – as X is complete, the sequence will converge.

We have for all $n = 0, 1, 2, \ldots$

$$d(x_n, x_{n+1}) \leqslant kd(x_{n-1}, x_n) \leqslant k^2 d(x_{n-2}, x_{n-1}) \leqslant \cdots \leqslant k^n d(x_0, x_1).$$

Now let $m > n$. Then

$$\begin{aligned} d(x_n, x_m) &\leqslant d(x_n, x_{n+1}) + d(x_{n+1}, x_{n+2}) + \cdots + d(x_{m-1}, x_m) \\ &\leqslant (k^n + k^{n+1} + \cdots + k^{m-1}) d(x_0, x_1) \\ &\leqslant [k^n + k^{n+1} + \ldots] d(x_0, x_1) \\ &= \frac{k^n}{1-k} d(x_0, x_1). \end{aligned}$$

Since $k < 1$, $\frac{k^n}{1-k} \to 0$ by the algebra of limits and 2.5(iii)(b). Hence, given $\varepsilon > 0$, there exists N such that whenever $m > n > N$,

$$d(x_n, x_m) \leqslant \frac{k^n}{1-k} d(x_0, x_1) < \varepsilon.$$

Thus (x_n) is Cauchy, and since X is complete, $x_n \to x$ for some $x \in X$.

Let's now see that x is a fixed point. Since the sequence (x_n) converges to x, the subsequence (x_{n+1}) also converges to x. But, as f is continuous (by 6.5),

$$x_{n+1} = f(x_n) \to f(x)$$

so $x = f(x)$.

Now we'll show that the fixed point is unique. Suppose that both x and x' are fixed points. Then

$$d(x, x') = d(f(x), f(x')) \leqslant kd(x, x').$$

As $k < 1$, the only way this can happen is if $d(x, x') = 0$, and so $x = x'$. $\square$

Remark 6.7. The proof of the theorem tells us how to find x to any degree of approximation and gives information on *how quickly* the sequence $(f^n(x_0))$, converges to x.

From the proof, we have, for $m > n$,

$$d(x_m, x_n) \leq \frac{k^n}{1-k} d(x_0, x_1).$$

Fixing n, letting $m \to \infty$ and applying Theorem 4.11 (or use Problem 10) we get $d(x_m, x_n) \to d(x, x_n)$. Applying Proposition 3.2 to $d(x_m, x_n) - \frac{k^n}{1-k} d(x_0, x_1)$, we get

$$d(x, x_n) \leq \frac{k^n}{1-k} d(x_0, x_1).$$

Example 6.8. Show that there exist unique $x, y \in \mathbb{R}$ such that

$$\begin{aligned} x &= \frac{\sin y}{4} \\ y &= \frac{\sin x}{5} + 2. \end{aligned}$$

Solution

Define

$$\begin{aligned} f \colon \mathbb{R}^2 &\to \mathbb{R}^2 \\ (x, y) &\longmapsto \left(\frac{\sin y}{4}, \frac{\sin x}{5} + 2 \right) \end{aligned}$$

A point $(x, y) \in \mathbb{R}^2$ is fixed by f if and only if

$$\begin{aligned} x &= \frac{\sin y}{4} \\ y &= \frac{\sin x}{5} + 2. \end{aligned}$$

We check that the function f is a contraction. We need to find $k < 1$ such that, for any two points (x, y) and (x', y') in $\mathbb{R}^2$, we have $d(f(x, y), f(x', y')) \leqslant kd((x, y), (x', y'))$. We use the fact, from

Example 6.1, that $|\sin x - \sin y| \leqslant |x - y|$ for all $x, y \in \mathbb{R}$. Now

$$\begin{aligned}
& d(f(x,y), f(x',y')) \\
= \;& d\left(\left(\frac{\sin y}{4}, \frac{\sin x}{5} + 2\right), \left(\frac{\sin y'}{4}, \frac{\sin x'}{5} + 2\right)\right) \\
= \;& \sqrt{\left(\frac{\sin y}{4} - \frac{\sin y'}{4}\right)^2 + \left(\frac{\sin x}{5} - \frac{\sin x'}{5}\right)^2} \\
\leqslant \;& \sqrt{\left(\frac{\sin y - \sin y'}{4}\right)^2 + \left(\frac{\sin x - \sin x'}{4}\right)^2} \\
\leqslant \;& \frac{1}{4}\sqrt{(y-y')^2 + (x-x')^2} \\
= \;& \frac{1}{4} d((x,y),(x',y')).
\end{aligned}$$

It follows that f is a contraction on $\mathbb{R}^3$ with contraction factor $k = \frac{1}{4}$. As $\mathbb{R}^2$ is complete, the Contraction Mapping Principle tells us that f has a unique fixed point $P = (x, y)$.

Here is an easy criterion for a differentiable real function to be a contraction.

Theorem 6.9 (Differential Criterion)**.** *Let $f\colon \mathbb{R} \to \mathbb{R}$ be differentiable. Then f is a contraction of $\mathbb{R}$ if and only if there exists $0 \le k < 1$ with $|f'(x)| \leqslant k$ for all $x \in \mathbb{R}$.*

Proof. Suppose f is a contraction. Then fix $x \in \mathbb{R}$, and let $h > 0$. We have

$$|f(x+h) - f(x)| \leqslant k|(x+h) - x| = k|h|,$$

and so $|\frac{f(x+h)-f(x)}{h}| \leqslant k$. If we let $h \to 0$, then this inequality becomes $|f'(x)| \leqslant k$ as required.

Conversely, suppose that $|f'(x)| \leqslant k$ for all $x \in \mathbb{R}$, and let $x > y \in \mathbb{R}$. By the mean value theorem (MAS221), there exists c between x and y such that

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

But $|f'(c)| \leqslant k$, so

$$\left|\frac{f(x) - f(y)}{x - y}\right| \leqslant k,$$

and hence f is a contraction. $\square$

Versions of Theorem 6.9 hold for the intervals $[a, b]$, $[a, \infty)$ and $(-\infty, b]$. For example if $f \colon [a, b] \to [a, b]$ is differentiable on (a, b). Then f is a contraction on $[a, b]$ if and only if there exists $k < 1$ with $|f'(x)| \leqslant k$ for all $x \in (a, b)$.

Example 6.10. Consider $f : \mathbb{R} \to \mathbb{R}$ or $f : [-\pi/2, \pi/2] \to [-\pi/2, \pi/2]$ given by $f(x) = \cos x$ so that $f'(x) = -\sin x$. This fails the criteria on $\mathbb{R}$ and $[-\pi/2, \pi/2]$ as, for any $k < 1$ there exists $x \in (-\pi/2, \pi/2)$ with $|f'(x)| > k$. So cos is not a contraction on $\mathbb{R}$ or on $[-\pi/2, \pi/2]$.

Now let $g : \mathbb{R} \to \mathbb{R}$ be given by $g(x) = f^2(x) = \cos(\cos x)$. Then $g'(x) = -\sin(\cos x)(-\sin x) = \sin x. \sin(\cos x)$ so

$$|g'(x)| = |\sin x|.|\sin(\cos x)| \leqslant |\sin(\cos x)|.$$

Note that $-1 \leqslant \cos x \leqslant 1$ for all x, and consequently, since the mapping $x \to \sin(x)$ is monotonic increasing on $[-1, 1]$, we have $-\sin 1 \leqslant \sin(\cos x) \leqslant \sin 1$. So $|\sin(\cos x)| \leqslant \sin 1 \approx 0.8415 < 1$. It follows that $|g'(x)| \leq \sin 1 < 1$. By Theorem 6.9, g is a contraction on $\mathbb{R}$.

The next result is crucial for the main application to the existence of solutions to differential equations.

Theorem 6.11. *Let X be a complete metric space, and let $f \colon X \to X$ have the property that, for some m, the iterate f^m is a contraction of X. Then f has a unique fixed point.*

Proof. Since f^m is a contraction and X is complete, f^m has a unique fixed point x, i.e., $f^m(x) = x$. Apply f to this relation, and we see that $f^{m+1}(x) = f(x)$, i.e., $f^m(f(x)) = f(x)$, so $f(x)$ is also a fixed point of f^m. By uniqueness, $f(x) = x$. Also if $f(y) = y$ then $f^m(y) = y$ so $y = x$. $\square$

Remark 6.12. Given a starting point x_0, the sequence

$$x_0, x_1, x_2, \ldots, x_m, x_{m+1}, x_{m+2}, \ldots, x_{2m}, x_{2m+1}, x_{2m+2}, \ldots,$$

is made up of m subsequences $x_0, x_m, x_{2m}, \ldots$ and $x_1, x_{m+1}, x_{2m+1}, \ldots$, etcetera, each obtained by iterating f^m. But these subsequences are iterations of f^m so they all converge to the unique fixed point x. So, in Theorem 6.11, the fixed point x can be approximated by iterating from any x_0 using f.

6.3 Applications to Differential Equations

The contraction mapping principle can be used to prove that many integral and differential equations have a solution. We look at an example and then establish a general result.

Example 6.13. Show that there is a unique solution in $C[0,1]$ of the differential equation

$$\frac{df}{dx} = (f(x) + x)x \qquad (0 \leqslant x \leqslant 1), \qquad f(0) = 0$$

Solution For $f \in C[0,1]$, let $T(f) = g$ where

$$g(x) = \int_0^x (f(u) + u)u \, du.$$

Then

$$\begin{aligned} T(f) = f \quad &\Leftrightarrow \quad f(x) = \int_0^x (f(u) + u)u \, du \\ &\Leftrightarrow \quad f'(x) = (f(x) + x)x \text{ and } f(0) = 0. \end{aligned}$$

So f is a fixed point for T if and only if it is a solution of the differential equation.

To apply the contraction mapping principle, we need to be working in a complete space, so $C[0,1]$ must be given the supremum metric.

Let's prove that T is a contraction.

$$\begin{aligned}
d_\infty(T(f_1), T(f_2)) &= \sup_{x\in[0,1]} |T(f_1)(x) - T(f_2)(x)| \\
&= \sup_{x\in[0,1]} \left| \int_0^x (f_1(u) + u)u\,du - \int_0^x (f_2(u) + u)u\,du \right| \\
&= \sup_{x\in[0,1]} \left| \int_0^x (f_1(u) - f_2(u))u\,du \right| \\
&\leqslant \sup_{x\in[0,1]} \int_0^x |f_1(u) - f_2(u)|u\,du \\
&= \int_0^1 |f_1(u) - f_2(u)|u\,du \\
&\leqslant \sup_{x\in[0,1]} |f_1(x) - f_2(x)| \cdot \int_0^1 u\,du \\
&= \sup_{x\in[0,1]} |f_1(x) - f_2(x)|/2 \\
&= d_\infty(f_1, f_2)/2.
\end{aligned}$$

So T is a contraction on the complete metric space $C([0,1])$ and the differential equation has a unique solution. To find it, we can start with any initial guess, $f_0(x) = 0$ for all $x \in [0,1]$, say, and then we can iterate the function T to get the limit. We find that $f_1 = T(f_0) = \frac{x^3}{3}$, $f_2 = T(f_1) = \frac{x^5}{15} + \frac{x^3}{3} \ldots$ and find the series solution

$$f(x) = \frac{x^3}{3} + \frac{x^5}{3.5} + \frac{x^7}{3.5.7} + \frac{x^9}{3.5.7.9} + \ldots,$$

which we can then check directly.

In fact, this method of proof works fairly generally, although we may need to use some iterate of T in general.

Theorem 6.14. *Consider the differential equation*

$$\frac{df}{dx} = \alpha(f(x), x), \quad a \leqslant x \leqslant b, \quad f(a) = \beta,$$

where $\alpha\colon \mathbb{R}\times[a,b] \to \mathbb{R}$ *is a continuous function of* y *and* x*. Suppose that there exists* $L \geq 0$ *such that*

$$|\alpha(y_1, x) - \alpha(y_2, x)| \leqslant L|y_1 - y_2|, \tag{4}$$

for all $y_1, y_2 \in \mathbb{R}$, and $x \in [a,b]$. Then the differential equation has a unique solution.

Proof. As in the example, we define $T\colon C[a,b] \to C[a,b]$ by

$$(T(f))(x) = \beta + \int_a^x \alpha(f(u), u)\, du.$$

Then f is a fixed point of T if and only if f is a solution of the differential equation satisfying the initial condition.

For two functions $f_1, f_2 \in C[a,b]$, and $x \in [a,b]$, then

$$\begin{aligned}
|(T(f_1))(x) - (T(f_2))(x)| &= \left|\left[\beta + \int_a^x \alpha(f_1(u), u)\, du\right] - \left[\beta + \int_a^x \alpha(f_2(u), u)\, du\right]\right| \\
&= \left|\int_a^x [\alpha(f_1(u), u) - \alpha(f_2(u), u)]\, du\right| \\
&\leqslant \int_a^x |\alpha(f_1(u), u) - \alpha(f_2(u), u)|\ du \\
&\leqslant \int_a^x L|f_1(u) - f_2(u)|\, du \\
&\leqslant \int_a^x L \cdot d_\infty(f_1, f_2)\, du \\
&= L \cdot d_\infty(f_1, f_2) \cdot (x - a).
\end{aligned}$$

Similarly, with $T(f_1)$, $T(f_2)$ replacing f_1, f_2:

$$\left|T^2(f_1)(x) - T^2(f_2)(x)\right| \leqslant \int_a^x L|T(f_1)(u) - T(f_2)(u)|\, du,$$

and so

$$\begin{aligned}
\left|T^2(f_1)(x) - T^2(f_2)(x)\right| &\leqslant \int_a^x L|T(f_1)(u) - T(f_2)(u)|\, du \\
&\leqslant \int_a^x L^2 d_\infty(f_1, f_2)(u - a)\, du \\
&= L^2 d_\infty(f_1, f_2) \cdot \frac{(x-a)^2}{2}.
\end{aligned}$$

Continuing in this way, we get the general result that

$$|T^m(f_1)(x) - T^m(f_2)(x)| \leqslant L^m d_\infty(f_1, f_2) \cdot \frac{(x-a)^m}{m!} \leqslant L^m d_\infty(f_1, f_2) \cdot \frac{(b-a)^m}{m!}$$

since $x \in [a, b]$. So

$$d_\infty(T^m(f_1), T^m(f_2)) \leqslant \frac{L^m(b-a)^m}{m!} d_\infty(f_1, f_2).$$

As $m \to \infty$, the coefficients $\frac{L^m(b-a)^m}{m!} \to 0$. Eventually, then, this coefficient is less than 1, and for this value of m, we conclude that T^m is a contraction, and the theorem follows from Theorem 6.11. $\square$

It turns out that the condition on α is almost automatic: one needs little more than that α is (partially) differentiable with respect to y. The theorem can be extended to simultaneous differential equations and to higher-order differential equations. For reasons of time, we will omit these applications.

Note. The constraint on the function α in (4) is called a *Lipschitz condition*. These play an important role in the theory of both ordinary and stochastic differential equations (see MAS352/452/6052 for the latter).

7 Compactness.

This chapter does not follow on from Chapter 6 but picks up earlier themes, in particular closed sets, convergence and completeness. Think back to our proof of Theorem 5.9, that $\mathbb{R}$ is complete: a Cauchy sequence is bounded, so it has a convergent subsequence, and so the whole sequence converges. The key step used the Bolzano-Weierstrass Theorem 5.8, which stated that any sequence in the interval $[a, b]$ has a convergent subsequence with limit in the set. Generalisations of this turn out to be so useful that we will isolate this property.

7.1 Compact Sets

Definition 7.1. Let $A \subseteq X$ be a subset of a metric space. We say that A is *compact* if every sequence in A has a subsequence that converges to a point of A.

Example 7.2. All closed intervals of the form $[a, b]$ are compact, by the Bolzano-Weierstrass Theorem 5.8. However, $\mathbb{R}$ is not compact because the sequence $0, 1, 2, 3, 4, \ldots$ has no convergent subsequence. Similarly intervals which are not bounded, such as $[a, \infty)$ or $(-\infty, b]$, are not compact because of the sequences $a, a+1, a+2, \ldots$ and $b, b-1, b-2, \ldots$.

Here's a way to find many compact sets:

Lemma 7.3. *Let $A \subseteq X$ be a closed subset of a compact space X. Then A is compact.*

Proof. Let (a_n) be a sequence in A. Then (a_n) is a sequence in X and, as X is compact, there is a subsequence (a_{n_k}) converging to a limit $a \in X$. But each $a_{n_k} \in A$ and A is closed so $a \in A$. Therefore A is compact. $\square$

Now we are going to prove several properties of compact sets, in order to get a feel for which sorts of set are compact. First, we've already noted, in Proposition 5.11, that complete sets are closed, so that the property of being complete is stronger than the property of being closed. The property of being compact is stronger still.

Proposition 7.4. *Let A be a compact set in a metric space. Then A is complete. In particular, A is closed.*

Proof. Let (a_n) be a Cauchy sequence in A. Since A is compact, there is a convergent subsequence (a_{n_k}), with limit $a \in A$. By Theorem 5.7, $a_n \to a$. Thus A is complete and, by Proposition 5.11, it is closed in X. $\square$

Definition 7.5. A subset A of a metric space (X, d) is *bounded* if there is a $D \geqslant 0$ such that $d(a, b) \leqslant D$ for all $a, b \in A$.

For example, in $\mathbb{R}$, $[a, b]$ is bounded with $D = b - a$.

Lemma 7.6. *A subset A of $\mathbb{R}^k$ is bounded in the sense of Definition 7.5 if and only if it is bounded in the usual sense that there exists $M > 0$ such that $d(a, 0) \leqslant M$ for all $a \in A$.*

Proof. If $d(a, 0) \leqslant M$ for all $a \in A$ then, for all $a, b \in A$,

$$d(a, b) \leq d(a, 0) + d(0, b) \leq 2M$$

so the condition of Definition 7.5 holds with $D = 2M$.

Suppose that the condition of Definition 7.5 holds, fix $b \in A$ (if A is empty the result is trivial), and let $m = d(b, 0)$. Then for all $a \in A$,

$$d(a, 0) \leqslant d(a, b) + d(b, 0) \leqslant D + m$$

so A is bounded in the usual sense with $M = D + m$. $\square$

Proposition 7.7. *Let A be a compact subset of a metric space (X, d). Then A is bounded.*

Proof. Suppose not. Then for every $D \geqslant 0$ the condition

$$d(a, b) \leqslant D \text{ for all } a, b \in A$$

fails. Thus, taking $D = 1, 2, \ldots$, we have sequences $a_1, a_2, \ldots$, $b_1, b_2 \ldots$ in A such that $d(a_i, b_i) > i$ for all $i = 1, 2, \ldots$. Since A

is compact, there is a convergent subsequence (a_{m_k}) of (a_n) converging to $a \in A$. Then (b_{m_k}) has a convergent subsequence (b_{n_k}), converging to $b \in A$. Thus (a_{n_k}), as a subsequence of (a_{m_k}), converges to a and (b_{n_k}) converges to $b \in A$. Now for all $k \in \mathbb{N}$,

$$n_k < d(a_{n_k}, b_{n_k}) \leq d(a_{n_k}, a) + d(a, b) + d(b_{n_k}, b).$$

Choosing $\epsilon = 1/2$, since (a_{n_k}) converges to a and (b_{n_k}) converges to b, we can find $K \in \mathbb{N}$ so that for all $k > K$, $d(a_{n_k}, a) < 1/2$ and $d(b_{n_k}, b) < 1/2$. Then for all $k > K$, $k \leqslant n_k < d(a, b) + 1$. This is impossible and we have a contradiction. □

We now know that compact subsets of a metric space X are closed and bounded. When $X = \mathbb{R}^k$, these two conditions are enough to guarantee that the set is compact:

Theorem 7.8 (Heine-Borel)**.** *A subset K of $\mathbb{R}^k$ with the Euclidean metric is compact if and only if it is closed and bounded.*

Proof. If K is compact then it is closed, by Proposition 7.4, and bounded, by Proposition 7.7. Conversely, suppose that K is closed and bounded. We shall show that K is compact. Let $(\mathbf{x_n})$ be a sequence in K and write $\mathbf{x_n} = (x_n^1, \ldots, x_n^k)$. Since K is bounded, there is, by Lemma 7.6, an $M > 0$ such that

$$|x_n^i| \leqslant \sqrt{(x_n^1)^2 + \cdots + (x_n^k)^2} \leqslant M$$

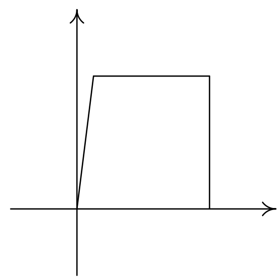
for each $i = 1, \ldots, k$ and for all $n \in \mathbb{N}$, and therefore each co-ordinate gives us a bounded sequence (x_n^i) of real numbers. Use Bolzano-Weierstrass successively in the following way: first find a subsequence of $(\mathbf{x_n})$ so that the first coordinates give a sequence converging to x^1, say. Then find a subsequence of this new sequence so that the second co-ordinates give a sequence converging to x^2, say. Continue until the k-th coordinate. This process gives us a subsequence $(\mathbf{x_{n_r}})$ with $x_{n_r}^i \to x^i$ for each $i = 1, \ldots, k$. By the d_2-version of Proposition 2.9, $\mathbf{x_{n_r}} \to \big(x^1, \ldots, x^k\big)$. As K is closed, $\big(x^1, \ldots, x^k\big) \in K$. Thus K is compact. □

It is worth noting that the same result is false for some other spaces:

Example 7.9. Let $A = \{f \in C[0,1]; f([0,1]) \subseteq [0,1]\}$ and use the supremum metric d_∞. Then, for $f, g \in A$, $d_\infty(f,g) \leq 1$. Thus A is bounded. Also A is closed because if $g_n \to g$ where each $g_n \in A$ then, for $x \in [0,1]$, $0 \leqslant g_n(x) \leqslant 1$ from which it follows that $0 \leqslant g(x) \leqslant 1$ and $g \in A$.

Consider the sequence of functions $f_n \in A$ defined by

$$f_n(t) = \begin{cases} 2^n t, & \text{if } 0 \leqslant t \leqslant \frac{1}{2^n} \\ 1, & \text{if } \frac{1}{2^n} \leqslant t \leqslant 1 \end{cases}$$



If $m > n$ then $f_m(\frac{1}{2^{n+1}}) = 1$, as $m \geq n+1$ so $\frac{1}{2^{n+1}} \geqslant \frac{1}{2^m}$, but $f_n(\frac{1}{2^{n+1}}) = \frac{1}{2}$, as $\frac{1}{2^{n+1}} \leqslant \frac{1}{2^n}$. So

$$d_\infty(f_n, f_m) \geq \left| f_n\left(\frac{1}{2^{n+1}}\right) - f_m\left(\frac{1}{2^{n+1}}\right) \right| = \frac{1}{2}$$

whenever $m \neq n$.

No subsequence of f_n can be Cauchy, because any two terms are at least $\frac{1}{2}$ apart, so no subsequence can be convergent. Hence A is not compact although it is closed and bounded.

7.2 Compactness and Continuity

We've already seen the definition of continuity. We'll now see how the property of being compact behaves in relation to continuous maps.

Theorem 7.10. *Let $f\colon X \to Y$ be a continuous map between metric spaces, and let $K \subseteq X$ be compact. Then $f(K)$ is compact.*

Proof. Let (y_n) be any sequence in $f(K)$. For each n, there is some $x_n \in K$ with $f(x_n) = y_n$. As K is compact, there is a subsequence $x_{n_r} \to x$ for some $x \in K$. As f is continuous, $y_{n_r} = f(x_{n_r})$ converges to $f(x) \in f(K)$. Thus (y_n) has a convergent subsequence, with limit in $f(K)$, and so $f(K)$ is compact. $\square$

Corollary 7.11. *A function f which is real-valued and continuous on a compact set K is bounded on K and attains its bounds.*

Proof. From Theorem 7.10, $f(K)$ is compact in $\mathbb{R}$, and so is closed and bounded. Thus f is a bounded function. Let $M = \sup\{f(x) : x \in K\}$. For each $n \in \mathbb{N}$, there exists $x_n \in K$ with $f(x_n) > M - \frac{1}{n}$. Then $f(x_n) \to M$ and, because $f(K)$ is closed, $M \in f(K)$. This means that there is some $x_M \in K$ with $f(x_M) = M$, so that the bound M is attained. A similar argument works for the lower bound $m = \inf\{f(x) : x \in K\}$. $\square$

For example, if $f\colon [a,b] \to \mathbb{R}$ is continuous, then the image of f is a closed interval $[c,d]$ for some $c, d \in \mathbb{R}$. Here $c = \inf f([a,b])$ and $d = \sup f([a,b])$ are attained and, by the Intermediate Value Theorem, all points in $[c,d]$ are attained.

The corollary implies that continuous real-valued functions on a compact set are automatically bounded. Note that this is false for non-compact sets. For example, if $X = \mathbb{R}$, then the function $f(x) = x$ is continuous, but is not bounded.

Uniform Continuity Recall the first definition of continuity. A function $f\colon X \to Y$ between two metric spaces (X, d_X) and (Y, d_Y) is *continuous* if for all $x \in X$ and for all $\epsilon > 0$, there exists $\delta > 0$ such that $f(B(x,\delta)) \subseteq B(f(x), \epsilon)$. We could demand more, namely that the same δ works for all $x \in X$.

Definition 7.12. A function $f\colon X \to Y$ between metric spaces (X, d_X) and (Y, d_Y) is *uniformly continuous* if for all $\epsilon > 0$, there exists $\delta > 0$ such that $f(B(x,\delta)) \subseteq B(f(x), \epsilon)$ for all $x \in X$.

Clearly any uniformly continuous function is automatically continuous. However, not every continuous function is uniformly continuous:

Example 7.13. Consider $f\colon (0,1] \to [1,\infty)$, $f(x) = \frac{1}{x}$. Then

$$|f(\tfrac{1}{2^{n+1}}) - f(\tfrac{1}{2^n})| = 2^{n+1} - 2^n = 2^n.$$

If we set $\epsilon = 1$, then for any choice of $\delta > 0$, choose n large enough so that $\frac{1}{2^{n+1}} < \delta$ and let $x = \frac{1}{2^n}$. Then $|\frac{1}{2^{n+1}} - x| = \frac{1}{2^{n+1}} < \delta$ but $|f(\frac{1}{2^{n+1}}) - f(x)| = 2^n \geq \epsilon$. Thus $f(B(x,\delta)) \not\subseteq B(f(x),\epsilon)$ and f is not uniformly continuous.

One of the good things about compact spaces is that any continuous function from a compact space must be uniformly continuous.

Theorem 7.14. *Let $f\colon X \to Y$ be a continuous function and suppose that X is compact. Then f is uniformly continuous.*

Proof. Suppose on the contrary that f is not uniformly continuous. Then there exists $\epsilon > 0$ such that, for each $\delta > 0$, we can find $a, b \in X$ with $d_X(a,b) < \delta$ but $d_Y(f(a), f(b)) \geqslant \epsilon$. Taking $\delta = 1/n$, we can thus find sequences $(a_n), (b_n)$ in X such that $b_n \in B(a_n, 1/n)$ but $f(b_n) \notin B(f(a_n), \epsilon)$.

As X is compact, as in the proof of Prop 7.7, we can find (n_k) such that $a_{n_k} \to a$ and $b_{n_k} \to b$ for some $a, b \in X$.

We must have $a = b$ because

$$\begin{aligned} d_X(a,b) &\leqslant d_X(a, a_{n_k}) + d_X(a_{n_k}, b_{n_k}) + d_X(b_{n_k}, b) \\ &< d_X(a, a_{n_k}) + 1/n_k + d_X(b_{n_k}, b) \to 0. \end{aligned}$$

Since f is continuous at a, we get $f(a_{n_k}) \to f(a)$ and $f(b_{n_k}) \to f(a)$. But

$$\begin{aligned} 0 &< \epsilon \\ &\leqslant d_Y(f(a_{n_k}), f(b_{n_k})) \\ &\leqslant d_Y(f(a_{n_k}), f(a)) + d_Y(f(a), f(b_{n_k})) \\ &\to 0, \end{aligned}$$

which is a contradiction. $\square$

7.3 Equivalent Formulations of Compactness

Definition 7.15. Let X be a metric space. A collection $\{U_i : i \in I\}$ of subsets of X is *a cover of* $E \subset X$, or *covers* $E \subset X$, if

$$E \subseteq \bigcup_{i \in I} U_i.$$

If the indexing set I is a finite set then $\{U_i : i \in I\}$ is a *finite cover.* If each of the U_i is an open set then the collection is an *open cover.* If $\{U_i : i \in I\}$ is a cover for E, then a finite collection $U_{i_1}, \dots, U_{i_n}$ with $i_1, \dots, i_n \in I$ is called a *finite subcover* of E if it is itself a finite cover, i.e. if $E \subseteq U_{i_1} \cup \dots \cup U_{i_n}$.

We can break up the closed interval $[0, 1]$ into two intervals of length a half, three intervals of length a third, and so on. Moving on to metric spaces we might wonder if the whole space can be covered by a finite number of balls of any fixed radius. Clearly this is not always going to be the case: There is no way of writing $\mathbb{R}$ as a finite union of intervals of length one. We isolate this property in the following definition.

Definition 7.16. A metric space (X, d) is *totally bounded* if for each $\epsilon > 0$ there is a finite collection of open balls of radius ϵ which cover X.

Proposition 7.17. *A compact metric space is totally bounded.*

Proof. We prove this by contradiction. So let (X, d) be a compact metric space which is not totally bounded. So there is a positive number ϵ such that no finite collection of open balls of radius ϵ covers X.

Take any element $x_1 \in X$. The open ball $B(x_1, \epsilon) \neq X$. So there exists $x_2 \in X$ such that $x_2 \notin B(x_1, \epsilon)$. Also $B(x_1, \epsilon) \cup B(x_2, \epsilon) \neq X$ so there exists $x_3 \in X$ with $x_3 \notin B(x_1, \epsilon) \cup B(x_2, \epsilon)$. Continuing this process, we get a sequence $x_1, x_2, \dots$ in X with the property that

$$x_{n+1} \notin B(x_1, \epsilon) \cup \dots \cup B(x_n, \epsilon)$$

for all $n \geqslant 1$. Whenever $m > n$, $d(x_m, x_n) \geqslant \epsilon$, because $x_m \notin B(x_n, \epsilon)$ and so the sequence (x_n) clearly cannot have a subsequence which is Cauchy. As all convergent sequences are Cauchy, the sequence (x_n) cannot have a convergent subsequence, and this contradicts compactness. $\square$

Definition 7.18. A metric space X is said to have the *Heine–Borel property* if every open cover of X has a finite subcover. That is, X has the Heine-Borel property if for any open cover $\{U_i : i \in I\}$ there are finitely many indices $i_1, \dots, i_n$ such that $U_{i_1} \cup \cdots \cup U_{i_n} = X$.

Proposition 7.19. *Let X be a totally bounded, complete metric space. Then X has the Heine–Borel property.*

Proof. Suppose, for a contradiction, X doesn't have the Heine–Borel property. So there is an open cover $\{U_i : i \in I\}$ of X without any finite subcover. For the purposes of this proof, call a subset E of X *nice* if it is contained in a finite union of the U_i's and *nasty* otherwise. So X is nasty. Let $\epsilon > 0$. Since X is totally bounded we must have $X = B(y_1, \epsilon) \cup \cdots \cup B(y_n, \epsilon)$ for some $y_1, \dots, y_n \in X$ depending on ϵ. Thus

$$E = E \cap X = (E \cap B(y_1, \epsilon)) \cup \cdots \cup (E \cap B(y_n, \epsilon)).$$

Now let E be nasty. If each $E \cap B(x, \epsilon)$ is nice for all $x \in X$, which of course includes the y_i's, then E is nice. Therefore $E \cap B(x, \epsilon)$ is nasty for some $x \in X$.

In particular, taking $E = X$ and $\epsilon = 1$, $X \cap B(x_1, 1) = B(x_1, 1)$ is nasty for some $x_1 \in X$. Then, taking $E = B(x_1, 1)$ and $\epsilon = \frac{1}{2}$, $B(x_1, 1) \cap B(x_2, \frac{1}{2})$ is nasty for some $x_2 \in X$. Continuing in this way, we get a sequence (x_n) in X such that

$$E_n := B(x_1, 1) \cap B(x_2, 1/2) \cap \cdots \cap B(x_n, 1/n)$$

is nasty. Clearly E_n is non-empty because $\emptyset$ is nice. So, for each $n \in \mathbb{N}$, we may choose $a_n \in E_n$.

Let $N \in \mathbb{N}$. For $m, n \geqslant N$ we have $a_m, a_n \in E_N \subset B(x_N, 1/N)$. So

$$d(a_m, a_n) \leqslant d(a_m, x_N) + d(a_n, x_N) \leqslant \frac{2}{N} \to 0 \text{ as } N \to \infty,$$

and so (a_n) is Cauchy. As X is complete, $a_n \to a$ for some $a \in X$.

Now $a \in U_i$ for some $i \in I$. As U_i is open, we can find $r > 0$ such that $B(a, r) \subseteq U_i$. Since $a_n \to a$, we must have $d(a_n, a) < r/2$ for n large enough. In fact we can find an n such that $2/n < r/2$ and $d(a_n, a) < r/2$. Now let $e \in E_n$. Then $d(x_n, e) < \frac{1}{n}$ and $d(a_n, x_n) < \frac{1}{n}$ so $d(e, a_n) < \frac{2}{n} < r/2$. As $d(a_n, a) < r/2$, $d(e, a) < r/2 + r/2 = r$ so $e \in B(a, r) \subseteq U_i$. Thus $E_n \subseteq U_i$ is nice, contradicting the earlier deduction that it is nasty. So the result is true. □

Theorem 7.20. *Let (X, d) be a metric space. The following are equivalent:*

a. X is compact.

b. X is totally bounded and complete.

c. X has the Heine–Borel property.

Proof. We shall show the equivalence by showing that $a \implies b \implies c \implies a$. The first implication $a \implies b$ is Proposition 7.17 and Proposition 7.4 put together. The implication $b \implies c$ is Proposition 7.19.

It remains to show $c \implies a$. So let (X, d) be a metric space possessing the Heine–Borel property, and let (a_n) be a sequence in X without any convergent subsequence. Let $x \in X$, and first suppose that for all $\delta > 0$, there are infinitely many values of n such that $a_n \in B(x, \delta)$. Then we can find a subsequence $n_1 < n_2 < n_3 < n_4 \dots$ such that each $a_{n_k} \in B(x, \frac{1}{k})$. But then $a_{n_k} \to x$, which is impossible. So for all $x \in X$, there exists $\delta_x > 0$ such that $a_n \in B(x, \delta_x)$ for only a finite number of values of n. The collection $B(x, \delta_x)$, for $x \in X$, gives an open cover of X and so must have a finite subcover $B(x_i, \delta_{x_i})$, $1 \leq i \leq m$ for some m. But each $B(x_i, \delta_{x_i})$ contains a_n for only finitely many values of n, so the finite cover would exclude a_n for infinitely many n. This is our desired contradiction. □