F1.3YE2/F1.3YK3

ALGEBRA AND ANALYSIS

Part 1: ANALYSIS.

THEORY OF METRIC SPACES

LECTURE NOTES AND EXERCISES

Contents

Chapter 1

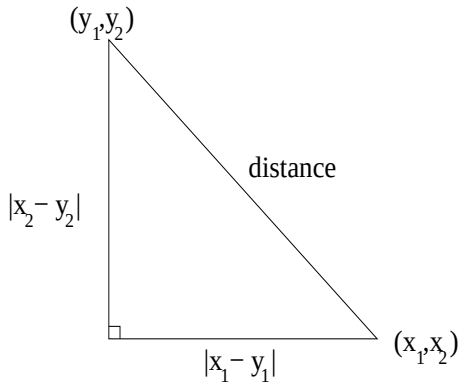
Introduction to Metric Spaces

1.1 Ways of measuring distance

This course is concerned with the notion of 'distance'. There are many different ways in which we can define and measure distance in different circumstances. Here are some examples.

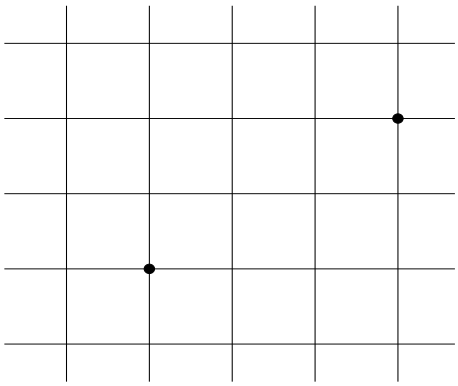
1. In the Euclidean plane $\mathbb{R}^2$, we can define the distance between two points $\underline{x} = (x_1, x_2)$ and $\underline{y} = (y_1, y_2)$ to be the length of the straight line segment joining these points, in other words (by Pythagoras' theorem)

$$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}.$$



   This is called the *usual* or *euclidean* distance on $\mathbb{R}^2$.

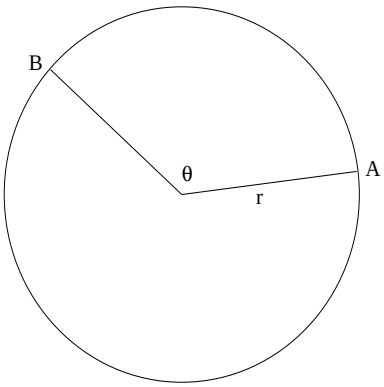
2. Euclidean distance may not always be the appropriate way to measure distance. For example, some cities are laid out in a rectangular 'Manhattan' grid. It is not feasible to travel in a straight line between two points, unless they are connected by a North-South or East-West street.



Thus, to reach a point 2 units to the North and 3 to the East of your starting point, you must travel $2+3=5$ units, rather than $\sqrt{2^2+3^2}=\sqrt{13}\sim 3.61$ units. More generally, we can define the 'Manhattan' or 'taxi-cab' distance between two points $\underline{x}=(x_1,x_2)$ and $\underline{y}=(y_1,y_2)$ in $\mathbb{R}^2$ to be

$$|x_1-y_1|+|x_2-y_2|.$$

3. Suppose that we are on a circular path around a pond. To reach another point on the path by a straight line, we would have to walk through water, which would ruin a perfectly good pair of shoes. Realistically, we should walk around the path, so the distance we will travel will be $r\theta$, where r is the radius of the pond, and $\theta\in[0,\pi]$ is the angle subtended by the two points concerned at the centre of the pond.



4. Data in computer memories, or on CD or DVD, is in binary format, that is long strings of binary digits 0 and 1:

$$1011\ 0110\ 1110\ 0010\ \cdots.$$

Sometimes data gets corrupted by electro-magnetic interference, for example while being transmitted along a wire from one computer to another. We can

measure the extent of the damage by counting the number of 'bits' that have changed from 0 to 1 or *vice versa*. This is an appropriate notion of 'distance' between two pieces of binary data.

For example, using the standard ASCII encoding system, a capital letter A is stored as the 8-bit string 01000001, B as 01000010, C as 01000011, etc. So the 'distance' between A and B is 2, while the 'distance' between A and C is 1.

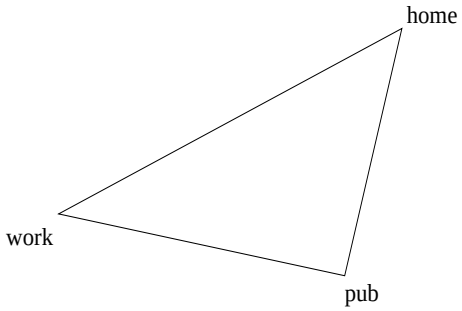
1.2 Metrics

These examples of distance functions all satisfy some obvious properties, which intuition suggests should be common to *all* distance functions:

1. The distance between two *distinct* points should be a positive real number. The distance from any point to itself should be 0.

2. Distance should be *symmetric* – that is, the distance from A to B should be the same as the distance from B to A.

3. The *triangle inequality* should hold: the distance from A to C should be no greater than the sum of the distances from A to B and from B to C. (It is quicker to go straight home from work than it is to go via the pub.)



We make these properties into a formal definition as follows.

Definitions. Let X be a set. A *metric*, or *distance function*, on X is a function $d : X \times X \to \mathbb{R}$ satisfying the following three axioms:

1. $d(x, y) \geq 0$ for all $x, y \in X$, with $d(x, y) = 0 \iff x = y$;

2. $d(y, x) = d(x, y)$ for all $x, y \in X$;

3. (The triangle inequality) $d(x, z) \leq d(x, y) + d(y, z)$ for all $x, y, z \in X$.

(The real number $d(x, y)$ should be interpreted as 'the distance from x to y.)

A *metric space* (X, d) consists of a set X together with a metric d on X.

1.3 Examples of metrics

1. The *euclidean* or *usual* metric on $\mathbb{R}$ is given by $d(x,y) = |x-y|$. Let us check the axioms for a metric:

   Firstly, for any $t \in \mathbb{R}$ we have $|t| \geq 0$ with $|t| = 0 \iff t = 0$. In particular, for any $x, y \in \mathbb{R}$, we have $|x-y| \geq 0$, with $|x-y| = 0 \iff x - y = 0 \iff x = y$.

   Secondly, $|y-x| = |-(x-y)| = |x-y|$ for any $x, y \in \mathbb{R}$.

   Finally, suppose $x, y, z \in \mathbb{R}$. If $x \leq y \leq z$ or $z \leq y \leq x$ then $|x-z| = |x-y| + |y-z|$. Otherwise $|x-z| < |x-y| + |y-z|$. Hence the triangle inequality is satisfied for all $x, y, z \in \mathbb{R}$.

2. The *euclidean metric* on $\mathbb{R}^2$ is defined by

$$d(\underline{x}, \underline{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2},$$

   where $\underline{x} = (x_1, x_2)$ and $\underline{y} = (y_1, y_2)$. Again, to prove that this is a metric, we should check the axioms.

   Firstly, $d(\underline{x}, \underline{y}) \geq 0$ by the convention that the square root of a positive real number is taken to be the positive square root unless we are explicitly told otherwise. Moreover, $d(\underline{x}, \underline{y}) = 0$ if and only if $(x_1 - y_1)^2 = (x_2 - y_2)^2 = 0$, which happens precisely when $x_1 = y_1$ and $x_2 = y_2$, in other words when $\underline{x} = \underline{y}$.

   Secondly, since $(y_1 - x_1)^2 = (x_1 - y_1)^2$ and $(y_2 - x_2)^2 = (x_2 - y_2)^2$, it follows that $d(\underline{y}, \underline{x}) = d(\underline{x}, \underline{y})$.

   The hard part is the triangle inequality. Suppose that $\underline{x} = (x_1, x_2)$, $\underline{y} = (y_1, y_2)$ and $\underline{z} = (z_1, z_2)$ are three points in $\mathbb{R}^2$. Let $\alpha = x_1 - y_1$, $\beta = y_1 - z_1$, $\gamma = x_2 - y_2$ and $\delta = y_2 - z_2$. Then

$$\begin{aligned} d(\underline{x}, \underline{z})^2 &= (x_1 - z_1)^2 + (x_2 - z_2)^2 = (\alpha + \beta)^2 + (\gamma + \delta)^2 \\ &= \alpha^2 + \beta^2 + \gamma^2 + \delta^2 + 2(\alpha\beta + \gamma\delta), \end{aligned}$$

   while

$$\begin{aligned} \big(d(\underline{x}, \underline{y}) + d(\underline{y}, \underline{z})\big)^2 &= d(\underline{x}, \underline{y})^2 + d(\underline{y}, \underline{z})^2 + 2d(\underline{x}, \underline{y})d(\underline{y}, \underline{z}) \\ &= \alpha^2 + \gamma^2 + \beta^2 + \delta^2 + 2\sqrt{(\alpha^2 + \gamma^2)(\beta^2 + \delta^2)}. \end{aligned}$$

   Hence the triangle inequality will follow if we can show that

$$(\alpha\beta + \gamma\delta)^2 \leq (\alpha^2 + \gamma^2)(\beta^2 + \delta^2).$$

   But

$$0 \leq (\alpha\delta - \beta\gamma)^2 = \alpha^2\delta^2 + \beta^2\gamma^2 - 2\alpha\beta\gamma\delta = (\alpha^2 + \gamma^2)(\beta^2 + \delta^2) - (\alpha\beta + \gamma\delta)^2.$$

   Hence d really is a metric on $\mathbb{R}^2$.

3. In a similar way, we can define the *euclidean metric* on $\mathbb{R}^n$ for any natural number n by

$$d(\underline{x}, \underline{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2},$$

where $\underline{x} = (x_1, \ldots, x_n)$ and $\underline{y} = (y_1, \ldots, y_n)$.

The proof that this is a metric follows the same pattern as the case $n = 2$ given in the previous example.

Note also that for $n = 1$ we have the metric defined in the first example, since $\sqrt{(x-y)^2} = |x - y|$.

4. The *taxi-cab metric*, or the *Manhattan metric* on $\mathbb{R}^n$ is defined by

$$d(\underline{x}, \underline{y}) = \sum_{i=1}^{n} |x_i - y_i|.$$

Once again, to prove that this is a metric, one needs to check the three axioms. The first two are easy, and I will omit the details. For the third, note that, for each i, we have $|x_i - z_i| \leq |x_i - y_i| + |y_i - z_i|$ (by the triangle inequality applied to the euclidean metric on $\mathbb{R}$). Summing up these inequalities for $i = 1, \ldots, n$ gives the triangle inequality for d.

5. The *max-metric* on $\mathbb{R}^n$ is defined by

$$d(\underline{x}, \underline{y}) = \max_i |x_i - y_i|.$$

Once again, I will check the triangle inequality, and leave the other two axioms as an exercise for you.

$$\max_i |x_i - z_i| \leq \max_i (|x_i - y_i| + |y_i - z_i|) \leq \max_{i,j} (|x_i - y_i| + |y_j - z_j|)$$

$$\leq \max_i |x_i - y_i| + \max_j |y_j + z_j| = d(\underline{x}, \underline{y}) + d(\underline{y}, \underline{z}).$$

6. The *discrete metric* on an arbitrary set X is defined by

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

Exercise Check the metric axioms for d.

7. Let $\mathbb{Z}_2 = \{0, 1\}$. Then $\mathbb{Z}_2^n$ is the set of strings of 0s and 1s of length n. The *Hamming metric* on $\mathbb{Z}_2^n$ is defined by

$$d(\underline{x}, \underline{y}) = \#\{k; x_k \neq y_k\},$$

in other words the number of places at which the strings $\underline{x}$ and $\underline{y}$ disagree. As usual, the first two axioms for a metric are easy to check. To check the triangle inequality, note that if $x_k \neq z_k$ then precisely one of the following statements is true: $x_k \neq y_k$; $y_k \neq z_k$. It follows that

$$\#\{k; x_k \neq z_k\} \leq \#\{k; x_k \neq y_k\} + \#\{k; y_k \neq z_k\},$$

in other words $d(\underline{x}, \underline{z}) \leq d(\underline{x}, \underline{y}) + d(\underline{y}, \underline{z})$ as required.

8. Let X be a set. We let $B(X)$ denote the set of all *bounded* real valued functions on X, in other words the set of functions $f : X \to \mathbb{R}$ such that there is a real number K satisfying $|f(x)| \leq K$ for all $x \in X$.

   The *sup-metric* on $B(X)$ is defined by

$$d(f, g) = \sup_{x \in X} |f(x) - g(x)|.$$

   This generalises the max-metric on $\mathbb{R}^n$ in the following sense. If X is a finite set with n elements $x_1, \ldots x_n$, then all functions from X to $\mathbb{R}$ are bounded, and we can identify $B(X)$ with $\mathbb{R}^n$ via the identification $f \leftrightarrow (f(x_1), \ldots, f(x_n))$. Then

$$\sup_{x \in X} |f(x) - g(x)| = \max_{i=1}^{n} |f(x_i) - g(x_i)|,$$

   so this identification translates the sup-metric on $B(X)$ to the max-metric on $\mathbb{R}^n$, and *vice versa*.

   The verification of the axioms for a metric is exactly the same as for the max-metric on $\mathbb{R}^n$.

9. Let (X, d) be a metric space, and let $Y \subset X$ be a subset of X. Then the *induced metric* or *subspace metric* on Y induced by d is just the restriction of d to $Y \times Y$:

$$d_Y(y, y') = d(y, y') \ \forall \ y, y' \in Y.$$

   Exercise Check the metric axioms for d_Y.

   The metric space (Y, d_Y) is said to be a *subspace* of the metric space (X, d).

1.4 Metric spaces of functions

One of our examples of a metric space was the space $B(X)$ of bounded real-valued functions on a set X, with respect to the sup-metric.

In this section we look more closely at this example in the case where X is a closed interval in the real line. To simplify, let us choose X to be the interval $[0,1] := \{t \in \mathbb{R};\ 0 \leq t \leq 1\}$.

Calculating the distance

By definition, the distance $d(f,g)$ between two functions $f, g : [0,1] \to \mathbb{R}$ (in the sup-metric) is given by the formula

$$d(f,g) = \sup_{0 \leq t \leq 1} |f(t) - g(t)|.$$

For arbitrary (badly behaved) functions, this distance may be difficult to calculate. However, for reasonable functions it is not at all difficult.

Lemma 1.1 *Let $h : [0,1] \to \mathbb{R}$ be a continuous function that is differentiable on $(0,1)$. Then there exists $x \in [0,1]$ such that either (i) $x \in \{0,1\}$ or (ii) $h'(x) = 0$, and such that $\sup_{0 \leq t \leq 1} |h(t)| = |h(x)|$.*

Proof. A theorem of real analysis says that a continuous real valued function on $[0,1]$ is bounded and attains its bounds. Hence there are elements $x_0, x_1 \in [0,1]$ such that $h(x_0) = \inf_t h(t)$, $h(x_1) = \sup_t h(t)$. So we can take $x = x_0$ or $x = x_1$ in the statement of the Lemma. We need to check that one of the conditions (i), (ii) holds. But if (i) does not hold then $x \in (0,1)$, and h is differentiable on $(0,1)$. Since x is either a local minimum or local maximum of h, we must have $h'(x) = 0$.

Corollary 1.2 *Suppose $f, g : [0,1] \to \mathbb{R}$ are differentiable functions, and that the turning points of $f - g$ in $(0,1)$ are $x_1, \ldots, x_k$. Then*

$$d(f,g) = \max\{|f(0) - g(0)|, |f(x_1) - g(x_1)|, \ldots, |f(x_k) - g(x_k)|, |f(1) - g(1)|\}.$$

Examples

1. Let $f(t) = t$, $g(t) = t^2$. Then $\frac{d}{dt}(f(t) - g(t)) = 1 - 2t$, so $f - g$ has a unique turning point at $t = \frac{1}{2}$. Now $|f(0) - g(0)| = 0 = |f(1) - g(1)|$, while $|f(\frac{1}{2}) - g(\frac{1}{2})| = \frac{1}{4}$, so $d(f,g) = \frac{1}{4}$.

2. Let $f(t) = \frac{4t}{5}$, $g(t) = t^2$. Then $\frac{d}{dt}(f(t) - g(t)) = \frac{4}{5} - 2t$, so $f - g$ has a unique turning point at $t = \frac{2}{5}$. Now $|f(0) - g(0)| = 0$, $|f(\frac{2}{5}) - g(\frac{2}{5})| = \frac{4}{25}$ and $|f(1) - g(1)| = \frac{1}{5}$, so $d(f,g) = \max\{0, \frac{4}{25}, \frac{1}{5}\} = \frac{1}{5}$.

Note the difference between these two examples: in one case the maximum difference between f and g happens at a turning point in $(0, 1)$, while in the other it occurs at one of the endpoints $0, 1$. It is important to remember that both possibilities can occur, so we must evaluate $|f - g|$ at 0 and 1 as well as the turning points of $f - g$ to be sure of getting the maximum value.

Now let $C([0, 1])$ denote the set of all continuous functions $[0, 1] \to \mathbb{R}$. Since continuous functions on $[0, 1]$ are bounded, this is a subset of $B([0, 1])$, so we may regard it as a subspace of $B([0, 1])$. In other words, the sup-metric $d(f, g) = \sup_t |f(t) - g(t)|$ gives us a metric on $C([0, 1])$ as well.

The sup-metric is not the only metric on $C([0, 1])$.

Exercise Check that the following definition also defines a metric on $C([0, 1])$:

$$d(f, g) := \int_0^1 |f(t) - g(t)| \, dt.$$

This and the sup-metric are only two of many ways to define metrics on $C([0, 1])$. However, for the purposes of this course, the only two spaces of functions that we shall seriously consider are $B([0, 1])$ and $C([0, 1])$, and in each case we work exclusively with the sup-metric.

1.5 Exercises on metric spaces

1. Without looking at the lecture notes, write down the definition of a **metric** on a set X. Now check with the lecture notes.

2. Let d be a metric on the set X and let $x, y, z \in X$. Prove that

$$d(x, y) \geq |d(x, z) - d(y, z)| .$$

3. Let d be a metric on the set X. For all x, y in X let

$$\begin{aligned} d_1(x, y) &= k\, d(x, y) \qquad \text{for some constant} \qquad k > 0 ; \\ d_2(x, y) &= \min\{1, d(x, y)\} ; \\ d_3(x, y) &= (d(x, y))^2 . \end{aligned}$$

Prove that d_1 and d_2 are metrics for X but that d_3 may or may not be a metric for X.

4. Define functions $d_1, d_2 \colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ by

$$d_1((x_1, x_2), (y_1, y_2)) = \max\{|x_1 - y_1|, |x_2 - y_2|\} ,$$

$$d_2((x_1, x_2), (y_1, y_2)) = |x_1 - y_1| + |x_2 - y_2| .$$

Prove that d_1 and d_2 are metrics on $\mathbb{R}^2$. (Fill in the gaps in the notes.)

If d_0 is the euclidean metric on $\mathbb{R}^2$, $\underline{x} = (1, 3)$ and $\underline{y} = (-2, -1)$, calculate $d_0(\underline{x}, \underline{y})$, $d_1(\underline{x}, \underline{y})$, and $d_2(\underline{x}, \underline{y})$.

5. Let d be the metric on $\mathbb{R}^2$ defined by $d((x_1, x_2), (y_1, y_2)) = \max\{|x_1 - y_1|, |x_2 - y_2|\}$. Calculate $d((1, 2), (3, 4))$. Describe the set $\{(x, y) \in \mathbb{R}^2 \mid d((x, y), (1, 0)) = 1\}$.

6. Let d be the metric on $\mathbb{R}^2$ defined by $d((x_1, x_2), (y_1, y_2)) = |x_1 - y_1| + |x_2 - y_2|$. Calculate $d((1, 2), (3, 4))$. Describe the set $\{(x, y) \in \mathbb{R}^2 \mid d((x, y), (1, 0)) = 1\}$.

7. Let $X = (\mathbb{Z}_2^4)$ and let d denote the Hamming metric on X. Find all $x \in X$ such that $d(x, 0101) = 2$.

8. Let d be the discrete metric on $\mathbb{R}$. What is $d(3, 4)$?
What is $\{x \in \mathbb{R} \mid d(x, 1) = \frac{1}{2}\}$?

9. Let d be the Hamming metric on $\mathbb{Z}_2^4$, and fix an element $\underline{x} \in \mathbb{Z}_2^4$. How many elements $y \in \mathbb{Z}_2^4$ satisfy

   (a) $d(x, y) = 0$?

   (b) $d(x, y) = 1$?

(c) $d(x, y) = 2$?

(d) $d(x, y) = 3$?

(e) $d(x, y) = 4$?

(f) $d(x, y) > 4$?

10. Determine $d(f_1, f_2)$, $d(f_3, f_4)$ and $d(f_5, f_6)$ in the metric space $(B[0, 1], d)$, where d is the supremum metric, and

- $f_1(x) = \frac{1}{2}x$, $f_2(x) = x^2$,
- $f_3(x) = \sin(\pi x)$, $f_4(x) = \cos(\pi x)$,
- $f_5(x) = x^2 - x + 1$, $f_6(x) = x^3$.

Chapter 2

Open Sets and Closed Sets

2.1 Open Balls

The open intervals $(a, b) := \{x \in \mathbb{R};\ a < x < b\}$ in the real line play an important rôle in real analysis. For example, look at the definition of continuity: a function $f : \mathbb{R} \to \mathbb{R}$ is *continuous* at $x_0 \in \mathbb{R}$ if, for every $\varepsilon > 0$, there is a $\delta > 0$ such that $|x - x_0| < \delta \Rightarrow |f(x) - f(x_0)| < \varepsilon$.

In other words, f maps the open interval $(x_0 - \delta, x_0 + \delta)$ into the open interval $(f(x_0) - \varepsilon, f(x_0) + \varepsilon)$.

Note that there is a metric involved in this definition: $|x - x_0|$ is the distance between x and x_0 in the usual metric on $\mathbb{R}$. The open intervals that are implicitly appearing in the definition can also be described in terms of the usual metric on $\mathbb{R}$. For example $(x_0 - \delta, x_0 + \delta)$ is the set of points whose distance from x_0 is strictly less than δ: $(x_0 - \delta, x_0 + \delta) = \{x \in \mathbb{R} :\ |x - x_0| < \delta\}$.

This leads to a natural analogue of an open interval in any metric space, called an *open ball*.

Definition Let (X, d) be a metric space, $x_0 \in X$ a point in X, and $r > 0$ a positive real number. Then the *open ball* of *radius* r with *centre* x_0 in (X, d) is the set

$$\{\ x \in X;\ d(x, x_0) < r\ \}.$$

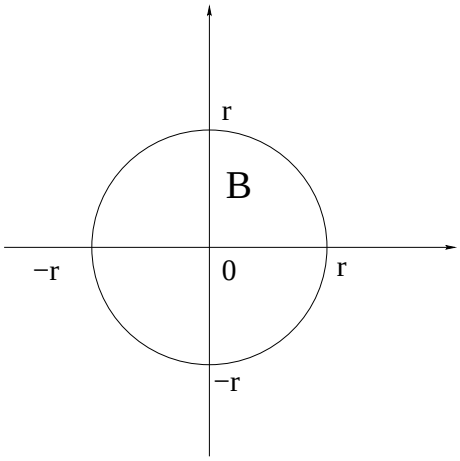
There are many common notations for this set. I will try to be consistent in using $B_r(x_0)$. $B_r(x_0)$ is sometimes also referred to as the *r-neighbourhood* of x_0 in (X, d).

Examples

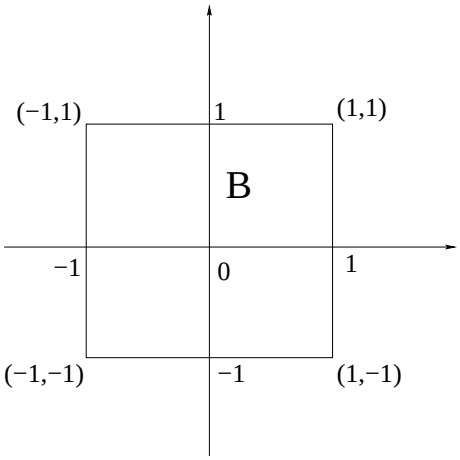
1. Let $X = \mathbb{R}$ and d the usual metric. Then, as mentioned above, the open ball $B_r(x_0)$ is just the open interval $(x_0 - r, x_0 + r)$.

2. Let $X = \mathbb{R}^2$, and d the usual metric. Let $\underline{0} = (0, 0)$. Then for any $r > 0$, $B_r(\underline{0}) = \{(x, y) \in \mathbb{R}^2;\ x^2 + y^2 < r^2\}$. This consists of the (inside of the) disc

of radius r with centre the origin $\underline{0}$, *not including* the circular edge $\{(x, y) \in \mathbb{R}^2;\ x^2 + y^2 = r^2\}$ of the disc.



3. Let $X = \mathbb{R}^2$, and let d be the max-metric, $d(\underline{x}, \underline{y}) = \max\{|x_1 - y_1|, |x_2 - y_2|\}$. What is $B_1(\underline{0})$? By definition, it is the set $\{\underline{x} \in \mathbb{R}^2;\ \max\{|x_1|, |x_2|\} < 1\}$. In other words, $\underline{x} \in B_1(\underline{0})$ if and only if $|x_1| < 1$ and $|x_2| < 1$. Thus $B_1(\underline{0})$ is the inside of the square in $\mathbb{R}^2$ with vertices $(\pm 1, \pm 1)$ (but not including the sides or vertices of the square).



4. In the discrete metric on X, open balls are very boring:

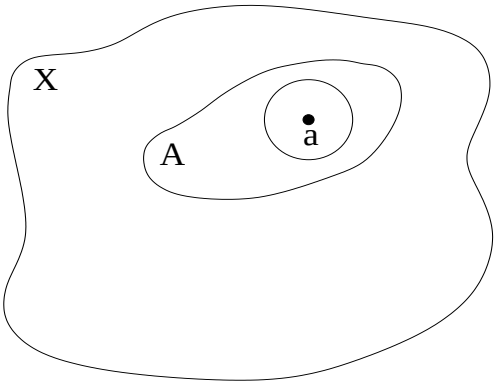
$$B_r(x_0) = \begin{cases} \{x_0\} & \text{if} \quad 0 < r \leq 1, \\ X & \text{if} \quad r > 1. \end{cases}$$

5. Let $X = \mathbb{Z}_2^5$, the set of strings of 0's and 1's of length 5, and let $\underline{0} = 00000$. Then, with respect to the Hamming metric on X, $B_1(\underline{0}) = \{\underline{0}\}$, while $B_2(\underline{0}) = \{00000, 00001, 00010, 00100, 01000, 10000\}$ (the set of strings of length 5 with fewer than two 1's).

2.2 Open Sets

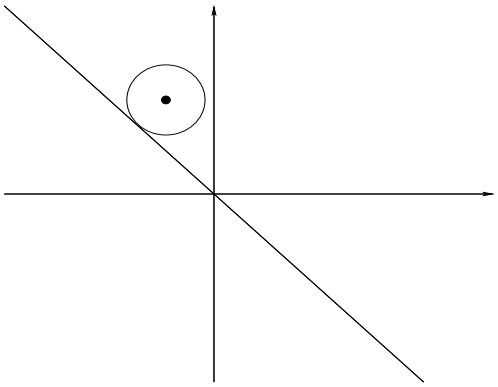
A set A in a metric space (X, d) is said to be *open* if it contains an open neighbourhood of each of its points. In other words, if

$$(\forall\ a \in A)\ (\exists\ r > 0)\ B_r(a) \subset A.$$



Examples

1. The set $A = \{(x_1, x_2) \in \mathbb{R}^2;\ x_1 + x_2 > 0\}$ is open in $\mathbb{R}^2$ (with the usual metric).



To prove this, we use the definition. Suppose that $\underline{a} = (a_1, a_2) \in A$. Then $a_1 + a_2 > 0$ by definition of A. Define $r = (a_1 + a_2)/\sqrt{2}$. If $\underline{b} = (b_1, b_2) \in B_r(\underline{a})$, then by definition of $B_r(\underline{a})$ we have $d(\underline{a}, \underline{b}) < r$. Writing the vector $\underline{a} - \underline{b}$ in polar coordinates (ρ, θ), we have $b_1 = a_1 - \rho\cos\theta$, $b_2 = a_2 - \rho\sin\theta$, where $0 \le \theta \le 2\pi$ and $0 \le \rho < r$. (The last inequality follows from $r^2 > d(\underline{a}, \underline{b})^2 = \rho^2$.)

Also, differentiating with respect to θ, we find that the minimum value of $b_1 + b_2 = (a_1 + a_2) - \rho(\sin\theta + \cos\theta)$ occurs at $\theta = \pi/4$, so that

$$b_1 + b_2 \ge (a_1 + a_2) - \rho\sqrt{2} > (a_1 + a_2) - r\sqrt{2} = 0.$$

Hence $B_r(\underline{a}) \subset A$. Since a was an arbitrary element of A, then every element of A is the centre of an open ball contained in A, so by definition A is an open set.

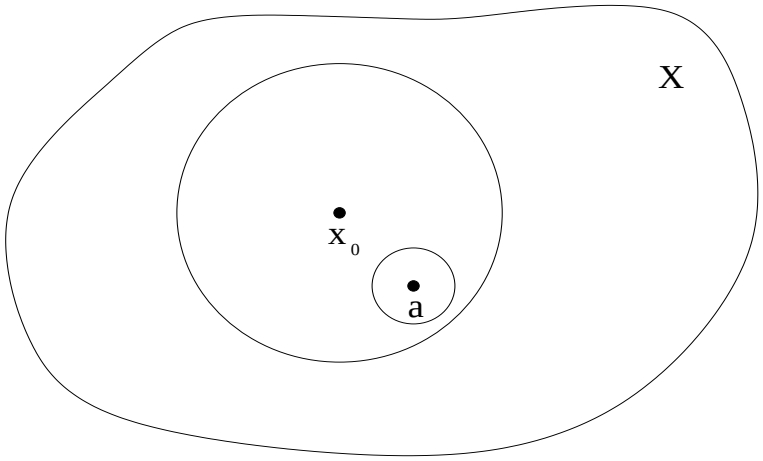
2. If d is the discrete metric on a set X, then *every* subset of X is open in (X, d). To see this, let $A \subset X$ and $a \in A$. We must show that some open ball centred at a is contained in A. But we have already seen that, for example, $B_1(a) = \{a\}$, and $\{a\} \subset A$ since $a \in A$. Hence A is open, as claimed.

Remark In the definition of open set, the radius r of the ball $B_r(a)$ that is contained in A will, in general, depend on the centre a. For each a, choose $r(a) > 0$ such that $B_{r(a)}(a) \subset A$. Then we can write A as a union of open balls thus:

$$A = \bigcup_{a \in A} B_{r(a)}(a).$$

It turns out that the converse of this statement is also true. This follows from some general properties of open sets that we will now prove.

Lemma 2.1 *Every open ball in (X, d) is an open set in (X, d).*



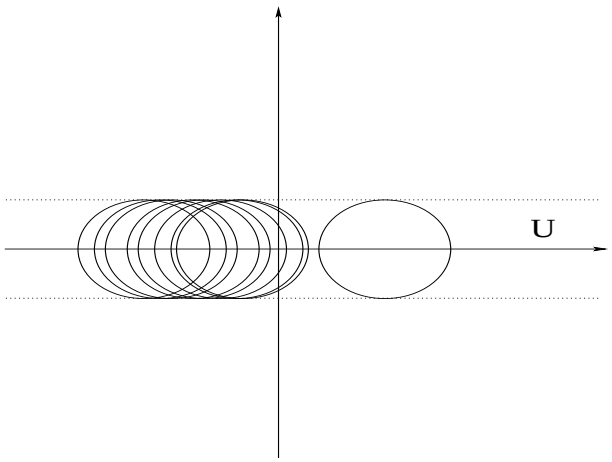
Proof. Let $A = B_r(x_0) \subset X$, and let $a \in A$. We must show that $B_s(a) \subset A$ for some $s > 0$. Since $a \in B_r(x_0)$, we have $d(a, x_0) < r$. Define $s = r - d(a, x_0) > 0$. Then $B_s(a) \subset A = B_r(x_0)$. To see this, let $y \in B_s(a)$. Then, by definition of $B_s(a)$, we have $d(a, y) < s$. By the triangle inequality, $d(y, x_0) \leq d(y, a) + d(a, x_0) < s + (r - s) = r$. Hence $y \in B_r(x_0) = A$. But y was an arbitrary element of $B_s(a)$, and so $B_s(a) \subset A$, as claimed. Hence A is an open set.

Theorem 2.2 *The union of any collection of open sets in (X, d) is an open set in (X, d).*

Proof. Let $\{U_\lambda;\ \lambda \in \Lambda\}$ be a collection of open sets in (X, d), and let $U = \bigcup_{\lambda \in \Lambda} U_\lambda$. We must show that U is open. So, let $u \in U$. Then $u \in U_\lambda$ for some $\lambda \in \Lambda$. But U_λ is an

open set, so by definition there is a positive real number r such that $B_r(u) \subset U_\lambda \subset U$. Hence U is open, as claimed.

Example Let $U = \{(x_1, x_2) \in \mathbb{R}^2 : \ |x_2| < 1\}$. For each $\underline{x} = (x_1, x_2) \in \mathbb{R}^2$, we have $\underline{x} \in U \Leftrightarrow d(\underline{x}, (x_1, 0)) = |x_2| < 1$. Hence $U = \bigcup_{t\in\mathbb{R}} B_1((t, 0))$ is a union of open balls, and so an open set, in $\mathbb{R}^2$ with respect to the usual metric.



Corollary 2.3 *A subset of X is open in (X, d) if and only if it is a union of open balls in (X, d).*

Proof. We have noted above that an open set can be expressed as a union of open balls. Conversely, every open ball is an open set, so a union of open balls is a union of open sets, and so an open set.

Theorem 2.4 *Let $U_1, \ldots, U_n$ be open sets in (X, d). Then $U_1 \cap \cdots \cap U_n$ is an open set in (X, d).*

Proof. Let $U = U_1 \cap \cdots \cap U_n$, and let $u \in U$. To show that U is an open set, we must show that $B_r(u) \subset U$ for some $r > 0$. Now $u \in U_i$ for each $i = 1, \ldots, n$, so there are real numbers $r(1), \ldots, r(n) > 0$ such that $B_{r(i)}(u) \subset U_i$ for each i. Now let $r = \min\{r(1), \ldots, r(n)\}$. Then, for each $i = 1, \ldots, n$, we have $B_r(u) \subset B_{r(i)}(u) \subset U_i$, and hence

$$B_r(u) \subset U_1 \cap \cdots \cap U_n = U.$$

Example The above theorem does not extend to an intersection of *infinitely* many open sets. For example, let $X = \mathbb{R}$ and let d be the usual metric. Then the open interval $U_n = (-\frac{1}{n}, \frac{1}{n}) = \{t \in \mathbb{R};\ |t| < \frac{1}{n}\}$ is an open set in $\mathbb{R}$ for each $n = 1, 2, 3, \ldots$. However, $\bigcap_{n=1}^{\infty} U_n = \{0\}$ is not an open set in $\mathbb{R}$, since $B_r(0) = (-r, r) \not\subset \{0\}$ for any $r > 0$.

2.3 Closed Sets

We define a set A in a metric space (X, d) to be *closed* if its complement $X \setminus A$ is open.

Examples

1. The closed interval $[0, 1] := \{t \in \mathbb{R};\ 0 \leq t \leq 1\}$ is a closed set in $\mathbb{R}$ with respect to the usual metric. Its complement is the union of the two (infinite) open intervals $(-\infty, 0)$ and $(1, \infty)$. Each of these is a union of an infinite collection of finite open intervals (for example $(-\infty, 0) = \bigcup_{n\in\mathbb{N}}(-n, 0)$), so is an open set. Hence $\mathbb{R} \setminus [0, 1]$ is open, so $[0, 1]$ is closed, by definition.

2. If (X, d) is any metric space, and $x_0 \in X$, then the set $\{x_0\}$ is closed. To see this, let $y \in X \setminus \{x_0\}$. In other words, $y \in X$ and $y \neq x_0$. Then $r := d(x_0, y) > 0$. By definition, $x_0 \notin B_r(y)$, since $d(x_0, y) \geq r$. In other words, $B_r(y) \subset X \setminus \{x_0\}$. Hence $X \setminus \{x_0\}$ is an open set in (X, d), so $\{x_0\}$ is a closed set in (X, d), as claimed.

3. If d is the discrete metric on X, then every subset A of X is closed in (X, d) (since its complement, $X \setminus A$, is a subset of X, and so open in (X, d) because every subset of X is open with respect to the discrete metric).

Using the De Morgan laws for unions and intersections of complements, we can translate the theorems about open sets in the previous section to theorems about closed sets, as follows.

Theorem 2.5 *The intersection of any collection of closed sets in a metric space (X, d) is closed in (X, d).*

Theorem 2.6 *If $A_1, \ldots, A_n$ are closed sets in the metric space (X, d) then $A_1 \cup \cdots \cup A_n$ is a closed set in (X, d).*

There is an alternative definition of closed sets, which I will now describe. The fact that the two definitions are equivalent is a theorem.

Definition Let (X, d) be a metric space and A a subset of X. An element $x \in X$ is a *limit point*, (or *cluster point*) of A if there are points of A arbitrarily close to x, excluding x itself. In symbols,

$$(\forall\ \varepsilon > 0)\ (\exists\ a \in A)\ 0 < d(a, x) < \varepsilon.$$

Example If $X = \mathbb{R}$, d is the usual metric, and $A = \{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots\}$, then 0 is a limit point of A. Given $\varepsilon > 0$ we can choose an integer $N > \frac{1}{\varepsilon}$. Then $\frac{1}{N} \in A$ with $0 < d(0, \frac{1}{N}) < \varepsilon$, as required.

Theorem 2.7 *Let (X, d) be a metric space and A a subset of X. Then A is closed if and only if A contains all its limit points.*

Proof. Suppose first that A is closed. Then $X \setminus A$ is open. Suppose that $x \in X$ is a limit point of A. If $x \notin A$ then $x \in X \setminus A$, so $B_r(x) \subset X \setminus A$ for some $r > 0$. In other words, there is no element $a \in A$ with $d(a, x) < r$, contradicting the definition of limit point. Hence $x \in A$.

Conversely, suppose that A contains all its limit points, and let $x \in X \setminus A$. Then $x \notin A$, so x is not a limit point of A. By definition, there is a $\varepsilon > 0$ such that there is no $a \in A$ with $0 < d(a, x) < \varepsilon$. Note that there is also no $a \in A$ with $d(a, x) = 0$, since $x \notin A$. Hence $A \cap B_\varepsilon(x) = \emptyset$, or in other words, $B_\varepsilon(x) \subset X \setminus A$. Hence $X \setminus A$ is open, so A is closed.

Definition The set $\overline{A}$ consisting of A together with all the limit points of A is called the *closure* of A.

Lemma 2.8 *Let A be a set in a metric space (X, d). Then $\overline{A}$ is the smallest closed set containing A. In other words, $\overline{A}$ is a closed set, $A \subset \overline{A}$, and if B is any closed set with $A \subset B$ then $\overline{A} \subset B$.*

Proof. By definition, $A \subset \overline{A}$. To see that $\overline{A}$ is closed, suppose that $x \in X$ is a limit point of $\overline{A}$. We must show that $x \in \overline{A}$. If $x \in A$ then $x \in \overline{A}$, so we can assume that $x \notin A$.

Let $\varepsilon > 0$. Then, since x is a limit point of $\overline{A}$, there is a point $a' \in \overline{A}$ with $0 < d(a', x) < \frac{\varepsilon}{2}$. If $a' \in A$ put $a = a'$ and note that $0 < d(a, x) < \frac{\varepsilon}{2} < \varepsilon$.

Otherwise, a' is a limit point of A, so there is a point $a \in A$ with $0 < d(a, a') < \frac{\varepsilon}{2}$. In this case $d(a, x) \leq d(a, a') + d(a', x) < \varepsilon$, by the triangle inequality, while $d(a, x) > 0$ since $x \notin A$ (and so $x \neq a$).

In either case, $a \in A$ with $0 < d(a, x) < \varepsilon$. Hence x is a limit point of A, so $x \in \overline{A}$. Thus $\overline{A}$ contains all its limit points, and so is closed.

Now suppose that B is a closed set, and that $A \subset B$. To show that $\overline{A} \subset B$, it suffices to show that every limit point of A is contained in B. Let x be a limit point of A. If $\varepsilon > 0$, there is a point $a \in A$ with $0 < d(a, x) < \varepsilon$. But $a \in B$ since $A \subset B$. Hence x is a limit point of B. But B is closed, so $x \in B$, as required.

Dual to the notion of closure is that of the *interior* of a set A in a metric space (X, d). The *interior* of A is defined to be the set $\mathrm{Int}(A) := \{a \in A;\ (\exists\ r > 0)\ B_r(a) \subset A\}$.

Exercise $\mathrm{Int}(A) = X \setminus \overline{(X \setminus A)}$.

Corollary 2.9 $\mathrm{Int}(A)$ *is the largest open set contained in A. In other words,* $\mathrm{Int}(A)$ *is an open set,* $\mathrm{Int}(A) \subset A$*, and if B is an open set with $B \subset A$, then $B \subset \mathrm{Int}(A)$.*

Example Let $X = \mathbb{R}$ with the usual metric, and let $A = (0, 1] = \{t \in \mathbb{R};\ 0 < t \leq 1\}$. Then $\mathrm{Int}(A) = (0, 1) = \{t \in \mathbb{R};\ 0 < t < 1\}$, while $\overline{A} = [0, 1] = \{t \in \mathbb{R};\ 0 \leq t \leq 1\}$.

2.4 Bounded Sets

Later in the course, we shall need to consider open sets, closed sets, and *bounded* sets in a metric space (X, d). To motivate the definition of bounded set, recall first what this means in the case $X = \mathbb{R}$.

Definition A subset $A \subset \mathbb{R}$ is *bounded* if $\exists\ K > 0$ such that $|x| \leq K$ for all $x \in A$.

There are several equivalent ways of stating this definition. For example:

1. $\exists\ K > 0$ such that $A \subset (-K, K) = B_K(0)$;
2. $\forall\ x \in \mathbb{R}$, $\exists\ K > 0$ such that $A \subset B_K(x)$;
3. $\exists\ K > 0$ such that, $\forall\ a, a' \in A$, $|a - a'| < K$.

These can be translated into criteria for boundedness in an arbitrary metric space (X, d).

Lemma 2.10 *Let (X, d) be a metric space. Then the following conditions are equivalent, for A a subset of X:*

(i) $(\forall\ x \in X)\ (\exists\ K > 0)\ A \subset B_K(x)$;

(ii) $(\forall\ x \in X)\ (\exists\ K > 0)\ (\forall\ a \in A)\ d(a, x) < K$;

(iii) $(\exists\ K > 0)\ (\forall\ a, a' \in A)\ d(a, a') < K$.

Provided $X \neq \emptyset$, the following condition is also equivalent to the above:

(iv) $(\exists\ x \in X)\ (\exists\ K > 0)\ A \subset B_K(x)$

Proof. **Exercise**.

Remark The constants K in the various conditions need not be the same.

Exercise Why is condition (iv) not equivalent to the other three conditions when $X = \emptyset$?

Definition If A is a set in a metric space (X, d) satisfying the equivalent conditions of the Lemma, then A is said to be *bounded*.

Definition The *diameter* of a (nonempty) bounded set A in a metric space (X, d) is

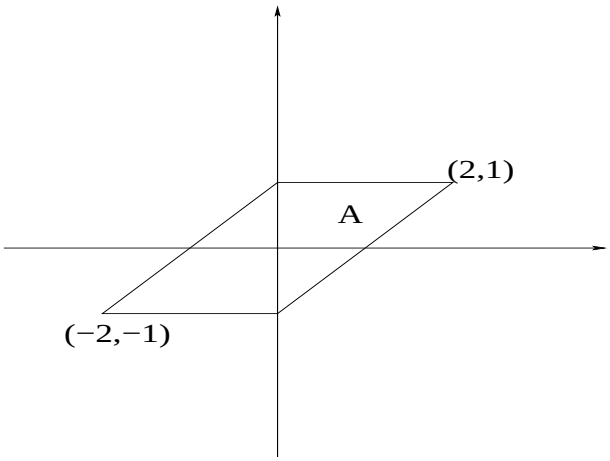
$$\operatorname{diam}(A) = \sup_{a, a' \in A} d(a, a').$$

(Note that this is a non-negative real number. It is less than ∞ because A is bounded.)

Examples

1. Let $a < b$ be real numbers, and A the open interval $(a, b) \subset \mathbb{R}$. Then A is bounded, and $\text{diam}(A) = b - a$ (with respect to the usual metric on $\mathbb{R}$).

2. Let A be the parallelogram $\{(x, y) \in \mathbb{R}^2;\ |y| \leq 1, |x - y| \leq 1\}$ in $\mathbb{R}^2$.



   Then

   (a) with respect to the euclidean metric on $\mathbb{R}^2$, A is bounded with diameter $\text{diam}(A) = d((-2, -1), (2, 1)) = 2\sqrt{5}$;

   (b) with respect to the Manhattan metric on $\mathbb{R}^2$, A is bounded with diameter $\text{diam}(A) = d((-2, -1), (2, 1)) = 6$;

   (c) with respect to the max-metric on $\mathbb{R}^2$, A is bounded with diameter $\text{diam}(A) = d((-2, -1), (2, 1)) = 4$;

3. Let d be the discrete metric on a set X, and let A be any nonempty subset of X. Then A is bounded with diameter (a) 0 if A has only one element; (b) 1 if A has more than one element.

2.5 Exercises on open sets, closed sets, bounded sets

1. Draw a picture of the open ball $B_3((1,1))$ in $\mathbb{R}^2$ with respect to

   (a) the euclidean metric;

   (b) the max-metric $d((x_1,x_2),(y_1,y_2)) = \max\{|x_1-y_1|,|x_2-y_2|\}$;

   (c) the Manhattan metric $d((x_1,x_2),(y_1,y_2)) = |x_1-y_1|+|x_2-y_2|$.

2. Show that if x and y are distinct points in a metric space, then there exist two open balls, one with centre x and the other with centre y, whose intersection is the empty set.

3. Let (X,d) be a metric space and let $a \in X$. Show that the set $\{x \in X : d(x,a) > r\}$ is open and hence deduce that the sets $\{x \in X : d(x,a) \leq r\}$ and $\{x \in X; d(x,a) = r\}$ are closed.

4. Show that every subset $\{x\}$ with a single element in a metric space (X,d) is closed. Deduce that every finite subset in (X,d) is closed.

5. Give an example of an infinite family of closed sets in $\mathbb{R}$ (with respect to the usual metric) whose union is not a closed set.

6. Let (X,d) be a metric space. If $B_r(x)$ is an open ball in X, show that $B_r(x)$ is a bounded set and that $\operatorname{diam}(B_r(x))) \leq 2r$. Show, by means of an example, that it is possible that $\operatorname{diam}(B_r(x)) < r$.

7. Let (X,d) be a metric space and suppose A and B are bounded sets in X.

   Prove that $A \cup B$ is also a bounded set.

   If $A \cap B \neq \emptyset$, prove

   $$\operatorname{diam}(A \cup B) \leq \operatorname{diam}(A) + \operatorname{diam}(B)\,.$$

   Give an example to show that this inequality need not hold if $A \cap B = \emptyset$.

Chapter 3

Sequences in Metric Spaces

3.1 Sequences and Limits

Definition A *sequence* in a metric space (X, d) is just a function $\mathbb{N} \to X$, $n \mapsto x_n$, where $\mathbb{N}$ denotes the set of all natural numbers ($\mathbb{N} = \{1, 2, 3, \ldots\}$).

In other words, a sequence is an infinite list $x_1, x_2, x_3, \ldots$ of elements of X (possibly with repetitions).

The notation $\{x_n\}$ is commonly used for the sequence $x_1, x_2, x_3, \ldots$.

Examples

1. The rule $x_n = \frac{1}{n}$ gives a sequence $1, \frac{1}{2}, \frac{1}{3}, \ldots$ in $\mathbb{R}$. We can also denote this sequence by $\{\frac{1}{n}\}$.

2. The rule $x_n = (\frac{1}{n}, \frac{1}{n^2})$ defines a sequence $(1, 1), (\frac{1}{2}, \frac{1}{4}), (\frac{1}{3}, \frac{1}{9}), \ldots$ in $\mathbb{R}^2$.

3. If (X, d) is any metric space and $a \in X$ is any element, then the rule $x_n = a \ \forall \ n \in \mathbb{N}$ defines a sequence $a, a, a, \ldots$. A sequence of this type is called a *constant* sequence.

Remarks

1. Strictly speaking, the metric d plays no rôle in the definition of a sequence, and one can just as easily consider sequences in an arbitrary set. The relevance of the metric will become apparent when we consider the notion of limits of sequences.

2. In the above examples, the sequences are all defined by (simple) rules. There is a good reason for this – it is the only way I can describe to you an infinite sequence using only a finite description. For this reason, any concrete example of a sequence that you ever encounter will also be described by some rule. However, one should realise that there exist many sequences that cannot be defined by rules at all. Indeed, it can be shown that the set of all possible rules that one can write

down is *countable* (has the same cardinality as $\mathbb{N}$), whereas the set of sequences in any set containing more than one element is *uncountable* (has cardinality strictly greater than that of $\mathbb{N}$).

Definition A *subsequence* of a sequence $\{x_n\}$ (in a metric space (X, d)) is a sequence of the form $\{y_n\}$, where $y_n = x_{m_n}$ for some sequence $\{m_n\}$ of natural numbers that is *strictly increasing*, in the sense that $m_n < m_{n'}$ whenever $n < n'$.

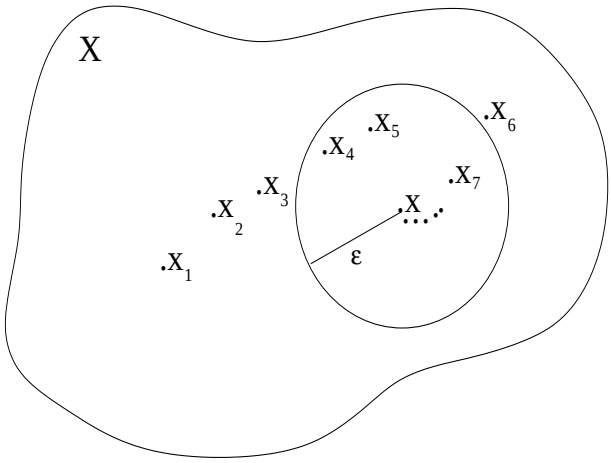
Example Suppose that $\{x_n\}$ is the sequence $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots$ defined by $x_n = \frac{1}{n}$, and $\{m_n\}$ is the increasing sequence of natural numbers $2, 4, 6, 8, \ldots$ defined by $m_n = 2n$. Then the resulting subsequence $\{y_n\} = \{x_{m_n}\}$ of $\{x_n\}$ is the sequence $\frac{1}{2}, \frac{1}{4}, \frac{1}{6}, \frac{1}{8} \ldots$ defined by $y_n = \frac{1}{2n}$.

Remark If we think of a sequence $\{x_n\}$ as a function $x : \mathbb{N} \to X$, then we can also think of the sequence $\{m_n\}$ of natural numbers as a function $m : \mathbb{N} \to \mathbb{N}$. Then the subsequence $\{x_{m_n}\}$ is just the composite function $x \circ m : \mathbb{N} \to X$.

Definition Let $\{x_n\}$ be a sequence in the metric space (X, d). We say that $\{x_n\}$ *converges* to a point $x \in X$, or that x is a *limit* of the sequence $\{x_n\}$, if

$$(\forall\ \varepsilon > 0)\ (\exists\ N \in \mathbb{N})\ (\forall\ n \geq N)\ d(x_n, x) < \varepsilon.$$

The condition $d(x_n, x) < \varepsilon$ can be rewritten as $x_n \in B_\varepsilon(x)$. Thus the condition $(\forall\ n \geq N)\ d(x_n, x) < \varepsilon$ means that the terms x_n of the sequence lie in the ball $B_\varepsilon(x)$, with only a finite number of exceptions.



No matter how small we choose the positive real number ε, this remains true (if we also choose N large enough).

Examples

1. The sequence $\{\frac{1}{n}\}$ in $\mathbb{R}$ converges to 0, with respect to the usual metric. This seems obvious, but let us prove it formally, using the definition of convergence. Suppose that $\varepsilon > 0$. Then $\frac{1}{\varepsilon}$ is a positive real number, so we can choose a natural number N with $N > \frac{1}{\varepsilon}$. This inequality can be rewritten in the equivalent form $\frac{1}{N} < \varepsilon$. Now if $n \geq N$, then $d(\frac{1}{n}, 0) = \frac{1}{n} \leq \frac{1}{N} < \varepsilon$.

2. The sequence $\{(\frac{1}{n}, \frac{1}{n^2})\}$ converges to $(0, 0)$ in $\mathbb{R}^2$ with respect to the max-metric. To see this, note that

$$d\left(\left(\frac{1}{n}, \frac{1}{n^2}\right), (0,0)\right) = \max\left\{\frac{1}{n}, \frac{1}{n^2}\right\} = \frac{1}{n}$$

   when $n \geq 1$, so we can use the same argument as in the previous example (choose $N > \frac{1}{\varepsilon}$).

3. The sequence $\{(\frac{1}{n}, \frac{1}{n^2})\}$ converges to $(0, 0)$ in $\mathbb{R}^2$ with respect to the euclidean metric. This time,

$$d\left(\left(\frac{1}{n}, \frac{1}{n^2}\right), (0,0)\right) = \sqrt{\frac{1}{n^2} + \frac{1}{n^4}} > \frac{1}{n},$$

   so the previous argument will not quite work. However,

$$d\left(\left(\frac{1}{n}, \frac{1}{n^2}\right), (0,0)\right) = \sqrt{\frac{1}{n^2} + \frac{1}{n^4}} < \frac{2}{n},$$

   so we can amend the argument slightly so that it does work: choose $N > \frac{2}{\varepsilon}$.

Remark A good way to think about proofs of this type is as a game played against an opponent, who is trying to trick us. We have control of any variables introduced by the $\exists$ symbol, our opponent has control of those introduced by the $\forall$ symbol, and the players choose in order from left to right in the definition. If we have a winning strategy, where we can force the final statement to be TRUE no matter what our opponent does, then the definition holds. (In this case, the given sequence really does converge to the given limit.) If our opponent has a winning strategy, then the definition fails – in this case the sequence does not converge to the given limit. In the above examples, the proof hinges on a sensible choice of the variable N. Fortunately, by the time we have to make that choice, our opponent has already chosen ε, which is then fixed for the rest of the game. This allows us to give N as a function of ε, if that helps us.

Example

To illustrate these ideas in the opposite direction, let us show that the sequence $\{n^2\}$ in $\mathbb{R}$ does not converge to any real number x. In logical symbols, non-convergence of $\{x_n\}$ to x can be expressed as

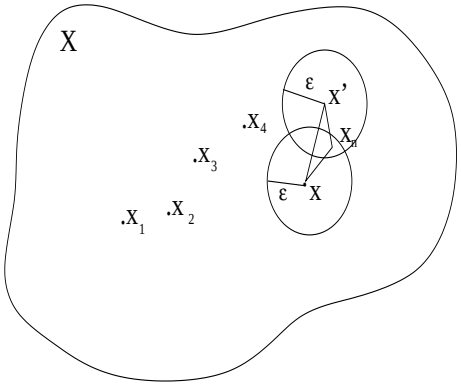
$$(\exists\ \varepsilon > 0)\ (\forall\ N \in \mathbb{N})\ (\exists\ n \geq N)\ d(x_n, x) \geq \varepsilon.$$

In other words, we have changed places with our opponent, and this time we get to choose $\varepsilon > 0$ and $n \geq N$. In this example, the choice of ε is not important, so let us arbitrarily decide that $\varepsilon = 1$. Now, given $N \in \mathbb{N}$, we must choose $n \geq N$ such that $|n^2 - x| = d(n^2, x) \geq \varepsilon = 1$. We can do this, for example, by choosing $n = \max\{N, t+1\}$, where t is the smallest integer greater than $\sqrt{|x|}$.

Hence $\{n^2\}$ has no limit in $\mathbb{R}$, as claimed.

Lemma 3.1 *Let (X, d) be a metric space, and $\{x_n\}$ a sequence in X. If $\{x_n\}$ converges to $x \in X$ and also to $x' \in X$, then $x = x'$.*

Proof. Choose $\varepsilon > 0$. Then, using the definition of convergence, there are natural numbers $N, N' \in \mathbb{N}$ such that $d(x_n, x) < \varepsilon$ whenever $n \geq N$, and $d(x_n, x') < \varepsilon$ whenever $n \geq N'$.



Now choose $n \geq \max(N, N')$. Then, by the triangle inequality,

$$d(x, x') \leq d(x, x_n) + d(x_n, x') = d(x_n, x) + d(x_n, x') < 2\varepsilon.$$

Since this is true for all $\varepsilon > 0$, we must have $d(x, x') \leq 0$, so $x = x'$.

To summarise, not every sequence has a limit, but if it does then the limit is unique. We will use the notation:

$$x = \lim_{n\to\infty} x_n, \qquad \text{or :} \qquad x_n \to x \quad \text{as} \quad n \to \infty,$$

to signify that the sequence $\{x_n\}$ converges to x (in (X, d)).

In a sense, if we understand convergence in $\mathbb{R}$ with respect to the usual metric, then we understand convergence in any metric space. This is explained in the following result.

Lemma 3.2 *Let $\{x_n\}$ be a sequence in a metric space (X, d), and let x be an element of X. Then $\{x_n\}$ converges to x in (X, d) if and only if the sequence $\{d(x_n, x)\}$ converges to 0 in $\mathbb{R}$ (with respect to the usual metric on $\mathbb{R}$).*

Proof. Exercise. [Use the definition of convergence together with the fact that the distance from $d(x_n, x)$ to 0 in the usual metric on $\mathbb{R}$ is just $|d(x_n, x) - 0| = d(x_n, x)$.]

The following result gives us an easy way of recognising that certain sequences are not convergent.

Definition A sequence $\{x_n\}$ in a metric space is *bounded* if its set of values $\{x_1, x_2, x_3, \ldots\}$ is a bounded subset in (X, d).

Lemma 3.3 *Every convergent sequence in a metric space (X, d) is bounded.*

Proof. Suppose that $x_n \to x$ in (X, d) as $n \to \infty$. Setting $\varepsilon = 1$ in the definition of convergence, there is a natural number $N \in \mathbb{N}$ such that $d(x_n, x) < 1$ for $n \geq N$. Now let $K = \max\{d(x_1, x), d(x_2, x), \ldots, d(x_N, x), 1\}$. Then $d(x_n, x) \leq K$ for all $n \in \mathbb{N}$, so $\{x_n\}$ is bounded.

Example The sequence $\{\sqrt{n}\}$ in $\mathbb{R}$ is not bounded, since for any $K \in \mathbb{R}$ we can find $n \in \mathbb{N}$ with $n > K^2$, so $|\sqrt{n}| > K$. Hence this sequence has no limit in $\mathbb{R}$.

The converse of the above lemma is false, however: not every bounded sequence is convergent.

Examples

1. The bounded sequence $\{(-1)^n\}$ has no limit in $\mathbb{R}$ (with respect to *any* metric). Let d be a metric on $\mathbb{R}$, and let $\varepsilon = d(-1, 1)/2 > 0$. Suppose that $(-1)^n \to x \in \mathbb{R}$ as $n \to \infty$ (with respect to the metric d). Then by definition there is a natural number $N \in \mathbb{N}$ such that $d((-1)^n, x) < \varepsilon$ for all $n \geq N$. In particular, $2\varepsilon = d((-1)^N, (-1)^{N+1}) \leq d((-1)^N, x) + d((-1)^{N+1}, x) < 2\varepsilon$, a contradiction.

2. $\{\frac{1}{n}\}$ is a bounded sequence in $X = (0, 2) \subset \mathbb{R}$ that converges (in $\mathbb{R}$, with respect to the usual metric) to 0. If this sequence converges to $x \in X$ with respect to the subspace metric on X, then it also converges to x in $\mathbb{R}$. But then uniqueness of limits says that $0 = x \in X = (0, 2)$, a contradiction.

3.2 Sequences in $\mathbb{R}^n$

We have considered three metrics on the plane $\mathbb{R}^2$, namely the euclidean metric

$$d_e(\underline{x}, \underline{y}) := \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2},$$

the max-metric

$$d_m(\underline{x}, \underline{y}) := \max\{|x_1 - y_1|, |x_2 - y_2|\},$$

and the Manhattan, or taxi-cab, metric

$$d_t(\underline{x}, \underline{y}) := |x_1 - y_1| + |x_2 - y_2|.$$

Note that these are linked by the following inequalities:

$$d_m^2 \leq d_e^2 \leq d_t^2 \leq 4d_m^2,$$

so

$$d_m \leq d_e \leq d_t \leq 2d_m.$$

It follows that a sequence in $\mathbb{R}^2$ that converges with respect to one of these metrics also converges (to the same limit) with respect to the other two metrics.

Moreover, to say that a sequence $\{(x_n, y_n)\}$ converges to (x, y) in $(\mathbb{R}^2, d_m)$ is to say that $\max\{|x_n - x|, |y_n - y|\} \to 0$ in $\mathbb{R}$ (with respect to the usual metric in $\mathbb{R}$), in other words $|x_n - x| \to 0$ in $\mathbb{R}$ and $|y_n - y| \to 0$ in $\mathbb{R}$. But this is equivalent to $x_n \to x$ and $y_n \to y$ in $\mathbb{R}$.

In conclusion, a sequence $\{(x_n, y_n)\}$ in $\mathbb{R}^2$ converges to $(x, y) \in \mathbb{R}^2$ with respect to any one of the the metrics d_e, d_m, d_t if and only if both the sequences of real numbers $\{x_n\}$, $\{y_n\}$ converge to x, y respectively, in $\mathbb{R}$ with respect to the usual metric.

A similar analysis applies in $\mathbb{R}^N$ for any natural number N. The euclidean, max and taxi metrics satisfy the inequalities

$$d_m \leq d_e \leq d_t \leq Nd_m.$$

It follows that convergence of sequences with respect to these three metrics are equivalent concepts. Again, convergence in $\mathbb{R}^N$ with respect to any one of these metrics is equivalent to simultaneous convergence of all N co-ordinate sequences in $\mathbb{R}$. We can summarise this as follows.

Theorem 3.4 *For each $i = 1, \ldots, N$, let $\{x_n^{(i)}\}$ be a sequence of real numbers, and let $\underline{x}_n = (x_n^{(1)}, \ldots, x_n^{(N)}) \in \mathbb{R}^N$ for each $n \in \mathbb{N}$. Then the following are equivalent:*

1. *The sequence $\{\underline{x}_n\}$ converges to $\underline{x} = (x^{(1)}, \ldots, x^{(N)})$ in $(\mathbb{R}^N, d_e)$.*
2. *The sequence $\{\underline{x}_n\}$ converges to $\underline{x} = (x^{(1)}, \ldots, x^{(N)})$ in $(\mathbb{R}^N, d_m)$.*
3. *The sequence $\{\underline{x}_n\}$ converges to $\underline{x} = (x^{(1)}, \ldots, x^{(N)})$ in $(\mathbb{R}^N, d_t)$.*
4. *For each $i = 1, \ldots, N$, the sequence $\{x_n^{(i)}\}$ converges to $x^{(i)}$ in $(\mathbb{R}, d)$ (where d is the usual metric on $\mathbb{R}$).*

Examples

1. The sequence $\{(\frac{1}{n}, \frac{1}{n^2})\}$ converges to $(0, 0)$ in $\mathbb{R}^2$ (with respect to each of the euclidean, max and taxi metrics).
2. The sequence $\{(\frac{1}{n}, -\frac{1}{n}, -n)\}$ does not converge in $\mathbb{R}^3$ (with respect to any of the metrics d_e, d_m, d_t).

3.3 Sequences of bounded functions

Consider the following two examples of sequences in the space $B(X)$ of bounded functions on X, with respect to the sup-metric.

Examples

1. $X = [0,1]$, $f_n : [0,1] \to \mathbb{R}$ given by $f_n(x) = \frac{x}{n}$. Then the sequence $\{f_n\}$ converges (in $B([0,1])$ with the sup-metric) to the zero function $z : [0,1] \to \mathbb{R}$ given by $z(x) = 0 \ \forall \ x$.

   To see this, let d be the sup-metric on $B([0,1])$ and calculate

   $$d(f_n, z) = \sup\{|f_n(x), z(x)|; \ x \in [0,1]\} = \sup\left\{\frac{x}{n}; \ 0 \leq x \leq 1\right\} = \frac{1}{n} \to 0 \text{ as } n \to \infty.$$

2. $X = \mathbb{R}$,

   $$f_n(x) = \begin{cases} 0, & x \leq n, \\ 1, & x > n \end{cases}$$

   Then the sequence $\{f_n\}$ has no limit in $B(\mathbb{R})$ with respect to the sup-metric.

   To see this, suppose that $f_n \to f$, where $f : \mathbb{R} \to \mathbb{R}$ is a bounded function. Then, for any $x \in \mathbb{R}$, we have

   $$|f(x)| = \lim_{n\to\infty} |f(x) - f_n(x)| \leq \lim_{n\to\infty} d(f_n, f) = 0,$$

   so $f(x) = 0$ for all $x \in \mathbb{R}$. But then

   $$d(f_n, f) = 1 \not\to 0 \text{ as } n \to \infty.$$

In the second example, the sequence $\{f_n(x)\}$ of real numbers converges (with respect to the usual metric on $\mathbb{R}$) for each $x \in \mathbb{R}$. This property is known as *pointwise convergence*.

Definition Let $\{f_n\}$ be a sequence in $B(X)$. We say that $\{f_n\}$ *converges pointwise* to a function $f \in B(X)$, or that f is a *pointwise limit* of $\{f_n\}$ if, for each $x \in X$, the sequence $\{f_n(x)\}$ converges to $f(x)$ in $\mathbb{R}$ (with respect to the usual metric).

In terms of logical symbols,

$$(\forall \ x \in X) \ (\forall \ \varepsilon > 0) \ (\exists \ N \in \mathbb{N}) \ (\forall \ n \geq N) \ |f_n(x) - f(x)| < \varepsilon.$$

How is this different from convergence in $B(X)$ with respect to the sup-metric? Well, to say that f_n converges to f in the sup-metric, we need to show that $d(f_n, f) < \varepsilon$ – that is, $\sup\{|f_n(x) - f(x)|; \ x \in X\} < \varepsilon$ – whenever n is large enough. In other words, $|f_n(x) - f(x)| < \varepsilon$ for all $x \in X$ and for all $n \geq N$. In other words, we must choose our $N \in \mathbb{N}$ depending on $\varepsilon > 0$, but without knowledge of $x \in X$. The same $N \in \mathbb{N}$ must work for every $x \in X$. In symbols:

$$(\forall\ \varepsilon > 0)\ (\exists\ N \in \mathbb{N})\ (\forall\ x \in X)\ (\forall\ n \geq N)\ |f_n(x) - f(x)| < \varepsilon.$$

This is a stronger property, which is called *uniform convergence*.

Definition Let $\{f_n\}$ be a sequence in $B(X)$. We say that $\{f_n\}$ *converges uniformly* to a function $f \in B(X)$, or that f is a *uniform limit* of $\{f_n\}$ if $f_n \to f$ in $B(X)$ with respect to the sup metric, as $n \to \infty$.

We have seen above that not every pointwise convergent sequence of bounded functions is uniformly convergent. However, the converse is true: every uniformly convergent sequence is pointwise convergent. Thus uniform convergence really is a strictly stronger property than pointwise convergence.

Lemma 3.5 *Let $\{f_n\}$ be a uniformly convergent sequence in $B(X)$, with uniform limit $f \in B(X)$. Then $\{f_n\}$ is also pointwise convergent, with pointwise limit f.*

(Note that we already effectively proved this in the case of the above example of a pointwise convergent sequence of functions in $B(\mathbb{R})$ that is not uniformly convergent. The general proof is the same.)

Proof. Let $x \in X$. Then we must show that $f_n(x) \to f(x)$ as $n \to \infty$. But $|f_n(x) - f(x)| \leq d(f_n, f) \to 0$ as $n \to \infty$. The result follows.

In practice, we can apply this result to any pointwise convergent sequence of bounded functions, as follows. The pointwise limit is the only candidate for a uniform limit. We can then examine whether it really is a uniform limit or not, to decide whether or not our sequence is uniformly convergent.

Example Define $f_n : [0, 1] \to \mathbb{R}$ by

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

This is the sequence of partial sums in the Taylor series for e^x, so f_n is pointwise convergent with pointwise limit e^x. Moreover, $e^x - f_n(x) = \frac{e^t}{(n+1)!}$ for some $t \in (0, 1)$, so $d(f_n(x), e^x) \leq \frac{e}{(n+1)!} \to 0$ as $n \to \infty$. Hence f_n converges uniformly to e^x on the interval $[0, 1]$.

3.4 Cauchy sequences

Consider the following example. Let $X = \mathbb{Q}$, the set of rational numbers, and let d be the subspace metric on $\mathbb{Q}$, thought of as a subset of $\mathbb{R}$ with the usual metric. For each

natural number n, let x_n be the decimal approximation to $\sqrt{2}$, correct to n decimal places. For example,

$$\begin{aligned} x_1 &= 1 \cdot 4 \\ x_2 &= 1 \cdot 41 \\ x_3 &= 1 \cdot 414 \\ x_4 &= 1 \cdot 4142 \\ x_5 &= 1 \cdot 41421 \\ x_6 &= 1 \cdot 414214 \end{aligned}$$

Note that $x_n \in \mathbb{Q}$, since x_n is an integer multiple of 10^{-n}. Moreover, $|x_n - \sqrt{2}| < 10^{-n} \to 0$ as $n \to \infty$, so $x_n \to \sqrt{2}$ in $\mathbb{R}$ as $n \to \infty$. However, $\sqrt{2} \notin \mathbb{Q}$, so $\{x_n\}$ is not convergent in $\mathbb{Q}$. (If x_n converges to a limit $x \in \mathbb{Q}$, then $\{x_n\}$ also converges to x in $\mathbb{R}$, since we are using the same metric in both cases. But limits are unique, so $x = \sqrt{2}$, contradicting $x \in \mathbb{Q}$.)

This example illustrates the idea of what is known as a *Cauchy sequence* (named after the 19th century French mathematician Augustin Louis Cauchy – also known for the Cauchy-Riemann equations in complex analysis). The sequence does not converge in the given space $\mathbb{Q}$, but does converge in a bigger space $\mathbb{R}$, of which $\mathbb{Q}$ is a subspace with the subspace metric. A useful way of thinking about Cauchy sequences is that they are convergent, but the limit is missing from the space.

How can we make this into a formal definition? The definition of convergent sequence:

$$(\forall\ \varepsilon > 0)\ (\exists\ N \in \mathbb{N})\ (\forall\ n \geq N)\ d(x_n, x) < \varepsilon$$

involves the limit x of the sequence, but only in one place. The definition says that the terms x_n of the sequence get closer and closer to the limit point x as n gets bigger. To avoid mention of the limit point, we say instead that the terms of the sequence get closer to one another as the indices get bigger.

Definition A sequence $\{x_n\}$ in a metric space (X, d) is *Cauchy* if

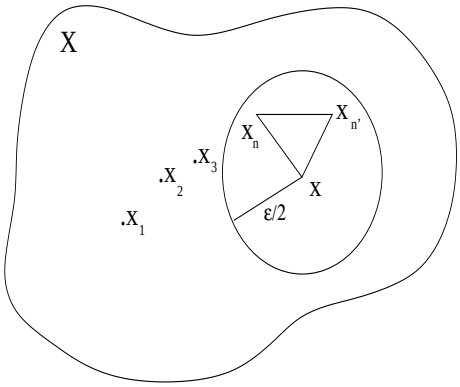
$$(\forall\ \varepsilon > 0)\ (\exists\ N \in \mathbb{N})\ (\forall\ n, n' \geq N)\ d(x_n, x_{n'}) < \varepsilon.$$

As required, this definition makes no mention of a limit, and makes sense independently of whether or not a limit exists.

If we have the right definition, then it should certainly include any sequences which are genuinely convergent. Let us check this!

Lemma 3.6 *In a metric space (X, d), every convergent sequence is Cauchy.*

Proof. Suppose that $\{x_n\}$ is a sequence in (X, d) that converges to $x \in X$. Let $\varepsilon > 0$. Then we can choose a natural number $N \in \mathbb{N}$ such that $d(x_n, x) < \frac{\varepsilon}{2}$ for all $n \geq N$.



By the triangle inequality, if $n, n' \geq N$ then

$$d(x_n, x_{n'}) \leq d(x_n, x) + d(x, x_{n'}) \leq \varepsilon.$$

Hence $\{x_n\}$ is Cauchy, as required.

As indicated above, there are many other examples of Cauchy sequences in addition to the convergent sequences.

Examples

1. Let x_n be the n decimal point approximation to $\sqrt{2}$. Then the sequence $\{x_n\}$ in $(\mathbb{Q}, d)$ is Cauchy (where d is the usual metric on $\mathbb{R}$, restricted to $\mathbb{Q}$). To see this, note that, if $n < n'$, then $|x_n - x_{n'}| < 10^{-n} < \varepsilon$ provided $n > -\log_{10}(\varepsilon)$. In the definition of Cauchy sequence, given $\varepsilon > 0$ we can choose any natural number $N > -\log_{10}(\varepsilon)$. Then $|x_n - x_{n'}| < \varepsilon$ for any $n, n' \geq N$, as required.

2. Let X be the open interval $(0, 2)$ in $\mathbb{R}$, with the subspace metric. Then the sequence $\{\frac{1}{n}\}$ in X is Cauchy. To see this, note that $|\frac{1}{n} - \frac{1}{n'}| < \frac{1}{N} < \varepsilon$ if $n, n' \geq N$ and $N > \frac{1}{\varepsilon}$. However, the sequence $\{\frac{1}{n}\}$ does not converge in X, since it converges to 0 in $\mathbb{R}$, and X is a subspace of $\mathbb{R}$ (with the subspace metric) and $0 \notin X$.

Remark In all the examples we have seen so far of Cauchy sequences which do not converge, there is a bigger space in which the sequence does converge, but the limit does not belong to the space under consideration. Am I cheating by showing you only examples of this type? Perhaps, but it cannot be helped. It turns out that *all* examples of non-convergent Cauchy sequences have that property. More about this later in the course.

Here is another example of how Cauchy sequences behave like convergent ones. Recall that every convergent sequence is bounded.

Lemma 3.7 *Every Cauchy sequence in a metric space (X, d) is bounded.*

The proof is almost entirely the same as that for convergent sequences. I have deliberately omitted it from the notes to leave as an exercise. (See the examples at the end of the chapter.)

The classes of convergent, Cauchy, and bounded sequences are thus linked as follows:

$$\text{convergent} \;\Rightarrow\; \text{Cauchy} \;\Rightarrow\; \text{bounded}.$$

Neither of these implications is reversible. We have already seen examples of Cauchy sequences that are not convergent. It is not difficult to construct examples of sequences that are bounded but not Cauchy. (See the examples at the end of the chapter.)

3.5 Sequences and closed sets

We have come across the word 'limit' earlier in this course, in a different context. In the previous chapter we defined the notion of a *limit point* or *cluster point* of a set A in a metric space (X, d), and showed that a set is closed if and only if it contains all its limit points.

I shall show below that limit points of A are essentially the same thing as limits (in X) of sequences in A. This gives rise to a new characterisation of closed sets in terms of sequences.

Recall first the definition: x is a limit point of A in (X, d) if

$$(\forall\, \varepsilon > 0)\ (\exists\, a \in A)\ 0 < d(a, x) < \varepsilon.$$

Lemma 3.8 *Let (X, d) be a metric space, A a subset of X, and $x \in X$. Then x is a limit point of A in (X, d) if and only if there is a sequence $\{x_n\}$ in A with $x_n \neq x$ for all n and $x_n \to x$ in (X, d) as $n \to \infty$.*

Proof. First of all, suppose that x is a limit point of A. For each $n \in \mathbb{N}$, taking $\varepsilon = \frac{1}{n}$ in the definition of limit point, we can choose an element $x_n \in A$ with $0 < d(x_n, x) < \frac{1}{n}$. In particular $x_n \neq x$. However, the sequence of real numbers $\{d(x_n, x)\}$ converges to 0 in $\mathbb{R}$ (with the usual metric), since $|d(x_n, x)| < \frac{1}{n} \to 0$. Hence $\{x_n\}$ converges to x in (X, d).

Conversely, suppose that $\{x_n\}$ is a sequence in (X, d) that converges to x, with $x_n \neq x$ and $x_n \in A$ for all $n \in \mathbb{N}$. Let $\varepsilon > 0$, and choose $N \in \mathbb{N}$ such that $d(x_n, x) < \varepsilon$ for all $n \geq N$. Then in particular $x_N \in A$ with $x_N \neq x$ and $d(x_N, x) < \varepsilon$. Hence x is a limit point of A.

Corollary 3.9 *A set A in a metric space (X, d) is closed if and only if, whenever $\{x_n\}$ is a convergent sequence in X with $x_n \in A\ \forall\ n \in \mathbb{N}$, then $\lim_n(x_n) \in A$.*

3.6 Exercises on sequences

1. Determine which of the following sequences are convergent in $\mathbb{R}^2$ with respect to the euclidean metric, and give the limit where it exists:

   (a) $\{(\frac{1}{n^2}, \frac{2}{n})\}$;

   (b) $\{(\frac{n-1}{n}, \frac{n^2-1}{n})\}$;

   (c) $\{(\frac{n-1}{n}, \frac{2n-1}{n})\}$.

2. Find the pointwise limit f of the sequence of functions $\{f_n\}$ defined on the real interval X in each of the following cases:

$$\text{(i)} \quad X = \mathbb{R}; \qquad f_n(x) = \frac{nx}{1 + n^2x^2}$$

$$\text{(ii)} \quad X = [0, 1]; \quad f_n(x) = \frac{x^n}{n}$$

$$\text{(iii)} \quad X = \mathbb{R} \qquad f_n(x) = \begin{cases} 1 & \text{if } -n \leq x \leq n \\ 0 & \text{if} |x| > n. \end{cases}$$

   In each case determine whether or not $\{f_n\}$ converges to f in the metric space $B(X)$.

3. Working in $(\mathbb{R}, d)$ where d is the usual metric give examples, one in each case, of

   (a) a bounded sequence which is not convergent;

   (b) a sequence with no convergent subsequence;

   (c) an unbounded sequence with two convergent subsequences which converge to different limits.

4. Let (X, d) be a metric space. Prove that every Cauchy sequence is bounded.

5. Give an example of a bounded sequence in $\mathbb{Q}$ (with the usual metric) that is not Cauchy.

6. Suppose $\{f_n\}$ and $\{g_n\}$ are sequences of bounded real-valued functions which are convergent in the metric space $B(X)$ to f and g, respectively. Prove that $\{f_n + g_n\}$ is convergent to $f + g$ in $B(X)$.

   [Here $+$ denotes the pointwise sum of two functions: $(f + g)(x) = f(x) + g(x)$.]

7. Let (X, d) be a discrete metric space. Find a simple necessary and sufficient condition for a sequence $\{x_n\}$ to be convergent in X.

Chapter 4

Continuity

4.1 Maps between metric spaces

Recall that a function $f : \mathbb{R} \to \mathbb{R}$ is said to be *continuous* at a point $a \in \mathbb{R}$ if points close to a are mapped close to $f(a)$. Formally, the condition is:

$$(\forall\ \varepsilon > 0)\ (\exists\ \delta > 0)\ |x - a| < \delta \Rightarrow |f(x) - f(a)| < \varepsilon.$$

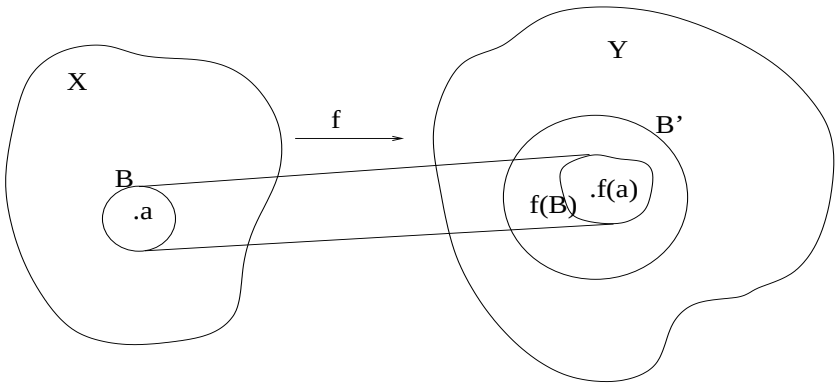
Since this is a metric condition, in other words it is defined in terms of the (usual) metric $d(x, y) = |x - y|$ on $\mathbb{R}$, we can generalise it directly to arbitrary metric spaces.

Definition Let (X, d) and (Y, ρ) be two metric spaces, $f : X \to Y$ a function, and $a \in X$. We say that f is *continuous* at a (with respect to the metrics d and ρ) if

$$(\forall\ \varepsilon > 0)\ (\exists\ \delta > 0)\ d(x, a) < \delta \Rightarrow \rho(f(x), f(a)) < \varepsilon.$$

Equivalently:

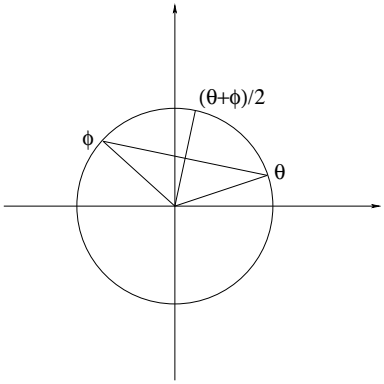
$$(\forall\ \varepsilon > 0)\ (\exists\ \delta > 0)\ f(B_\delta(a)) \subseteq B_\varepsilon(f(a)).$$



Definitions If $f : X \to Y$ is not continuous at $a \in X$, then we say that f is *discontinuous* at a, or that f has a *discontinuity* at a. If $f : X \to Y$ is continuous at a for every $a \in X$, then we say that f is *continuous* on X, or just that f is *continuous*.

Examples

1. $f : \mathbb{R} \to \mathbb{R}^2$, $f(\theta) = (\cos\theta, \sin\theta)$ is continuous with respect to the usual metric on $\mathbb{R}$ and the euclidean metric on $\mathbb{R}^2$. To see this, let $\theta, \phi \in \mathbb{R}$ and consider the distance $d(f(\theta), f(\phi))$ in the euclidean metric d. Elementary trigonometry shows that
$$d(f(\theta), f(\phi)) = \left|2\sin\left(\frac{\theta-\phi}{2}\right)\right| \leq |\theta - \phi|.$$
(The straight-line path between two points on the circle is shorter than any path round the circle between these points.)



In the definition of continuity, suppose we are given $\varepsilon > 0$. Then we can set $\delta = \varepsilon$. If $|\theta - \phi| < \delta = \varepsilon$, then $d(f(\theta), f(\phi)) < \varepsilon$, as required.

2. $f : \mathbb{R}^2 \to \mathbb{R}$,
$$f(x,y) = \begin{cases} \dfrac{x^3+y^3}{2x^2+y^2} & (x,y) \neq (0,0) \\ \\ 0 & (x,y) = (0,0) \end{cases}$$
is continuous at $(0,0)$ with respect to the euclidean metric on $\mathbb{R}^2$ and the usual metric on $\mathbb{R}$.

To see this, use polar coordinates (r, θ) on $\mathbb{R}^2$, where $x = r\cos\theta$ and $y = r\sin\theta$, so that $r = d((x,y),(0,0)$. We need to show that $|f(r\cos\theta, r\sin\theta)|$ is small when

r is small (independently of θ). Now

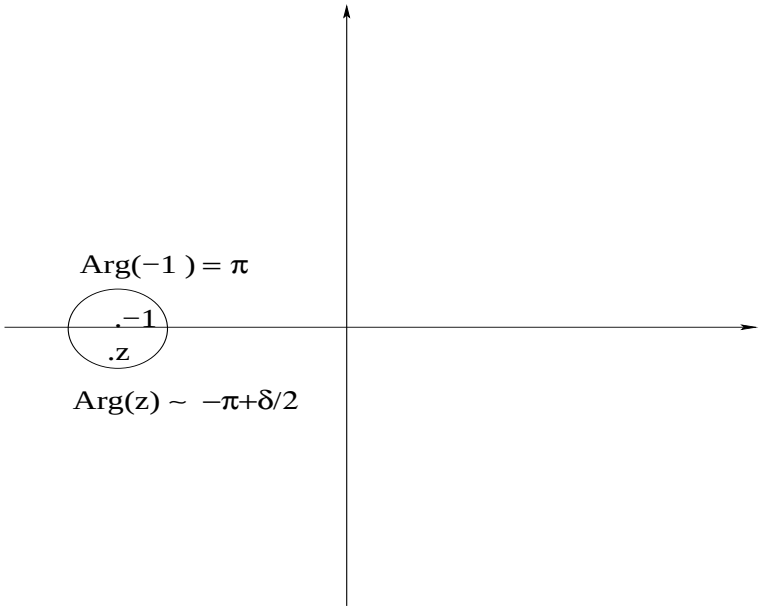
$$\begin{aligned} |f(r\cos\theta, r\sin\theta)| &= \left|\frac{r^3(\cos^3\theta + \sin^3\theta)}{r^2(2\cos^2\theta + \sin^2\theta)}\right| \\ &= r\frac{|\cos^3\theta + \sin^3\theta|}{1+\cos^2\theta} \\ &\leq r|\cos^3\theta + \sin^3\theta| < 2r. \end{aligned}$$

Thus, given $\varepsilon > 0$ (as in the definition of continuity), we can put $\delta = \frac{\varepsilon}{2}$. Then, if $r = d((x,y),(0,0)) < \delta$, it follows that $|f(x,y)| < 2r < \varepsilon$, as required.

3. $f : \mathbb{C} \setminus \{0\} \to \mathbb{R}$, $f(z) = \mathrm{Arg}(z)$ (the unique angle $\theta \in (-\pi, \pi]$ such that $z = |z|e^{i\theta}$), is discontinuous at all negative real numbers (with respect to the usual metrics $d(x,y) = |x-y|$ on $\mathbb{R}$ and $\mathbb{C}$).

   For example, $\mathrm{Arg}(-1) = \pi$, but $\mathrm{Arg}(-1 - i\delta) = -\pi + \tan^{-1}(\delta)$ if $\delta > 0$.

   To see that f is not continuous at -1, choose $\varepsilon = \pi$ in the definition of continuity. Given any $\delta > 0$, put $z = -1 - \frac{i\delta}{2}$. Then $d(z,-1) = \frac{\delta}{2} < \delta$, but $|f(z) - f(-1)| = 2\pi - \tan^{-1}(\frac{\delta}{2}) > \pi = \varepsilon$.



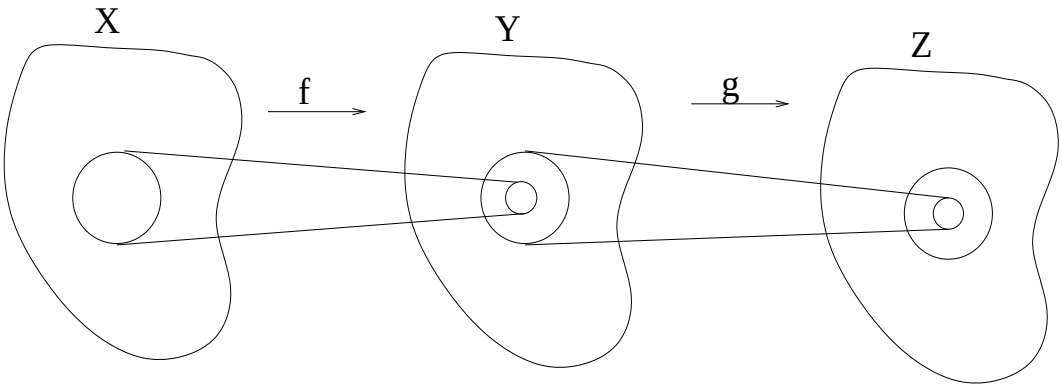
4. Let d be the sup-metric on $B(X)$, and let x_0 be an element of X. Then $\alpha : B(X) \to \mathbb{R}$, $\alpha(f) = f(x_0)$, is continuous on $B(X)$ with respect to d and the usual metric on $\mathbb{R}$.

To see this, note that, for any $f, g \in B(X)$,

$$\begin{aligned} |\alpha(f) - \alpha(g)| &= |f(x_0) - g(x_0)| \\ &\leq \sup_{x \in X} |f(x) - g(x)| \\ &= d(f, g). \end{aligned}$$

Given $\varepsilon > 0$ in the definition of continuity, put $\delta = \varepsilon$. Then $d(f, g) < \delta = \varepsilon \Rightarrow |\alpha(f) - \alpha(g)| < \varepsilon$, as required.

Theorem 4.1 (Composite of continuous maps is continuous) *Suppose that (X, d), (Y, ρ) and (Z, σ) are metric spaces, $f : X \to Y$ is continuous (with respect to the metrics d, ρ) at a point $x \in X$, $g : Y \to Z$ is continuous (with respect to the metrics ρ, σ) at $f(x) \in Y$, and $h = g \circ f : X \to Z$ is their composite. Then h is continuous at x with respect to the metrics d, σ.*



Proof. Suppose that $\varepsilon > 0$. Since g is continuous at $f(x)$, we can find $\delta > 0$ such that $g(B_\delta(f(x))) \subset B_\varepsilon(g(f(x)))$.

Now, since f is continuous at x, we can find $\gamma > 0$ such that $f(B_\gamma(x)) \subset B_\delta(f(x))$. Hence

$$h(B_\gamma(x)) = g(f(B_\gamma(x))) \subset g(B_\delta(f(x)) \subset B_\varepsilon(g(f(x))) = B_\varepsilon(h(x)).$$

By definition, h is continuous at x.

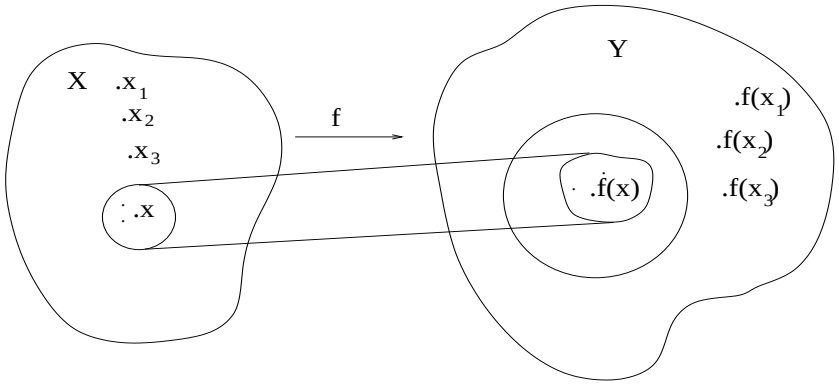
Examples

1. If $h : \mathbb{R}^2 \to \mathbb{R}^2$ is defined by $h(x, y) = (\cos(x + y), \sin(x + y))$, then h is the composite of two continuous functions $f : \mathbb{R}^2 \to \mathbb{R}$ ($f(x, y) = x + y$) and $g : \mathbb{R} \to \mathbb{R}^2$ ($g(\theta) = (\cos\theta, \sin\theta)$), with respect to the usual metrics on $\mathbb{R}$ and $\mathbb{R}^2$. Hence h is continuous with respect to the usual metric on $\mathbb{R}^2$.

2. $x \mapsto \frac{1}{4}\cosh(\exp(\sin(2x^2 + x + 3)))$ is a composite of 5 continuous functions $\mathbb{R} \to \mathbb{R}$ with respect to the usual metric, so is also a continuous function $\mathbb{R} \to \mathbb{R}$ with respect to the usual metric.

4.2 Continuity and sequences

Intuitively, a continuous function is one which maps nearby points to nearby points. A convergent sequence is one whose terms are ultimately arbitrarily close to the limit. Hence applying a continuous function to a convergent sequence, we should get a sequence whose terms are ultimately arbitrarily close to the image (under the function) of the limit.



In fact, it turns out that this property is an alternative way of characterising continuous function. The precise formal statement is as follows.

Theorem 4.2 *Let (X, d) and (Y, ρ) be metric spaces, and $x \in X$. Then f is continuous at x (with respect to d, ρ) if and only if, for every sequence $\{x_n\}$ that converges to x in (X, d), the sequence $\{f(x_n)\}$ converges to $f(x)$ in (Y, ρ).*

Proof. First suppose that f is continuous at x, and let $\{x_n\}$ be a sequence converging to x in (X, d).

Continuity of f means

$$(\forall\ \varepsilon > 0)\ (\exists\ \delta > 0)\ d(x', x) < \delta \Rightarrow \rho(f(x'), f(x)) < \varepsilon. \tag{4.1}$$

Convergence of the sequence means

$$(\forall\ \delta > 0)\ (\exists\ N \in \mathbb{N})\ (\forall\ n \geq N)\ d(x_n, x) < \delta. \tag{4.2}$$

Combining (4.1) and (4.2) (with the same δ, and putting $x' = x_n$) gives

$$(\forall\ \varepsilon > 0)\ (\exists\ N \in \mathbb{N})\ (\forall\ n \geq N)\ \rho(f(x_n), f(x)) < \varepsilon.$$

In other words, the sequence $\{f(x_n)\}$ converges in (Y, ρ) to $f(x)$, as required.

Conversely, suppose that f is discontinuous at x. Then there is a positive real number ε such that, for all $\delta > 0$, there is an $x' \in X$ with $d(x', x) < \delta$ but $\rho(f(x'), f(x)) > \varepsilon$.

For each $n \in \mathbb{N}$ put $\delta = \frac{1}{n}$ in the above, and let x_n be a suitable choice of x'. In other words, we have a sequence $\{x_n\}$ in X with $d(x_n, x) < \frac{1}{n}$ for all n, but $\rho(f(x_n), f(x)) > \varepsilon$ for all n.

It follows that $\{x_n\}$ converges to x in (X, d), but that $\{f(x_n)\}$ does not converge to $f(x)$.

[Note: it is possible that the sequence $\{f(x_n)\}$ does converge in (Y, ρ), but if so its limit is not $f(x)$.]

Examples

1. Define $f : \mathbb{R}^2 \to \mathbb{R}$ by

$$f(x,y) = \begin{cases} \dfrac{x^2y + xy^2}{x^2 + y^2} & (x,y) \neq (0,0) \\ \\ 0 & (x,y) = (0,0). \end{cases}$$

   Then f is continuous (with respect to the euclidean metrics on $\mathbb{R}^2$ and $\mathbb{R}$).

   At $(x, y) \neq (0, 0)$ it is clearly continuous because f is defined by a quotient of polynomials with the denominator nonzero. We need only show that f is continuous at $(0, 0)$.

   Intuitively, this is true because the numerator $x^2y + xy^2$ in the formula is a homogeneous polynomial of degree 3, and so converges to 0 more rapidly then the denominator $x^2 + y^2$, which is homogeneous of degree 2.

   To prove it properly, we show that, for any sequence $\{(x_n, y_n)\}$ in $\mathbb{R}^2$ that converges to $(0, 0)$, the corresponding sequence $\{f(x_n, y_n)\}$ converges to 0 in $\mathbb{R}$.

   Note that $\left|\frac{x^2y}{x^2+y^2}\right| \leq \left|\frac{x^2y}{x^2}\right| = |y|$. Similarly $\left|\frac{xy^2}{x^2+y^2}\right| \leq |x|$, so if $x^2 + y^2 = r^2$ then $|f(x, y)| \leq |x| + |y| \leq 2r \to 0$ as $r \to 0$.

   If $(x_n, y_n) \to (0, 0)$ in $\mathbb{R}^2$, and $\varepsilon > 0$, then there is an $N \in \mathbb{N}$ such that, for all $n \geq N$, $x_n^2 + y_n^2 < \frac{\varepsilon^2}{4}$, and so $|f(x_n, y_n)| < \varepsilon$. Hence $\{f(x_n, y_n)\}$ converges in $\mathbb{R}$ (with the usual metric) to $0 = f(0, 0)$. Hence f is continuous at $(0, 0)$.

2. The function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x,y) = \begin{cases} \dfrac{2xy}{x^2 + y^2} & (x,y) \neq (0,0) \\ \\ 0 & (x,y) = (0,0) \end{cases}$$

   has a discontinuity at $(0, 0)$.

   This time, the reason is because the numerator and denominator are both homogeneous polynomials of the same degree.

To prove that f is discontinuous at $(0,0)$, we need to find a sequence $\{(x_n, y_n)\}$ converging to $(0,0)$ in $\mathbb{R}^2$, such that the corresponding sequence $\{f(x_n, y_n)\}$ does not converge to 0 in $\mathbb{R}$. Note that we need only find *one* sequence with this property to prove discontinuity. We do not need to prove the property for all sequences.

Some care is needed in the choice of sequence. For example, if we try $\{(\frac{1}{n}, 0)\}$, we find that $f(\frac{1}{n}, 0) = 0$ for all n. However, a different choice, $\{(\frac{1}{n}, \frac{1}{n})\}$, will work, since $f(\frac{1}{n}, \frac{1}{n}) = 1$ for all n.

3. Define $f : \mathbb{R} \to \mathbb{R}$ by

$$f(x) = \begin{cases} \cos(\frac{1}{x}) & x \neq 0 \\ 0 & x = 0. \end{cases}$$

Then f is discontinuous at 0. For example, if $x_n = \frac{1}{n\pi}$, then $x_n \to 0$ as $n \to \infty$, but $f(x_n) = (-1)^n \not\to 0$ as $n \to \infty$.

4.3 Continuity and open sets

Suppose that $f : X \to Y$ is a continuous map (with respect to metrics d on X and ρ on Y).

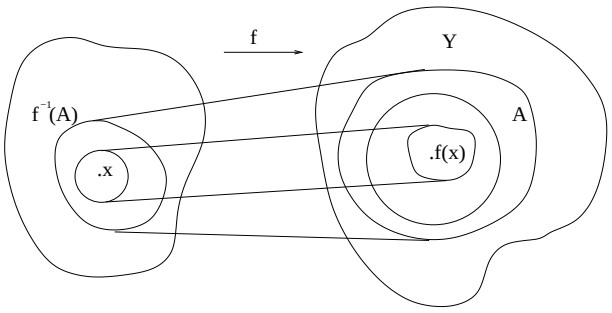
Suppose that $A \in Y$ is an open set,, $x \in X$ and $f(x) \in A$.

Then the definition of open set says

$$(\exists\ \varepsilon > 0)\ B_\varepsilon(f(x)) \subset A.$$

Now the definition of continuity says

$$(\exists\ \delta > 0)\ f(B_\delta(x)) \subset B_\varepsilon(f(x)) \subset A.$$



This suggests a criterion for continuity in terms of open sets

Definition If $f : X \to Y$ is a map and $A \subset Y$, then the *preimage* or the *inverse image* of A under f is the subset

$$f^{-1}(A) := \{x \in X;\ f(x) \in A\}$$

of X.

Example Let $X = Y = \mathbb{R}$, $A = (0, \infty)$ and $f : X \to Y$ be $f(x) = \sin(x)$. Then

$$f^{-1}(A) = \{x \in \mathbb{R};\ \sin(x) > 0\} = \bigcup_{n \in \mathbb{Z}} (2n\pi, (2n+1)\pi).$$

Theorem 4.3 *Let (X, d) and (Y, ρ) be metric spaces and $f : X \to Y$ a map. Then f is continuous (with respect to d and ρ) if and only if, for every open set A in the metric space (Y, ρ), its preimage $f^{-1}(A)$ is an open set in the metric space (X, d).*

Proof. Suppose first that f is continuous. Let A be an open set in (Y, ρ). We must show that $f^{-1}(A)$ is open in (X, d).

Suppose that $x \in f^{-1}(A)$. Then, by definition, $f(x) \in A$. By the remarks above, there is a positive real number δ such that $f(B_\delta(x)) \subset A$. In other words, if $y \in B_\delta(x)$ then $f(y) \in A$, and so $y \in f^{-1}(A)$. Thus $B_\delta(x) \subset f^{-1}(A)$. It follows that $f^{-1}(A)$ is open in (X, d), as required.

Conversely, suppose that $f^{-1}(A)$ is open in (X, d) whenever A is open in (Y, ρ). We must show that f is continuous at all $x \in X$.

So, let $x \in X$ and $\varepsilon > 0$. The open ball $B_\varepsilon(f(x))$ in (Y, ρ) is an open set, so by hypothesis $f^{-1}(B_\varepsilon(f(x)))$ is open in (X, d). Also $x \in f^{-1}(B_\varepsilon(f(x)))$, since $f(x) \in B_\varepsilon(f(x))$. By definition of open set, there is a $\delta > 0$ such that $B_\delta(x) \subset f^{-1}(B_\varepsilon(f(x)))$. If $x' \in X$ with $d(x, x') < \delta$, then $x' \in B_\delta(x) \subset f^{-1}(B_\varepsilon(f(x)))$. In other words, $f(x') \in B_\varepsilon(f(x))$, so $\rho(f(x'), f(x)) < \varepsilon$.

This proves that f is continuous at x, as required.

Example The function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = x + y$ is continuous. The set $A = (0, \infty) \subset \mathbb{R}$ is open in $\mathbb{R}$. Hence the set $f^{-1}(A) = \{(x, y) \in \mathbb{R}^2 :\ x + y > 0\}$ is open in $\mathbb{R}^2$. (All this is with respect to the usual metric on $\mathbb{R}^2$ and $\mathbb{R}$.)

Using the fact that closed sets are the complements of open sets, together with the observation that, for any map $f : X \to Y$ and subset $A \subset Y$,

$$f^{-1}(Y \setminus A) = \{x \in X;\ f(x) \notin A\} = X \setminus f^{-1}(A),$$

we can translate all of the above into a criterion for continuity in terms of closed sets.

Theorem 4.4 *Let (X, d) and (Y, ρ) be metric spaces and $f : X \to Y$ a map. Then f is continuous (with respect to d and ρ) if and only if, for every closed set A in the metric space (Y, ρ), its preimage $f^{-1}(A)$ is a closed set in the metric space (X, d).*

Proof. If f is continuous and $A \subset Y$ is closed, then $Y \setminus A$ is open, so $X \setminus f^{-1}(A) = f^{-1}(Y \setminus A)$ is open, so $f^{-1}(A)$ is closed.

Conversely, if the preimage of every closed set is closed and $A \subset Y$ is open, then $Y \setminus A$ is closed, so $X \setminus f^{-1}(A) = f^{-1}(Y \setminus A)$ is closed, so $f^{-1}(A)$ is open. It follows that f is continuous.

Example $\{(x, y) \in \mathbb{R}^2 : \ x + 2y \leq 1\}$ is a closed subset of $\mathbb{R}^2$ (with respect to the usual metric), since $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = x + 2y$, is continuous, and $\{(x, y) \in \mathbb{R}^2 : \ x + 2y \leq 1\}$ is the preimage under f of the closed set $(-\infty, 1] \subset \mathbb{R}$.

Summary

We have proved the following, which give several different criteria for continuity of a function between metric spaces.

Theorem 4.5 *Let (X, d) and (Y, ρ) be metric spaces, and $f : X \to Y$ a function. Then the following are equivalent:*

(i) *f is continuous (with respect to d and ρ);*

(ii) *for every convergent sequence $\{x_n\}$ in (X, d), the sequence $\{f(x_n)\}$ is convergent in (Y, ρ), with $\lim_{n\to\infty} f(x_n) = f(\lim_{n\to\infty} x_n)$;*

(iii) *for every open set A in (Y, ρ), $f^{-1}(A)$ is open in (X, d);*

(iv) *for every closed set A in (Y, ρ), $f^{-1}(A)$ is closed in (X, d).*

4.4 Homeomorphisms and equivalent metrics

In this section I want to discuss what it means for two metric spaces to be 'the same'. Recall, for example, that two vector spaces (over $\mathbb{R}$) are called *isomorphic*, and regarded as being the same, if there is a bijective map between them that is linear (meaning that it preserves all the structure inherent in a vector space, namely addition and scalar multiplication). The analogous concept for metric spaces is a bijective map preserving the metric space structure (that is, the metric). Such a map is called an *isometry*.

Definition Let (X, d) and (Y, ρ) be metric spaces. A map $f : X \to Y$ is called an *isometry* if it is bijective (that is, $f^{-1}(\{y\})$ contains precisely one element, for each $y \in Y$) and satisfies:

$$\rho(f(x), f(x')) = d(x, x') \ \ \forall \ x, x' \in X.$$

Two metric spaces are said to be *isometric* if there exists an isometry between them.

It is not difficult to show that the notion of isometry is an equivalence relation on metric spaces. (The identity function $X \to X$ is an isometry; the inverse of an isometry is an isometry; the composite of two isometries is an isometry.) We regard isometric metric spaces as being 'the same'.

Exercise Show that every isometry $X \to Y$ is continuous.

There is a weaker equivalence relation than isometry, defined on metric spaces using continuous maps as follows.

Definition Let (X, d) and (Y, ρ) be metric spaces. A map $f : X \to Y$ is called a *homeomorphism* if it is bijective and both $f : X \to Y$ and $f^{-1} : Y \to X$ are continuous (with respect to d and ρ). In this case the spaces (X, d) and (Y, ρ) are said to be *homeomorphic*.

Homeomorphic spaces are not in general isometric. Nevertheless, many of their properties are similar: for example convergence of sequences, open sets, and closed sets.

If X is a fixed set with two metrics d_1 and d_2, one can consider the identity map $X \to X$ as a map between the two metric spaces (X, d_1) and (X, d_2). If this is a homeomorphism from (X, d_1) to (X, d_2), then we say that the metrics d_1 and d_2 on X are *equivalent*. Equivalent metrics give rise to the same collection of open sets (the *topology*) in X. They also, of course, give rise to the same collection of closed sets, and to the same collection of convergent sequences. Moreover, a convergent sequence converges to the same limit with respect to the two metrics.

Examples

1. The max-metric, the Manhattan metric, and the euclidean metric on $\mathbb{R}^n$ are all equivalent.

2. If d is a metric on X and λ is a positive real number, then $\lambda \cdot d$ is a metric equivalent to d.

3. If X is a finite set, then every metric on X is equivalent to the discrete metric.

   To see this, note that any one-element set $\{x\}$ is closed, so any subset of X is a union of finitely many closed sets, so is closed.

   This is true for all metrics on X, so the identity map on X is continuous with respect to any two metrics d, d' on X, and so a homeomorphism between (X, d) and (X, d').

4.5 Exercises on continuous functions

1. If $f : \mathbb{R} \to \mathbb{R}$ is defined by

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

show that f is not continuous at $x = 0$ (with respect to the usual metric).

2. If $f : \mathbb{R}^2 \to \mathbb{R}$ is defined by

$$f(x, y) = \begin{cases} \dfrac{x^2 y}{x^4 + y^2} & \text{if } (x, y) \neq (0, 0) \\ \\ 0 & \text{if } (x, y) = (0, 0) \end{cases}$$

show that f is not continuous at $(0, 0)$ (with respect to the usual metrics).

3. If $f : \mathbb{R} \to \mathbb{R}$ is defined by

$$f(x) = \begin{cases} x & \text{if } x \text{ is irrational} \\ 1 - x & \text{if } x \text{ is rational} \end{cases}$$

show that f is continuous at $\dfrac{1}{2}$ and discontinuous at every other point of $\mathbb{R}$ (with respect to the usual metric).

4. Suppose $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ are continuous on $\mathbb{R}$. Show that $h : \mathbb{R} \to \mathbb{R}^2$ such that $h(x) = (f(x), g(x))$ is continuous on $\mathbb{R}$ (with respect to the usual metrics).

5. Suppose $f : \mathbb{R}^2 \to \mathbb{R}$ is continuous. Show that $g : \mathbb{R}^2 \to \mathbb{R}$ such that $g(x, y) = f(x + y, x - y)$ is continuous on $\mathbb{R}^2$ (with respect to the usual metrics).

6. Let X be a set and let $x_0 \in X$. Define a map $\theta : B(X) \to \mathbb{R}$ by $\theta(f) = f(x_0)$. Show that θ is continuous with respect to the sup-metric on $B(X)$ and the usual metric on $\mathbb{R}$.

7. Let (X, d) be a discrete metric space and let (Y, ρ) be any metric space. Prove that any function $f : X \to Y$ is continuous.

8. Let (X, d) be a metric space and let $x_0 \in X$. Show that the function $f: X \to \mathbb{R}$ defined by $f(x) = d(x, x_0)$ is continuous on X.

Chapter 5

Compactness and completeness

5.1 Compact sets in metric spaces

Closed intervals $[a, b]$ in the real line have many nice properties. A good example is that a continuous, real-valued function on $[a, b]$ is bounded and attains its bounds. Other subsets of $\mathbb{R}$ do not in general have these properties.

For example, on the set $(0, 1)$, the function $x \mapsto \frac{1}{x}$ is continuous but unbounded, since $\frac{1}{x} \to \infty$ as $x \to 0$. On $[0, \infty)$, the function $x \mapsto e^{-x}$ is bounded ($0 < e^{-x} \leq 1$) but does not attain its lower bound 0.

Compact sets in general metric spaces share many of the nice properties of closed intervals in $\mathbb{R}$, and so are useful in metric space theory.

Definition A set Z in a metric space (X, d) is *compact* if if every sequence in A has a subsequence that converges (in (X, d)) to a limit that belongs to A.

Examples

1. The set $\mathbb{R} \subset \mathbb{R}$ is not compact (with respect to the usual metric). For example, the sequence $\{n\}$ has no bounded subsequence, and hence no convergent subsequence. (Remember: convergent sequences are bounded.)

2. The set $(0, 1] \subset \mathbb{R}$ is not compact (with respect to the usual metric on $\mathbb{R}$). For example, the sequence $\{\frac{1}{n}\}$ converges in $\mathbb{R}$ to 0. It follows that every subsequence also converges to 0. By uniqueness of limit, no subsequence can converge to an element of $(0, 1]$.

3. $[0, 1] \subset \mathbb{R}$ is compact (with respect to the usual metric on $\mathbb{R}$).

   Proof. Suppose that $\{x_n\}$ is a sequence in $[0, 1]$. By the definition of compactness, we have to select a subsequence of $\{x_n\}$ which converges to a limit in $[0, 1]$. We will construct the limit x of such a subsequence as an infinite decimal $0 \cdot i_1 i_2 i_3 \ldots$, where $i_j \in \{0, 1, \ldots, 9\}$ for each $j \in \mathbb{N}$.

We do this construction by induction. At the same time, at the n-th inductive step, we select the term x_{m_n} in the subsequence. (In other words, we select the natural number m_n in the increasing sequence $\{m_n\}$.)

To begin the process, regard the sequence as a function

$$x : \mathbb{N} \to [0, 1].$$

The union of the 10 subintervals $[0 \cdot i, 0 \cdot i + 0.1]$ $(i = 0, 1, \ldots, 9)$ is the whole of $[0, 1]$, so the union of their preimages

$$A_i := x^{-1}([0 \cdot i, 0 \cdot i + 0 \cdot 1] = \{n \in \mathbb{N};\ 0 \cdot i \leq x \leq 0 \cdot i + 0 \cdot 1\}$$

is the whole of $\mathbb{N}$. In particular, $\bigcup_{i=0}^{9} A_i$ is infinite, so at least one of the A_i is infinite. Choose i_1 such that A_{i_1} is infinite. In particular, $A_{i_1} \neq \emptyset$, so we can choose $m_1 \in A_{i_1}$. Hence $x_{m_1} \in [0 \cdot i_i, 0 \cdot i_1 + 0 \cdot 1]$.

Suppose, by way of induction, that we have chosen $i_1, \ldots, i_k \in \{0, 1, \ldots, 9\}$ and $m_1 < m_2 < \cdots < m_k \in \mathbb{N}$ such that

$$x^{-1}([0 \cdot i_1 i_2 \ldots i_{k-1} i_k,\ 0 \cdot i_1 i_2 \ldots i_{k-1} i_k + 10^{-k}])$$

is infinite, and

$$x_{m_j} \in [0 \cdot i_1 i_2 \ldots i_{j-1} i_j,\ 0 \cdot i_1 i_2 \ldots i_{j-1} i_j + 10^{-j}]$$

for each $j = 1, \ldots, k$.

Then at least one of the sets

$$B_i = x^{-1}([0 \cdot i_1 i_2 \ldots i_{k-1} i_k i,\ 0 \cdot i_1 i_2 \ldots i_{k-1} i_k i + 10^{-(k+1)}]$$

is infinite. Choose i_{k+1} such that $B_{i_{k+1}}$ is infinite, and choose m_{k+1} such that $m_{k+1} \in B_{i_{k+1}}$ and $m_{k+1} > m_k$.

By induction, we have chosen a sequence $i_n \in \{0, 1, \ldots, 9\}$ of digits and a strictly ascending sequence m_n of natural numbers, such that

$$0 \cdot i_1 i_2 \ldots i_n \leq x_{m_n} \leq 0 \cdot i_1 i_2 \ldots i_n + 10^{-n}$$

for each n. Define x to be the real number given by the infinite decimal expression

$$x = 0 \cdot i_1 i_2 \ldots i_n i_{n+1} \ldots$$

Then, for each $n \in \mathbb{N}$, $|x - x_{m_n}| \leq 10^{-n} \to 0$ as $n \to \infty$. It follows that the sequence $\{x_{m_n}\}$ converges to x.

Theorem 5.1 *Every compact set in a metric space (X, d) is closed and bounded.*

Proof. Suppose that A is a compact set in the metric space (X, d). Suppose that $x \in X$ is a limit point of A. We show that $x \in A$, from which it follows that A is closed.

There is a sequence $\{a_n\}$ in A which converges in (X, d) to x. By definition of compactness, there is a subsequence $\{a_{m_n}\}$ of $\{a_n\}$ that converges to a point $a \in A$. Now every subsequence of a convergent sequence converges to the same limit as the original sequence, so $\{a_{m_n}\}$ converges to x. By the uniqueness of limits, $a = x$, and so $x \in A$ as claimed.

Now suppose that A is unbounded. Then, in particular A is nonempty, and so we can choose a point $a \in A$. For each $n \in \mathbb{N}$, $A \not\subset B_n(a)$, since A is unbounded. Hence we can choose $a_n \in A$ with $d(a_n, a) > n$. This defines a sequence $\{a_n\}$ in A. By hypothesis, A is compact, so there is a subsequence $\{a_{m_n}\}$ which is convergent, and hence bounded.

But $d(a, a_{m_n}) > m_n \geq n$ for all $n \in \mathbb{N}$, so the subsequence $\{a_{m_n}\}$ is unbounded. This is a contradiction, so the set A must be bounded, as claimed.

The converse of this theorem is false in general. For example, if X is a discrete metric space then every set in X is closed and bounded, but a subset A of X is compact if and only if it is finite. (See the examples at the end of this chapter.)

On the other hand, for some nice metric spaces there is a converse to the theorem, as follows.

Theorem 5.2 *Let $X = \mathbb{R}^n$, with the euclidean metric. Then a subset $A \subset X$ is compact if and only if A is closed and bounded.*

Proof. We have seen above that the subset $[0, 1]$ of $\mathbb{R} = \mathbb{R}^1$ is compact. A similar argument shows that any closed interval $[a, b]$ in $\mathbb{R}$ is compact. In fact, these two sets are homeomorphic: the map $f : [0, 1] \to [a, b]$, $f(t) = a + t(b - a)$, is a continuous bijection, whose inverse $g : [a, b] \to [0, 1]$, $g(x) = \frac{x-a}{b-a}$, is also continuous.

Now suppose that $\{x_n\}$ is a sequence in $[a, b]$. Then $\{g(x_n)\}$ is a sequence in $[0, 1]$. Since $[0, 1]$ is compact, this sequence has a convergent subsequence $\{g(x_{m_n})\}$ say, with limit $t \in [0, 1]$.

Since f is continuous, the sequence $\{x_{m_n}\} = \{f(g(x_{m_n}))\}$ is also convergent, and moreover its limit is $f(t) \in [a, b]$.

We now show by induction on k that any subset of the form $A = [a_1, b_1] \times \cdots \times [a_k, b_k]$ of $\mathbb{R}^k$ (with the euclidean metric) is compact. We have just seen that this is true for $k = 1$. Suppose then that $k > 1$, and that the statement holds in dimensions less than k.

Write $a = a_k$, $b = b_k$, and $B = [a_1, b_1] \times \cdots \times [a_{k-1}, b_{k-1}]$. Then $A = B \times [a, b]$, and by inductive hypothesis we know that B and $[a, b]$ are both compact.

Let $\{(b_n, x_n)\}$ be a sequence in $A = B \times [a, b]$. Since B is compact, there is a subsequence $\{b_{m_n}\}$ of $\{b_n\}$ that converges in B (to a limit $c \in B$, say).

Since $[a, b]$ is compact, the sequence $\{x_{m_n}\}$ has a subsequence $\{x_{m_{p_n}}\}$ that converges in $[a, b]$ (to a limit $d \in [a, b]$, say).

Now $\{b_{m_{p_n}}\}$ is a subsequence of the convergent sequence $\{b_{m_n}\}$, so it converges to the same limit $c \in B$.

Hence $\{(b_{m_{p_n}}, x_{m_{p_n}})\}$ is a subsequence of the given sequence $\{(b_n, x_n)\}$, and it converges in $B \times [a, b]$ to (c, d).

This shows that $A = B \times [a, b]$ is compact, as claimed.

Now any bounded subset of $\mathbb{R}^n$ is contained in $[-K, K]^n$ for some $K > 0$. To complete the proof of our theorem, we use the following.

Lemma 5.3 *Any closed set in a compact metric space is compact (with respect to the subspace metric).*

Proof. Let (X, d) be a compact metric space. (This means that the subset $X \subset X$ is compact in the metric space (X, d).) Let A be a closed set in (X, d). Let $\{x_n\}$ be a sequence in A. We must show that some subsequence of $\{x_n\}$ converges to a limit in A.

We can also regard $\{x_n\}$ as a sequence in the larger space X. Since X is compact, there is a convergent subsequence $\{x_{m_n}\}$, with limit $x \in X$. But the terms of the sequence $\{x_{m_n}\}$ all belong to A, and A is a closed set in (X, d) by hypothesis. Hence the limit x of this convergent subsequence is also in A.

We have shown that the sequence $\{x_n\}$ has a subsequence $\{x_{m_n}\}$ that converges to a limit $x \in A$. Hence A is compact, as claimed.

The following is a useful property of compact sets.

Theorem 5.4 *Let (X, d) and (Y, ρ) be metric spaces, A a compact set (X, d), and $f : X \to Y$ a map that is continuous with respect to d and ρ. Then $f(A)$ is a compact set in (Y, ρ).*

Proof. Let $\{y_n\}$ be a sequence in $f(A)$.

For each $n \in \mathbb{N}$, choose $x_n \in A$ such that $f(x_n) = y_n$.

Then $\{x_n\}$ is a sequence in A.

Since A is compact, we can choose a subsequence $\{x_{m_n}\}$ that converges in (X, d) to a limit $x \in A$.

By continuity, the subsequence $\{y_{m_n}\} = \{f(x_{m_n})\}$ of $\{y_n\}$ converges in (Y, ρ) to $f(x) \in f(A)$.

Hence each sequence in $f(A)$ has a subsequence that converges in (Y, ρ) to a limit in $f(A)$. In other words, $f(A)$ is compact.

Corollary 5.5 *If A is a nonempty, compact set in (X, d) and $f : X \to \mathbb{R}$ is continuous, then $f(X)$ is a nonempty, closed, bounded subset of $\mathbb{R}$. In particular, there are points $a, a' \in A$ such that:*

$$f(a) = \inf\{f(x) : \ x \in A\};$$
$$f(a') = \sup\{f(x) : \ x \in A\}.$$

Example The functions $f, g : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x, y) = x^2 + y^2, \qquad g(x, y) = x^3 + y^3$$

are both continuous with respect to the usual metrics on $\mathbb{R}^2$ and $\mathbb{R}$.

The set $A = \{(x, y) \in \mathbb{R}^2 : \ x^2 + y^2 \leq 1\} = f^{-1}([0, 1])$ is closed in $\mathbb{R}^2$, since $[0, 1]$ is closed in $\mathbb{R}$.

Also, $A \subset B_2(\underline{0})$ so A is bounded.

Being a closed, bounded subset of $\mathbb{R}^2$, A is compact.

By the corollary, g is bounded on A and attains its bounds.

In fact, using polar coordinates (r, θ) and differentiating with respect to θ, it is easy to see that g achieves its minimum value -1 on A at $(-1, 0)$ and $(0, -1)$, and its maximum value $+1$ at $(1, 0)$ and $(0, 1)$.

5.2 Complete metric spaces

Recall that there are subspaces of $\mathbb{R}$ (for example, $\mathbb{Q}$ and $(0, \infty)$) that contain Cauchy sequences that do not converge to limits in the subspace, but do converge to limits in $\mathbb{R}$. In fact, it turns out that every Cauchy sequence in $\mathbb{R}$ actually converges to a limit in $\mathbb{R}$. We use the term 'complete' to describe a metric space with this property.

Definition A metric space (X, d) is *complete* if every Cauchy sequence in (X, d) converges to a limit in (X, d).

Examples

1. The subspaces $\mathbb{Q}$ and $(0, \infty)$ are not complete.

2. Any discrete metric space is complete. (See the examples at the end of the chapter.) One can prove this by showing that every Cauchy sequence is eventually constant ($x_n = x_{n+1}$ for all sufficiently large n).

As indicated above, $\mathbb{R}$ with the usual metric is a complete space. We will shortly prove this using the following result.

Theorem 5.6 *Every compact metric space is complete.*

Proof. Let (X, d) be a compact metric space, and let $\{x_n\}$ be a Cauchy sequence in (X, d). We must show that $\{x_n\}$ is convergent.

Since (X, d) is compact, there is a convergent subsequence $\{x_{m_n}\}$, with limit $x \in X$. We will show that, in fact, x is a limit of the original sequence.

Let $\varepsilon > 0$. The fact that $\{x_n\}$ is Cauchy means that there is a natural number $N_1 \in \mathbb{N}$ such that $d(x_n, x_{n'}) < \frac{\varepsilon}{2}$ for all $n, n' \geq N_1$.

The fact that $\{x_{m_n}\}$ converges to x means that there is a natural number N_2 such that $d(x_{m_n}, x) < \frac{\varepsilon}{2}$ for all $n > N_2$.

Let $N = \max(N_1, N_2)$, and let $n \geq N$. Then $m_n \geq n \geq N \geq N_1$, so $d(x_{m_n}, x_n) < \frac{\varepsilon}{2}$. Also, $m_n \geq n \geq N \geq N_2$, so $d(x_{m_n}, x) < \frac{\varepsilon}{2}$.

By the triangle inequality,

$$d(x_n, x) \leq d(x_n, x_{m_n}) + d(x_{m_n}, x) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Hence $\{x_n\}$ converges to x in (X, d), as claimed.

Corollary 5.7 *$\mathbb{R}^k$ (with the euclidean metric) is complete for all $k \geq 1$.*

Proof. Let $\{\underline{x}_n\}$ be a Cauchy sequence in $\mathbb{R}^k$. Then $\{\underline{x}_n\}$ is bounded, and so contained in the subset $[-K, K]^k$ of $\mathbb{R}^k$, for some $K > 0$.

But $[-K, K]$ is a closed, bounded subset of $\mathbb{R}^k$, so it is compact. By the theorem, it is also complete.

Hence the Cauchy sequence $\{\underline{x}_n\}$ converges to a limit $\underline{x}$ in $[-K, K]^k$. But $\underline{x} \in \mathbb{R}^k$, so $\{\underline{x}_n\}$ also converges to $\underline{x}$ in $\mathbb{R}^k$.

We now have several examples of complete metric spaces: all compact spaces, all discrete spaces, and the euclidean spaces $\mathbb{R}^k$. Here are some more examples.

Theorem 5.8 *Let Y be a nonempty set. Then the space $B(Y)$ of bounded functions on Y, with the sup-metric, is complete.*

Proof. Let $\{f_n\}$ be a Cauchy sequence in $B(Y)$. In other words, given $\varepsilon > 0$, there is a natural number $N \in \mathbb{N}$ such that, whenever $n, n' \geq N$,

$$\sup\{|f_n(y) - f_{n'}(y)|;\ y \in Y\} = d(f_n, f_{n'}) < \varepsilon.$$

For a fixed $y_0 \in Y$,

$$|f_n(y_0) - f_{n'}(y_0)| \leq \sup\{|f_n(y) - f_{n'}(y)|;\ y \in Y\} < \varepsilon,$$

so $\{f_n(y_0)\}$ is a Cauchy sequence in $\mathbb{R}$.

But $\mathbb{R}$ is complete, so the sequence $\{f_n(y_0)\}$ has a limit $f(y_0)$ in $\mathbb{R}$ for every $y_0 \in Y$. In other words, the sequence f_n converges pointwise to $f : Y \to \mathbb{R}$.

We would like f to be a limit to $\{f_n\}$ in $B(Y)$. First, let us check that $f \in B(Y)$. In other words, that f is bounded.

The sequence $\{f_n\}$ is a Cauchy sequence, so it is bounded. There is a real number $K > 0$ such that $f_n \in B_K(z)$ for all n, where z is the zero function $z(y) = 0\ \forall\ y \in Y$.

For any $y_0 \in Y$ and $n \in \mathbb{N}$,

$$|f_n(y_0)| \leq \sup\{|f_n(y)|;\ y \in Y\} = d(f_n, z) \leq K.$$

Hence $f(y_0) = \lim_{n\to\infty} f_n(y_0) \in [-K, K]$ for all y_0. Hence f is bounded, as claimed.

Now we show that $f_n \to f$ in $B(Y)$. Let $\varepsilon > 0$.

Since $\{f_n\}$ is Cauchy, we can find $N \in \mathbb{N}$ such that $d(f_n, f_{n'}) < \frac{\varepsilon}{2}$ whenever $n, n' \geq N$. In particular, $d(f_n, f_N) < \frac{\varepsilon}{2}$ whenever $n \geq N$.

For any $y \in Y$, it follows that $|f_n(y) - f_N(y)| < \frac{\varepsilon}{2}$ whenever $n \geq N$.

Since $f(y) = \lim_n f_n(y)$, $|f(y) - f_N(y)| \leq \frac{\varepsilon}{2}$.

Taking the supremum over all $y \in Y$, we get $d(f, f_N) \leq \frac{\varepsilon}{2}$.

By the triangle inequality, $d(f_n, f) \leq d(f_n, f_N) + d(f_N, f) < \varepsilon$ whenever $n \geq N$.

Hence $f_n \to f$ in $B(Y)$, as claimed.

Lemma 5.9 *Any closed subset of a complete space is complete with respect to the subspace metric.*

Proof. Let (X, d) be a complete metric space, and let A be a closed set in (X, d). Let $\{x_n\}$ be a Cauchy sequence in A. Then $\{x_n\}$ is convergent in (X, d), with limit x (say), since (X, d) is complete.

But A is closed, so $x \in A$.

Corollary 5.10 *The space $C([0, 1])$ of continuous real-valued functions on $[0, 1]$ is complete with respect to the sup-metric.*

Proof. We know that continuous functions on $[0, 1]$ are bounded (since $[0, 1]$ is compact). Hence $C([0, 1])$ is a subset of $B([0, 1])$, which is complete. It is enough to show that $C([0, 1])$ is closed as a subset of $B([0, 1])$ with the sup-metric.

Let $f : [0, 1] \to \mathbb{R}$ be a limit point of $C([0, 1])$ in $B([0, 1])$. We must show that f is continuous.

Let $x_0 \in [0, 1]$ and $\varepsilon > 0$. Since f is a limit point of $C([0, 1])$, we can find $g \in C([0, 1])$ with $d(f, g) < \frac{\varepsilon}{3}$.

Then g is continuous at x_0, so there is a $\delta > 0$ such that $|g(x) - g(x_0)| < \frac{\varepsilon}{3}$ for any $x \in [0, 1]$ such that $|x - x_0| < \delta$.

For any such x, it follows from the triangle inequality that

$$|f(x) - f(x_0)| \leq |f(x) - g(x)| + |g(x) - g(x_0)| + |g(x_0) - f(x_0)| < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon.$$

Hence f is continuous, as required.

5.3 Completion of a metric space

Suppose that (X, d) is a metric space that is not complete. There is a natural way to construct a 'completion' of X, a larger metric space which is complete and contains X as a subspace. The example to have in mind is the case $X = \mathbb{Q}$, where the completion turns out to be (isometric to) $\mathbb{R}$.

I will give a brief description of this construction here. More details can be found in the Appendix to W. A. Sutherland, *Introduction to metric and topological spaces*, Oxford University Press, 1977.

To motivate the general construction, consider how we might use the rational numbers $\mathbb{Q}$ to construct the real numbers $\mathbb{R}$. Every real number is the limit of a sequence of rational numbers. Such a sequence is necessarily Cauchy. So we might think of using the set of all Cauchy sequences in $\mathbb{Q}$ as a model for $\mathbb{R}$.

A problem with this approach is that the choice of Cauchy sequence is not unique. Two Cauchy sequences in $\mathbb{Q}$ may converge to the same limit. (As an example, consider $\{\frac{1}{n}\}$ and $\{-\frac{1}{n}\}$, both of which converge to 0.)

However, if $x_n \to z$ and $y_n \to z$, then $|x_n - y_n| \to |z - z| = 0$. It turns out that the converse is also true: if $\{x_n\}$ and $\{y_n\}$ are Cauchy sequences in $\mathbb{R}$ with $|x_n - y_n| \to 0$, then $\lim_n x_n = \lim_n y_n$.

These remarks suggest the following construction. We define a relation $\sim$ on the set of Cauchy sequences in (X, d) by $\{x_n\} \sim \{y_n\}$ if $d(x_n, y_n) \to 0$ in $\mathbb{R}$. We check that $\sim$ is an equivalence relation on the set of Cauchy sequences, and then define $\widehat{X}$ to be the set of equivalence classes of Cauchy sequences under the equivalence relation $\sim$.

We define a metric $\widehat{d}$ on $\widehat{X}$ by $\widehat{d}([\{x_n\}], [\{y_n\}]) = \lim_{n\to\infty} d(x_n, y_n)$, where $[\{x_n\}]$ denotes the equivalence class of the sequence $\{x_n\}$. (Of course, we need to check that this is well-defined and satisfies the metric axioms.)

The key property of this space is that it is complete. To check this, one has to consider a Cauchy sequence of (equivalence classes of) Cauchy sequences, and identify the limit as an equivalence class of Cauchy sequences. I will omit the details.

We also check that X is isometric to a subspace of $\widehat{X}$ by identifying $x \in X$ with the equivalence class of the constant sequence $\{x\}$.

It is not difficult to check that X is *dense* in $\widehat{X}$ (which means that the closure of X in $\widehat{X}$ is the whole of $\widehat{X}$).

Finally, we check that $(\widehat{X}, \widehat{d})$ is uniquely determined by these properties, in the following sense. If Y is a complete metric space, and f is an isometry from X onto a dense subspace of Y, then f extends (uniquely) to an isometry from $\widehat{X}$ to Y.

5.4 Exercises on compactness and completeness

1. By producing appropriate sequences show that the following sets are not compact:
   (a) $\mathbb{Q}$ in the metric space $\mathbb{R}$ with the usual metric;
   (b) $(-2, 2)$ in the metric space $\mathbb{R}$ with the usual metric;
   (c) $\bigcup_{n=1}^{\infty}[3n, 3n+1]$ in the metric space $\mathbb{R}$ with the usual metric;
   (d) $B_1((0,0)) = \{(x, y) : x^2 + y^2 < 1\}$ in the metric space $\mathbb{R}^2$ with the usual metric;
   (e) The x-axis $\mathbb{R} \times \{0\}$ in $\mathbb{R}^2$ with the usual metric.

2. Suppose that (X, d) is a discrete metric space such that X contains infinitely many distinct points. Prove that X is closed and bounded but that X is not compact.

3. Prove that every discrete metric space is complete.

4. Show that the metric space $(\mathbb{R}^2, d)$, where d is the metric
$$d((x, y), (a, b)) = |x - a| + |y - b|,$$
is complete.

5. By producing appropriate sequences show that the following sets are not complete:
   (a) $[0, 2)$ in the metric space $\mathbb{R}$ with the usual metric;
   (b) $B_1((0,0)) = \{(x, y) : x^2 + y^2 < 1\}$ in the metric space $\mathbb{R}^2$ with the usual metric;
   (c) $\mathbb{R} \times \mathbb{Q}$ in $\mathbb{R}^2$ with the usual metric.

6. Let (X, d) be a metric space and let $\{x_n\}$ and $\{y_n\}$ be Cauchy sequences in X. Prove that $\{d(x_n, y_n)\}$ is a convergent sequence in $\mathbb{R}$.

7. Let (X, d) be a metric space. Prove that if every closed, bounded subset of X is compact then X is complete.

Chapter 6

Contraction mappings

6.1 The Contraction Mapping Theorem

Entering a complicated expression such as $\sqrt{537}$, or e^{143}, or $\cos(83.25^\circ)$ into a calculator instantly produces an answer, usually expressed as a decimal ($23 \cdot 173260453$ for $\sqrt{537}$) or in scientific notation ($1 \cdot 270898632\mathrm{e}+62$ – meaning $1 \cdot 270898632 \times 10^{62}$ – for e^{143}).

It is worth making two comments about these answers. Firstly, they are not correct – or very seldom correct. The reason for this is that the correct answer will usually be an irrational number, which the calculator is not capable of displaying.

The second comment is that, nevertheless, the displayed result is very accurate, and has been computed very quickly. In principle, there is no limit to the accuracy with which a number like this can be calculated. For example, the MAPLE command

```
> evalf(sqrt(537),70);
```

very quickly gives the answer correct to 70 significant figures:

$$23.17326045251293506464009809549485010009571216665445574023047514715231 0.$$

Question How do they do that?

The answer is that they use iterative processes which produce sequences of approximate values which converge rapidly to the correct answer. Thus after a reasonably small number of iterations, we have an answer that, while not completely correct, is accurate to within our chosen tolerance level.

Example In the Taylor expansion

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \cdots,$$

the sequence of partial sums $a_n = \sum_{k=0}^{n} \frac{x^k}{k!}$ converges to $a = e^x$. Moreover, a realistic upper bound for $|a - a_n|$ can be computed in terms of n and x, so it is easy to find a value of n such that a_n will be an accurate approximation to a.

One tool to ensure the convergence of sequences (in suitable metric spaces), and hence the accuracy of computations, is the idea of a contraction mapping.

Definition Let (X, d) be a metric space. A *contraction mapping* on (X, d) is a map $f : X \to X$ such that, for some $K \in (0, 1)$,

$$(\forall\ x, y \in X)\ d(f(x), f(y)) \leq K \cdot d(x, y).$$

The constant K is called a *contraction factor* for f.

Examples

1. $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \dfrac{x}{2}$ is a contraction mapping on $\mathbb{R}$ (with respect to the usual metric), with contraction factor $\frac{1}{2}$:

$$|f(x) - f(y)| = \frac{1}{2}|x - y|.$$

2. $f : \mathbb{R}^2 \to \mathbb{R}^2$, $f(x, y) = \left(\dfrac{y}{2}, \dfrac{x}{3}\right)$, is a contraction mapping on $\mathbb{R}^2$ (with respect to the euclidean metric), with contraction factor $\frac{1}{2}$:

$$d\left(f(x, y), f(x', y')\right) = \sqrt{\frac{(y - y')^2}{4} + \frac{(x - x')^2}{9}} \leq \frac{1}{2} d\left((x, y), (x', y')\right).$$

3. Let $T : C([0, 1]) \to C([0, 1])$ be defined by

$$T(u)(x) = \frac{1}{2} \int_0^x u(s) ds$$

for all $u \in C([0, 1])$ and for all $x \in [0, 1]$.

So, for example, if $u(x) = x + x^2$, then $T(u)(x) = \dfrac{x^2}{4} + \dfrac{x^3}{6}$.

Then T is a contraction mapping on $C([0, 1])$ with respect to the sup-metric, with

contraction factor $\frac{1}{2}$:

$$\begin{aligned} d(T(u),T(v)) &= \sup\{|T(u)(x)-T(v)(x)|;\ x\in[0,1]\} \\ &= \sup\left\{\frac{1}{2}\left|\int_0^x (u(s)-v(s))ds\right|;\ x\in[0,1]\right\} \\ &\leq \frac{1}{2}\int_0^1 |u(s)-v(s)|ds \\ &\leq \tfrac{1}{2}\sup\{|u(s)-v(s)|; s\in[0,1]\} \\ &= \tfrac{1}{2}d(u,v). \end{aligned}$$

Remark Contraction mappings are always continuous.

Exercise Prove this.

Definition A *fixed point* of a map $f : X \to X$ is an element $x \in X$ such that $f(x) = x$.

Examples

1. Let $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \dfrac{x}{2}$. Then $x \in \mathbb{R}$ is a fixed point of f if and only if $x = \dfrac{x}{2}$, that is, if and only if $x = 0$.

2. Let $f : \mathbb{R}^2 \to \mathbb{R}^2$, $f(x,y) = \left(\dfrac{y}{2}, \dfrac{x}{3}\right)$. Then (x,y) is a fixed point of f if and only if $x = \dfrac{y}{2}$ and $y = \dfrac{x}{3}$, which holds if and only if $(x,y) = (0,0)$.

3. Let $T : C([0,1]) \to C([0,1])$ be defined by
$$T(u)(x) = \frac{1}{2}\int_0^x u(s)ds$$
for all $u \in C([0,1])$ and for all $x \in [0,1]$.
Then a fixed point of T is a solution of the *integral equation*:
$$u(x) = \frac{1}{2}\int_0^x u(s)ds.$$
Differentiating with respect to x, this becomes a linear differential equation
$$\frac{du}{dx} = \frac{1}{2}u(x)$$
with initial condition $u(0) = 0$.
The differential equation has general solution $u(x) = Ae^{x/2}$ where A is a constant. From the initial condition we get $A = 0$. Hence T has a unique fixed point, namely the zero function $u(x) = 0\ \forall\ x$.

In each of the above examples, the functions are contraction mappings. It is no accident that each has a unique fixed point.

Theorem 6.1 (The Contraction Mapping Theorem) *Let (X, d) be a nonempty complete metric space, and let $f : X \to X$ be a contraction mapping on (X, d). Then f has a unique fixed point.*

Proof. The proof of uniqueness is easy. Suppose that $x, y \in X$ with $f(x) = x$ and $f(y) = y$. If $K < 1$ is a contraction factor for f, then

$$d(x, y) = d\left(f(x), f(y)\right) \leq K \cdot d(x, y).$$

Since $K < 1$ and $d(x, y) \geq 0$, this is possible only if $d(x, y) = 0$, that is if $x = y$.

Now let us prove that a fixed point always exists.

By hypothesis, $X \neq \emptyset$, so we can choose an element $x_1 \in X$.

Define a sequence $\{x_n\}$ in X by $x_2 = f(x_1)$, $x_3 = f(x_2)$, and so on. We will show that the sequence $\{x_n\}$ is Cauchy.

Since (X, d) is complete, the Cauchy sequence $\{x_n\}$ converges to a limit $x \in X$. Now f is continuous, so

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

In other words, $x = \lim_n x_n$ is a fixed point for f.

So all we have to do is prove that $\{x_n\}$ is Cauchy. We need to understand the behaviour of $d(x_m, x_n)$ when m, n are large. Suppose that $m \leq n$. Then

$$\begin{aligned} d(x_m, x_n) &= d(f(x_{m-1}), f(x_{n-1})) \\ &\leq K \cdot d(x_{m-1}, x_{n-1}) \\ &\leq K^2 \cdot d(x_{m-2}, x_{n-2}) \\ &\leq \cdots \\ &\leq K^{m-1} \cdot d(x_1, x_{n-m+1}). \end{aligned}$$

Now, for any $t \in \mathbb{N}$,

$$\begin{aligned} d(x_1, x_t) &\leq d(x_1, x_2) + d(x_2, x_3) + \cdots + d(x_{t-1}, x_t) \\ &\leq (1 + K + \cdots + K^{t-2}) \cdot d(x_1, x_2) \\ &\leq \frac{d(x_1, x_2)}{1 - K}. \end{aligned}$$

Hence, if $N \leq m \leq n$ then

$$d(x_m, x_n) \leq \frac{K^{N-1} \cdot d(x_1, x_2)}{1 - K} \to 0 \text{ as } N \to \infty.$$

Thus $\{x_n\}$ is Cauchy, as required.

Remark Note the following addendum to the Contraction Mapping Theorem, which follows immediately from the proof:

If x_1 is any element of X, then the sequence $\{x_n\}$, defined inductively by

$$x_{n+1} = f(x_n) \ \forall \ n \in \mathbb{N},$$

converges to the fixed point of f.

6.2 Applications

A key condition in the Contraction Mapping Theorem is that the underlying metric space be complete. We are familiar with a number of complete metric spaces: $\mathbb{R}$, $\mathbb{R}^k$, $B(X)$, $C([a,b])$. I will describe some applications in the cases $X = \mathbb{R}$ and $X = C([a,b])$.

6.2.1 Approximate solutions to algebraic equations in $\mathbb{R}$

Given an algebraic equation, the idea is to rearrange the equation in the form $x = f(x)$ for some contraction mapping f on a closed subset of $\mathbb{R}$. Since closed subsets are complete, there is a unique solution, namely the fixed point of f. Moreover, given any starting point x we get a sequence $x, f(x), f(f(x)), \ldots$ that converges to the solution.

Example An iterative procedure to calculate $\sqrt{2}$:

Define $f : [1, \frac{3}{2}] \to [1, \frac{3}{2}]$ by $f(x) = 1 + x - \dfrac{x^2}{2}$.

First, let us check that $f\left([1, \frac{3}{2}]\right) \subset [1, \frac{3}{2}]$.

$f'(x) = 1 - x \in [-\frac{1}{2}, 0]$ for $x \in [0, \frac{3}{2}]$.

Thus f is a decreasing function on this interval, with $f(1) = \frac{3}{2}$, and $f(\frac{3}{2}) = \frac{11}{8} \in [1, \frac{3}{2}]$.

So $f\left([1, \frac{3}{2}]\right) \subset [1, \frac{3}{2}]$, as required.

Moreover, for $x, y \in [1, \frac{3}{2}]$, we have

$$\begin{aligned} |f(x) - f(y)| &= \left| \int_x^y f'(t)dt \right| \\ &\leq |x - y| \cdot \sup \left\{ |f'(t)| \, ; \ t \in \left[1, \frac{3}{2}\right] \right\} \\ &\leq \frac{1}{2} \cdot |x - y|, \end{aligned}$$

so f is a contraction mapping on $[1, \frac{3}{2}]$, with contraction factor $\frac{1}{2}$.

The unique fixed point of f in $[1, \frac{3}{2}]$ is the solution of the algebraic equation $x = 1 + x - \frac{x^2}{2}$, namely $\sqrt{2}$.

6.2.2 Integral equations

Example Solve the integral equation

$$u(x) = \int_0^x \frac{u(s) - 1}{2} ds$$

for $u \in C([0, 1])$.

Answer A solution to this equation is the same thing as a fixed point of the map $T : C([0, 1]) \to C([0, 1])$ defined by

$$T(u)(x) = \int_0^x \frac{u(s) - 1}{2} ds.$$

Using the sup-metric d on $C([0, 1])$, we have

$$\begin{aligned} d(T(u), T(v)) &= \sup_{0 \le x \le 1} \left| \int_0^x \frac{u(s) - v(s)}{2} ds \right| \\ &\le \frac{1}{2} \cdot \sup_{0 \le s \le 1} |u(s) - v(s)| \\ &= \frac{1}{2} \cdot d(u, v), \end{aligned}$$

so T is a contraction mapping on $C([0, 1])$ with contraction factor $\frac{1}{2}$.

Since $C([0, 1])$ is complete, there is a unique fixed point of T, and hence a unique solution of the equation. Indeed, the Contraction Mapping Theorem gives us an iterative procedure for approximating this solution.

First choose an arbitrary element of $C([0, 1])$, for example the constant function $u_0(x) = 0 \ \forall \ x$.

Then repeatedly apply T to get a sequence $\{u_n\}$ of functions:

$$u_1(x) = T(u_0)(x) = \int_0^x -\frac{1}{2} ds = -\frac{x}{2},$$

$$u_2(x) = T(u_1)(x) = \int_0^x \left(-\frac{s}{4} - \frac{1}{2} \right) ds = -\frac{x^2}{8} - \frac{x}{2},$$

$$u_3(x) = T(u_2)(x) = \int_0^x \left(-\frac{s^2}{16} - \frac{s}{4} - \frac{1}{2} \right) ds = -\frac{x^3}{48} - \frac{x^2}{8} - \frac{x}{2},$$

and so on.

An easy induction argument shows that the n-th term of this sequence is

$$u_n(x) = \sum_{k=1}^{n} -\frac{(x/2)^k}{k!},$$

so the limit is

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

We can also confirm this result theoretically, by differentiating our integral equation to yield a differential equation with initial condition:

$$u'(x) = \frac{1}{2}\left(u(x) - 1\right), \quad u(0) = 0,$$

and solving.

Here is a general result on the existence and uniqueness of solutions to certain integral equations, that can be proved using the Contraction Mapping Theorem.

Theorem 6.2 *Let $\theta : [a,b] \to \mathbb{R}$ and $K : [a,b] \times [a,b] \to \mathbb{R}$ be continuous functions. Then there exists $\varepsilon > 0$ such that, for any $\lambda \in (-\varepsilon, \varepsilon)$, there is a unique solution $\psi(x) \in C([a,b])$ of the integral equation*

$$\psi(x) = \lambda \int_a^b K(x,y)\psi(y)dy + \theta(x).$$

Proof. Define a map $T : C([a,b]) \to C([a,b])$ by

$$T(u)(x) = \lambda \int_a^b K(x,y)u(y)dy + \theta(x).$$

Then a solution $\psi(x)$ in $C([a,b])$ to our equation is a fixed point of T, so we need to prove that T is a contraction mapping for small values of $|\lambda|$.

Now K is a bounded function, since it is continuous and $[a,b] \times [a,b]$ is compact. Let $C = \sup\{|K(x,y)|\,;\ x,y \in [a,b]\}$. Hence

$$\begin{aligned} d(T(u),T(v)) &= \sup\left\{|\lambda| \cdot \left|\int_a^b K(x,y)\left(u(y) - v(y)\right)dy\right|\,;\ x \in [a,b]\right\} \\ &\leq |\lambda| \cdot C \cdot |b-a| \cdot \sup\{|u(y) - v(y)|\,;\ y \in [a,b]\} \\ &= |\lambda| \cdot C \cdot |b-a| \cdot d(u,v). \end{aligned}$$

Hence, provided $|\lambda| \cdot C \cdot |b-a| < 1$, it follows that T is a contraction mapping, as required.

6.2.3 Differential equations

A first order differential equation, together with an initial condition, can often be transformed to an integral equation by integrating. If the above theorem applies to the resulting integral equation, then we have a result about existence and uniqueness of solutions to the differential equation we started with.

For example:

Theorem 6.3 (Picard's Theorem)[1] *Let U be an open subset of $\mathbb{R}^2$, and $f : U \to \mathbb{R}$ a continuous function satisfying an inequality*

$$(\forall\ x, y_1, y_2 \text{ with } (x, y_1),\ (x, y_2) \in U)\ |f(x, y_1) - f(x, y_2)| \leq K \cdot |y_1 - y_2|$$

for some constant $K > 0$.

Then, given $(x_0, y_0) \in U$, there is a positive real number $a > 0$, and a unique differentiable function $g : [x_0 - a, x_0 + a] \to \mathbb{R}$ such that

$$g'(x) = f(x, g(x))\ \forall\ x \in (x_0 - a, x_0 + a), \quad \text{and} \quad g(x_0) = y_0.$$

Proof. Integrate the differential equation to get an integral equation

$$g(x) = y_0 + \int_{x_0}^{x} f(t, g(t))dt.$$

A solution on an interval $[x_0 - a, x_0 + a]$ is therefore the same as a fixed point of the transformation

$$T : C([x_0 - a, x_0 + a]) \to C([x_0 - a, x_0 + a]), \quad T(g)(x) := y_0 + \int_{x_0}^{x} f(t, g(t))dt.$$

Since U is open and $(x_0, y_0) \in U$, there is a positive ε such that

$$(x_0 - \varepsilon, x_0 + \varepsilon) \times (y_0 - \varepsilon, y_0 + \varepsilon) \subset U.$$

Put $b = \frac{\varepsilon}{2}$. Then

$$B := [x_0 - b, x_0 + b] \times [y_0 - b, y_0 + b] \subset U.$$

Since B is compact, f is bounded on B. Let $C = \sup\{|f(x, y)|;\ (x, y) \in B\}$. Now

$$\begin{aligned} d(T(g), T(h)) &= \sup\left\{\left|\int_{x_0}^{x} (f(t, g(t)) - f(t, h(t))dt\right|;\ x \in [x_0 - a, x_0 + a]\right\} \\ &\leq K \cdot |x - x_0| \cdot \sup\{|g(t) - h(t)|;\ t \in [x_0 - a, x_0 + a]\} \\ &\leq aK \cdot d(g, h), \end{aligned}$$

[1] No, not Captain Jean-Luc of the Starship Enterprise – the French mathematician Charles Emile. See `http://www-history.mcs.st-andrews.ac.uk/Mathematicians/Picard_Emile.html`

provided $(t, g(t)), (t, h(t)) \in U$ for all $t \in [x_0 - a, x_0 + a]$.

Finally, choose $a > 0$ small enough so that $aK < 1$ and $aC \leq b$, and define X to be the set of continuous functions $g : [x_0 - a, x_0 + a] \to \mathbb{R}$ such that

$$g(x_0) = y_0 \text{ and } g(x) \in [y_0 - b, y_0 + b] \ \forall \ x \in [x_0 - a, x_0 + a]\}.$$

Then it is easy to check that:

1. X is a closed subset of $C([x_0 - a, x_0 + a])$, and hence complete;
2. $T(X) \subset X$; and
3. T is a contraction mapping on X, with contraction factor $aK < 1$.

It follows that T has a unique fixed point in X, and hence our equation has a unique solution in X. To complete the proof, notice that for any solution g of the equation, we have

$$g(x_0) = y_0 + \int_{x_0}^{x_0} f(t, g(t))dt = y_0,$$

and

$$|g(x) - y_0| \leq |x - x_0| \sup\{|f(t, g(t))|; \ t \in [x_0 - a, x_0 + a]\} \leq aC \leq b$$

for all $x \in [x_0 - a, x_0 + a]$, so that $g \in X$.

6.3 Exercises on contraction mappings

1. Show that the map $f : [1, \frac{5}{2}] \to [1, \frac{5}{2}]$ defined by

$$f(x) = 1 + x - \frac{x^3}{10}$$

is a contraction mapping on $[1, \frac{5}{2}]$. What is the fixed point of f? Evaluate the first few terms of the iterative sequence $\{f^n(2)\}$ (i.e., $x_0 = 2$, $x_{n+1} := f(x_n)$ for $n \geq 0$) and so determine the decimal approximation to this fixed point, correct to two decimal places.

2. Show that $T : B([0,1]) \to B([0,1])$, defined by

$$T(u)(x) = \frac{u(x) + u(1-x)}{3},$$

is a contraction mapping, and find its unique fixed point.

3. Show that $T : B([0,1]) \to B([0,1])$, defined by

$$T(u)(x) = \frac{u(x) + x^2}{2},$$

is a contraction mapping, and find its unique fixed point.

4. Show that $T : C[0,1] \to C[0,1]$ is a contraction mapping where

$$T(f)(x) = \int_0^x (x-t)f(t)\,dt\,, \qquad x \in [0,1]\,, \qquad f \in C[0,1]\,.$$

Find the fixed point of T.

5. Let $T : X \to X$ be a contraction mapping, where X is a discrete metric space. Show that T is a constant function.

6. Let $T : X \to X$ be a function such that $T^2 = T \circ T : X \to X$ is a contraction mapping. Show that

   (a) T has a unique fixed point $x \in X$.

   (b) Given any $x_0 \in X$, the sequence $\{x_n\}$ defined inductively by $x_{n+1} := T(x_n)$ converges to x.

7. If $h \in C[0,a]$, define $F : C[0,a] \to C[0,a]$ by

$$F(u)(x) = \int_0^x u(s)\,ds + h(x)\,, \qquad x \in [0,a], \qquad u \in C[0,a]\,.$$

Show that F is a contraction map in $C[0,a]$ if and only if $a < 1$.

If h is differentiable, write down a differential equation with initial condition, for which the fixed point of F is the unique solution in $C[0,a]$.