Supplementary material for Chapter 2

Since Chapter 2 contains just notation and terminology, there may not seem to be much need for supplementary material. The topics considered here are:

The definition of function (2) Here are a few thoughts about the importance of specifying the domain and the receiving set in the definition of 'function' used in modern mathematics. In particular, without specifying the domain and the receiving set, the notions of *injective* and *onto* for a function would make no sense.

To illustrate this, let us consider a familiar function from school mathematics, $y = x^2$, given just by a formula without specifying a domain X or a receiving set Y. By contrast, here are four distinct functions in the precise modern sense:

Let $\mathbb{R}_+ = \{x \in \mathbb{R} : x \geqslant 0\}$. Define functions

$$f_1 : \mathbb{R} \to \mathbb{R}, \quad f_2 : \mathbb{R}_+ \to \mathbb{R}, \quad f_3 : \mathbb{R} \to \mathbb{R}_+, \quad f_4 : \mathbb{R}_+ \to \mathbb{R}_+$$

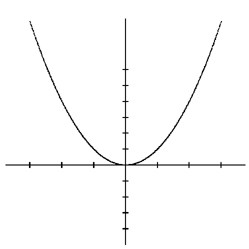
by the formulas $f_1(x) = x^2$, $f_2(x) = x^2$, $f_3(x) = x^2$, $f_4(x) = x^2$. The following facts are straightforward to check (see also the graphs below):

f_1 is neither injective nor onto,

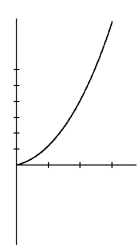
f_2 is injective but not onto,

f_3 is onto but not injective,

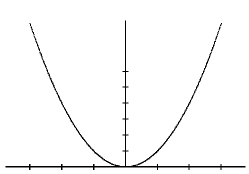
f_4 is both injective and onto.



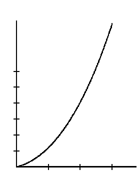
(a) Graph of f_1



(b) Graph of f_2



(a) Graph of f_3



(b) Graph of f_4

Thus we see that without the precise definition of a function, the concepts of injective and onto would not have any meaning.

Negating mathematical statements containing quantifiers (2) Next we comment on something the reader may well have met already in real analysis: the manipulation of logical symbols when negating complicated mathematical statements. It may not seem appropriate to attach this to Chapter 2, but there is no perfect place to include it.

There are two phrases in ordinary English which often feature in modern mathematics, particularly in analysis: *for all* and *there exists*. The usual symbols for these are $\forall$ and $\exists$; they are called *quantifiers*. In addition to these symbols we use $\neg$ for *not*. There is much consideration of these symbols and their interactions in mathematical logic. Here we restrict consideration to their use in analysis to negate some complicated mathematical statements such as the following definition of continuity of a function $f : \mathbb{R} \to \mathbb{R}$ at a point $a \in \mathbb{R}$: for all $\varepsilon > 0$, there exists $\delta > 0$ such that for all x satisfying $|x - a| < \delta$ we have $|f(x) - f(a)| < \varepsilon$. Of course, to negate this, we can just put a 'not' at the front: it is not the case that for all $\varepsilon > 0$, there exists $\delta > 0$ such that for all x satisfying $|x - a| < \delta$ we have $|f(x) - f(a)| < \varepsilon$. However, often we would like to pass the 'not' right through the statement to get some equivalent statement which ends up '... $|f(x) - f(a)| \geqslant \varepsilon$'.

There is a purely mechanical way to get this right, based on the two rules

$$\neg\forall \;\equiv\; \exists\neg,$$
$$\neg\exists \;\equiv\; \forall\neg.$$

We are persuaded that these rules should hold by common-sense examples, although we shall express them in language that may not seem to be quite every-day English: for instance to say that 'not all apples are red' is equivalent to 'there exists an apple which is not red'; while to say 'there does not exist a black rose' is equivalent to 'given any rose, it is not black'. The phrases 'for all' and 'given any' are equivalent, as are 'there exists' 'there exists some' and just 'there is'.

Now let us apply this to negate the above definition that a function $f : \mathbb{R} \to \mathbb{R}$ is continuous at a point $a \in \mathbb{R}$. We take it in easy stages: the definition begins 'Given any $\varepsilon > 0$', and goes on to say that some property P holds true. To negate this we say 'There exists some $\varepsilon > 0$ such that the property P does not hold', i.e. 'There exists some $\varepsilon > 0$ such that $\neg P$'. Next we focus on what P is: it says 'there exists $\delta > 0$ such that some property Q holds', so by our rules, $\neg P$ is equivalent to saying 'for all $\delta > 0$, the property Q does not hold', or equivalently '$\forall\delta > 0, \neg Q$'.

Up to this point then, we have moved 'not' past two quantifiers to get 'There exists some $\varepsilon > 0$ such that for all $\delta > 0$, Q fails'. Next we focus on Q: it says 'for all x satisfying $|x - a| < \delta$, a certain property R holds. To negate this according to our rules, we say 'there exists an x satisfying $|x - a| < \delta$ such that R fails'. Finally, R is the property $|f(x) - f(a)| < \varepsilon$, so to negate this we simply write $|f(x) - f(a)| \geqslant \varepsilon$. Putting all this together, we see that to say that f is *not* continuous at a is equivalent to: there exists some $\varepsilon > 0$ such that for all $\delta > 0$ there exists an x with $|x - a| < \delta$ and yet $|f(x) - f(a)| \geqslant \varepsilon$. (Notice that the 'yet' here is inserted just for emotional emphasis; it has no mathematical significance.) In symbols, using just a dot for the phrase 'such that' we may write this as:

$$\exists \varepsilon > 0 . \forall \delta > 0, \exists x . |x - a| < \delta \text{ and } |f(x) - f(a)| \geqslant \varepsilon.$$

We note that the 'not' has passed all the way through the statement, and indeed has disappeared in the replacement of $\neg <$ by $\geqslant$.

This procedure may look lengthy, but after a little practice it becomes very quick and easy. You might like to try negating for example the definition that a sequence (x_n) converges to a point x, or (one stage harder) to negate the statement that (x_n) converges.

Countability (2) At various stages in the book and in the following supplementary material we mention properties of countability. We now consider facts about countability that we use later.

The intuitive idea of a countable set is one that can be labelled by the natural numbers, or by a subset $\{1, 2, 3, \ldots, n\}$. In connection with this latter alternative, we mention that there is no general agreement whether to consider a finite set as countable, or whether to insist that a countable set should be in one-one correspondence with $\mathbb{N}$. We shall allow finite sets to be called countable, which seems intuitively correct. With this convention, the following definition is convenient:

Definition S.2.1 A set X is *countable* if there is an injective function $f : X \to \mathbb{N}$. If a set is countable but not finite, we say it is *countably infinite*.

The next result reconciles this definition with the alternative one that says a set is countably infinite if it can be put in one-one correspondence with $\mathbb{N}$. The proof we give is intuitive; we are getting rather close to areas of mathematics in which set theory and logic should really be used.

Proposition S.2.2 A set X is countable iff it can be put in one-one correspondence either with $\{1, 2, \ldots, n\}$ for some $n \in \mathbb{N}$ or with $\mathbb{N}$.

Proof If X can be put in one-one correspondence either with $\{1, 2, \ldots, n\}$ for some $n \in \mathbb{N}$, or with all of $\mathbb{N}$, then it is clear that there is an injective function $f : X \to \mathbb{N}$ (in the first case we compose the one-one correspondence with the inclusion of $\{1, 2, \ldots, n\}$ into $\mathbb{N}$, and in the second case we already have an injection).

Conversely suppose we are given an injective function $f : X \to \mathbb{N}$. The subset $f(X)$ of $\mathbb{N}$ has a least element n_1, and there is a unique x_1 in X such that $f(x_1) = n_1$. We begin defining a new function $g : X \to \mathbb{N}$ by setting $g(x_1) = 1$. Now suppose we have partially defined g, from a subset $\{x_1, x_2, \ldots, x_n\}$ of X bijectively to the set $\{1, 2, \ldots, n\}$. If $\{x_1, x_2, \ldots, x_n\} = X$ then we're done. Otherwise let m be the least integer in $f(X) \setminus f(\{x_1, x_2, \ldots, x_n\})$, let x_{n+1} be the unique element of X which is mapped by f to m, and put $g(x_{n+1}) = n+1$. The process continues, and if it does not stop at any finite stage then it constructs a one-one correspondence g of X with $\mathbb{N}$. [This last sentence is the hairy bit of the proof, concerning infinite processes to which we should really apply set theory and logic.] □

The next result follows easily from Definition S.2.1.

Proposition S.2.3 Any subset A of a countable set X is countable.

Proof We know there is an injection $f : X \to \mathbb{N}$, and if we follow the inclusion of A to X by f we get an injection of A to $\mathbb{N}$. □

Another easy consequence of the definition is:

Proposition S.2.4 The set $\mathbb{Z}$ of all integers is countable.

Proof We may define an injection of $\mathbb{Z}$ into $\mathbb{N}$ (indeed, a one-one correspondence) by defining $f(n) = 2n + 1$ if $n \geqslant 0$ and $f(n) = -2n$ if $n < 0$. □

The next result is rather central.

Proposition S.2.5 If X and Y are countable sets then so is their Cartesian product $X \times Y$.

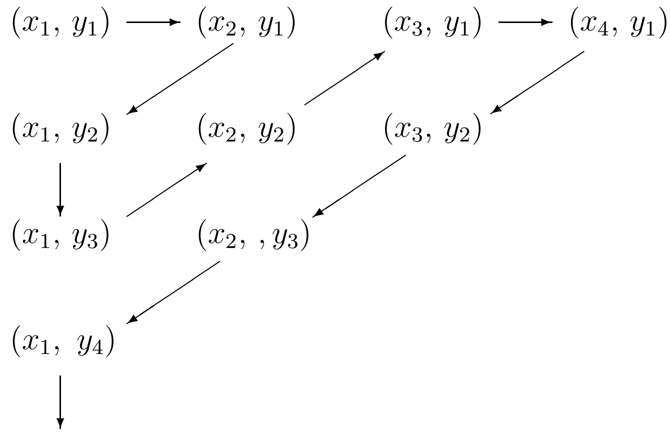
We offer two proofs of this. The first is geometric and visually appealing, but not very rigorous. Let us suppose for simplicity that X and Y are both countably infinite. We may suppose we are given one-one correspondences of each set with $\mathbb{N}$ - such a correspondence comes from Proposition S.2.2. Thus we may list the elements of X as $x_1, x_2, \ldots, x_n, \ldots$ and the elements of Y as $y_1, y_2, \ldots, y_n, \ldots$. Now consider the elements of $X \times Y$ set out in the following array:

$$\begin{array}{lllll}
(x_1, y_1), & (x_2, y_1), & \ldots & (x_n, y_1), & \ldots \\
(x_1, y_2), & (x_2, y_2), & \ldots & (x_n, y_2), & \ldots \\
\ldots & \ldots & \ldots & \ldots & \ldots \\
(x_1, y_n), & (x_2, y_n), & \ldots & (x_n, y_n), & \ldots \\
\ldots & \ldots & \ldots & \ldots & \ldots
\end{array}$$

Now we proceed to list all the elements in this array ‘by diagonals’. So the list begins

$$(x_1,\,y_1),\,(x_2,\,y_1),\,(x_1,\,y_2)\,(x_1,\,y_3),\,(x_2,\,y_2),\,(x_3,\,y_1),\,(x_4,\,y_1),\,(x_3,\,y_2),\,(x_2,\,y_3),\,(x_1,\,y_4),\ldots$$

With a little effort, we could work out formulae to tell us which ordered pair $(x_1,\,y_j)$ is supposed to come at the nth position on this list, but the essence of this proof is to provide visual persuasion. A pictorial version of what we are doing here is indicated below.



This is supposed to persuade one that $X \times Y$ is countable.

The second (rigorous) proof depends on unique factorization of integers into prime powers. We are given injections $f : X \to \mathbb{N}$ and $g : Y \to \mathbb{N}$. Define $h : X \times Y \to \mathbb{N}$ by

$$h(x,\,y) = 2^{f(x)}3^{g(y)}.$$

Then h is an injection, for if $h(x,\,y) = h(x',\,y')$ then by unique factorization $f(x) = f(x')$ and $g(y) = g(y')$, so since f and g are injections, $x = x'$ and $y = y'$, i.e. $(x,\,y) = (x',\,y')$. □

Note that the second proof applies equally well when one or both of the sets X, Y is finite.

Proposition S.2.6 The set $\mathbb{Q}$ of all rational numbers is countable.

Proof We may define an injection k of $\mathbb{Q}$ into $\mathbb{Z} \times \mathbb{Z}$ as follows: write each rational number as m/n where m, n are integers with no common factor and $n > 0$. Then $m/n \mapsto (m,\,n)$ defines a suitable injection k. But by Propositions S.2.4 and S.2.5 there is an injection j of $\mathbb{Z} \times \mathbb{Z}$ into $\mathbb{N}$, hence the composition $j \circ k$ is an injection of $\mathbb{Q}$ into $\mathbb{N}$, showing $\mathbb{Q}$ is countable. □

Corollary S.2.7 If there is an injection l from a set S into $\mathbb{Q}$ then S is countable.

Proof By Proposition S.2.6 there is an injection $i : \mathbb{Q} \to \mathbb{N}$. The composition $i \circ l$ is an injection from S to $\mathbb{N}$ as required. □

The next result may be paraphrased 'a countable union of countable sets is countable'.

Proposition S.2.8 Suppose that I is either the set $\{1,\,2,\,\ldots,\,n\}$ for some $n \in \mathbb{N}$, or $I = \mathbb{N}$. Suppose for each integer $i \in I$ we are given a countable set A_i. Then $\displaystyle\bigcup_{i \in I} A_i$ is countable.

Proof Since we wish to define an injection from the union of all the A_i, it is convenient first to replace the A_i by sets B_i which are pairwise disjoint but have the same union as the A_i. To do this, put $B_1 = A_1$, and suppose inductively that we have defined pairwise disjoint sets $B_1, B_2, \ldots, B_r$ such that $B_i \subseteq A_i$ for each $i = 1, 2, \ldots, r$ and for each $s = 1, 2, \ldots, r$ we have $\bigcup_{i=1}^{s} B_i = \bigcup_{i=1}^{s} A_i$. Put $B_{r+1} = A_{r+1} \setminus \bigcup_{i=1}^{r} B_i$, or equivalently $B_{r+1} = A_{r+1} \setminus \bigcup_{i=1}^{r} A_i$. This inductive procedure replaces the A_i by pairwise disjoint B_i having the same union as the A_i, as desired. Each B_i is countable by Proposition S.2.3 since $B_i \subseteq A_i$, so for each $i \in I$ there is an injection $g_i : B_i \to \mathbb{N}$. Now define $h : \bigcup_{i \in I} B_i \to \mathbb{N} \times \mathbb{N}$ by $h(b) = (i, g_i(b))$ where $i \in I$ is the unique integer such that $b \in B_i$. Suppose $h(b) = h(b')$, say $(i, g_i(b)) = (j, g_j(b'))$. Then $j = i$ and since then $g_i(b) = g_i(b')$ we get also $b' = b$ since g_i is an injection. So h is an injection. We already know from Proposition S.2.5 that $\mathbb{N} \times \mathbb{N}$ is countable, so we may follow h by an injection k of $\mathbb{N} \times \mathbb{N}$ into $\mathbb{N}$. The composition $k \circ h$ is an injection of $\bigcup_{i \in I} A_i = \bigcup_{i \in I} B_i$ into $\mathbb{N}$, so $\bigcup_{i \in I} A_i$ is countable as required. □

Note that Proposition S.2.8 applies in particular to a countable union of finite sets (each A_i could be finite.)

Armed with the above results, we can get practical results about countability. As an illustration, we prove countability of what are known as the algebraic numbers.

Definition S.2.9 An *algebraic number* is the root of a polynomial equation with integer coefficients.

Proposition S.2.10 The set A of all algebraic numbers is countable.

Proof Let $\mathcal{P}_n$ be the set of distinct polynomials with integer coefficients and of degree at most n. Then $\mathcal{P}_n$ is in one-one correspondence with the product of $n+1$ copies of $\mathbb{Z}$. (Note that a polynomial of degree n has $n+1$ coeffients, counting the constant term. Of course, if the leading coefficient is 0 then the degree of the polynomial drops.) By Propositions S.2.4, S.2.5 and induction, the set $\mathcal{P}_n$ is countable. The set of all algebraic numbers which are roots of a polynomial equation $P(x) = 0$ with $P \in \mathcal{P}_n$ will be written A_n. For a given polynomial equation of degree n with integer coefficients, by elementary algebra the number of distinct roots is at most n. Since there are only countably many equations $P(x) = 0$ with $P \in \mathcal{P}_n$, A_n is countable by Proposition S.2.8 (specifically, the case of Proposition S.2.8 saying that a countable union of finite sets is countable). Finally, we get A by taking the union of the A_n over all $n \in \mathbb{N}$. This is a countable union of countable sets, so A is countable by Proposition S.2.8. □

Although our interest in the book is mainly in positive results about countability, we conclude with the famous method of Cantor for establishing the existence of uncountable sets.

Proposition S.2.11 The set $\mathbb{R}$ of all real numbers is uncountable.

Proof (Cantor) It is enough to show that the set of real numbers in $[0, 1]$ is uncountable, since any subset of a countable set is countable. We proceed by contradiction. Suppose that the real numbers in $[0, 1]$ is countable, and list them all in decimal form, attempting to label them all uniquely by positive integers. The list will look something like:

(1) $0.a_{11}a_{12}a_{13}\ldots a_{1n}\ldots$

(2) $0.a_{21}a_{22}a_{23}\ldots a_{2n}\ldots$

(3) $0.a_{31}a_{32}a_{33}\ldots a_{3n}\ldots$

....

Now comes the cunning bit. The list cannot include all possible decimal numbers in $[0, 1]$ since we can construct such a number $0.b_1b_2b_3\ldots b_n\ldots$ which is not on the above list as follows: if $a_{nn} \leqslant 5$ then take $b_n = 6$ and if $a_{nn} > 5$ then take $b_n = 4$. Then $0.b_1b_2\ldots b_n\ldots$ differs from the nth number on the list in the nth decimal place in such a way that $0.b_1b_2\ldots b_n\ldots$ cannot equal the nth number on the list due to any trick involving recurring decimals. This completes the proof. $\square$

Supplementary material for Chapter 3

This proves two results in Chapter 3.

Proof of Proposition 3.7 (1) We look at the parts of Proposition 3.7 not proved in the book. We wish to prove that if $f : X \to Y$ is a map and for each i in some indexing set I we are given a subset A_i of X and a subset C_i of Y, then

$$f\left(\bigcup_{i\in I} A_i\right) = \bigcup_{i\in I} f(A_i), \qquad f\left(\bigcap_{i\in I} A_i\right) \subseteq \bigcap_{i\in I} f(A_i), \qquad f^{-1}\left(\bigcup_{i\in I} C_i\right) = \bigcup_{i\in I} f^{-1}(C_i).$$

Let $y \in f\left(\bigcup_{i\in I} A_i\right)$. Then $y = f(x)$ for some $x \in \bigcup_{i\in I} A_i$, so $x \in A_{i_0}$ for some $i_0 \in I$. Then $y = f(x) \in f(A_{i_0})$ so $y \in \bigcup_{i\in I} f(A_i)$. This argument may be run backwards, so we get the first equality.

Now let $y \in f\left(\bigcap_{i\in I} A_i\right)$. Then $y = f(x)$ for some $x \in \bigcap_{i\in I} A_i$. So $x \in A_i$ for every $i \in I$, hence $y \in f(A_i)$ for every $i \in I$, and the second identity is proved.

Finally, $x \in X$ is in $f^{-1}\left(\bigcup_{i\in I} C_i\right)$ iff $f(x) \in \bigcup_{i\in I} C_i$ iff $f(x) \in C_{i_0}$ for some $i_0 \in I$ iff $x \in f^{-1}(C_{i_0})$ for some $i_o \in I$ iff $x \in \bigcup_{i\in I} f^{-1}(C_i)$. So the third identity also holds. □

Proof of Proposition 3.9 (1) We recall the setting for Proposition 3.9: $f : X \to Y$ is a map, A, B are subsets of X, and C, D are subsets of Y. We want to prove that $f(A\backslash B) \supseteq f(A)\backslash f(B)$ and $f^{-1}(C \setminus D) = f^{-1}(C) \setminus f^{-1}(D)$.

Let $y \in f(A) \setminus f(B)$. Then $y = f(a)$ for some $a \in A$, but $y \neq f(b)$ for any $b \in B$. Hence we must have $a \in A \setminus B$, so $y \in f(A \setminus B)$ as required.

Suppose that $x \in f^{-1}(C \setminus D)$. Then $f(x) \in C \setminus D$. So $f(x) \in C$ but $f(x) \notin D$. Hence $x \in f^{-1}(C)$ but $x \notin f^{-1}(D)$, so $x \in f^{-1}(C)\backslash f^{-1}(D)$. This proves $f^{-1}(C\backslash D) \subseteq f^{-1}(C)\backslash f^{-1}(D)$.

Suppose $x \in f^{-1}(C) \setminus f^{-1}(D)$. Then $x \in f^{-1}(C)$ but $x \notin f^{-1}(D)$. So $f(x) \in C$ but $f(x) \notin D$. Hence $f(x) \in C \setminus D$, so $x \in f^{-1}(C \setminus D)$. This proves $f^{-1}(C) \setminus f^{-1}(D) \subseteq f^{-1}(C \setminus D)$.

These two together give $f^{-1}(C) \setminus f^{-1}(D) = f^{-1}(C \setminus D)$ as required. □

Supplementary material for Chapter 4

Here is a list of supplementary topics for Chapter 4.

The square root of 2 (2) The set $S = \{x \in \mathbb{Q} : x^2 < 2\}$ is non-empty (for example $1 \in S$) and bounded above (for example 2 is an upper bound for S: if $x \in S$ then $x < 2$, since if $x \geqslant 2$ then $x^2 \geqslant 4 > 2$). Hence by Proposition 4.4 there is a real number u which is the least upper bound sup S, and $u \geqslant 1$. We shall show that $u^2 = 2$, which is what we mean by '$u = \sqrt{2}$'. The proof proceeds by showing that both $u^2 < 2$ and $u^2 > 2$ lead to contradictions.

First suppose that $u^2 > 2$. Then $(u^2 - 2)/2u > 0$, so by Proposition 4.6 there is some integer n such that $0 < 1/n < (u^2 - 2)/2u$, with the consequence that $2u/n < u^2 - 2$. Then

$$(u - 1/n)^2 = u^2 - 2u/n + 1/n^2 > u^2 - 2u/n > u^2 - (u^2 - 2) = 2.$$

Hence if $x \in S$ we have $x^2 < 2 < (u - 1/n)^2$ so $x < u - 1/n$. This says that $u - 1/n$ is an upper bound for S, contradicting leastness of u.

Now suppose that $u^2 < 2$. Choose an integer n such that $0 < 1/n < (2 - u^2)/4u$ (with the consequence that $4u/n < 2 - u^2$) and also such that $1/n < 2u$ (with the consequence that $1/n^2 < 2u/n$). Then

$$(u + 1/n)^2 = u^2 + 2u/n + 1/n^2 < u^2 + 2u/n + 2u/n < u^2 + 2 - u^2 = 2.$$

Now (using Corollary 4.7) choose a rational number q such that $u < q < u + 1/n$. Then $q^2 < (u + 1/n)^2 < 2$ so $q \in S$, and since $q > u$ this contradicts the fact that u is an upper bound for S.

From these two contradictions we see that $u^2 = 2$. $\square$

We could go on and prove the existence of other square roots, and of cube roots etc. similarly. But our techniqe here is later replaced by much more efficient methods (the intermediate value theorem) which make the detailed work above redundant.

Triangle inequality (1) By definition, for $x \in \mathbb{R}$

$$|x| = \begin{cases} x & \text{if } x \geqslant 0, \\ -x & \text{if } x < 0. \end{cases}$$

From this various properties easily follow: $x^2 = |x|^2$, $|x| = +\sqrt{x^2}$ and $x \leqslant |x|$, $-x \leqslant |x|$ for all $x \in \mathbb{R}$, and $|xy| = |x||y|$ for all $x, y \in \mathbb{R}$. So for any $x, y \in \mathbb{R}$,

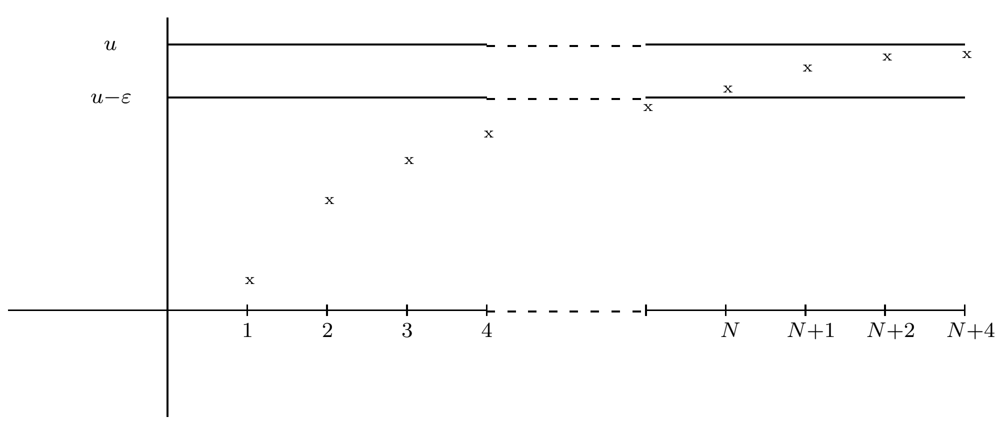
$$|x+y|^2 = (x+y)^2 = x^2 + y^2 + 2xy = |x|^2 + |y|^2 + 2xy \leqslant |x|^2 + |y|^2 + 2|xy|$$

$$= |x|^2 + |y|^2 + 2|x||y| = (|x| + |y|)^2.$$

Now taking positive square roots gives the triangle inequality. □

Reverse triangle inequality (1) We now prove Corollary 4.10. For any real numbers x, y the triangle inequality gives $|x| = |x - y + y| \leqslant |x - y| + |y|$, so $|x| - |y| \leqslant |x - y|$. Similarly $|y| - |x| \leqslant |y - x| = |x - y|$. Hence $\big|\,|x| - |y|\,\big| \leqslant |x - y|$. □

Bounded monotonic sequences converge (1) Next we prove Theorem 4.16, that any bounded monotonic sequence of real numbers converges. First we consider a monotonic increasing sequence (x_n) of real numbers which is bounded above. The set of members of the sequence is a non-empty set of real numbers which is bounded above, hence has a least upper bound which we write u. We shall prove that (x_n) converges to u. The idea of the proof is indicated in the diagram below.



Proof Let $\varepsilon > 0$. Since u is the *least* upper bound of the set of members of the sequence (x_n), there must exist an integer N with $x_N > u - \varepsilon$. The idea is that, because of monotonicity, once x_n has got above $u - \varepsilon$ 'it can never get back down again': formally, for any $n \geqslant N$ we have $x_n \geqslant x_N > u - \varepsilon$. On the other hand, u is *an* upper bound for the set $\{x_n : n \in \mathbb{N}\}$, so $x_n \leqslant u$ for all $n \in \mathbb{N}$. This tells us that $u - \varepsilon < x_n \leqslant u$ for all integers $n \geqslant N$, so in particular $|u - x_n| < \varepsilon$ for all such n, and (x_n) converges to u as claimed.

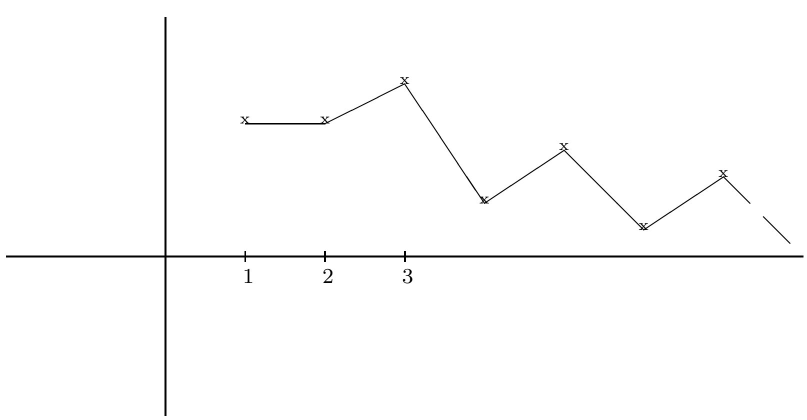
When (x_n) is monotonic decreasing and bounded below, we can either give an entirely similar argument to the above, or deduce it from the increasing case by a useful trick: put $y_n = -x_n$. Then (y_n) is monotonic increasing since (x_n) is monotonic decreasing, and (y_n) is bounded above since (x_n) is bounded below. So by the previous argument, (y_n) converges, say to y, and it is easy to check that then (x_n) converges to $-y$. $\square$

Bolzano-Weierstrass theorem (1) Recall the form of this theorem stated in Theorem 4.19: every bounded sequence of real numbers has at least one convergent subsequence. There are least two ways of proving this which are worth knowing about.

Method 1 In this we first prove the remarkable result that *any* sequence of real numbers has a monotonic subsequence. If we begin with a bounded sequence then any monotonic subsequence (provided by this remarkable result) will be bounded also, so it is convergent by Theorem 4.16, which says that a monotonic bounded sequence converges.

Proposition S.4.1 Given any sequence (x_n) of real numbers, it has either a monotonic increasing subsequence or a monotonic decreasing subsequence.

Proof The proof hinges on the definition of a particular kind of point in (x_n), which is called a 'terrace point' in Hart (2001). We shall use the similar name 'scenic viewpoint'. The integer n is called a *scenic viewpoint* (a *sv*.) for (x_n) if $x_m \leqslant x_n$ for all $m \geqslant n$. The idea is illustrated by drawing the points (n, x_n) and joining them up to form a kind of graph of the sequence: if you stand at (n, x_n) where n is a scenic viewpoint for (x_n) then you can 'see to infinity' out to the right. In the diagram below, 3 is intended to be a *sv*.



Now either (x_n) has an infinite number of *sv*.s or not. If $n_1 < n_2 < \ldots < n_r < \ldots$ is an infinite sequence of *sv*.s for (x_n) then, by definition of *sv*., (x_{n_r}) is a monotonic decreasing subsequence of (x_n). Now suppose there are only finitely many *sv*.s for (x_n), let N be the largest of them, and set $n_1 = N + 1$ (or set $n_1 = 1$ if (x_n) has no *sv*.s at all). Then since n_1 is not a *sv*. for (x_n) there must exist an integer $n_2 > n_1$ such that $x_{n_1} < x_{n_2}$. Suppose inductively that we have chosen integers $n_1 < n_2 < \ldots n_r$ such that $x_{n_1} < x_{n_2} < \ldots x_{n_r}$. Since n_r is not a *sv*. for (x_n) there exists $n_{r+1} > n_r$ such that $x_{n_r} < x_{n_{r+1}}$. In this way we construct inductively a monotonic increasing subsequence of (x_n), completing the proof of Proposition S.4.1. The Bolzano-Weierstrass theorem now follows as described at the outset. □

Method 2 More directly we can use the bisection method: if the orginal sequence (x_n) is contained in an interval $[a, b]$ then by successive bisections we find shorter and shorter subintervals of $[a, b]$ each containing a whole subsequence of (x_n), and this leads to the desired conclusion.

Since (x_n) is bounded, it is contained in some interval $[a, b]$. At least one of the intervals $[a, (a+b)/2]$, $[(a+b)/2, b]$ must contain x_n for infinitely many values of n. Label one that does $[a_1, b_1]$ (if both do, choose the left-hand one for definiteness). Inductively suppose we have already chosen real numbers $a_1, b_1, a_2, b_2, \ldots, a_m, b_m$ such that

(1) $a \leqslant a_1 \leqslant a_2 \leqslant \ldots \leqslant a_m < b_m \leqslant \ldots \leqslant b_2 \leqslant b_1 \leqslant b$,

(2) for each $i = 1, 2, \ldots m$ the interval $[a_i, b_i]$ has length $(b-a)/2^i$,

(3) for each $i = 1, 2, \ldots, m$ the interval $[a_i, b_i]$ contains x_n for infinitely many values of n.

Since $[a_m, b_m]$ contains x_n for infinitely many values of n, at least one of $[a_m, (a_m+b_m)/2]$, $[(a_m+b_m)/2, b_m]$ must have the same property. Label one that does $[a_{m+1}, b_{m+1}]$ (if both do, choose the left-hand one for definiteness). Then the inductive hypotheses are fulfilled by the numbers $a_1, b_1, a_2, b_2, \ldots, a_{m+1}, b_{m+1}$.

This construction provides a monotonic increasing sequence (a_m) which is bounded above - in fact b_i is an upper bound for each integer i. Hence (a_m) converges to a limit x satisfying $x \leqslant b_i$ for all integers i. Similarly (b_m) is a monotonic decreasing sequence which is bounded below by x, so (b_m) converges to some real number $y \geqslant x$. Since for each integer m we have $a_m \leqslant x \leqslant y \leqslant b_m$, and $b_m - a_m = (b-a)/2^m$, it follows that $x = y$.

Now construct a convergent subsequence of (x_n) as follows. Let x_{n_1} be any point in $[a_1, b_1]$. Inductively suppose that for some integer $r \geqslant 1$ integers $n_1 < n_2 < \ldots < n_r$ have been chosen so that $x_{n_i} \in [a_i, b_i]$ for each $i = 1, 2, \ldots, r$. Since $[a_{r+1}, b_{r+1}]$ contains x_n for infinitely many values of n, we may choose $n_{r+1} > n_r$ such that $x_{n_{r+1}} \in [a_{r+1}, b_{r+1}]$. This subsequence (x_{n_r}) of (x_n) converges to x, since both x_{n_r} and x are in $[a_{n_r}, b_{n_r}]$, so $|x_{n_r} - x| < (b-a)/2^r$. □

★ **Subsequences of subsequences: notation (3)** In Chapter 4 the formal definition of a sequence of objects in a set S is given as a map $s : \mathbb{N} \to S$. We then point out that traditionally we set $s(n) = s_n$ and think of a sequence in S as an infinite ordered string $s_1,\ s_2,\ s_3 \ldots$. Formally, a subsequence of a sequence $s : \mathbb{N} \to S$ is the composition $s \circ j : \mathbb{N} \to S$ where $j : \mathbb{N} \to \mathbb{N}$ is some strictly increasing function. In informal notation, a subsequence of (s_n) is usually written (s_{n_r}) where the $n_1,\ n_2,\ n_3\ \ldots$ form a strictly increasing sequence of positive integers. In the formal setting there is no serious problem with considering subsequences of subsequences and so on, since the composition of a finite number of strictly increasing functions from $\mathbb{N}$ to itself is another such function. The informal notation (s_{n_r}) is fine for considering subsequences, but when we want to look at subsequences of subsequences and so on, the multiple subscript notation $(s_{n_{r_s}})$ has no future beyond the second stage at most. The remedy we shall adopt is to talk about a subsequence $(s_{n(r,\,1)})$ of (s_n) and a subsequence $(s_{n(r,\,2)})$ of $(s_{n(r,\,1)})$ and so on. In general $(s_{n(r,\,i)})$ is a subsequence of $(s_{n(r,\,i-1)})$; the terms of $(s_{n(r,\,i)})$ are $s_{n(1,\,i)},\ s_{n(2,\,i)},\ s_{n(3,\,i)} \ldots$ where $n(1,\,i),\ n(2,\,i),\ n(3,\,i),\ \ldots$ is an increasing sequence of positive integers. So in the term $s_{n(r,\,i)}$ the 'i' tells you that this is a term in a 'level i' subsequence, and the r tells you how far along that sequence you've gone. We use subsequences of subsequences only in starred sections.★

Algebraic properties of limits of sequences (2) We prove Proposition 4.20, which says that limits of sequences are well-behaved under algebraic operations. The proofs are very similar to those in the book for continuity of real-valued functions of a real variable (Proposition 4.31). Suppose that sequences (s_n), (t_n) of real numbers converge to s, t respectively. We first prove that $(s_n + t_n)$ converges to $s + t$. Let $\varepsilon > 0$. By convergence of (s_n) to s, there exists an integer N_1 such that $|s_n - s| < \varepsilon/2$ for all $n \geqslant N_1$. Similarly there exists an integer N_2 such that $|t_n - t| < \varepsilon/2$ for all $n \geqslant N_2$. Put $N = \max\{N_1,\ N_2\}$. Then for any $n \geqslant N$,

$$|(s_n + t_n) - (s + t)| \leqslant |s_n - s| + |t_n - t| < \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

So $(s_n + t_n)$ converges to $s + t$.

Next we prove that $(s_n t_n)$ converges to st. As in the proof of the corresponding property in Proposition 4.31, to design this proof we begin at the end, using the identity

$$s_n t_n - st = s_n(t_n - t) + t(s_n - s).$$

We know that $|t_n - t|$ and $|s_n - s|$ are small when n is large; we just have to deal with the multipliers s_n and t. Here is the proof written in the forwards direction.

Let $\varepsilon > 0$. By convergence of (s_n) to s, there exists an integer N_1 such that $|s_n - s| < \varepsilon/2(1+|t|)$ whenever $n \geqslant N_1$. Since (s_n) converges to s, there exists an integer N_2 such that $|s_n - s| < 1$, and hence $|s_n| < 1 + |s|$, for all $n \geqslant N_2$. Finally, by convergence of (t_n) to t, there exists an integer N_3 such that $|t_n - t| < \varepsilon/2(1 + |s|)$ for all $n \geqslant N_3$. Let $N = \max\{N_1, N_2, N_3\}$. Then for all $n \geqslant N$,

$$|s_n t_n - st| = |s_n(t_n - t) + t(s_n - s)| \leqslant |s_n(t_n - t)| + |t(s_n - s)| = |s_n|\,|t - t_n| + |t|\,|s_n - s|$$

$$< \frac{(1+|s|)\varepsilon}{2(1+|s|)} + \frac{|t|\varepsilon}{2(1+|t|)} < \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

Hence $(s_n t_n)$ converges to st.

Finally we prove that if $t \neq 0$ then $(1/t_n)$ converges to $1/t$. Again, to understand how the proof is built up, we recommend the proof of the corresponding part of Proposition 4.31. Let $\varepsilon > 0$. We may take "ε" to be $|t|/2$ in the definition that (t_n) converges to t, to get that there exists an integer N_1 such that $|t_n - t| < |t|/2$ for all $n \geqslant N_1$. So using the reverse triangle inequality, $|t_n| = |t - (t - t_n)| \geqslant |t| - |t_n - t| > |t|/2 > 0$ for all $n \geqslant N_1$. In particular $1/t_n$ is well-defined for $n \geqslant N_1$. Also, again by convergence of (t_n) to t, there exists an integer N_2 such that $|t_n - t| < |t|^2\varepsilon/2$ for all $n \geqslant N_2$. Put $N = \max\{N_1, N_2\}$. Then for all $n \geqslant N$,

$$\left|\frac{1}{t_n} - \frac{1}{t}\right| = \frac{|t_n - t|}{|t_n||t|} < \frac{2|t|^2\varepsilon}{2|t|^2} = \varepsilon.$$

So $(1/t_n)$ converges to $1/t$. □

More on limits of functions (2) Recall that for the limit of a function f at a point a to exist, $f(a)$ need not be defined, and if $f(a)$ *is* defined then in general its value is irrelevant both to the existence of $\lim_{x\to a} f(x)$ and also to the value of this limit if it exists. In particular Example 4.22 showed that the limit of a function f at a point a may not equal the value $f(a)$. This may seem to be slightly mysterious, so here are more examples.

First here is another example in which the limit exists but does not equal $f(a)$.

Example S.4.2 Define $f : [0, 1] \to \mathbb{R}$ by:

$$f(x) = \begin{cases} 1 & \text{if } x \neq 1/2 \\ 0 & \text{if } x = 1/2 \end{cases}$$

Then $\lim_{x\to 1/2} f(x)$ exists and is 1, which is not equal to $f(1/2)$. To prove $\lim_{x\to 1/2} f(x) = 1$, let $\varepsilon > 0$. Choose any $\delta > 0$ and for all $x \in [0, 1]$ such that $0 < |x - 1/2| < \delta$ we have $|f(x) - 1| = 0 < \varepsilon$.

Here is almost the same example, this time with $f(a)$ not defined.

Example S.4.3 Define $f : [0, 1] \setminus \{1/2\} \to \mathbb{R}$ by $f(x) = 1$ for all $x \in [0, 1] \setminus \{1/2\}$. Then $\lim_{x\to 1/2} f(x)$ exists and is 1 by the same proof as for Example S.4.2.

Next here is another example in which $f(a)$ is undefined yet $\lim_{x\to a} f(x)$ exists.

Example S.4.4 Define $f : [0, 1] \setminus \{1/2\} \to \mathbb{R}$ by $f(x) = x$ for any $x \in [0, 1] \setminus \{1/2\}$. Then $\lim_{x\to 1/2} f(x) = 1/2$ although $f(1/2)$ is undefined. If you sketch the graph of f this is intuitively obvious. For a formal proof, let $\varepsilon > 0$. Let $\delta = \varepsilon$. Now suppose that $x \in [0, 1] \setminus \{1/2\}$ and $0 < |x - 1/2| < \delta$. Then $|f(x) - 1/2| = |x - 1/2| < \delta = \varepsilon$.

There is a wide class of situations similar to the above, in which a function is not defined at a point yet its limit exists as we approach that point. In fact the whole of differential calculus is based on this. For given a function $f : \mathbb{R} \to \mathbb{R}$, in order to define differentiability of f at a point $a \in \mathbb{R}$ we define, for any real number $h \neq 0$, the *Newton quotient* $\dfrac{f(a+h) - f(a)}{h}$. This is defined for $h \in \mathbb{R} \setminus \{0\}$, but not when $h = 0$. But the limit as $h \to 0$ often exists, and gives the value of the derivative $f'(a)$ of f at a.

On the other hand there are also examples where the limit of a function fails to exist at a point a even when the function is defined at a. We can get an example of this from Exercise 4.14 in the book: define $g(x)$ to be $\sin(1/x)$ when $x \neq 0$ and $g(0) = 0$. Then by Exercise 4.14, $\lim_{x\to 0} g(x)$ does not exist even although $g(0)$ does.

Proof of the conversion lemma (1) This refers to Lemma 4.25, which allows us to convert results about limits of sequences into results about limits of functions. Recall the lemma says the following are equivalent:

(i) $\lim_{x\to a} f(x) = l$,

(ii) if (x_n) is any sequence such that (x_n) converges to a but for all $n \in \mathbb{N}$ we have $x_n \neq a$, then $(f(x_n))$ converges to l.

Proof Suppose first that (i) holds, and let (x_n) be a sequence of the kind described in (ii). Let $\varepsilon > 0$. Then by (i) there exists $\delta > 0$ such that $|f(x) - l| < \varepsilon$ whenever $0 < |x - a| < \delta$. Also, there exists an integer N such that $0 < |x_n - a| < \delta$ for all $n \geqslant N$. So $|f(x_n) - l| < \varepsilon$ for all $n \geqslant N$, which says that $(f(x_n))$ converges to l.

Conversely suppose that (ii) holds. We show that $\lim_{x\to a} f(x) = l$ by contradiction. So suppose this conclusion is false. Using the technique illustrated in the supplementary material for Chapter 2 we write down the statement that $\lim_{x\to a} f(x) = l$ is false: there exists $\varepsilon > 0$ such that for any $\delta > 0$ there exists x with $0 < |x - a| < \delta$ and yet $|f(x) - l| \geqslant \varepsilon$. In particular we think about this with δ taken to be $1/n$ for a given integer n, and we give the name x_n to a point satisfying $0 < |x_n - a| < 1/n$ yet $|f(x_n) - l| \geqslant \varepsilon$. This provides a sequence (x_n) which has the properties described in (ii), yet $(f(x_n))$ does not converge to l, since it 'never gets closer than ε to l' (formally, $|f(x_n) - l| \geqslant \varepsilon$ for all $n \in \mathbb{N}$). □

Hints for starred exercises These are Exercises 4.6, 4.16, 4.17, 4.18.

4.6 Suppose that $m/n = (r/s)^2$ where r, s are mutually prime integers. Deduce from this that $ms^2 = nr^2$, hence that $r^2|m$ (r^2 divides m), say $m = r^2k$. Then note that $k|m$ and $k|n$, and use the fact that m, n have highest common factor 1 to get $r = m^2$ and $s = n^2$.

4.16 For discontinuity at a non-zero rational number a, where $f(a) = 1/q$, for any $\delta > 0$ use the existence of an irrational number between a and $a + \delta$. For continuity at $a = 0$ or a an irrational number, given any $\varepsilon > 0$ choose $q \in \mathbb{N}$ with $q \geqslant 1/\varepsilon$, and note that there are only finitely many (non-zero) rational numbers, say in $(a - 1,\, a + 1)$, whose denominators do not exceed q. Then take δ to be the distance from a to the nearest of this finite number. If m/n is a rational number within distance δ of a we must have $n > q$ so $f(m/n) = 1/n < 1/q \leqslant \varepsilon$.

4.17 For a geometric solution, note that the graph of a convex function is convex in the usual geometric sense; for any real numbers x, y the straight line segment joining the points $(x,\, f(x))$ and $(y,\, f(y))$ lies above the part of the graph between x and y. Now to see that f is continuous at $a \in \mathbb{R}$, choose $b < a$ and $c > a$, and think about the straight lines L_1 through $(a,\, f(a))$ and $(c,\, f(c))$ and L_2 through $(a,\, f(a))$ and $(b,\, f(b))$. Then you can see that the graph of f is trapped between L_1 and L_2, and since the angle between each of these lines and the x-axis is less than $\pi/2$ you may deduce continuity at a.

4.18 There are two ways in which f can have a simple jump discontinuity at $a \in \mathbb{R}$: either the left- and right-hand limits of f at a differ, or they are equal but differ from $f(a)$. One can show that each of the sets at which these two possibilities occur is countable.

To deal with the first kind of point, for each $n \in \mathbb{N}$ let D_n be the set of points $a \in \mathbb{R}$ such that $\left|\lim_{x\to a-} f(x) - \lim_{x\to a+} f(x)\right| \geqslant 1/n$. Then the points where the left- and right-hand limits of f differ is the countable union of D_n over $n \in \mathbb{N}$. With some effort one can prove that D_n is countable (the key is to show that if $a \in D_n$ then there exists $\delta_a > 0$ such that there is no other point of D_n within δ_a of a. Then we can choose a rational number in each interval $(a,\, a + \delta_a)$, get an injective function from D_n to $\mathbb{Q}$, and apply Corollary S.2.7 to show that D_n is countable.) Since a countable union of countable sets is countable, the first kind of jump discontinuities form a countable set.

The proof that the other kind of discontinuities form a countable set is similar, replacing

$$\left|\lim_{x\to a-} f(x) - \lim_{x\to a+} f(x)\right| \quad \text{by} \quad \left|f(a) - \lim_{x\to a+} f(x)\right|.$$

Supplementary material for Chapter 5

Here are the topics supplementary to Chapter 5. You may wish to skip the starred sections at a first reading.

Normed vector spaces $(1,3)$ This section assumes the reader knows what a vector space is, and just discusses the normed aspect. Normed vector spaces (NVS) are intermediate in generality between Euclidean spaces and metric spaces: as we shall see, every NVS gives rise to a metric space, but not every metric space arises from a NVS. The norm of a vector is a kind of measure of its length: for example in the Euclidean space $\mathbb{R}^n$ the Euclidean norm of any vector $x = (x_1, x_2, \ldots, x_n)$ is $||x|| = \sqrt{\sum_{i=1}^{n} x_i^2}$. This norm has properties which can be abstracted to give the following definition.

Definition S.5.1 A *normed vector space* (NVS) consists of a vector space V over $\mathbb{R}$ or $\mathbb{C}$, together with a function $V \to \mathbb{R}$ written $v \mapsto ||v||$ and called a *norm* on V, such that

(N1) for all $v \in V$, $||v|| \geqslant 0$; and $||v|| = 0$ iff $v = 0$;

(N2) for any $v \in V$ and any scalar α, $||\alpha v|| = |\alpha|\,||v||$;

(N3) (subadditivity) for all $u, v \in V$, $||u + v|| \leqslant ||u|| + ||v||$.

We immediately relate this to metric spaces.

Proposition S.5.2 Any NVS $(V, ||\ ||)$ gives rise to a metric space (X, d) where, as sets, $X = V$ and for $x, y \in V$, $d(x, y) = ||x - y||$.

Proof The following proof is fairly easy.

(M1) For any $x, y \in X$, we know $d(x, y) \geqslant 0$ since $||x-y|| \geqslant 0$; also, $d(x, y) = 0$ iff $||x-y|| = 0$ iff $x - y = 0$ by (N1), i.e. iff $x = y$.

(M2) Symmetry of d follows immediately from (N2):

for any $x,\, y \in X$, $d(y,\, x) = ||y-x|| = ||-(x-y)|| = |-1|\, ||x-y|| = ||x-y|| = d(x,\, y)$.

(M3) For any $x,\, y,\, z \in X$, using (N3) we get

$d(x,\, z) = ||x-z|| = ||x-y+y-z|| \leqslant ||x-y|| + ||y-z|| = d(x,\, y) + d(y,\, z)$. □

We now explore which metrics in Chapter 5 arise this way, and at the same time introduce examples of NVS. First, here are some norms on $\mathbb{R}^n$. The first is the Euclidean norm which we have already seen, sometimes called the l_2 norm: if $x = (x_1,\, x_2,\, \ldots,\, x_n) \in \mathbb{R}^n$, we let $||x||_2 = \sqrt{\sum x_i^2}$. It is easy to see that this satisfies (N1) and (N2). For (N3) we need Cauchy's inequality: for if also $y = (y_1,\, y_2,\, \ldots,\, y_n)$ then to prove $||x+y|| \leqslant ||x|| + ||y||$ it is equivalent to prove the same inequality after squaring both sides, which amounts to proving

$$\sum x_i y_i \leqslant \sqrt{\sum x_i^2}\sqrt{\sum y_i^2}.$$

This again is equivalent to the same inequality squared, which is Cauchy's inequality

$$\left(\sum x_i y_i\right)^2 \leqslant \left(\sum x_i^2\right)\left(\sum y_i^2\right).$$

We prove this in the next section. The metric arising from the l_2 norm is the Euclidean metric d_2 of Example 5.4 in the book.

Next we consider the l_1 and l_∞ norms on $\mathbb{R}^n$. These give rise to the metrics d_1 and d_∞ of Example 5.7 in the book. For $x \in \mathbb{R}^n$ as before, define

$$||x||_1 = \sum_{i=1}^{n} |x_i|, \quad ||x||_\infty = \max\{|x_1|,\, |x_2|,\, \ldots,\, |x_n|\}.$$

We now check that these are norms for $\mathbb{R}^n$. It is straightforward to check that (N1) and (N2) hold for each of them. Let us check (N3) for $||\ ||_1$. Given $x,\, y \in \mathbb{R}^n$ as before,

$$||x+y||_1 = |x_1+y_1| + |x_2+y_2| + \ldots + |x_n+y_n| \leqslant |x_1| + |y_1| + \ldots + |x_n| + |y_n| = ||x||_1 + ||y||_1.$$

This completes the proof that $||\ ||_1$ is a norm for $\mathbb{R}^n$.

Now let us check (N3) for $||\ ||_\infty$. With $x,\, y \in \mathbb{R}^n$ as before, for each $i = 1,\, 2,\, \ldots,\, n$ we have $|x_i + y_i| \leqslant |x_i| + |y_i| \leqslant ||x||_\infty + ||y||_\infty$, so

$$||x+y||_\infty = \max\{|x_1+y_1|,\, |x_2+y_2|,\, \ldots,\, |x_n+y_n|\} \leqslant ||x||_\infty + ||y||_\infty$$

as required.

In the section on sequence spaces we shall look at infinite-dimensional analogues of these norms on $\mathbb{R}^n$. In the meantime, we consider Examples 5.13, 5.14 and 5.16 in the book. In the first of these, X is the set of all bounded real-valued functions $f : [a, b] \to \mathbb{R}$, and we consider it as a metric space with the sup metric, so for any $f, g \in X$,

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

This metric arises from a norm on X: first, we can give X the structure of a real vector space by defining addition and scalar multiplication as follows: for any $f, g \in X$ and $\alpha \in \mathbb{R}$,

$$(f + g)(x) = f(x) + g(x); \quad (\alpha f)(x) = \alpha f(x) \text{ for all } x \in [a, b].$$

The zero vector here is the zero function on $[a, b]$. This is logically slightly sophisticated, since 'vectors' in X are now whole functions. But the check that all the vector space axioms hold follows quickly from the analogous properties for $\mathbb{R}$. For example to see that addition (of functions) is associative, we need to check, for any three functions $f, g, h : [a, b] \to \mathbb{R}$, that $(f + g) + h = f + (g + h)$ as functions. This means $(f(x) + g(x)) + h(x) = f(x) + (g(x) + h(x))$ for all $x \in [a, b]$, which follows immediately from associativity of $\mathbb{R}$. The other vector space axioms are equally straightforward to check, so we omit them and press on to define the 'sup norm' on X: for any $f \in X$ let $||f||_\infty = \sup_{x \in [a, b]} |f(x)|$. This exists since f is bounded. Again (N1) and (N2) are straightforward to check, and we concentrate on (N3). For any $f, g \in X$ and any $x \in [a, b]$ we have $|f(x) + g(x)| \leqslant |f(x)| + |g(x)| \leqslant ||f||_\infty + ||g||_\infty$, so

$$||f + g||_\infty = \sup_{x \in [a, b]} |f(x) + g(x)| \leqslant ||f||_\infty + ||g||_\infty,$$

as required. It is immediate that the metric this norm gives rise to is the sup metric on X.

Now we consider the space X of all continuous functions $f : [a, b] \to \mathbb{R}$ as in Example 5.14 in the book. This again has a vector space structure with the definitions of addition and scalar multiplication as in the above example (in fact, since all continuous real-valued functions on $[a, b]$ are bounded, we have a vector subspace of the previous vector space). Let us define 'the L^1 norm' on X by $||f||_1 = \int_a^b |f(t)| \mathrm{d}t$. In order to check (N1) in this case, we need Lemma 5.15 from the book, which says that a non-negative continuous function whose integral over $[a, b]$ is zero must be the zero function. This shows that $||f||_1 = 0$ implies $f = 0$. The rest of (N1) is straightforward to check, as is (N2). For (N3), suppose that f, g are continuous real-valued functions on $[a, b]$. Then $|f(t) + g(t)| \leqslant |f(t)| + |g(t)|$ for each $t \in [a, b]$, and integrating over $[a, b]$ we get $||f + g||_1 \leqslant ||f||_1 + ||g||_1$ as required. Again it is clear that the L^1 norm gives rise to the L^1 metric of Example 5.14.

Recall that in the final example of this trio, Example 5.16, the space X is still the space of continuous functions $f : [a,\, b] \to \mathbb{R}$, but it now has the L^2 (instead of the L^1) metric. Again we can see that this metric arises from a norm: for any $f \in X$, define

$$||f||_2 = \left\{ \int_a^b (f(t))^2 \right\}^{\frac{1}{2}}.$$

The checks that (N1) and (N2) hold are similar to those for the L^1 norm. To see that (N3) holds we need the Cauchy-Schwarz-Bunyakovsky inequality,

$$\int_a^b (f(t))^2 \mathrm{d}\, t \int_a^b (g(t))^2 \mathrm{d}\, t \geqslant \left\{ \int_a^b f(t)g(t) \mathrm{d}\, t \right\}^2.$$

(This will be proved in the next section.) Given this inequality, the check that (N3) holds here is entirely analogous to the check that (N3) holds in the case of the norm $||\ ||_2$ on $\mathbb{R}^n$. Again, it is clear that the L^2 norm gives rise to the L^2 metric of Example 5.16.

To complete this section, we show that not every metric arises from a norm. In particular the discrete metric on a set with more than one point in it does not arise from a norm. For suppose that X is a set and d is the discrete metric on X. In order for d to arise from a norm, we would require that X can be given the structure of a real or complex vector space. This already rules out any finite set with more than one point in it. (A real or complex vector space is infinite unless it has dimension zero, in which case it contains only one point.) If X can be given the structure of a real or complex vector space, there is still no norm on X which gives rise to the discrete metric. For suppose there were such a norm $||\ \ ||$, and let $x \neq 0$ in X. Then $d(x,\, 0) = 1$, so we must have $||x|| = d(x,\, 0) = 1$. But then the norm of $x/2$ would necessarily, by (N2), be $1/2$, so $||x/2 - 0|| = 1/2$, whereas in the discrete metric $d(x/2,\, 0) = 1$.

Another metric which cannot arise from a norm is the p-adic metric of Example 5.11. For $\mathbb{Z}$ does not admit the structure of a real or complex vector space on grounds of cardinality - it is a countable set, whereas, as we have already noted, any real or complex vector space is uncountable, unless it it 0-dimensional in which case it contains only one point. Similarly the word metric on a finitely generated group G in Example 5.12 does not arise from a norm, since again such a group is countable.

We shall return to NVS in later supplementary material - see the next four sections of S.5, the last section of S.6 and the first section of S.17.

In the meantime we note that there is a definition of equivalence of norms analogous to what we call 'Lipschitz equivalence' of metrics in Chapter 6. Again, the terminology is not universal.

Definition S.5.3 Norms $|| \; ||$ and $|| \; ||'$ on a vector space V are called *Lipschitz equivalent* if there exist positive real numbers h, k such that

$$h||v||' \leqslant ||v|| \leqslant k||v||' \text{ for all } v \in V.$$

It is easy to check that Lipschitz equivalent norms give rise to Lipschitz equivalent metrics.

Inner product spaces and Cauchy's inequality $(1,3)$ We have explored one 'boundary' of the NVS concept by showing that not all metrics arise from norms. We now explore a stronger condition than existence of a norm on a real or complex vector space, namely the existence of an *inner product* structure. The Euclidean norm on $\mathbb{R}^n$ arises from an inner product on $\mathbb{R}^n$: for any $x \in \mathbb{R}^n$ we have $||x||_2 = \sqrt{\langle x, x\rangle}$. We shall see that not all norms arise from inner products.

Recall that an inner product on a real vector space is a generalization of the 'scalar' or 'dot' product on $\mathbb{R}^3$: $\langle x, y\rangle = x_1y_1 + x_2y_2 + x_3y_3$ for vectors $x = (x_1, x_2, x_3)$ and $y = (y_1, y_2, y_3)$.

We abstract the properties of this dot product to make the next definition. In general, an inner product on a real vector space V is a function $V \times V \to \mathbb{R}$ written $(u, v) \mapsto \langle u, v\rangle$ satisfying:

(RI1) (Positivity) $\langle u, u\rangle \geqslant 0$ for any $u \in V$, and $\langle u, u\rangle = 0$ iff $u = 0$.

(RI2) (Symmetry) $\langle u, v\rangle = \langle v, u\rangle$ for all $u, v \in V$.

(RI3) (Bilinearity) For all $u_1, u_2, v_1, v_2 \in V$ and real numbers α, β both the following hold:

$$\langle \alpha u_1 + \beta u_2, v_1\rangle = \alpha\langle u_1, v_1\rangle + \beta\langle u_2, v_1\rangle \text{ and } \langle u_1, \alpha v_1 + \beta v_2\rangle = \alpha\langle u_1, v_1\rangle + \beta\langle u_1, v_2\rangle$$

The second part of (RI3) could be omitted, since it follows from the first part of (RI3) and symmetry.

★ To some extent it is just introducing another complication to look at the complex case, and at a first reading it is probably wise to concentrate on the real case, but in analysis one is often interested in the complex case (for the usual reasons of algebraic completeness) so we include a mention of the complex case here.

We may define complex inner product spaces by abstracting the properties of the function $\mathbb{C}^3 \times \mathbb{C}^3 \to \mathbb{C}$ given by $((z_1, z_2, z_3), (w_1, w_2, w_3)) \mapsto z_1\overline{w}_1 + z_2\overline{w}_2 + z_3\overline{w}_3$.

A *complex inner product space* is a complex vector space V and a function $V \times V \to \mathbb{C}$ satisfying:

(CI1) $\langle u, u\rangle \geqslant 0$ for any $u \in V$, and $\langle u, u\rangle = 0$ iff $u = 0$.

(CI2) $\langle u, v\rangle = \overline{\langle v, u\rangle}$ for all $u, v \in V$.

(CI3) $\langle \alpha u_1 + \beta u_2, v\rangle = \alpha\langle u_1, v\rangle + \beta\langle u_2, v\rangle$ for all $u_1, u_2, v \in V$ and all $\alpha, \beta \in \mathbb{C}$.

We note that from (CI2) each $\langle u,\, u\rangle$ is real, so (CI1) makes sense. This time we have been more economical in the third property; (CI3) and (CI2) give $\langle u,\, \alpha v_1 + \beta v_2\rangle = \overline{\alpha}\langle u,\, v_1\rangle + \overline{\beta}\langle u,\, v_2\rangle$ for all $u,\, v_1,\, v_2 \in V$ and $\alpha,\, \beta \in \mathbb{C}$. The combination of this and (CI3) is called 'sesquilinearity' rather than bilinearity; it is linear in the first variable but conjugate linear in the second. Some textbooks interchange the roles of the first and second factors here, modelling complex inner products on $\mathbb{C}^3 \times \mathbb{C}^3 \to \mathbb{C}$ given by $((z_1,\, z_2,\, z_3),\, (w_1,\, w_2,\, w_3)) \mapsto \overline{z_1}w_1 + \overline{z_2}w_2 + \overline{z_3}w_3$. ★

We want to show that any inner product on a real or complex vector space V gives rise to a norm on V. For this we shall need a slightly abstract form of Cauchy's inequality; the more concrete forms we need elsewhere follow easily from this abstract version. We state it first in the real case.

Proposition S.5.4 Let $(V,\, \langle\ \rangle)$ be a real inner product space, and for any $u \in V$ define $||u||$ by $||u|| = \sqrt{\langle u, u\rangle}$. Then for any $u,\, v \in V$,

$$|\langle u,\, v\rangle| \leqslant ||u||\, ||v||.$$

Proof For any real number x, the inner product $\langle xu{+}v,\, xu{+}v\rangle = ||xu{+}v||^2 \geqslant 0$. This says that $x^2||u||^2{+}2x\langle u,\, v\rangle{+}||v||^2 \geqslant 0$ for all $x \in \mathbb{R}$. Think of this as a quadratic expression in x. The fact that is it always non-negative means that the quadratic equation $x^2||u||^2 + 2x\langle u,\, v\rangle + ||v||^2 = 0$ has either no real roots, or else one repeated real root; for otherwise the graph of the quadratic function would cross the real axis twice and would be negative for some values of x. Hence the discriminant "$b^2 - 4ac$" of the quadratic does not exceed zero. This says $\langle u,\, v\rangle^2 \leqslant ||u||^2||v||^2$, from which the result follows. □

★ The version of Cauchy's inequality in Proposition S.5.4 holds also in the complex case, where again we define $||u|| = \sqrt{\langle u,\, u\rangle}$; the proof has an added complication, which we now explain. We first note that any non-zero complex number w may be written in polar form, $w = |w|\mathrm{e}^{i\theta}$ for some (real) $\theta \in [0,\ 2\pi)$. We also note that Cauchy's inequality is trivially true if $\langle u,\, v\rangle = 0$, so we may assume that $\langle u,\, v\rangle \neq 0$, and as above we may write it as $|\langle u,\, v\rangle|\mathrm{e}^{i\theta}$ for some $\theta \in [0,\, 2\pi)$. For any complex number z, the inner product $\langle zu + v,\, zu + v\rangle \geqslant 0$ by (CI1). Now

$$\langle zu + v,\, zu + v\rangle = |z|^2||u||^2 + z\langle u,\, v\rangle + \overline{z}\langle v,\, u\rangle + ||v||^2 = |z|^2||u||^2 + z\langle u,\, v\rangle + \overline{z\langle u,\, v\rangle} + ||v||^2$$

$$= |z|^2||u||^2 + 2\,\mathrm{Re}(z\langle u,\, v\rangle) + ||v||^2.$$

Let x be any real number x and set $z = x\mathrm{e}^{-i\theta}$. Then $|z|^2 = x^2$ and $\mathrm{Re}(z\langle u,\, v\rangle) = x|\langle u,\, v\rangle|$, since $\langle u,\, v\rangle = |\langle u,\, v\rangle|\mathrm{e}^{i\theta}$. The upshot is that

$$x^2||u||^2 + 2x|\langle u,\, v\rangle| + ||v||^2 \geqslant 0 \ \text{ for all } x \in \mathbb{R},$$

and Cauchy's inequality follows just as in the case of a real inner product space. ★

We can now prove

Proposition S.5.5 Given a real or complex inner product space $(V, \langle\ \rangle)$, we may define a norm $||\ ||$ for V by the formula $||u|| = \sqrt{\langle u, u\rangle}$.

Proof We prove this in the real case first, to avoid complications. (N1) and (N2) are straightforward to check.

(N3) We want to check that $||u+v|| \leqslant ||u|| + ||v||$ for any $u, v \in V$. It is equivalent to prove this after squaring both sides, so it is enough to prove

$$\langle u+v, u+v\rangle \leqslant ||u||^2 + 2||u||\,||v|| + ||v||^2.$$

Now $\langle u+v, u+v\rangle = ||u||^2 + 2\langle u, v\rangle + ||v||^2$, so it is enough to prove $\langle u, v\rangle \leqslant ||u||\,||v||$, which follows from Cauchy's inequality, Proposition S.5.4. □

★ In the complex case, for checking (N2) we note that

$$\langle \alpha u, \alpha u\rangle = \alpha\overline{\alpha}\langle u, u\rangle = |\alpha|^2\langle u, u\rangle, \text{ so taking square roots } ||\alpha u|| = |\alpha|\,||u||.$$

To check (N3) in the complex case, we note that for any complex number w we have $\mathrm{Re}\, w \leqslant |w|$. When we expand $\langle u+v, u+v\rangle$ in the complex case, we get $||u||^2+||v||^2+2\mathrm{Re}\langle u, v\rangle$ and as above $\mathrm{Re}\langle u, v\rangle \leqslant |\langle u, v\rangle|$ Also, $|\langle u, v\rangle| \leqslant ||u||\,||v||$ by Cauchy's inequality. So $\mathrm{Re}\langle u, v\rangle \leqslant ||u||\,||v||$. The proof of (N3) is completed as in the real case. ★

In particular, Proposition S.5.4 applies to $\mathbb{R}^n$ with its usual inner product, and this gives the classical form of Cauchy's inequality: for vectors $u = (a_1, a_2, \ldots, a_n)$ and $v = (b_1, b_2, \ldots, b_n)$ in $\mathbb{R}^n$ we have

$$\left(\sum_{i=1}^{n} a_i b_1\right)^2 \leqslant \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right).$$

This is used in Chapter 5 to prove that the Euclidean d_2 is a metric on $\mathbb{R}^n$.

It is worth mentioning a more elementary proof of the classical Cauchy inequality above. We first check that for real numbers $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ the following Lagrange identity holds

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 = \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right) - \sum_{1\leqslant i<j\leqslant n} (a_i b_j - a_j b_i)^2.$$

Since the third summation on the right-hand side is always non-negative, being a sum of squares, Cauchy's inequality follows.

Another case in which Proposition S.5.5 applies is Example 5.16: we can define a (real) inner product on the space $\mathcal{C}[a, b]$ of all continuous functions $f : [a, b] \to \mathbb{R}$ by the formula

$$\langle f, g\rangle = \int_a^b f(t)g(t)\,\mathrm{d}\,t \quad \text{for all } f, g \in \mathcal{C}[a, b].$$

As before, we need Lemma 5.15 of the book to check positivity, but the other checks that this is an inner product are straightforward. The abstract form of Cauchy's inequality in Proposition S.5.4 gives in this case the Cauchy-Schwarz-Bunyakovsky inequality: for any $f,\, g \in \mathcal{C}[a,\, b]$,

$$\left\{\int_a^b f(t)g(t)\mathrm{d}\,t\right\}^2 \leqslant \int_a^b (f(t))^2 \mathrm{d}\,t \int_a^b (g(t))^2 \mathrm{d}\,t.$$

The resulting norm is the L^2-norm discussed in the previous section.

★ We can give a complex version of the above example by letting X be the set of all continuous complex-valued functions on $[a,\, b]$ and defining

$$\langle f,\, g\rangle = \int_a^b f(t)\overline{g(t)}\mathrm{d}\,t \quad \text{for all } f,\, g \in X.$$

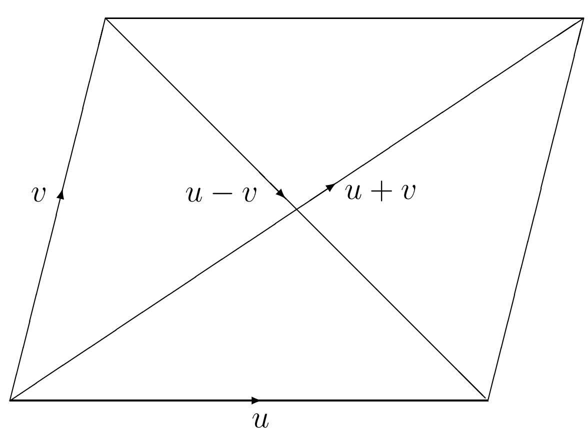
One can check that this gives a complex inner product space and (using Cauchy's inequality for a complex inner product space) that this gives rise to a normed complex vector space. ★

The norms which come from an inner product have a more geometric feel than the others. This impression is reinforced by the next result.

Proposition S.5.6 For any inner product space V, the norm arising from the inner product satisfies the 'parallelogram equality': for any $u,\, v \in V$

$$||u+v||^2 + ||u-v||^2 = 2(||u||^2 + ||v||^2).$$

Before proving this we indicate in a diagram how it generalizes an elementary geometric property in the Euclidean plane, that the sum of the squares of the lengths of the diagonals in a parallelogram equals the sum of the squares of the lengths of the sides (a fact which can be deduced from Pythagoras' theorem).



Proof of Proposition S.5.6 The proof is the same in the real and complex cases: $||u+v||^2 + ||u-v||^2 = \langle u+v,\, u+v\rangle + \langle u-v,\, u-v\rangle = 2(\langle u,\, u\rangle + \langle v,\, v\rangle) = 2(||u||^2 + ||v||^2)$. □

Proposition S.5.6 enables us to see easily that certain norms cannot arise from any inner product. For example consider the l_1 norm on $\mathbb{R}^n$. Let $u = (1, 0, 0, \ldots, 0)$, $v = (0, 1, 0, 0, \ldots, 0)$. Then

$$||u+v||_1^2 + ||u-v||_1^2 = 8, \text{ but } 2(||u||_1^2 + ||v||_1^2) = 4.$$

Similarly for the l_∞ norm on $\mathbb{R}^n$ take $u = (1, 1, 0, 0, \ldots, 0)$, $v = (1, -1, 0, 0, \ldots, 0)$. We get

$$||u+v||_\infty^2 + ||u-v||_\infty^2 = 8, \text{ but } 2(||u||_\infty^2 + ||v||_\infty^2) = 4.$$

(In fact it can be shown that the parallelogram equality is sufficient as well as necessary for a norm to arise from an inner product, but we do not include the proof.)

Reason for choice of subscripts $\mathbf{d_1}$, $\mathbf{d_2}$, $\mathbf{d_\infty}$ (1) The reason is slightly technical. The point is that for every $r \in \mathbb{N}$ there is a metric on $\mathbb{R}^n$ called d_r, associated to the r-norm which we define by

$$||(x_1, x_2, \ldots, x_n)||_r = \left(\sum_{i=1}^{n} |x_i|^r\right)^{1/r}.$$

In the cases $r = 1$ and $r = 2$ this gives the 1-norm and the 2-norm and hence the metrics d_1 and d_2 that we have already met. In fact more generally for any real number p with $p \geqslant 1$ in place of r there is a norm, defined by the same formula. In spite of the danger that the notation p suggests a prime integer, we stick with the traditional terminology of "p-norm". The check that $||\ ||_p$ satisfies sub-additivity (for any $p \geqslant 1$) involves a generalization of Cauchy's inequality called Minkowski's inequality, proved below, and this is where things become slightly technical. But before discussing Minkowski's inequality, here is why the name for $||(x_1, x_2, \ldots, x_n)||_\infty$ is reasonable.

Recall from Exercise 4.11 (with a slight change of notation) that given non-negative real numbers $a_1, a_2, \ldots, a_n$ we have

$$\lim_{r\to\infty} (a_1^r + a_2^r + \ldots + a_n^r)^{1/r} = \max\{a_1, a_2, \ldots, a_n\}.$$

Taking $a_i = |x_i|$ for each $i = 1, 2, \ldots, n$, we see that in a fairly precise sense, $||x||_r \to ||x||_\infty$ as $r \to \infty$. (In fact, the same is true for the p-norm as $p \to \infty$ through real values.)

★ We now tackle Minkowski's inequality (sub-additivity for $||\ ||_p$ when $p \geqslant 1$).

We have already proved sub-additivity for $p = 1$ so let p be a real number with $p > 1$. We define q ('the conjugate exponent' to p) by the equation $1/p + 1/q = 1$. It is easy to check that then $1/(p-1) = q - 1$, so for positive real numbers s, t we have $s = t^{p-1}$ iff $t = s^{q-1}$.

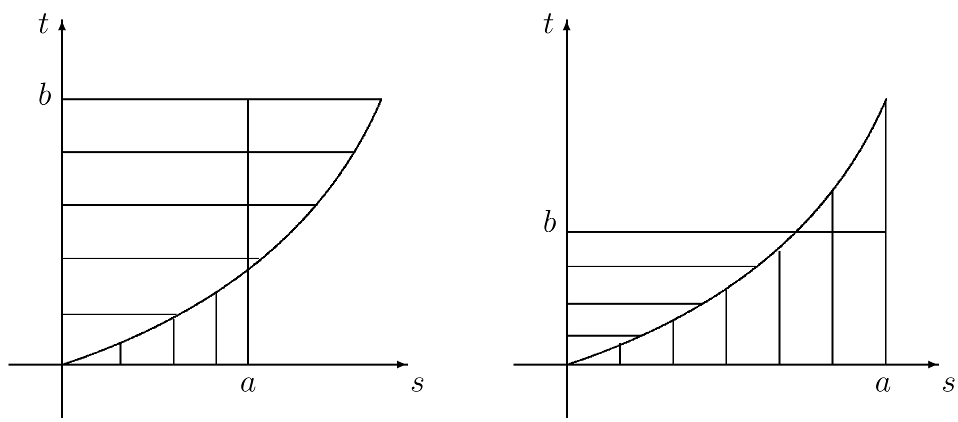
Proposition S.5.7 (Young's inequality) With p, q as above and any positive real numbers a, b,

$$ab \leqslant a^p/p + b^q/q.$$

Proof We shall show that

$$ab \leqslant \int_0^a t^{p-1}\mathrm{d}\,t + \int_0^b s^{q-1}\mathrm{d}\,s = a^p/p + b^q/q.$$

The right-hand equality comes from elementary integration, while the left-hand inequality is seen in terms of area in the diagram below. The idea is that the first integral is the area shaded vertically while the second is the area shaded horizontally. The graph of $t \mapsto t^{p-1}$ as t goes from 0 to a along the horizontal axis, matches the graph of $s \mapsto s^{q-1}$ as s goes from 0 to b along the vertical axis, because $s = t^{p-1}$ iff $t = s^{q-1}$. The left-hand diagram is for the case $a \leqslant b$ and the right-hand one is for $b \leqslant a$.



As already noted, in each diagram the horizontally shaded area is $\int_0^b t^{q-1}\mathrm{d}\,t$ while the vertically shaded area is $\int_0^a s^{p-1}\mathrm{d}\,s$. In each case the sum of the shaded areas is at least as big as the area ab of the rectangle with corners $(0, 0)$, $(a,\ 0)$, $(0, b)$ and $(a,\ b)$. This gives the left-hand inequality at the beginning, and hence completes the proof. □

The proof using area is the traditional one. Here is an alternative. Let $f(x) = ax - a^p/p - x^q/q$. Then $f'(x) = a - x^{q-1}$ which is zero iff $x = a^{1/(q-1)} = a^{p-1}$. This is a local maximum since $f''(x) = (1-q)x^{q-2}$, and $f''(a^{p-1}) < 0$ since $1 - q < 0$. Now $f(a^{p-1}) = a^p - a^p/p - a^p/q = 0$. Also, $f(0) < 0$, and $f(x) \to -\infty$ as $x \to \infty$, since $q > 1$. Hence $f(x) \leqslant 0$ for all $x \in [0, \infty)$, and in particular $f(b) \leqslant 0$, which gives Young's inequality. □

Next we deduce Hölder's inequality.

Proposition S.5.8 For real numbers $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ and any $p, q \in [1, \infty)$ with $1/p + 1/q = 1$,

$$\sum_{i=1}^{n} |a_i b_i| \leqslant \left(\sum_{i=1}^{n} |a_i|^p \right)^{1/p} \left(\sum_{i=1}^{n} |b_i|^q \right)^{1/q}.$$

Proof First we reduce the problem by noting that the target inequality is homogeneous in the sense that if it holds for given vectors $a = (a_1, a_2, \ldots, a_n)$ and $b = (b_1, b_2, \ldots, b_n)$, then it holds also for the vectors λa and μb where λ and μ are any real numbers, since the move to these vectors simply introduces a factor of $|\lambda\mu|$ on each side. It is therefore enough to assume that

$$\sum_{i=1}^{n} |a_i|^p = 1 = \sum_{i=1}^{n} |b_i|^q \text{ and prove that } \sum_{i=1}^{n} |a_i b_i| \leqslant 1.$$

(Then in the general case for vectors $a' = (a'_1, a'_2, \ldots, a'_n)$ and $b' = (b'_1, b'_2, \ldots, b'_n)$ we can take $\lambda = 1/(\sum |a'_i|^p)^{1/p}$ and $\mu = 1/(\sum |b'_i|^q)^{1/q}$, and put $a_i = \lambda a'_i$, $b_i = \mu b'_i$, which gives $\sum |a_i|^p = 1 = \sum |b_i|^q$. Then $\sum |a_i b_i| \leqslant 1$ gives $\sum |a'_i b'_i| \leqslant (\sum |a'_i|^p)^{1/p} (\sum |b'_i|^q)^{1/q}$ as required.)

So suppose that $\sum |a_i|^p = 1 = \sum |b_i|^q$. For each i we see that $|a_i b_i| \leqslant a_i^p/p + b_i^q/q$ by Young's inequality, Proposition S.5.7. Summing over i we get

$$\sum_{i=1}^{n} |a_i b_i| \leqslant \left(\sum_{i=1}^{n} a_i^p \right) /p + \left(\sum_{i=1}^{n} b_i^q \right) /q = 1/p + 1/q = 1, \quad \text{as required.} \qquad \square$$

We are now ready to prove Minkowski's inequality.

Proposition S.5.9 Given vectors $a = (a_1, a_2, \ldots, a_n)$ and $b = (b_1, b_2, \ldots, b_n)$ in $\mathbb{R}^n$ and a real number $p \geqslant 1$,

$$||a + b||_p \leqslant ||a||_p + ||b||_p, \quad \text{or equivalently} \quad \left(\sum_{i=1}^{n} |a_i + b_i| \right)^{1/p} \leqslant \left(\sum_{i=1}^{n} |a_i| \right)^{1/p} + \left(\sum_{i=1}^{n} |b_i| \right)^{1/p}.$$

Proof We have already proved this for $p = 1$ so we assume that $p > 1$. We use Hölder's inequality in a slightly quirky way. Let us write c_i for $a_i + b_i$. Multiplying the inequality $|c_i| \leqslant |a_i| + |b_i|$ by $|c_i|^{p-1}$ and then summing over i we get

$$\sum_{i=1}^{n} |c_i|^p \leqslant \sum_{i=1}^{n} |a_i||c_i|^{p-1} + \sum_{i=1}^{n} |b_i||c_i|^{p-1}.$$

We apply Hölder's inequality to each sum on the right-hand side. For the first sum we get

$$\sum_{i=1}^{n} |a_i|\,|c_i|^{p-1} \leqslant \left(\sum_{i=1}^{n} |a_i|^p \right)^{1/p} \left(\sum_{i=1}^{n} (|c_i|^{p-1})^q \right)^{1/q} = \left(\sum_{i=1}^{n} |a_i|^p \right)^{1/p} \left(\sum_{i=1}^{n} (|c_i|^p \right)^{1/q}$$

since $(|c_i|^{p-1})^q = |c_i|^{(p-1)q} = |c_i|^p$, using $(p-1)q = p$. Similarly,

$$\sum_{i=1}^{n} |b_i|\,|c_i|^{p-1} \leqslant \left(\sum_{i=1}^{n} |b_i|^p \right)^{1/p} \left(\sum_{i=1}^{n} (|c_i|^p \right)^{1/q}.$$

Putting these two inequalities together we get

$$\sum_{i=1}^{n} |c_i|^p \leqslant \left\{ \left(\sum_{i=1}^{n} |a_i|^p \right)^{1/p} + \left(\sum_{i=1}^{n} |b_i|^p \right)^{1/p} \right\} \left(\sum_{i=1}^{n} |c_i|^p \right)^{1/q} .$$

Finally we divide by the last factor here, and use $1 - 1/q = 1/p$ to get the result. □ ★

★**Brief mention of l_p (3)** We have already mentioned that for each $p \in [1, \infty)$ a norm, called the l_p norm, may be put on $\mathbb{R}^n$: for $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ we define

$$||x||_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} .$$

The norm properties (N1) and (N2) are easily checked for $||\ ||_p$, and subadditivity (N3) follows from Minkowski's inequality, Proposition S.5.9. ★

Sequence spaces $(1, 3)$ By a sequence space we mean a space in which the points are sequences, usually of real or complex numbers. (The terminology 'sequence space' is not universal.) Sequence spaces are examples of normed vector spaces which are much studied in modern analysis. They are special cases of function spaces, since for example a sequence of real numbers is formally a function $S : \mathbb{N} \to \mathbb{R}$.

Sequence spaces do have applications - for example Hilbert space is used in quantum mechanics - but here we give just a few examples of sequence spaces and ideas about them. In the supplementary material for later chapters we shall return to the examples introduced here and certain subspaces of them.

We look at the sequence spaces usually called $\mathbf{l^1}$, $\mathbf{l^2}$, $\mathbf{l^\infty}$. There are real and complex versions; analysts tend to prefer the complex versions. Here we concentrate on the real versions. These are the infinite-dimensional analogues of $\mathbb{R}^n$ with the metrics d_1, d_2, d_∞ (or the norms $||\ ||_1$, $||\ ||_2$, $||\ ||_\infty$). So $\mathbf{l^2}$ is 'the most geometric', being the infinite-dimensional analogue of $\mathbb{R}^n$ with the euclidean metric (or norm). It is called *Hilbert space* after the German mathematician Hilbert, who is famous for many things, including the list of problems which he put forward in 1900 as the most important ones unsolved at that time - many but not all have subsequently been solved, and all of them have given rise to interesting mathematics created to tackle them.

However, there's a big difference between $\mathbb{R}^n$ with the different norms mentioned above and its infinite-dimensional analogues - in finite dimensions these three norms give rise to metrics which are all Lipschitz equivalent in the sense of Chapter 6. But the infinite-dimensional analogues are not equivalent in any reasonable sense, as we shall see in the supplementary material associated with Chapter 6.

$\mathbf{l}^\infty$ This is the space of bounded sequences of real numbers. It is a vector space, with co-ordinate-wise addition and scalar multiplication, that is for (x_n), $(y_n) \in \mathbf{l}^\infty$ and $c \in \mathbb{R}$,

$$(x_n) + (y_n) = (x_n + y_n), \quad c(x_n) = (cx_n).$$

To see that this makes sense we need to check that the sum of two bounded sequences is bounded and so is the scalar product of a bounded sequence by a number. These are easy to prove. The other vector space axioms follow co-ordinatewise from the corresponding properties for real numbers. We then define the sup norm: $||\boldsymbol{x}||_\infty = \sup_{n\in\mathbb{N}} |x_n|$ where $\boldsymbol{x}$ is the sequence (x_n). Let us check that this is indeed a norm.

(N1) For any $\boldsymbol{x} = (x_n) \in \mathbf{l}^\infty$, we have $||\boldsymbol{x}||_\infty = \sup_{n\in\mathbb{N}} |x_n| \geqslant 0$. Also, for any such $\boldsymbol{x}$, $||\boldsymbol{x}||_\infty = 0$ iff $\boldsymbol{x} = (0, 0, \ldots, 0, \ldots,)$.

(N2) For any $\boldsymbol{x} = (x_n) \in \mathbf{l}^\infty$ and real number c we have

$$||c\boldsymbol{x}||_\infty = ||(cx_n)||_\infty = \sup_{n\in\mathbb{N}} |cx_n| = |c| \sup_{n\in\mathbb{N}} |x_n| = |c| ||\boldsymbol{x}||_\infty.$$

(N3) For any $\boldsymbol{x} = (x_n)$, $\boldsymbol{y} = (y_n) \in \mathbf{l}^\infty$, from the triangle inequality for $\mathbb{R}$ and the definition of $||\ ||_\infty$ we have

$$|x_n + y_n| \leqslant |x_n| + |y_n| \leqslant ||\boldsymbol{x}||_\infty + ||\boldsymbol{y}||_\infty,$$

and since this is true for every $n \in \mathbb{N}$, we get

$$||\boldsymbol{x} + \boldsymbol{y}||_\infty = \sup_{n\in\mathbb{N}} |x_n + y_n| \leqslant ||\boldsymbol{x}||_\infty + ||\boldsymbol{y}||_\infty.$$

The metric arising from this norm is the sup metric, $d_\infty(\boldsymbol{x}, \boldsymbol{y}) = \sup_{n\in\mathbb{N}} |x_n - y_n|$.

There are interesting subspaces of $\mathbf{l}^\infty$ such as the one often labelled $\mathbf{c}$, the space of all convergent sequences (of real numbers). A convergent sequence is bounded, so $\mathbf{c} \subseteq \mathbf{l}^\infty$, and the sum of two convergent sequences is convergent as is the scalar product of a convergent sequence by a constant number. So $\mathbf{c}$ is a vector subspace of $\mathbf{l}^\infty$ and we take it with the sup norm.

$\mathbf{l}^2$ The (real) Hilbert space $\mathbf{l}^2$ consists of all sequences (x_n) of real numbers such that $\sum_{n=1}^{\infty} x_n^2$ converges. The norm is defined by $||(x_n)||_2 = \sqrt{\sum_{n=1}^{\infty} x_n^2}$. It is actually an inner product space, unlike l^1 and l^∞, and it has geometric notions such as orthonormal bases and the like. Recall that this is 'the infinite-dimensional generalization of $\mathbb{R}^n$'. We now prove some of these claims.

We first show that $\mathbf{l^2}$ is a real vector space under coordinate-wise addition and scalar multiplication. It will be enough to show that if $(x_i),\,(y_i)\in\mathbf{l^2}$ and $a,\,b\in\mathbb{R}$ then $(ax_i+by_i)\in\mathbf{l^2}$. For any $n\in\mathbb{N}$ we have

$$\sum_{i=1}^{n}(ax_i+by_i)^2=a^2\sum_{i=1}^{n}x_i^2+b^2\sum_{i=1}^{n}y_i^2+2ab\sum_{i=1}^{n}x_iy_i.$$

By Cauchy's inequality in $\mathbb{R}^n$,

$$\sum_{i=1}^{n}x_iy_i\leqslant\left(\sum_{i=1}^{n}x_i^2\right)^{1/2}\left(\sum_{i=1}^{n}y_i^2\right)^{1/2}.$$

Hence

$$\begin{aligned}\sum_{i=1}^{n}(ax_i+by_i)^2&\leqslant a^2\sum_{i=1}^{n}x_i^2+b^2\sum_{i=1}^{n}y_i^2+2|ab|\left(\sum_{i=1}^{n}x_i^2\right)^{1/2}\left(\sum_{i=1}^{n}y_i^2\right)^{1/2}\\&=\left\{|a|\left(\sum_{i=1}^{n}x_i^2\right)^{1/2}+|b|\left(\sum_{i=1}^{n}y_i^2\right)^{1/2}\right\}^2\\&\leqslant\left\{|a|\left(\sum_{i=1}^{\infty}x_i^2\right)^{1/2}+|b|\left(\sum_{i=1}^{\infty}y_i^2\right)^{1/2}\right\}^2.\end{aligned}$$

This shows that the sequence of partial sums $\left(\sum_{i=1}^{n}(ax_i+by_i)^2\right)$ is bounded above, and since it is also monotonic increasing, it converges, so the sequence (ax_i+by_i) is in $\mathbf{l^2}$.

The metric defined by this norm is $d_2((x_n),\,(y_n))=\sqrt{\sum_{i=1}^{\infty}(x_n-y_n)^2}$.

We can define a real inner product on $\mathbf{l^2}$ by the formula

$$\langle(x_i),\,(y_i)\rangle=\sum_{i=1}^{\infty}x_iy_i\ \text{ for any }(x_i),\,(y_i)\in\mathbf{l^2}.$$

We first need to check that the sum on the right converges. But it is absolutely convergent, since for any $n\in\mathbb{N}$ (using Cauchy's inequality),

$$\sum_{i=1}^{n}|x_iy_i|\leqslant\left(\sum_{i=1}^{n}x_i^2\right)^{1/2}\left(\sum_{i=1}^{n}y_i^2\right)^{1/2}\leqslant\left(\sum_{i=1}^{\infty}x_i^2\right)^{1/2}\left(\sum_{i=1}^{\infty}y_i^2\right)^{1/2},$$

which says that the monotonic increasing sequence of partial sums $(\sum_{i=1}^{n}|x_iy_i|)$ is bounded above; hence it converges.

We now check that $\langle\ \rangle$ defined above is a real inner product on $\mathbf{l^2}$. Positivity is clear -

$$\langle\boldsymbol{x},\,\boldsymbol{x}\rangle=\sum_{i=1}^{\infty}x_i^2=0\ \text{ iff }\boldsymbol{x}\text{ is the zero vector in }\mathbf{l^2}.$$

Symmetry is immediate, and bilinearity follows from $(ax_i+by_i)z_i = ax_iz_i+by_iz_i$ for real numbers, since we can sum this for i going from 1 to ∞, to get for any vectors $\boldsymbol{x},\ \boldsymbol{y},\ \boldsymbol{z} \in \mathbf{l^2}$ and any real numbers a, b that

$$\langle a\boldsymbol{x} + b\boldsymbol{y},\, \boldsymbol{z}\rangle = a\langle \boldsymbol{x},\, \boldsymbol{z}\rangle + b\langle \boldsymbol{y},\, \boldsymbol{z}\rangle.$$

Hence $\mathbf{l^2}$ is a real inner product space.

From Proposition S.5.5 we know that the formula $||\,\boldsymbol{x}\,||_2 = \sqrt{\langle \boldsymbol{x},\, \boldsymbol{x}\rangle}$ gives a norm on $\mathbf{l^2}$.

$\mathbf{l^1}$ This has to do with absolute convergence. We take $\mathbf{l^1}$ to be the set of all real sequences (x_i) such that $\sum_{1=1}^{\infty} |x_i|$ converges. We have to check that this is a vector space under coordinate-wise sum and scalar product. So suppose that $\boldsymbol{x} = (x_i)$ and $\boldsymbol{y} = (y_i)$ are in $\mathbf{l^1}$ and $a,\, b \in \mathbb{R}$. For each $i \in \mathbb{N}$ we have $|ax_i + by_i| \leqslant |a||x_i| + |b||y_i|$. So for each $n \in \mathbb{N}$,

$$\sum_{i=1}^{n} |ax_i + by_i| \leqslant |a|\sum_{i=1}^{n} |x_i| + |b|\sum_{i=1}^{n} |y_i| \leqslant |a|\sum_{i=1}^{\infty} |x_i| + |b|\sum_{i=1}^{\infty} |y_i|.$$

As before this shows that $\sum_{i=1}^{\infty} |ax_i + by_i|$ converges, so $a\boldsymbol{x} + b\boldsymbol{y} \in \mathbf{l^1}$, and this shows that the latter is a vector space.

We define a norm on $\mathbf{l^1}$ by $||\boldsymbol{x}||_1 = \sum_{n=1}^{\infty} |x_n|$. The properties (N1) and (N2) are easily checked. For (N3), suppose that $\boldsymbol{x} = (x_i)$ and $\boldsymbol{y} = (y_i)$ are in $\mathbf{l^1}$. For each $i \in \mathbb{N}$, $|x_i + y_i| \leqslant |x_i| + |y_i|$. Summing this from 1 to n, we get

$$\sum_{i=1}^{n} |x_i + y_i| \leqslant \sum_{i=1}^{n} |x_i| + \sum_{i=1}^{n} |y_i| \leqslant \sum_{i=1}^{\infty} |x_i| + \sum_{i=1}^{\infty} |y_i| = ||\boldsymbol{x}||_1 + ||\boldsymbol{y}||_1.$$

Since this holds for all $n \in \mathbb{N}$ it follows that $||\boldsymbol{x} + \boldsymbol{y}||_1 \leqslant ||\boldsymbol{x}||_1 + ||\boldsymbol{y}||_1$ as required.

The metric arising from this norm is $d_1((x_n),\, (y_n)) = \sum_{i=1}^{\infty} |x_n - y_n|$.

Just as in the finite-dimensional case, we can show that $||\ ||_1$, $||\ ||_\infty$ do not arise from inner products on $\mathbf{l^1}$, $\mathbf{l^\infty}$. Also as in the finite-dimensional case, there are complex analogues of $\mathbf{l^1}$, $\mathbf{l^2}$ and $\mathbf{l^\infty}$. Their story is very similar to the real case, give or take a conjugate or two.

Finally we show that $\mathbf{l^1} \subseteq \mathbf{l^2} \subseteq \mathbf{l^\infty}$, but both inclusions are strict.

If $(x_n) \in \mathbf{l^2}$ then $\{x_n^2 : n \in \mathbb{N}\}$ is bounded and hence $\{|x_n| : n \in \mathbb{N}\}$ is bounded, so $\mathbf{l^2} \subseteq \mathbf{l^\infty}$. The reverse inclusion does not hold, since if $\boldsymbol{x} = (1,\, 1,\, \ldots,\, 1\, \ldots)$, then $\boldsymbol{x} \in \mathbf{l^\infty}$ but $\boldsymbol{x} \notin \mathbf{l^2}$.

Next $\mathbf{l^1} \subseteq \mathbf{l^2}$, since if $\sum |x_n|$ converges then $|x_n|$ is bounded, so there exists $K \in \mathbb{R}$ such that $|x_n| \leqslant K$ for all $n \in \mathbb{N}$, so $x_n^2 \leqslant K|x_n|$, and by comparison $\sum x_n^2$ converges. The reverse inclusion does not hold, for if $\boldsymbol{x} = (1,\, 1/2,\, 1/3,\, \ldots,\, 1/n,\, \ldots)$, then by elementary theory of series, $\boldsymbol{x} \in \mathbf{l^2}$ but $\boldsymbol{x} \notin \mathbf{l^1}$.

General open sets in $\mathbb{R}$ (2) We have seen in Example 5.33 that an 'open interval' (a, b) in $\mathbb{R}$ is open in $\mathbb{R}$ according to Definition 5.32. In this section we ask what a general open set in $\mathbb{R}$ looks like. The language of Chapter 12 can be used to give a more streamlined account, as we shall mention in S.12.

Proposition S.5.10 A subset X of $\mathbb{R}$ is open in $\mathbb{R}$ iff X is the disjoint union of a countable collection of open intervals.

Proof Since a single open interval is open in $\mathbb{R}$, any union of open intervals is open in $\mathbb{R}$.

Conversely, suppose that $X \subseteq \mathbb{R}$ is open in $\mathbb{R}$. First we define an equivalence relation on X: if $x, y \in X$ then $x \sim y$ iff for some a, b with $a < b$ we have $x, y \in (a, b) \subseteq X$. This is an equivalence relation since:

Reflexivity follows from the definition of X being open in $\mathbb{R}$: given any $x \in X$, there exists $\varepsilon > 0$ such that $(x - \varepsilon, x + \varepsilon) \subseteq X$. So $x \sim x$.

Symmetry follows from the definition of $\sim$.

Transitivity: suppose that $x, y, z \in X$ and that $x \sim y$ and $y \sim z$. Then there are intervals (a, b), (c, d) such that $x, y \in (a, b) \subseteq X$ and $y, z \in (c, d) \subseteq X$. From $a < y < b$ and $c < y < d$ we get $a < d$ and $c < b$. I claim that $(a, b) \cup (c, d)$ is an open interval (u, v) with $x, z \in (u, v) \subseteq X$. This is one of these 'obvious' facts which really needs a proof. First, if either $(a, b) \subseteq (c, d)$ or $(c, d) \subseteq (a, b)$ then the conclusion really is obvious. Suppose neither holds. Then

either $c < a$ or $d > b$ (otherwise $(c, d) \subseteq (a, b)$) and also
either $a < c$ or $b > d$ (otherwise $(a, b) \subseteq (c, d)$).
So we must have either
(i) $c < a$ and $b > d$ or
(ii) $a < c$ and $d > b$.

In case (i), $(a, b) \cup (c, d) = (c, b)$. To see that the left-hand side is contained in the right-hand side, note that $c < a$ implies $(a, b) \subseteq (c, b)$ and $b > d$ implies $(c, d) \subseteq (c, b)$. Conversely if $t \in (c, b)$ then either $c < t \leqslant a < d$ or $a < t < b$, so $t \in (a, b) \cup (c, d)$.

In case (ii), $(a, b) \cup (c, d) = (a, d)$. The proof is entirely similar to the proof in (i).

Thus $\sim$ is an equivalence relation. I claim that the corresponding equivalence classes are (disjoint) open intervals. For suppose that E is an equivalence class of X under $\sim$, and let $x \in E$. Consider the set S of $y \in \mathbb{R}$ such that $(x, y) \subseteq X$. Then S is non-empty since X is open so $(x, x + \varepsilon) \subseteq (x - \varepsilon, x + \varepsilon) \subseteq X$ for some $\varepsilon > 0$.

Case (1) If S is not bounded above, then for any $y > x$ there is some $z > y$ with $(x, z) \subseteq X$, so $y \in E$. Thus $(x, \infty) \subseteq E$.

Case (2) If S is bounded above let b be its least upper bound. Then whenever $x < z < b$ there is some $y \in S$ with $z < y$, so $(x, y) \subseteq X$ and in particular $z \in X$. So $(x, b) \subseteq X$. Thus $(x, b) \subseteq E$. Moreover, in this case there is no $y > b$ such that $(x, y) \subseteq X$ since that would contradict the fact that b is the sup of S. So there is no $y > b$ with $y \in E$.

By an entirely similar proof we can show that either $(-\infty, x) \subseteq E$ or for some $a < x$ we have $(a, x) \subseteq E$ and there is no $y < a$ such that $(y, x) \subseteq E$.

The upshot is that E is one of the following: $(-\infty, \infty)$, $(-\infty, b)$, (a, ∞), (a, b). Thus E is an open interval as claimed.

It is of the nature of equivalence classes that distinct ones are disjoint. To see that there are only countably many such equivalence classes, choose a rational nunber in each. This gives an injection from the set of equivalence classes into $\mathbb{Q}$ and shows that the former is countable (see Corollary S.2.7). This completes the proof of Proposition S.5.10. $\square$

Supplementary material for Chapter 6

Here is a list of supplementary topics for Chapter 6.

More questions about closed sets (3) There are several examples of closed sets in Chapter 6. Here we discuss two subspaces of sequence spaces, one of which is closed and the other not, and two examples of closed subsets of function spaces.

★ Recall that $\mathbf{l}^\infty$ is the vector space of all bounded sequences of real numbers, with the sup norm: for $\boldsymbol{x} = (x_n)$ we define $||\boldsymbol{x}||_\infty = \sup\{|x_n| : n \in \mathbb{N}\}$. Recall too that any NVS becomes a metric space via the formula $d(x, y) = ||x - y||$.

Let $\mathbf{c}$ be the vector subspace consisting of all convergent sequences of real numbers. Since any convergent sequence is bounded, $\mathbf{c} \subseteq \mathbf{l}^\infty$. Also, it is easy to check that if (x_n), (y_n) are sequences of real numbers converging to x, y respectively, and $a, b \in \mathbb{R}$, then $(ax_n + by_n)$ converges (to $ax + by$). So $\mathbf{c}$ is a vector subspace of $\mathbf{l}^\infty$. We give it the same norm $||\ ||_\infty$ as $\mathbf{l}^\infty$.

Example S.6.1 With the above notation, $\mathbf{c}$ is closed in $\mathbf{l}^\infty$.

We offer two similar proofs, with the tentative thought that the first may be preferred by analysts and the second by topologists.

Proof 1 of S.6.1 It is enough by Exercise 6.26 to show that every sequence in $\mathbf{c}$ which converges in $\mathbf{l}^\infty$ in fact converges to a point of $\mathbf{c}$. The notation for sequences in a sequence space is a bit confusing since each point in the space is itself a sequence: we let $(\boldsymbol{x}^{(\boldsymbol{n})})$ be a sequence in $\mathbf{c}$ which converges to a point $\boldsymbol{x} \in \mathbf{l}^\infty$, where $\boldsymbol{x}^{(\boldsymbol{n})} = (x_1^{(n)}, x_2^{(n)}, \ldots, x_i^{(n)} \ldots)$ for each $n \in \mathbb{N}$ and $\boldsymbol{x} = (x_1, x_2, \ldots, x_i, \ldots)$. (The brackets in the superscripts are meant to avoid confusion with powers.) We need to show that $\boldsymbol{x} \in \mathbf{c}$, in other words that (x_i) converges. It is enough to prove that (x_i) is Cauchy, since then it converges by Theorem 4.18. Let $\varepsilon > 0$. Since $(\boldsymbol{x}^{(\boldsymbol{n})})$ converges to $\boldsymbol{x}$, there is some $N \in \mathbb{N}$ such that $||\boldsymbol{x}^{(\boldsymbol{n})} - \boldsymbol{x}||_\infty < \varepsilon/3$ whenever $n \geqslant N$. By definition of $||\ ||_\infty$ this shows that $|x_i^{(n)} - x_i| < \varepsilon/3$ for any $i \in \mathbb{N}$ as long as $n \geqslant N$. We use this just for $n = N$ to give $|x_i^{(N)} - x_i| < \varepsilon/3$ for all $i \in \mathbb{N}$. Now $(x_i^{(N)})$ converges (as a sequence in i) since $(x_i^{(N)}) \in \mathbf{c}$. Hence $(x_i^{(N)})$ is a Cauchy sequence. So there exists $I \in \mathbb{N}$ such that $|x_i^{(N)} - x_j^{(N)}| < \varepsilon/3$ whenever $j > i \geqslant I$. Now for $j > i \geqslant I$, we have $|x_i - x_j| \leqslant |x_i - x_i^{(N)}| + |x_i^{(N)} - x_j^{(N)}| + |x_j^{(N)} - x_j| < \varepsilon$. Thus (x_i) is a Cauchy sequence as required. □

Proof 2 We show that the complement of $\mathbf{c}$ in $\mathbf{l}^\infty$ is open in $\mathbf{l}^\infty$. So suppose $\boldsymbol{x} = (x_n) \in \mathbf{l}^\infty$ and $\boldsymbol{x} \notin \mathbf{c}$, which is to say that (x_n) does not converge. Hence (x_n) is not a Cauchy seqence. Hence there exists some $\varepsilon > 0$ such that for any $N \in \mathbb{N}$ there are integers m, n with $m > n \geqslant N$ such that $|x_m - x_n| \geqslant \varepsilon$. We shall prove that $B_{\varepsilon/3}(\boldsymbol{x}) \subseteq \mathbf{l}^\infty \setminus \mathbf{c}$, showing that this complement is open in $\mathbf{l}^\infty$ as required.

Let $\boldsymbol{y} \in B_{\varepsilon/3}(\boldsymbol{x})$. Then $||\boldsymbol{y} - \boldsymbol{x}||_\infty < \varepsilon/3$, and by definition of $||\ ||_\infty$ we get $|y_n - x_n| < \varepsilon/3$ for all $n \in \mathbb{N}$. We now show that (y_n) is not Cauchy and hence not convergent, which proves that $B_{\varepsilon/3}(\boldsymbol{x}) \subseteq \mathbf{l}^\infty \setminus \mathbf{c}$ as required. For let $N \in \mathbb{N}$. Then there exist integers m, n with $m > n \geqslant N$ such that $|x_m - x_n| \geqslant \varepsilon$. Since $|y_n - x_n| < \varepsilon/3$ and $|y_m - x_m| < \varepsilon/3$ we get $|y_m - y_n| \geqslant \varepsilon/3$, for otherwise $|x_m - x_n| \leqslant |x_m - y_m| + |y_m - y_n| + |y_n - x_n| < \varepsilon$, contradicting $|x_m - x_n| \geqslant \varepsilon$. Hence (y_n) is not Cauchy, so not convergent. (If (y_n) were Cauchy, then for some $N \in \mathbb{N}$ we would have $|y_m - y_n| < \varepsilon/3$ for all $m > n \geqslant N$.) □

Example S.6.2 Let $F \subseteq \mathbf{l}^\infty$ be the set of all sequences of real numbers (x_n) such that for some $N \in \mathbb{N}$ we have $x_n = 0$ whenever $n > N$. Thus F is the set of 'eventually zero' sequences. (Note that the integer N varies with the sequence in F.) Then F is not closed in $\mathbf{l}^\infty$.

Proof It is easy to see that F is closed under addition and scalar multiplication: for if (x_n) and (y_n) are sequences of real numbers with $x_n = 0$ for all $n \geqslant N_1$ and $y_n = 0$ for $n > N_2$, then for any $a, b \in \mathbb{R}$ we have for the sequence $(ax_n + by_n)$ that $ax_n + by_n = 0$ for all $n > \max\{N_1, N_2\}$. So F is a vector subspace of $\mathbf{l}^\infty$. However, F is not closed in $\mathbf{l}^\infty$. For consider the sequence $(\boldsymbol{x}^{(n)})$ in F, where $\boldsymbol{x}^{(n)} = (1, 1/2, 1/3, \ldots, 1/n, 0, 0, \ldots)$. Then $(\boldsymbol{x}^{(n)})$ converges in the $||\ ||_\infty$ norm to $\boldsymbol{x} = (1, 1/2, 1/3, \ldots, 1/n, 1/(n+1), \ldots)$: for $||\boldsymbol{x}^{(n)} - \boldsymbol{x}||_\infty = 1/(n+1) \to 0$ as $n \to \infty$. But $\boldsymbol{x} \notin F$, so by Corollary 6.30 F is not closed in $\mathbf{l}^\infty$. □

Example S.6.3 Recall that $(\mathcal{B}([a, b], \mathbb{R}), d_\infty)$ is the metric space of all the bounded functions $f : [a, b] \to \mathbb{R}$ with the sup metric d_∞. Let $c \in [a, b]$, $d \in \mathbb{R}$, and let $F(c, d) \subseteq \mathcal{B}([a, b], \mathbb{R})$ consist of all the bounded functions $f : [a, b] \to \mathbb{R}$ such that $f(c) = d$.

Proposition S.6.4 With the above notation, $F(c, d)$ is closed in $(\mathcal{B}([a, b], \mathbb{R}), d_\infty)$.

Proof We shall prove that $\mathcal{B}([a, b], \mathbb{R}) \setminus F(c, d)$ is open. Let $f \in \mathcal{B}([a, b], \mathbb{R}) \setminus F(c, d)$. Then $f(c) \neq d$. Set $r = |f(c) - d|$. We shall prove that $B_r(f) \subseteq \mathcal{B}([a, b], \mathbb{R}) \setminus F(c, d)$, which will prove that this complement is open as required. So let $g \in B_r(f)$. If $g(c) = d$ then $|g(c) - f(c)| = |d - f(c)| = r$, so $||g - f||_\infty \geqslant r$, contradicting $g \in B_r(f)$. Hence $g(c) \neq d$ and $g \in \mathcal{B}([a, b], \mathbb{R}) \setminus F(c, d)$ as required. □

Now let $C \subset [a, b]$ be any subset of $[a, b]$ and let $h : C \to \mathbb{R}$ be any bounded function. Let $F(C, h) \subseteq \mathcal{B}([a, b], \mathbb{R})$ be the subset of bounded functions $f : [a, b] \to \mathbb{R}$ such that F agrees with h on C, that is $f|C = h$.

Corollary S.6.5 With notation as above, $F(C, h)$ is closed in $(\mathcal{B}([a, b], \mathbb{R}), d_\infty)$.

Proof $F(C, h) = \bigcap_{c \in C} F(c, h(c))$, an intersection of sets closed in $(\mathcal{B}([a, b], \mathbb{R}), d_\infty)$. □★

Denseness (2) Recall that a subset A of a metric space X is said to be *dense in* X if the closure of A in X is X. We prove Example 6.10, that both $\mathbb{Q}$ and $\mathbb{R} \setminus \mathbb{Q}$ are dense in $\mathbb{R}$. It is enough to show that for any $x \in \mathbb{R}$ and any $\varepsilon > 0$ the sets $B_\varepsilon(x) \cap \mathbb{Q}$ and $B_\varepsilon(x) \cap (\mathbb{R} \setminus \mathbb{Q})$ are non-empty. This follows from Corollary 4.7 and Exercise 4.8, which say that between any two real numbers (for example x and $x + \varepsilon$) there is a rational number and also an irrational number.

Separability (3) There is a concept closely related to denseness which we now define.

Definition S.6.6 A metric space is said to be *separable* if it contains a countable dense set.

We later (in S.8) relate this to another concept, second countability.

Example S.6.7 The real line, with its usual metric, is separable. For $\mathbb{Q}$ is a countable dense subset.

Example S.6.8 The real line with the discrete metric is not separable: for no proper subset A of $\mathbb{R}$ is dense in $\mathbb{R}$ when it is given the discrete metric d_0; take any $x \in \mathbb{R} \setminus A$; then $B_1^{d_0}(x) = \{x\}$, so it has empty intersection with A, which tells us that x is not in the closure of A in the discrete metric. The same argument shows that a discrete metric space X is separable iff it is countable. (If X is a countable discrete space then X itself is a countable dense subset.)

Proposition S.6.9 The product of two separable metric spaces is separable.

Proof Suppose that A is a countable dense set in a metric space (X, d_X) and that B is a countable dense set in a metric space (Y, d_Y). It follows from Proposition S.2.5 that $A \times B$ is countable. We shall prove that it is dense in $X \times Y$, where this product is given the metric d_2 of Example 5.10, which we recall is defined by

$$d_2((x_1, y_1), (x_2, y_2)) = \sqrt{(d_X(x_1, x_2))^2 + (d_Y(y_1, y_2))^2}.$$

Consider any open ball $B_\varepsilon^{d_2}((x, y))$ in $X \times Y$. It is enough to show that there is a point of $A \times B$ in this open ball, for then every point of $X \times Y$ is in the closure of $A \times B$ in $X \times Y$. Now by denseness of A in X, there exists a point $a \in A$ such that $a \in B_{\varepsilon/2}^{d_X}(x)$. Similarly there exists a point $b \in B \cap B_{\varepsilon/2}^{d_Y}(y)$. Now $(a, b) \in (A \times B) \cap B_\varepsilon^{d_2}((x, y))$, for clearly $(a, b) \in A \times B$ and

$d_2((x, y), (a, b)) = \sqrt{(d_X(x, a))^2 + (d_Y(y, b))^2} < \sqrt{\varepsilon^2/4 + \varepsilon^2/4} < \varepsilon$, so $(a, b) \in B_\varepsilon^{d_2}((x, y))$. □

Now as a corollary of Example S.6.7, Proposition S.6.9 and finite induction, $\mathbb{R}^n$ with the Euclidean metric is separable for any n. (Alternatively we could just prove directly that the subset $\mathbb{Q}^n$ is countable and dense in $\mathbb{R}^n$.)

★ We now examine the sequence spaces we studied in the supplementary material for Chapter 5 to see whether they are separable. Recall that any NVS becomes a metric space via the formula $d(x, y) = ||x - y||$.

Proposition S.6.10 $\mathbf{l}^\infty$ is not separable.

Proof Consider the set $S \subseteq \mathbf{l}^\infty$ of all sequences whose entries are all zeros and ones. This is an uncountable set, for if we supposed it listed in a countable way, say $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \ldots$ then we could construct the 'awkward' element of S whose nth entry is 0 if the nth entry in $\boldsymbol{x}_n$ is 1 and vice-versa - this element is not on our list, so S can't be countable. Also, if $\boldsymbol{x}, \boldsymbol{y}$ are distinct elements of S then $||\boldsymbol{x} - \boldsymbol{y}||_\infty = 1$ (since $\boldsymbol{x}$ and $\boldsymbol{y}$ differ by 1 in at least one entry, but do not differ by more than 1 in any entry). We now prove that there cannot be a countable dense set $C \subseteq \mathbf{l}^\infty$. For suppose that C is such a countable set. Since C is dense, for each point $\boldsymbol{x} \in S$ there must exist a point $c(\boldsymbol{x}) \in C$ within distance $1/2$ of $\boldsymbol{x}$. But for distinct $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{l}^\infty$, we must have $c(\boldsymbol{x}) \neq c(\boldsymbol{y})$, since if these were equal then

$$||\boldsymbol{x} - \boldsymbol{y}||_\infty \leqslant ||\boldsymbol{x} - c(\boldsymbol{x})||_\infty + ||c(\boldsymbol{x}) - \boldsymbol{y}||_\infty < 1.$$

This would define an injection c from the uncountable set S into the countable set C, which is impossible, since if we followed c by an injection of C into $\mathbb{N}$ we would get an injection of S into $\mathbb{N}$ and S would be countable. □

Proposition S.6.11 Both $\mathbf{l^1}$ and $\mathbf{l^2}$ are separable.

Proof Let $C \subset \mathbf{l^1}$ be the set of sequences which are eventually zero (hence are absolutely convergent) and all of whose entries are in $\mathbb{Q}$. This is a countable set: to see this we note that if $C_m \subseteq C$ is the set of sequences (x_n) such that $x_n = 0$ for $n > m$ and $x_n \in \mathbb{Q}$ for all $n \in \mathbb{N}$, then C_m is in one-one correspondence with $\mathbb{Q}^m$ and hence is countable. Now C is the union over all $m \in \mathbb{N}$ of the C_m, a countable union of countable sets, hence countable. (For facts about countability we refer to S.2.)

We shall prove that C is dense in $\mathbf{l^1}$. For let $\boldsymbol{x} = (x_n) \in \mathbf{l^1}$ and let $\varepsilon > 0$. We need to show that there is a point $\boldsymbol{c} \in C$ such that $||\boldsymbol{x} - \boldsymbol{c}||_1 < \varepsilon$. Since $\sum |x_n|$ is convergent, there exists $N \in \mathbb{N}$ such that $\displaystyle\sum_{n=N+1}^{\infty} |x_n| < \varepsilon/2$. For each $n = 1, 2, \ldots, N$, by Corollary 4.7 there exists a rational number r_n such that $|x_n - r_n| < \varepsilon/(2N)$. Let $\boldsymbol{c} = (r_1, r_2, r_3, \ldots, r_N, 0, 0, \ldots)$. Then $\boldsymbol{c} \in C$ and

$$||\boldsymbol{c} - \boldsymbol{x}||_1 = \sum_{n=1}^{N} |r_n - x_n| + \sum_{n=N+1}^{\infty} |x_n| < \varepsilon/2 + \varepsilon/2 = \varepsilon.$$ □

The proof for $\mathbf{l^2}$ is very similar, using the same countable set C, which is also contained in $\mathbf{l^2}$. Let $\boldsymbol{x} = (x_n) \in \mathbf{l^2}$ and let $\varepsilon > 0$. By convergence of $\sum x_n^2$, there exists an $N \in \mathbf{N}$ such that $\sum_{n=N+1}^{\infty} x_n^2 < \varepsilon^2/2$. For each $n = 1, 2, 3, \ldots, N$ by Corollary 4.7 there exists a rational number r_n with $|r_n - x_n| < \varepsilon/\sqrt{2N}$. Put $\boldsymbol{c} = (r_1, r_2, r_3, \ldots, r_N, 0, 0, \ldots)$. Then

$$||\boldsymbol{c} - \boldsymbol{x}||_2 = \left(\sum_{n=1}^{N}(r_n - x_n)^2 + \sum_{n=N+1}^{\infty} x_n^2\right)^{1/2} < (\varepsilon^2/2 + \varepsilon^2/2)^{1/2} = \varepsilon. \qquad \square$$

Non-equivalence of norms on sequence spaces (3) Since the underlying sets in $\mathbf{l^1}$, $\mathbf{l^2}$, $\mathbf{l^\infty}$ are different, you may wonder how we can compare their norms. However, $\mathbf{l^1} \subseteq \mathbf{l^2} \subseteq \mathbf{l^\infty}$, so comparisons are possible.

Proposition S.6.12 The norms $||\ ||_1$ and $||\ ||_\infty$ are not Lipschitz equivalent on $\mathbf{l^1}$; neither are the norms $||\ ||_2$ and $||\ ||_\infty$ on $\mathbf{l^2}$. The norms $||\ ||_1$ and $||\ ||_2$ are not Lipschitz equivalent on $\mathbf{l^1}$.

Proof To prove the first assertion, it is enough to show that there is no positive constant k such that $||\boldsymbol{x}||_1 \leqslant k||\boldsymbol{x}||_\infty$ for all $\boldsymbol{x} \in \mathbf{l^1}$. But given any $k > 0$ we can choose an integer $n > k$ and let $\boldsymbol{x} = (1, 1, \ldots, 1, 0, 0, \ldots)$ with n non-zero entries. Then $\boldsymbol{x} \in \mathbf{l^1}$ and $||\boldsymbol{x}||_\infty = 1$ while $||\boldsymbol{x}||_1 = n > k = k||\boldsymbol{x}||_\infty$.

An entirely similar proof shows that $||\ ||_2$ and $||\ ||_\infty$ are not Lipschitz equivalent on $\mathbf{l^2}$.

Finally, consider $\boldsymbol{x}^{(n)} = (1, 2, 3, \ldots, n, 0, 0, \ldots)$. Then $\boldsymbol{x}^{(n)} \in \mathbf{l^1}$. Also, from elementary formulas we have $||\boldsymbol{x}^{(n)}||_1 = n(n+1)/2$, $||\boldsymbol{x}^{(n)}||_2 = \sqrt{n(n+1)(2n+1)/4} \leqslant (n+1)^{3/2}$. Now given any $k > 0$ we may choose an integer n such that $k(n+1)^{3/2} < n(n+1)/2$, since the power of n on the left is less than on the right. This gives $k||\boldsymbol{x}^{(n)}||_2 < ||\boldsymbol{x}^{(n)}||_1$. Hence there is no constant $k > 0$ such that $||\boldsymbol{x}||_1 \leqslant k||\boldsymbol{x}||_2$ for all $\boldsymbol{x} \in \mathbf{l^1}$. This shows that $||\ ||_1$ and $||\ ||_2$ are not Lipschitz equivalent on $\mathbf{l^1}$. $\square$

Remark In fact the metrics these norms give rise to are not equivalent even in the sense of Definition 6.31. (Recall that Lipschitz equivalent norms give rise to Lipschitz equivalent metrics and hence to topologically equivalent metrics.) To be precise, we know $\mathbf{l^1} \subseteq \mathbf{l^2} \subseteq \mathbf{l^\infty}$, but the topology induced by $\mathbf{l^\infty}$ on $\mathbf{l^1}$ (in the sense of Chapter 10) does not coincide with the topology on $\mathbf{l^1}$ arising from its norm (in the sense of Chapter 7), and similarly for the inclusions $\mathbf{l^1} \subseteq \mathbf{l^2}$ and $\mathbf{l^2} \subseteq \mathbf{l^\infty}$. The proofs are omitted.

Supplementary material for Chapter 7

Here is a list of supplementary topics for Chapter 7.

Mathematical induction: limitations (2) Ordinary mathematical induction proves things 'for any integer n'; it says that if a statement is true for some 'base' integer n_0 (such as $n_0 = 1$) and if whenever it is true for an integer n then it is also true for $n+1$, then it is true for all $n \in \mathbb{N}$ with $n \geqslant n_0$. This is unlikely to give any difficulties when the statement is about integers - for example the statement that the sum of the first n integers is $n(n+1)/2$. But when we have a statement about unions or intersections of sets there is a temptation to extrapolate statements which are true for finite intersections or unions 'by induction' to infinite unions and intersections. Mathematical induction does not allow you to go from finite cases to the infinite case. Examples may help to show why this does not work.

Example S.7.1 Consider the collection of intervals in $\mathbb{R}$ of the form $[n, \infty)$. The intersection of any finite number of these is again such an interval - for the intersection of $[n_1, \infty), [n_2, \infty), \ldots, [n_r, \infty)$ is just $[N, \infty)$ where $N = \max\{n_1, n_2, \ldots, n_r\}$. In particular it is non-empty. But $\bigcap_{n=1}^{\infty}[n, \infty) = \emptyset$, for suppose that $x \in \bigcap_{n=1}^{\infty}[n, \infty)$. Then $x \geqslant n$ for all $n \in \mathbb{N}$, and there is no such real number x. Thus it would be invalid to argue here that since the intersection of any n of the sets is non-empty, 'by induction' the intersection of all of them is non-empty.

Example S.7.2 If X is a set and $A_n \subseteq X$ is a finite subset of X for each $n \in \mathbb{N}$, then the union of any finite subcollection of the A_n is finite, but $\bigcup_{n\in\mathbb{N}} A_n$ is not in general finite - for example if $X = \mathbb{N}$ and $A_n = \{n\}$ then the union over all n is $\mathbb{N}$ which is not finite.

This discussion is relevant to the axioms for a topological space. From (T2), the axiom that says the intersection of two open sets is open, it follows by mathematical induction that the intersection of any finite number of open sets is open, but it does not follow 'by induction' that the intersection of *any* family of open sets is open - Example 5.40 describes the family of open intervals $(-1/n, 1/n)$ in $\mathbb{R}$; here the intersection of any finite subfamily is just the smallest interval in that subfamily, but the intersection of the whole family is $\{0\}$ which is not open in $\mathbb{R}$. On the other hand axiom (T3) says explicitly that the union of *any* family of open sets is open.

Infinite intersections of open sets are not usually open (2) We have already seen in Example 5.40 that the intersection of the open intervals $(-1/n, 1/n)$ for all $n \in \mathbb{N}$ is $\{0\}$ which is not open in $\mathbb{R}$. This is typical. In any metric space (X, d) any singleton set $\{x\}$ is the intersection I of all the open sets which contain x: for if $y \neq x$ is any other point of X then we may let $r = d(x, y)$ and then the open ball $B_r(x)$ is an open set containing x and not y, so y cannot be in I. Such a singleton set $\{x\}$ in a metric space X is not generally open in X.

There *are*, however, situations in which an infinite intersection of open sets is open: for example this is true in any discrete space X, since there *any* subset of X is open in X. For example in a discrete metric space any singleton set is open. Also, if X is a finite topological space then given any infinite family of open sets, there are only finitely many distinct sets in the family, so the intersection of the infinite family of open sets is the same as the intersection of some finite family of open sets, which is open by induction from (T2).

Sierpinski space and computer science: a reference (3) In computer science, continuity corresponds closely to computability; Sierpinski space, and Exercise 8.3, are relevant. For more details about this, see for example 'Synthetic topology of data types and classical spaces' by Martín Escardó, Elsevier 2004 (Electronic Notes in Theoretical Computer Science Vol. 87).

Hints for Exercise 7.1(b) It is a non-trivial task to list all the possible topologies even for a set with only three points in it. Exercise 7.1 (b) asks you to do this for the set $\{0, 1, 2\}$. One hint which may be helpful is: you should find 29 distinct topologies.

Here is one approach to tackling the problem in a systematic fashion. First of all note that by (T1) any topology on $X = \{0, 1, 2\}$ must contain $\emptyset$ and X itself. Now proceed by the number of singletons in the topology. For example, if this number is zero, then the number of open subsets with two points in them is 0 or 1: for if there are two distinct open sets with two points in each, then their intersection is open and is a singleton. There is just one topology (the indiscrete topology) with no singleton open sets and no open sets with two points in them. There are exactly three topologies with no open singletons and just one 2-point open set, the topology $\{X, \emptyset, \{0, 1\}\}$ and two similar topologies. Now proceed to consider topologies with exactly one singleton open set. Suppose this is $\{0\}$ (there will be two other sets of topologies with $\{1\}$ or $\{2\}$ as the sole singleton open set). Then there may be (a) no 2-point open sets (giving just one topology, namely $\{X, \emptyset, \{0\}\}$) or (b) exactly one 2-point open set (one of $\{0, 1\}$, $\{0, 2\}$ or $\{1, 2\}$ so three topologies) or (c) two 2-point open sets, which must be $\{0, 1\}$ and $\{0, 2\}$ in order to obey (T2). This gives five topologies with $\{0\}$ the only singleton open set, so altogether fifteen topologies with precisely one singleton open set. Next consider the case of precisely two singleton open sets; in this case you should get altogether nine topologies. Finally, with three singleton sets we get just the discrete topology. This gives 29 distinct topologies in all.

Supplementary material for Chapter 8

Here is a list of supplementary topics for Chapter 8.

Continuous v. open for maps (3) Definition 8.1 says that a map $f : X \to Y$ of topological spaces is continuous if the inverse image under f of any open subset of Y is open in X; in other words, U open in Y implies that $f^{-1}(U)$ is open in X. An *open map* on the other hand is a map of topological spaces $f : X \to Y$ such that the forwards image of any open subset of X is open in Y; in other words, U open in X implies that $f(U)$ is open in Y. To help emphasize the difference, here are two examples.

Example S.8.1 Suppose that $f : \mathbb{R} \to \mathbb{R}$ is a constant map, say with value c. Then f is continuous but not open. To see that f is continuous, let U be any open subset of $\mathbb{R}$. If $c \in U$ then $f^{-1}(U) = \mathbb{R}$ and if $c \notin U$ then $f^{-1}(U) = \emptyset$. Since both $\mathbb{R}$ and $\emptyset$ are open in $\mathbb{R}$ this shows that f is continuous. To see that f is not open, take any non-empty open subset $U \subseteq \mathbb{R}$. Then $f(U) = \{c\}$ which is not open in $\mathbb{R}$, so f is not open.

Example S.8.2 Define $f : \mathbb{R} \to D$, where $D = \{0,\, 1\}$ with the discrete topology, by

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

Then f is not continuous, since if U is the open subset $\{1\}$ of D then $f^{-1}(U) = [0,\, \infty)$ which is not open in $\mathbb{R}$. But f is open since any subset of D is open in D.
In fact there do exist examples of maps $f : \mathbb{R} \to \mathbb{R}$ (in which the second space is $\mathbb{R}$) which are open but not continuous; however, these are rather weird and will be omitted.

Examples of homeomorphisms (2) We shall explore homeomorphisms further in S.12. In the meantime, let us look at the topologies which arise in answer to Exercise 7.1. Most of the work here is combinatorial rather than topological in nature.

First recall that the topologies on $X = \{0,\, 1\}$ are four in number: the discrete topology, the indiscrete topology and the topologies $\{X,\, \{0\}\}$ and $\{X,\, \{1\}\}$. If $\mathcal{T}_1$ and $\mathcal{T}_2$ are any two of these topologies then any homeomorphism $f : (X, \mathcal{T}_1) \to (X, \mathcal{T}_2)$ is a bijection, so it permutes the points 0, 1. Also, f must permute the open sets, since it satisfies $f^{-1}(U) \in \mathcal{T}_1$ iff $U \in \mathcal{T}_2$. From this we see that there are three distinct topological types here: the only two distinct spaces which are homeomorphic are $(X, \{X, \{0\}\})$ and $(X, \{X, \{1\}\})$ via the map $f : X \to X$ which interchanges 0 and 1.

⋆ We now tackle the more complicated case arising from the 29 topological spaces in the answer to Exercise 7.1 (b) (see also S.7). How many different topological types (homeomorphism classes) are there? Let $X = \{0, 1, 2\}$. Again any homeomorphism $f : (X, \mathcal{T}_1) \to (X, \mathcal{T}_2)$ is a bijection, so it permutes the three points. It must also permute the open sets, since it satisfies $f^{-1}(U) \in \mathcal{T}_1$ iff $U \in \mathcal{T}_2$. Since it is a permutation and also preserves open sets, f must permute the singleton open sets and also the 2-point open sets. Also, if $\mathcal{T}_1$ contains a singleton open set A and a 2-point open set B such that $A \subseteq B$ then any topology $\mathcal{T}_2$ which is homeomorphic to $\mathcal{T}_1$ must contain a singleton open set A' and a 2-point open set B' such that $A' \subseteq B'$. All this implies that $(X, \mathcal{T}_1)$ and $(X, \mathcal{T}_2)$ are not homeomorphic when $\mathcal{T}_1$, $\mathcal{T}_2$ are distinct topologies from the following 9:

$$\{X, \emptyset\},\ \{X, \emptyset, \{0\}\},\ \{X, \emptyset, \{0, 1\}\},\ \{X, \emptyset, \{0\}, \{0, 1\}\},\ \{X, \emptyset, \{0\}, \{1, 2\}\},$$

$$\{X, \emptyset, \{0\}, \{0, 1\}, \{0, 2\}\},\ \{X, \emptyset, \{0\}, \{1\}, \{0, 1\}\},\ \{X, \emptyset, \{0\}, \{1\}, \{0, 1\}, \{0, 2\}\}$$

and the discrete topology. We shall show that any topological space with three points is homeomorphic to one of these 9.

First, the three spaces with exactly one singleton open set and no 2-point open sets have topologies $\{X, \emptyset, \{0\}\}$, $\{X, \emptyset, \{1\}\}$, $\{X, \emptyset, \{2\}\}$. These give homeomorphic spaces, via bijections interchanging two points and leaving the third point fixed.

Next consider the three different topologies with no singleton open sets and one 2-point open set. These are $\{X, \emptyset, \{0, 1\}\}$, $\{X, \emptyset, \{0, 2\}$, $\{X, \emptyset, \{1, 2\}\}$. Again these give homeomorphic spaces via bijections interchanging two points.

Next, the following topologies give six spaces which are all homeomorphic:

$$\{X, \emptyset, \{0\}, \{0, 1\}\},\ \{X, \emptyset, \{0\}, \{0, 2\}\},\ \{X, \emptyset, \{1\}, \{0, 1\}\},$$

$$\{X, \emptyset, \{1\}, \{1, 2\}\},\ \{X, \emptyset, \{2\}, \{1, 2\}\},\ \{X, \emptyset, \{2\}, \{0, 2\}\}.$$

The first is equivalent to the second by the bijection of X interchanging 1 and 2, to the third by the bijection interchanging 0 and 1, and to the fifth by the bijection interchanging 0 and 2. The third is equivalent to the fourth by the bijection interchanging 0 and 2, and to the sixth by the bijection interchanging 1 and 2. Since topological equivalence is an equivalence relation, it follows that all six topologies are equivalent.

To complete the story on spaces with one singleton open set and one 2-point open set: the topologies $\{X, \emptyset, \{0\}, \{1, 2\}\}$, $\{X, \emptyset, \{1\}, \{0, 2\}\}$, $\{X, \emptyset, \{2\}, \{0, 1\}\}$ are all equivalent via bijections of X interchanging two points.

We now look at topologies with one singleton open set and two 2-point open sets. There are three such topologies:

$$\{X, \emptyset, \{0\}, \{0, 1\}, \{0, 2\}\},\ \{X, \emptyset, \{1\}, \{1, 0\}, \{1, 2\}\},\ \{X, \emptyset, \{2\}, \{0, 2\}, \{1, 2\}\}.$$

These give homeomorphic spaces, the first two via the bijection of X interchanging 0 and 1, the second and third via the bijection interchanging 1 and 2. Since homeomorphism is an equivalence relation it follows that all three topologies are equivalent.

Next we consider topologies on X with two singleton open sets and one 2-point open set:

$$\{X, \emptyset, \{0\}, \{1\}, \{0, 1\}\},\ \{X, \emptyset, \{0\}, \{2\}, \{0, 2\}\},\ \{X, \emptyset, \{1\}, \{2\}, \{1, 2\}\}.$$

These give homeomorphic spaces. Again it is enough to note that the first and second are homeomorphic via the bijection of X interchanging 1 and 2, the second and third via the bijection interchanging 0 and 1.

Next topologies with two singleton open sets and two 2-point open sets. Omitting X, $\emptyset$ these are:

$$\{\{0\}, \{1\}, \{0, 1\}, \{0, 2\}\},\ \{\{0\}, \{2\}, \{0, 1\}, \{0, 2\}\},\ \{\{1\}, \{2\}, \{1, 2\}, \{1, 0\}\},$$

$$\{\{0\}, \{1\}, \{0, 1\}, \{1, 2\}\},\ \{\{0\}, \{2\}, \{0, 2\}, \{1, 2\}\},\ \{\{1\}, \{2\}, \{1, 2\}, \{0, 2\}\}.$$

These six give homeomorphic spaces: the first is equivalent to the second via the bijection of X interchanging 1 and 2 and to the fourth via the bijection interchanging 0 and 1. The second is equivalent to the third via the bijection interchanging 0 and 1, and to the fifth via the bijection interchanging 0 and 2. The third is equivalent to the sixth via the bijection interchanging 1 and 2. As before it follows that all six topologies are equivalent.

Finally, each of the discrete topology and the indiscrete topology is equivalent only to itself.

Overall we see that the 29 topological spaces with underlying set $\{0, 1, 2\}$ fall into precisely 9 topological equivalence classes. $\star$

$\star$ **Bases and proto-bases (3)** We have defined a *basis* for a topological space to be a subfamily $\mathcal{B}$ of the open sets such that any open set is a union of sets from $\mathcal{B}$ (Definition 8.9). This was called an *analytic basis* in the first edition of the book: the idea of that name is that we already have a topology and we analyse the open sets to find a subfamily of them in terms of which any open set can be expressed. In the first edition we introduced also the concept called a *synthetic basis*: the idea is that we begin with just a set, and we want to build up (synthesize) a topology on the set.

Experience shows that these two closely related concepts are confusing to many students, who want to know 'What is a basis really, then?' Hence in the second edition we have stressed the first version of the concept, and called that just 'a basis'. It is the more similar to the familiar concept of a basis in a vector space. However, the other concept is sometimes useful too. To help distinguish the two concepts, we give the second version a non-standard name.

Definition S.8.3 Given a set X, a collection $\mathcal{B}$ of subsets of X is a *proto-basis* for X when the following conditions hold:

(B1) X is a union of sets from $\mathcal{B}$.

(B2) if $B_1,\, B_2 \in \mathcal{B}$ then $B_1 \cap B_2$ is a union of sets from $\mathcal{B}$.

The point of this definition is explained by the next proposition.

Proposition S.8.4 Suppose that $\mathcal{B}$ is a proto-basis for a non-empty set X. Then the family $\mathcal{T}$ of all unions of sets from $\mathcal{B}$ is a topology for X.

Proof (T1) By (B1), $X \in \mathcal{T}$. Also, under the heading 'all unions of sets' we allow ourselves to take the union of *no* sets from $\mathcal{B}$, thus getting $\emptyset \in \mathcal{T}$.

(T2) If $U = \bigcup_{i \in I} B_{i1},\ V = \bigcup_{j \in J} B_{j2}$ for some indexing sets I, J, where all B_{i1} and all B_{j2} are in $\mathcal{B}$, then

$$U \cap V = \bigcup_{(i,\,j) \in I \times J} (B_{i1} \cap B_{j2}),$$

(see Exercise 2.6) which is a union of unions of sets from $\mathcal{B}$ by (B2), hence is a union of sets from $\mathcal{B}$ as required.

(T3) A union of unions of sets from $\mathcal{B}$ is again a union of sets from $\mathcal{B}$. □

We essentially used this when we came to products of topological spaces in Chapter 10. For, given topological spaces $(X,\, \mathcal{T}_X)$, $(Y,\, \mathcal{T}_Y)$, the product topology on the set $X \times Y$ is the topology which has as proto-basis the collection $\mathcal{B} = \{U \times V : U \in \mathcal{T}_X,\, V \in \mathcal{T}_Y\}$. We check that this collection satisfies conditions (B1) and (B2) of Definition S.8.3:

(B1) $X \times Y$ is itself in $\mathcal{B}$ since $X \in \mathcal{T}_X$ and $Y \in \mathcal{T}_Y$.

(B2) If $U_1 \times V_1,\ U_2 \times V_2 \in \mathcal{B}$ then $(U_1 \times V_1) \cap (U_2 \times V_2) = (U_1 \cap U_2) \times (V_1 \cap V_2)$ and this is again in $\mathcal{B}$ since $U_1 \cap U_2 \in \mathcal{T}_X$ and $V_1 \cap V_2 \in \mathcal{T}_Y$.

Bases and proto-bases are compatible in the sense of the next proposition, which follows readily from the definitions.

Proposition S.8.5 If $\mathcal{T}$ is the topology on a non-empty set X arising from a proto-basis $\mathcal{B}$ for X as in Proposition S.8.4, then $\mathcal{B}$ is a basis for $(X,\, \mathcal{T})$ in the sense of Definition 8.9. If $(X,\, \mathcal{T})$ is a topological space and $\mathcal{B}$ is a basis for it in the sense of Definition 8.9, then

$\mathcal{B}$ is a proto-basis for the set X and the topology it gives rise to, as in Proposition S.8.4, is $\mathcal{T}$. $\square$

To recapitulate: if a topological space $(X, \mathcal{T})$ is already given, we use Definition 8.9 when talking about a basis for X (more precisely, for $(X, \mathcal{T})$). If we want to define a topology on a set X, we may sometimes use Definition S.8.3.

So bases are for use when you are given a topological space. For example in Proposition 8.12 we proved the following application of bases: to check that a map $f : X \to Y$ of topological spaces $(X, \mathcal{T}_X)$ and $(Y, \mathcal{T}_Y)$ is continuous, it is sufficient to prove that for each open set B in some basis for $\mathcal{T}_Y$, the inverse image $f^{-1}(B)$ is open in X.

On the other hand, proto-bases are used when you want to build up a topology on a set which does not yet have a topology - as a way of specifying a topology on a set, we may say 'Let $\mathcal{T}$ be the topology with the sets ... as proto-basis'. In this case, the named sets should satisfy (B1) and (B2). This is the case for the product topology as discussed above.

Given the compatibility described in Proposition S.8.5 it is traditional to refer loosely to both concepts as 'bases'. But the confusion this can cause encourages us to be clear about the distinction, hence the non-standard name 'proto-basis' (indicating 'something that gives rise to a basis') for one of them.⋆

⋆ **Sub-bases (3)** There is another concept similar to, but more general then, bases, namely that of sub-bases.

Definition S.8.6 A *sub-basis* for a topological space $(X, \mathcal{T})$ is a subfamily $\mathcal{S} \subseteq \mathcal{T}$ such that any open set is the union of finite intersections of sets from $\mathcal{S}$.

Alternatively we could define a sub-basis as a subfamily $\mathcal{S}$ of $\mathcal{T}$ such the the set of all intersections of finite families of sets from $\mathcal{S}$ forms a basis for $\mathcal{T}$. A sub-basis tends to be even more economical than a basis.

Example S.8.7 In $\mathbb{R}$ the set of all open intervals is a basis for the usual topology. For a sub-basis it is enough to take all open intervals of the special form $(-\infty, b)$ or (a, ∞), since an open interval of the form (a, b) is the (finite) intersection $(a, b) = (-\infty, b) \cap (a, \infty)$
A slightly similar example concerns the topological product of spaces $(X, \mathcal{T}_X)$, $(Y, \mathcal{T}_Y)$. The following collection of sets is a sub-basis for the product topology: the collection $\{U \times Y : U \in \mathcal{T}_X\} \cup \{X \times V : V \in \mathcal{T}_Y\}$. (Since an open set in the product topology is a union of sets of the form $U \times V = (U \times Y) \cap (X \times V)$ where $U \in \mathcal{T}_X$ and $V \in \mathcal{T}_Y$.)

An analogue of Proposition 8.12 holds for sub-bases:

Proposition S.8.8 To check that a map $f : X \to Y$ of topological spaces $(X, \mathcal{T}_X)$ $(Y, \mathcal{T}_Y)$ is continuous, it is enough to check that for each open set S in some sub-basis $\mathcal{S}$ for $\mathcal{T}_Y$, the inverse image $f^{-1}(S)$ is open in X.

Proof We use Proposition 8.12. Suppose it is known that $f^{-1}(S)$ is open in X for each $S \in \mathcal{S}$. For sets $S_1, S_2, \ldots, S_r$ in $\mathcal{S}$, we have

$$f^{-1}(S_1 \cap S_2 \cap \ldots \cap S_r) = f^{-1}(S_1) \cap f^{-1}(S_2) \cap \ldots \cap f^{-1}(S_r). \qquad (*)$$

We know that each $f^{-1}(S_i)$ is open in X, so the finite intersection on the right-hand side of $(*)$ is also open in X. Now the set of all finite intersections such as $S_1 \cap S_2 \cap \ldots \cap S_r$ of sets from $\mathcal{S}$ forms a basis $\mathcal{B}$ for $\mathcal{T}_Y$. Thus by $(*)$ $f^{-1}(B)$ is open in X for each set B in a certain basis for $(Y, \mathcal{T}_Y)$, and by Proposition 8.12, f is continuous. □

There is a concept relating to sub-basis in the way that proto-basis relates to basis, but we do not give this a name. In fact any collection $\mathcal{S}$ of subsets of a non-empty set X will play this role, provided we make the convention that taking the intersection of *no* sets in a set of subsets of X gives X. So with this convention, the set of all finite intersections of sets from $\mathcal{S}$ is a topology for X, called 'the topology generated by $\mathcal{S}$'. ⋆

Hints for Exercise 8.7 The solution is in two parts; we need to show that the given set is countable, and that it is a basis. It is countable because it is the countable union of countable sets (see S.2.8): for each $(q_1, q_2) \in \mathbb{Q}$ the set of all $B_q((q_1, q_2))$ as q ranges over positive rationals is countable (it is in one-one correspondence with the positive rational numbers) and there are just countably many points in $\mathbb{Q}^2$, since $\mathbb{Q}$ is countable and the product of two countable sets is countable by S.2.5.

To see that $\mathcal{B}$ is a basis, we have to show that given any open subset U of $\mathbb{R}^2$ and any point $(x, y) \in U$ there are rational numbers q_1, q_2, q such that $(x, y) \in B_q((q_1, q_2)) \subseteq U$. First choose $\varepsilon > 0$ in $\mathbb{R}$ such that $B_{3\varepsilon}((x, y)) \subseteq U$. Then choose a rational number q such that $\varepsilon < q < 2\varepsilon$, and rational numbers q_1, q_2 such that $|q_1 - x| < \varepsilon/\sqrt{2}$ and $|q_2 - y| < \varepsilon/\sqrt{2}$, and show that these choices work.

Separable metric is second countable Exercise 8.7 generalises not only to $\mathbb{R}^n$ for any positive integer n but to any separable metric space (X, d). Recall that X being separable means that it has a countable dense subset $\{x_n\}$. Let $\mathcal{B} = \{B_q(x_n) : n \in \mathbb{N}, q \in \mathbb{Q}, q > 0\}$. Then $\mathcal{B}$ is a countable basis for open sets. For suppose that $U \subseteq X$ is open in X. For any $x \in U$ there exists $\varepsilon > 0$ such that $B_{3\varepsilon}(x) \subseteq U$. Now by denseness of $\{x_n\}$ there exists an integer n such that $x_n \in B_\varepsilon(x)$, so also $x \in B_\varepsilon(x_n)$. Choose a rational number q such that $\varepsilon < q < 2\varepsilon$. Then $B_q(x_n) \subseteq U$ since if $d(y, x_n) < q$ then $d(y, x) \leqslant d(y, x_n) + d(x_n, x) < q + \varepsilon < 3\varepsilon$. Also, $x \in B_\varepsilon(x_n) \subseteq B_q(x_n) \subseteq U$, and $x \in B_q(x_n) \subseteq U$ is enough to show that $\mathcal{B}$ is a basis for X. Also, as a countable union (namely over all $n \in \mathbb{N}$) of the countable sets $\{B_q(x_n) : q \in \mathbb{Q}, q > 0\}$, $\mathcal{B}$ is a countable family by Proposition S.2.8.

Supplementary material for Chapter 9

Here is a list of supplementary topics for Chapter 9.

Details of proof of 9.4 (1) For (C1), we note that by (T1) both $\emptyset$ and X are open in X, so by Definition 9.1, both X and $\emptyset$ are closed in X.

(C2) Suppose that V_1, V_2 are closed in X. By Definition 9.1, $X \setminus V_1$ and $X \setminus V_2$ are open in X, hence by (T2) so too is $(X \setminus V_1) \cap (X \setminus V_2)$. But by De Morgan's laws (Chapter 2) this says that $X \setminus (V_1 \cup V_2)$ is open in X. Hence by Definition 9.1, $V_1 \cup V_2$ is closed in X.

(C3) Suppose that I is an indexing set and for every $i \in I$ we have a subset V_i which is closed in X. Then by Definition 9.1, each $X \setminus V_i$ is open in X. By (T3) then $\bigcup_{i \in I} X \setminus V_i$ is open in X. By De Morgan's laws this says that $X \setminus \bigcap_{i \in I} V_i$ is open in X, so by Definition 9.1, $\bigcap_{i \in I} V_i$ is closed in X. $\square$

Details of proof of 9.5 (1) Suppose first that $f : X \to Y$ is continuous, and let $V \subset Y$ be closed in Y. Then $Y \setminus V$ is open in Y, so by continuity $f^{-1}(Y \setminus V)$ is open in X. But by Proposition 3.8, this says that $X \setminus f^{-1}(V)$ is open in X. Now by Definition 9.1, $f^{-1}(V)$ is closed in X as required.

Conversely suppose that $f : X \to Y$ is a map such that $f^{-1}(V)$ is closed in X whenever V is closed in Y. We want to prove that f is continuous. So let $U \subseteq Y$ be open in Y. By Definition 9.1, $Y \setminus U$ is closed in Y, so by the given condition, $f^{-1}(Y \setminus U)$ is closed in X. But by Proposition 3.8 this says that $X \setminus f^{-1}(U)$ is closed in X. Now by Definition 9.1, $f^{-1}(U)$ is open in X, and this proves that f is continuous as required. $\square$

Examples for Chapter 9 (2) Next we consider examples of the various concepts in topological spaces that arise in Chapter 9. Since we have already looked at examples of the corresponding concepts in metric spaces, we shall take as our first theme an infinite space X with the co-finite topology. (Example 11.6 shows that X is not metrizable.) So for the next five examples let X be any infinite set with the co-finite topology, which recall means that the open sets are the subsets $U \subseteq X$ such that either $U = \emptyset$ or $X \setminus U$ is finite.

Examples of dense subsets of X Recall from Exercise 9.2 (or rather, your solution of it) that the closed subsets of X are the finite subsets together with X. Hence the closure of any infinite subset A of X, which by Proposition 9.10 (f) is the smallest closed subset of X containing A, is X itself. This says that any infinite subset of X is dense in X.

Example in X of an infinite union of closed sets which is not closed For each i in some infinite indexing set I we take $A_i \subseteq X$ to be the singleton set $\{x_i\}$ for some point $x_i \in X$. Then each A_i is closed in X, but the union of the A_i over all $i \in I$ is infinite hence not closed in X unless it is all of X. For a concrete example here, take $X = \mathbb{N}$, let $I = \mathbb{N}$ and for each $i \in I$ let $A_i = \{2i\}$. Then the union of all the A_i is just the set of even integers, which is not closed in $\mathbb{N}$ with the co-finite topology.

Example of failure of 9.13 for an infinite union This is closely related to the previous example. Let $A_i = \{x_i\}$. Then each A_i is closed in X so $\overline{A_i} = A_i$. Hence the union over all $i \in \mathbb{N}$ of these closures is just the union of the A_i. But the union of all the A_i is infinite, so the closure in X of this union is X itself, which fails to equal the union of the closures unless X is the union over all $i \in \mathbb{N}$ of the A_i. Again, the same specific example works as in the previous example.

Examples of interior in X If the complement of a subset A in X is finite then A is open in X so $\mathring{A} = A$. Suppose now that the complement of A in X is infinite (in particular this holds if A is finite). Let $U \subseteq A$. Then $X \setminus U \supseteq X \setminus A$ is infinite, so U fails to be open unless it is empty. Hence $\mathring{A}$, which is the largest open subset of X contained in A, is empty.

Examples of boundary in X If $A \subset X$ is finite, then it is closed so $\overline{A} = A$. On the other hand we have seen that $\mathring{A} = \emptyset$. So in this case $\partial A = A$. If $A \subseteq X$ is infinite with finite complement, then A is open in X so $\mathring{A} = A$. But as we have seen above, $\overline{A} = X$. So in this case $\partial A = X \setminus A$. If $A \subseteq X$ is infinite and so is $X \setminus A$, then as we have seen above $\mathring{A} = \emptyset$, while $\overline{A} = X$, so in this case $\partial A = X$.

To conclude this section we look more briefly at two other examples of non-metrizable spaces, those in Exercises 7.4 and 7.6.

Recall that Exercise 7.4 looked at a topology on $\mathbb{N}$ in which the open sets are $\emptyset, \mathbb{N}$ and $\{1, 2, \ldots, n\}$ for each $n \in \mathbb{N}$. Hence apart from the empty set, the closed sets are all infinite, of the form $\{m \in \mathbb{N} : m \geqslant n\}$. Consider the closure of any singleton set $\{n\}$ in this space. If $i < n$, then $\{1, 2, \ldots, i\}$ is an open set containing i which has empty intersection with $\{n\}$, so $i \notin \overline{\{n\}}$. But if $m \geqslant n$ then any open set containing m also contains n, so $m \in \overline{\{n\}}$. This says that $\overline{\{n\}} = \{n,\, n+1,\, n+2,\, \ldots\}$. Similarly the closure of any subset S of $\mathbb{N}$ consists of all integers greater than or equal to the least integer in S. In particular the closure of the open set $\{1,\, 2,\, \ldots,\, n\}$ is $\mathbb{N}$.

Exercise 7.6 gives another weird topology on $\mathbb{R}$: the open sets are $\emptyset$, $\mathbb{R}$ and all intervals of the form $(-\infty, b)$ for $b \in \mathbb{R}$. So the closed sets are $\mathbb{R}$, $\emptyset$, and all intervals of the form $[b,\, \infty)$ for $b \in \mathbb{R}$. The closure of a singleton set $\{a\}$ is $[a,\, \infty)$, for similar reasons to those in Exercise 7.4. If $S \subseteq \mathbb{R}$ is not bounded below, then $\overline{S} = \mathbb{R}$, since given any $x \in \mathbb{R}$ there exists an $a \in S$ with $a < x$, and the smallest closed set containing $\{a\}$ is $[a,\, \infty)$, which contains x, so $\mathbb{R} \subseteq \overline{S}$. In particular the open set $(-\infty, a)$ has closure $\mathbb{R}$. If the non-empty set S is bounded below, and $b = \inf S$ then $\overline{S} = [b,\, \infty)$; for any $x < b$ is not in the closure of S since $(-\infty, (b+x)/2) \cap S = \emptyset$, and $(-\infty, (b+x)/2)$ is open. This proves that $\overline{S} \subseteq [b,\, \infty)$. But if $x \geqslant b$ than any open set containing x is either $\mathbb{R}$ or an interval $(-\infty, c)$ with $c > b$, and either of these has non-empty intersection with S. (For $\mathbb{R}$ contains S, and if $c > b$ then there is a point of S in $[b,\, c)$ since b is the *greatest* lower bound of S). We have now proved that $\overline{S} = [b,\, \infty)$.

Neighbourhoods (3) Topological spaces can be defined in terms of neighbourhoods (Definition 9.22) just as well as open sets, and there is some advantage in the greater flexibility of this concept. Recall that a neighbourhood of a point x in a topological space X is a set N which contains an open set containing x. [Some textbooks insist that neighbourhoods should be open sets.] We write $\mathcal{N}_x$ for the set of all neighbourhoods of x, and call this the *neighbourhood system of X at x*. Here we explain how neighbourhoods provide an approach to topological spaces equivalent to that via open sets, and explore some concepts in Chapter 9 in terms of neighbourhoods. First we prove some properties of neighbourhoods, given a topological space defined in terms of open sets as usual.

Proposition S.9.1 For any point x in a topological space X the neighbourhood system at x satisfies

($\mathcal{N}1$) If $N \in \mathcal{N}_x$ then $x \in N$.

($\mathcal{N}2$) If $N_1,\, N_2 \in \mathcal{N}_x$ then $N_1 \cap N_2 \in \mathcal{N}_x$.

($\mathcal{N}3$) If $N \in \mathcal{N}_x$ then there is a $V \in \mathcal{N}_x$ such that $V \subseteq N$ and $V \in \mathcal{N}_v$ for every $v \in V$.

($\mathcal{N}4$) If $N \in \mathcal{N}_x$ and $N' \supseteq N$ then $N' \in \mathcal{N}_x$.

Proof ($\mathcal{N}1$) is true by definition.

($\mathcal{N}2$) Let $N_1,\, N_2 \in \mathcal{N}_x$. Then there is some open U_1 such that $x \in U_1 \subseteq N_1$. Also there is an open U_2 such that $x \in U_2 \subseteq N_2$. Then by (T2) $U_1 \cap U_2$ is open, and $x \in U_1 \cap U_2 \subseteq N_1 \cap N_2$. Hence $N_1 \cap N_2 \in \mathcal{N}_x$ as required.

($\mathcal{N}3$) Take V to be an open set with $x \in V \subseteq N$. (We know there exists such an open set V by definition of neighbourhood.) Then for any $v \in V$ the set V is an open set containing v, so $V \in \mathcal{N}_v$. Also, $V \in \mathcal{N}_x$ since $x \in V$ and V is open in X.

($\mathcal{N}4$) If $N \in \mathcal{N}_x$ and N' is a subset of X with $N' \supseteq N$, then there exists an open set U with $x \in U \subseteq N$, so also $x \in U \subseteq N'$. Hence $N' \in \mathcal{N}_x$ as required. $\square$

Next we prove a property which will enable us to turn things around and define open sets in terms of neighbourhoods.

Proposition S.9.2 In a topological space X, a subset $U \subseteq X$ is open in X iff U contains a neighborhood of each of its points.

Proof Suppose that U is open in X. Then for each $x \in U$, the set U itself is in $\mathcal{N}_x$.

Conversely suppose $U \subseteq X$ and U contains a neighborhood N_x of each point $x \in U$. Then for each $x \in U$ there is some open set U_x such that $x \in U_x \subseteq N_x \subseteq U$. Now U itself is open in X, by Proposition 7.2. $\square$

★ Now we turn things around and see how neighbourhoods can be made the fundamental concept and open sets a subsidiary idea defined in terms of neighbourhoods.

Proposition S.9.3 Suppose that in a set X there is assigned to each $x \in X$ a non-empty family $\mathcal{N}_x$ of subsets of X such that $\mathcal{N}_x$ satisfies ($\mathcal{N}1$) and ($\mathcal{N}2$) for each $x \in X$. Let $\mathcal{T}$ be the family of those subsets $U \subseteq X$ such that for each $x \in U$ there is some $N_x \in \mathcal{N}_x$ such that $N_x \subseteq U$. Then $\mathcal{T}$ is a topology for X.

Proof Let $\mathcal{T}$ be constructed as described from the $\{\mathcal{N}_x : x \in X\}$.

(T1) For any $x \in X$, $\mathcal{N}_x$ is non-empty. Take $N_x \in \mathcal{N}_x$. Then $x \in N_x$ by ($\mathcal{N}1$), and $N_x \subseteq X$. So $X \in \mathcal{T}$. Also, $\emptyset \in \mathcal{T}$ since there is no $x \in \emptyset$ to check anything for.

(T2) Suppose that $U, U' \in \mathcal{T}$, and let $x \in U \cap U'$. Then there exists $N_x \in \mathcal{N}_x$ such that $N_x \subseteq U$ and there exists $N'_x \in \mathcal{N}_x$ such that $N'_x \subseteq U'$. Then $N_x \cap N'_x \in \mathcal{N}_x$ by ($\mathcal{N}2$), and $x \in N_x \cap N'_x \subseteq U \cap U'$. Hence $U \cap U' \in \mathcal{T}$.

(T3) Suppose $U_i \in \mathcal{T}$ for all i in some indexing set I. Let $x \in \bigcup_{i \in I} U_i$. Then $x \in U_{i_0}$ for some $i_0 \in I$, so by construction of $\mathcal{T}$ there exists $N_x \in \mathcal{N}_x$ such that

$$N_x \subseteq U_{i_0} \subseteq \bigcup_{i \in I} U_i, \text{ and } \bigcup_{i \in I} U_i \in \mathcal{T} \text{ as required.}$$ $\square$

We note that ($\mathcal{N}3$) and ($\mathcal{N}4$) play no role here. But for the next result, which says that the approaches to topological spaces via open sets and via neighbourhoods are equivalent, our neighbourhood systems need to satisfy all of ($\mathcal{N}1$), ($\mathcal{N}2$), ($\mathcal{N}3$) and ($\mathcal{N}4$) for part (b) of the proposition to hold.

Proposition S.9.4 (a) Suppose $(X, \mathcal{T})$ is a topological space defined in terms of open sets, and for each point $x \in X$ let $\mathcal{N}_x$ be the set of neighbourhoods of x associated wih $\mathcal{T}$ as in Definition 9.22. Then the topology constructed in Proposition S.9.3 from $\{\mathcal{N}_x : x \in X\}$ coincides with $\mathcal{T}$.

(b) Suppose that in a non-empty set X there is assigned to each $x \in X$ a non-empty family $\mathcal{N}_x$ of subsets of X satisfying $(\mathcal{N}1)$, $(\mathcal{N}2)$, $(\mathcal{N}3)$ and $(\mathcal{N}4)$, and that we construct from it a topology $\mathcal{T}$ on X as in Proposition S.9.3. Then the set of neighbourhoods of $x \in X$ defined in terms of $\mathcal{T}$ as in Definition 9.22 coincides with $\mathcal{N}_x$.

Proof (a) Suppose that $U \in \mathcal{T}$. Then U is a neighbourhood of each of its points, so it is in the topology constructed in Proposition S.9.3 from the set of neighbourhoods determined by $\mathcal{T}$.

Conversely suppose that U is in the topology constructed as in Proposition S.9.3 from the neighbourhood system associated with $\mathcal{T}$. Then for each $x \in U$ there is some neighbourhood N_x of x with $N_x \subseteq U$. Now by Definition 9.22 there is some open set U_x with $U_x \subseteq N_x$, so $x \in U_x \subseteq U$ and U is in $\mathcal{T}$ by Proposition 7.2.

(b) Suppose first that N'_x is a neighbourhood of $x \in X$ in the topology $\mathcal{T}$ constructed as in Proposition S.9.3 from $\{\mathcal{N}_x : x \in X\}$ where $\mathcal{N}_x$ satisfies $(\mathcal{N}1)$, $(\mathcal{N}2)$, $(\mathcal{N}3)$ and $(\mathcal{N}4)$ for each $x \in X$. Then for some $U \in \mathcal{T}$ we have $x \in U \subseteq N'_x$. But by construction of $\mathcal{T}$ there is some $N_x \in \mathcal{N}_x$ such that $N_x \subseteq U$. Now $N_x \subseteq N'_x$ and since the system $\{\mathcal{N}_x : x \in X\}$ has the property $(\mathcal{N}4)$ it follows that N'_x is in $\mathcal{N}_x$.

Conversely suppose that $N \in \mathcal{N}_x$ for some $x \in X$. We wish to show that N is a neighbourhood of x according to Definition 9.22 applied to the topology $\mathcal{T}$. So we want to see that there is a $V \in \mathcal{T}$ such that $x \in V \subseteq N$. The system $\{\mathcal{N}_x : x \in X\}$ satisfies $(\mathcal{N}3)$, so there is a $V \in \mathcal{N}_x$ such that $V \subseteq N$ and $V \in \mathcal{N}_v$ for every $v \in V$. Now V is in $\mathcal{T}$, since for every $v \in V$ we have $v \in V \in \mathcal{N}_v$. Finally $x \in V$ by $(\mathcal{N}1)$ since $V \in \mathcal{N}_x$. Since $x \in V \subseteq N$ it follows that N is in the neighbourhood system associated with $\mathcal{T}$. □

As mentioned in the book, the concept of neighbourhood is particularly useful in dicussing the topology 'around a point'. For this, we don't really need the whole of $\mathcal{N}_x$.

Definition S.9.5 A *neighborhood base* at a point x in a topological space X is a subfamily $\mathcal{B}_x \subseteq \mathcal{N}_x$ such that for each $N_x \in \mathcal{N}_x$ there exists a set $B_x \in \mathcal{B}_x$ such that $B_x \subseteq N_x$.

Example S.9.6 In any metric space (X, d), the set $\{B_\varepsilon(x) : \varepsilon > 0\}$ is a neighbourhood base at x. The same is true for the set $\{B_q(x) : q \in \mathbb{Q}^+\}$, the open ball neighbourhoods of x with positive rational radii.

We conclude this section by considering how the concepts of closure and interior can be expressed in terms of neighbourhoods. When we are given a neighbourhood base at each point, we refer to the neighbourhoods in it as 'basic neighbourhoods'.

Proposition S.9.7 Let A be a subset of a topological space X in which we have chosen a neighbourhood base at each point.

(a) A point $x \in X$ is in $\overline{A}$ iff every basic neighborhood B_x of x has $B_x \cap A \neq \emptyset$.

(b) A point $x \in X$ is in $\mathring{A}$ iff some basic neighbourhood of x is contained in A.

Proof (a) First suppose that $B_x \cap A \neq \emptyset$ whenever B_x is a basic neighbourhood of x. Let U be any open set containing x. Then U contains a neighbourhood N_x of x, and N_x contains some basic neighbourhood B_x of x by Definition S.9.5. Hence $U \cap A \supseteq B_x \cap A \neq \emptyset$, and x is in $\overline{A}$ by Definition 9.6. Conversely suppose that $x \in \overline{A}$, and let B_x be a basic neighbourhood of x. Then B_x is a neighbourhood of x, so there is some open set U such that $x \in U \subseteq B_x$. Then $B_x \cap A \supseteq U \cap A \neq \emptyset$ as required.

(b) First, if some basic neighbourhood B_x of x is contained in A then since B_x is a neighbourhood of x there is some open set U such that $x \in U \subseteq B_x$, so $x \in U \subseteq A$ and $x \in \mathring{A}$. Conversely suppose that $x \in \mathring{A}$. Then $x \in U \subseteq A$ for some open set U. Then by Proposition S.9.2 there exists a neighbourhood N_x of x with $N_x \subseteq U$. By Definition S.9.5 there is some basic neighbourhood B_x of x with $B_x \subseteq N_x$. So $B_x \subseteq A$ as required. □

If 'neighbourhoods' are taken as the fundamental concept for defining topological spaces, Proposition S.9.7 can be turned around to give definitions of closure and interior. ★

Supplementary material for Chapter 10

Here is a list of supplementary topics for Chapter 10.

Inevitability of the product topology (1) In this section we keep the promise made just before Proposition 10.10.

Proposition S.10.1 Let $(X, \mathcal{T}_X)$ and $(Y, \mathcal{T}_Y)$ be topological spaces. The product topology on $X \times Y$ is the only one such that for all topological spaces Z and maps $f : Z \to X \times Y$, the map f is continuous iff $p_X \circ f$ and $p_Y \circ f$ are continuous, where p_X, p_Y are the projections of $X \times Y$ onto X, Y respectively.

Proof We have already proved in Proposition 10.11 that the product topology $\mathcal{T}_{X \times Y}$ does have this property. Suppose that $\mathcal{T}$ is any topology on $X \times Y$ that has the same property. We shall prove that $\mathcal{T} = \mathcal{T}_{X \times Y}$.

First take $Z = X \times Y$ and f_1 the identity map, and consider f_1 as a map from $(X \times Y, \mathcal{T}_{X \times Y})$ to $(X \times Y, \mathcal{T})$. Then $p_X \circ f_1 = p_X$, $p_Y \circ f_1 = p_Y$, and we know that p_X, p_Y are continuous when regarded as maps from $(X \times Y, \mathcal{T}_{X \times Y})$ to $(X, \mathcal{T}_X)$, $(Y, \mathcal{T}_Y)$. Hence f_1 is continuous, by the assumed property of $\mathcal{T}$. This says that $\mathcal{T} \subseteq \mathcal{T}_{X \times Y}$.

Now take $Z = X \times Y$ and f_2 the identity again, and consider f_2 as a map from $(X \times Y, \mathcal{T})$ to itself. Then f_2 is continuous (by Proposition 8.6(a)), so $p_X = p_X \circ f_2$ and $p_Y = p_Y \circ f_2$ are continuous when thought of as maps from $(X \times Y, \mathcal{T})$ to $(X, \mathcal{T}_X)$, $(Y, \mathcal{T}_Y)$ respectively, by the assumed property of $\mathcal{T}$.

Finally we take the same choice of Z and $f_3 = f_2$, but now consider f_3 as a map from $(X \times Y, \mathcal{T})$ to $(X \times Y, \mathcal{T}_{X \times Y})$. Since we have seen that $p_X = p_X \circ f_3$ and $p_Y = p_Y \circ f_3$ are continuous when thought of as maps from $(X \times Y, \mathcal{T})$ to $(X, \mathcal{T}_X)$, $(Y, \mathcal{T}_Y)$ respectively, it follows from Proposition 10.11 that f_3 is continuous when thought of as a map from $(X \times Y, \mathcal{T})$ to $(X \times Y, \mathcal{T}_{X \times Y})$. This says that $\mathcal{T}_{X \times Y} \subseteq \mathcal{T}$.

We have now proved that $\mathcal{T} = \mathcal{T}_{X \times Y}$. □

★ **Products and weak topologies (2)** The topology on a set X generated by a family $\mathcal{S}$ of subsets of X as in S.8 may also be thought of as the topology having the fewest open sets (called the *weakest* topology) containing $\mathcal{S}$. Equivalently this is the intersection of all topologies on X which contain $\mathcal{S}$. The topology having the fewest open sets compatible with some given property (in the above, the property is 'containing $\mathcal{S}$') is referred to as the *weak topology* with regard to that property.

In functional analysis this is sometimes used in the following context. Suppose we are given a set X and an indexing set I. Suppose also that for each $i \in I$ we are given a topological space $(Y_i, \mathcal{T}_i)$ and a map $f_i : X \to Y_i$. Then the weak topology on X with respect to the family $\{f_i : i \in I\}$ is the weakest topology on X such that all the maps f_i are continuous. This is the same as the topology generated by the family of subsets $\{f_i^{-1}(U_i) : i \in I,\, U_i \in \mathcal{T}_i\}$. Thus in the terminology of S.8, this family is a sub-basis for the weak topology with respect to $\{f_i : i \in I\}$.

Example S.10.2 Let $(Y_1, \mathcal{T}_1)$, $(Y_2, \mathcal{T}_2)$ be topological spaces. Consider the set $Y_1 \times Y_2$. Take as indexing set $I = \{1, 2\}$ and let $\pi_1 : Y_1 \times Y_2 \to Y_1$, $\pi_2 : Y_1 \times Y_2 \to Y_2$ be the projection maps. Then the weak topology on $Y_1 \times Y_2$ with respect to $\{\pi_1, \pi_2\}$ is the product topology. For $\pi_1^{-1}(U) = U \times Y_2$ and $\pi_2^{-1}(V) = Y_1 \times V$ for each open set U in Y_1 and each open set V in Y_2, and we have already seen in Example S.8.7 that the collection of all such $U \times Y_2$ and $Y_1 \times V$ generates the product topology.

Weak topologies with respect to families of functions have a general property which is similar to a property of the product topology.

Proposition S.10.3 Suppose that a topological space X has the weak topology associated with the indexed family of functions $\{f_i : X \to Y_i : i \in I\}$, where $(Y_i, \mathcal{T}_i)$ is a topological space for each $i \in I$. Then a map $g : Z \to X$ from another topological space Z is continuous iff $f_i \circ g$ is continuous for every $i \in I$.

Proof If g is continuous so is every $f_i \circ g$, since the weak topology on X makes f_i continuous. Suppose that $f_i \circ g$ is continuous for every $i \in I$. Then for every open set in the sub-basis $\{f_i^{-1}(U) : i \in I,\, U \in \mathcal{T}_i\}$, the set $g^{-1}(f_i^{-1}(U)) = (f_i \circ g)^{-1}(U)$ is open in Z by continuity of $f_i \circ g$. Hence by Proposition S.8.8, g is continuous. □ ★

Hints for Exercise 10.20 First suppose that the topology on X is discrete. Apply Exercise 10.11 to see that the product topology on $X \times X$ is also discrete, so any subset of $X \times X$, in particular Δ, is open in $X \times X$.

For the converse, suppose that Δ is open in $X \times X$. Let $x \in X$ and apply the criterion Proposition 10.20 to the point $(x, x) \in \Delta$ to see that there are open subsets U, V of X such that $(x, x) \in U \times V \subseteq \Delta$. Now deduce that $U = V = \{x\}$ so $\{x\}$ is open in X. Since this is true for every $x \in X$ the topology on X is discrete.

Supplementary material for Chapter 11

Here is a list of supplementary topics for Chapter 11.

Convergence in the co-finite topology (2) The problems arising with convergence in a non-Hausdorff space are not confined to indiscrete topologies. Suppose that X is an infinite set with the co-finite topology, and let $x \in X$. Suppose (x_n) is a sequence in X in which no point is repeated infinitely often. Then (x_n) must converge to x: for any open set $U \ni x$ is non-empty so its complement is finite, say $X \setminus U = \{y_1, y_2, \ldots y_m\}$. For each $i = 1, 2, \ldots, m$ the point y_i does not appear as an entry in the sequence (x_n) infinitely often, so there is an integer N_i such that $x_n \neq y_i$ for $n \geqslant N_i$. Then for $n \geqslant \max\{N_1, N_2, \ldots, N_m\}$ the entry x_n in the sequence does not equal any of $y_1, y_2, \ldots, y_m$ so x_n must be in U. Hence (x_n) converges to x.

Sub-Hausdorff conditions (3) There are conditions which may hold for a given topology which are weaker than the Hausdorff condition but are still of interest.

Definition S.11.1 A topological space X is said to be T_0 if given any two distinct points in X, there exists an open set which contains one and not the other.

Example S.11.2 The topology of Exercise 7.4 on $\mathbb{N}$, in which the open sets are $\emptyset$, $\mathbb{N}$ and $\{1, 2, \ldots, n\}$ for each $n \in \mathbb{N}$, gives a T_0 space; for if $m, n \in \mathbb{N}$ are not equal, then either $m > n$ or $n > m$. In the first case, $\{1, 2, \ldots, n\}$ is an open set containing n but not m, and in the second case $\{1, 2, \ldots, m\}$ is an open set containing m but not n. Notice however that in the first case there is no open set containing m but not n, and in the second case there is no open set containing n but not m. There is a lack of symmetry here.

Definition S.11.3 A topological space X is said to be T_1 if given any two distinct points $x, y \in X$ there is an open set which contains x but not y and also an open set which contains y but not x.

Example S.11.4 An infinite set X with the co-finite topology is a T_1 space: for given distinct points $x, y \in X$, the set $X \setminus \{y\}$ is an open set containing x but not y, while $X \setminus \{x\}$ is an open set containing y but not x.

Clearly Hausdorff implies T_1 and T_1 implies T_0. In this hierarchy, another name for Hausdorff is T_2. Notice that the examples are chosen to show that a T_0 space need not be T_1 and a T_1 space need not be T_2. An indiscrete space with more than one point in it is not even T_0. This hierarchy extends further, but we do not pursue it.

★ **Urysohn's lemma (3)** Recall from Exercise 11.9 that given two disjoint closed sets A, B in any metric space X, there is a continuous real-valued function g on X such that $g(x) = 0$ for any $x \in A$ and $g(x) = 1$ for any $x \in B$. Urysohn's lemma gives a similar result when X is a normal space (and without a metric it is harder to see where a real-valued function is coming from).

As a preliminary, we describe a property equivalent to normality.

Proposition S.11.5 A topological space X is normal iff whenever $V \subseteq U \subseteq X$ with V closed in X and U open in X there exists an open subset W of X such that $V \subseteq W$ and $\overline{W} \subseteq U$.

Proof Suppose first that X is normal and that V, U are respectively closed, open subsets of X with $V \subseteq U$. Consider the sets V, $X \setminus U$. These are closed in X and disjoint since $V \subseteq U$. By normality, there are disjoint sets W, Y both open in X and such that $V \subseteq W$, $X \setminus U \subseteq Y$. Then $X \setminus Y$ is closed and contains W (since $W \cap Y = \emptyset$) hence $\overline{W} \subseteq X \setminus Y$ by Proposition 9.10(f). Hence $\overline{W} \subseteq X \setminus Y \subseteq X \setminus (X \setminus U) = U$ as required.

Conversely suppose that X is a topological space and that whenever $V \subseteq U \subseteq X$ with V, U respectively closed, open in X there is an open subset W of X with $V \subseteq W$ and $\overline{W} \subseteq U$. We shall prove that X is normal. So let V, Y be disjoint closed subsets of X. Then $V \subseteq X \setminus Y$ and V is closed, $X \setminus Y$ is open, in X. So there exists an open subset W of X with $V \subseteq W$ and $\overline{W} \subseteq X \setminus Y$. Then $X \setminus \overline{W}$ is open in X and $Y \subseteq X \setminus \overline{W}$. Moreover, W and $X \setminus \overline{W}$ are disjoint since $W \subseteq \overline{W}$. So X is normal. □

In the proof of Urysohn's lemma we are going to use the set D of rational numbers r in $(0, 1)$ of the form $i/2^n$ for positive integers i and n (so for $1 \leqslant i \leqslant 2^n - 1$). These have denseness properties similar to those of $\mathbb{Q}$. In particular for real numbers y, z with $0 \leqslant y < z \leqslant 1$ there is an $r \in D$ such that $y < r < z$. For choose n large enough that $1/2^n < z - y$. Then if i is the least integer such that $i/2^n > y$ we have also $i/2^n < z$, for $(i-1)/2^n \leqslant y$ so if $i/2^n \geqslant z$ then $1/2^n \geqslant z - y$, contradicting the choice of n. Hence $y < i/2^n < z$. (The set D is called 'the dyadic rationals in $(0, 1)$'.)

We are now ready for Urysohn's lemma. (Its proof is harder than those of some results we have called 'theorems'.)

Urysohn's lemma A space X is normal iff whenever A, B are disjoint closed subsets of X there is a continuous function $f : X \to [0,\, 1]$ such that $f(A) = 0$ and $f(B) = 1$.

Proof One way around is easy: if for any pair of disjoint closed subsets A, B such a function f exists then $f^{-1}([0,\, 1/2)$ and $f^{-1}(1/2,\, 1]$ are disjoint open sets containing A, B, so X is normal.

Conversely suppose that X is normal and that A, B are disjoint closed subsets of X. We are going to get a family of open sets U_r, one for each $r \in D$, such that $A \subset U_r$, $U_r \cap B = \emptyset$ for all $r \in D$ and $\overline{U}_r \subseteq U_s$ whenever $r,\, s \in D$ with $r < s$. This will help define a suitable function f.

By Proposition S.11.5 there is an open set W such that $A \subseteq W$ and $\overline{W} \subseteq X \setminus B$, so $\overline{W} \cap B = \emptyset$. Let us give W the new name $U_{1/2}$. Applying Proposition S.11.5 to the pairs $A, U_{1/2}$ and $\overline{U}_{1/2},\, X \setminus B$, we get open sets which we call $U_{1/4},\, U_{3/4}$ such that

$$A \subseteq U_{1/4}, \quad \overline{U}_{1/4} \subseteq U_{1/2}, \quad \overline{U}_{1/2} \subseteq U_{3/4}, \quad \overline{U}_{3/4} \cap B = \emptyset, \text{ so } \overline{U}_{3/4} \subseteq X \setminus B.$$

Suppose inductively that for some integer $n \geqslant 1$ we have already got open subsets $U_{i/2^n}$ for every $i \in \{1,\, 2,\, \ldots,\, 2^n - 1\}$ such that

$$A \subseteq U_{1/2^n}, \quad \ldots \quad, \overline{U}_{i/2^n} \subseteq U_{(i+1)/2^n}, \quad \overline{U}_{(i+1)/2^n} \subseteq U_{(i+2)/2^n}, \quad \ldots \quad, \overline{U}_{(2^n-1)/2^n} \subseteq X \setminus B.$$

Then applying Proposition S.11.5 to each pair $A, U_{1/2^n}, \ldots \overline{U}_{(2^n-1)/2^n}, X \setminus B$, we can continue to get the next stage of the induction. This gives an open set U_r for each $r \in D$, such that we have $A \subseteq U_r$ and $\overline{U}_r \subseteq X \setminus B$, so $\overline{U}_r \cap B = \emptyset$ for every $r \in D$, and $\overline{U}_r \subseteq U_s$ whenever $r,\, s \in D$ with $r < s$.

We now define $f : X \to [0,\, 1]$ by

$$f(x) = \begin{cases} 1 & \text{if } x \text{ does not belong to any of the above } U_r, \\ \inf\{t \in D : x \in U_t\} & \text{otherwise.} \end{cases}$$

Notice that $f(x)$ does not have to be in D - any number in $[0,\, 1]$ is of the form $\inf S$ for some $S \subseteq D$.

We now show that f has the required properties. First note

(a) $x \in \overline{U}_r$ implies $f(x) \leqslant r$ (since then $x \in U_t$ for all $t > r$, so $\inf\{t \in D : x \in U_t\} \leqslant r$). The contrapositive follows: if $f(x) > r$ then $x \notin \overline{U}_r$.

(b) $x \notin U_s$ implies $f(x) \geqslant s$ (since then $x \notin U_t$ for any $t < s$). The contrapositive follows: if $f(x) < s$ then $x \in U_s$.

Note $f(B) = 1$ since if $x \in B$ then $x \notin U_r$ for any $r \in D$ so by definition $f(x) = 1$.

Secondly, $f(A) = 0$ since if $x \in A$ then $x \in U_r$ for all $r \in D$, and $\inf D = 0$.

It remains to show that f is continuous on X. We prove continuity first at a point x_0 such that $f(x_0) \in (0, 1)$. Let $\varepsilon > 0$ (where we may choose ε so that $f(x_0)-\varepsilon,\ f(x_0)+\varepsilon \in (0, 1)$). Choose $r, s \in D$ such that $f(x_0) - \varepsilon < r < f(x_0) < s < f(x_0) + \varepsilon$. Consider the open set $U = U_s \setminus \overline{U}_r$. For any $x \in U$ we have $x \in U_s \subseteq \overline{U}_s$ so by (a), $f(x) \leqslant s < f(x_0) + \varepsilon$, and also $f(x) \notin \overline{U}_r$ so $f(x) \notin U_r$ and by (b) $f(x) \geqslant r > f(x_0) - \varepsilon$. This shows that $f(U) \subseteq (f(x_0) - \varepsilon,\ f(x_0) + \varepsilon)$. Moreover $x_0 \in U$ since $f(x_0) > r$ so $x_0 \notin \overline{U}_r$ by contrapositive (a), and $f(x_0) < s$ so $x_0 \in U_s$ by contrapositive (b). This proves continuity of f at x.

Next we prove continuity of f at a point x_0 such that $f(x_0) = 0$. Let $\varepsilon > 0$ and choose $r \in D$ such that $r \in (0, \varepsilon)$. Consider the open set U_r. Then $f(x_0) < r$ so $x_0 \in U_r$ by contrapositive (b). Also, if $x \in U_r$ then $x \in \overline{U}_r$ so $f(x) \leqslant r < \varepsilon$ by (a); hence $0 \leqslant f(x) < \varepsilon$. This shows that $x_0 \in U_r$ and $f(U_r) \subseteq [0, \varepsilon)$. So f is continuous at x_0.

Finally we prove that f is continuous at any point x_0 with $f(x_0) = 1$. Let $\varepsilon > 0$ and choose $r \in D$ such that $1-\varepsilon < r < 1$. Let $U = X\setminus\overline{U_r}$. Then $f(x_0) > r$ so $x_0 \notin \overline{U}_r$ by contrapositive (a). Hence $x_0 \in U$. Also, for any $x \in U$ we have $x \notin \overline{U}_r$ hence $x \notin U_r$ and by (b) $f(x) \geqslant r > 1 - \varepsilon$. Thus $x_0 \in U$ and $f(U) \subseteq (1 - \varepsilon, 1]$. This shows that f is continuous at x_0, and completes the proof of Urysohn's lemma. □

Remarks (1) It is tempting to think that there might be a result for regular spaces (in which a closed set A and a point b not in A can be separated by disjoint open sets) similar to Urysohn's lemma. But thinking about an analogous proof shows that it will not work for a regular space - we could choose $U_{1/2}$ such that $A \subset U_{1/2}$ and $b \notin U_{1/2}$, but at the next stage, choosing $U_{1/4}$ such that $A \subseteq U_{1/4}$ and $\overline{U}_{1/4} \subseteq U_{1/2}$ we need the full force of normality - regularity is not enough.

(2) In general there is no guarantee that $f^{-1}(0) = A$, $f^{-1}(1) = B$ - there may for example be points $x \notin A$ such that $f(x) = 0$. For example if $X = [0, 1]$ and $A = \{0\}$, , $B = \{1\}$ a Urysohn function is given by $f(x) = 0$ for $x \in [0, 1/2]$, and $f(x) = 2x - 1$ for $x \in [1/2, 1]$.

(3) A Urysohn function may well take on every value between 0 and 1 as in the previous example. But this is not necessarily the case in general; for example we might have $X = \{0, 1\}$ with the discrete topology, $A = \{0\}$, $B = \{1\}$ and a Urysohn function $f : X \to [0, 1]$ takes on the values 0, 1 only.

For the next proof it is convenient to have a simple modification of Urysohn's lemma.

Corollary S.11.6 If A and B are disjoint closed subsets of a normal space X and c, d are real numbers with $c < d$ then there is a continuous function $f : X \to [c, d]$ with $f(A) = c$, $f(B) = d$.

Proof Let $f = g \circ h$ where $h : X \to [0, 1]$ is a continuous function given by Urysohn's lemma with $h(A) = 0$, $h(B) = 1$, and $g : [0, 1] \to [c, d]$ is the continuous function defined by $g(x) = c + x(d - c)$ for all $x \in [0, 1]$. Then $f(A) = g(h(A)) = g(0) = c$ and likewise $f(B) = d$. □ ★

★ **Tietze's extension theorem (3)** Consider any continuous function $f : [a, b] \to \mathbb{R}$. This may be extended to a continuous function g defined on all of $\mathbb{R}$ by defining

$$g(x) = \begin{cases} f(x) & \text{if } x \in [a, b] \\ f(a) & \text{if } x \leqslant a \\ f(b) & \text{if } x \geqslant b \end{cases}$$

Tietze's theorem generalises this result to any closed subspace of a normal space.

Tietze's extension theorem Given any closed subset A of a normal space X and a continuous function $f : A \to \mathbb{R}$, there is a continuous map $g : X \to \mathbb{R}$ such that $g|A = f$. We call such a function g 'a continuous extension of f to X'.

In fact normality may be characterized by this extension property. For if A, B are disjoint closed subsets of a space X and $f : A \cup B \to \mathbb{R}$ is defined by $f(x) = 0$ when $x \in A$ and $f(x) = 1$ when $x \in B$, then an extension of f to a continuous map $g : X \to \mathbb{R}$ shows that X is normal, by the 'easy' part of Urysohn's lemma.

Proof of Tietze's extension theorem This uses Urysohn's lemma. It is easier first to prove the analogous result for maps into $[-1, 1]$. So suppose $f : A \to [-1, 1]$ is continuous. The idea of the proof is to construct a uniformly convergent sequence (see Chapter 16) of continuous functions (h_n) where each $h_n : X \to \mathbb{R}$, and as n increases $h_n|A$ is a progressively better approximation to f. The limit h of (h_n) is continuous by uniform convergence, and $h|A = f$.

Let $B = f^{-1}([-1, -1/3])$, $C = f^{-1}([1/3, 1])$. Then B and C are closed subsets of A and hence of X since A is closed in X. By Corollary S.11.6 there is a continuous function g_1 from X to $[-1/3, 1/3]$ such that

$$g_1(x) = \begin{cases} -1/3 & \text{for } x \in B \\ 1/3 & \text{for } x \in C \end{cases}$$

Now $|f(a) - g_1(a)| \leqslant 2/3$ for all $a \in A$. (If $a \in B$ then $f(a) \in [-1, -1/3]$ while $g_1(a) = -1/3$, if $a \in C$ then $f(a) \in [1/3, 1]$ while $g_1(a) = 1/3$, and finally if $a \in A \setminus (B \cup C)$ then both $f(a)$ and $g_1(a)$ are in $[-1/3, 1/3]$.) We set $h_1 = g_1$.

Now consider $f_1 = f - g_1 : A \to [-2/3, 2/3]$. We repeat the above process: this means we divide $[-2/3, 2/3]$ into three equal subintervals; let $B_2 = f_1^{-1}([-2/3, -2/9])$, $C_2 = f_1^{-1}([2/9, 2/3])$. By

Corollary S.11.6 there exists a continuous function $g_2 : X \to [-2/9,\, 2/9]$ such that $g_2(x) = -2/9$ for $x \in B_2$ and $g_2(x) = 2/9$ for $x \in C_2$. Then $|f(a) - (g_1(a) + g_2(a))| = |f_1(a) - g_2(a)| \leqslant (2/3)^2$ for all $a \in A$. (For when $a \in B_2$ we have $f_1(a) \in [-2/3, -2/9]$ and then $g_2(a) = -2/9$, so $|f_1(a) - g_2(a)| \leqslant 2/3 - 2/9 = (2/3)^2$; when $a \in C_2$ similarly $|f_1(a) - g_2(a)| \leqslant (2/3)^2$, and when $a \in A \backslash (B_2 \cup C_2)$, both $f_1(a)$ and $g_2(a)$ are in $[-2/9,\, 2/9]$, hence $|f_1(a) - g_2(a)| \leqslant 4/9 = (2/3)^2$.)

Inductively suppose that for each $i \in \{1, 2, \ldots, n\}$ we have defined a continuous function $g_i : X \to [-2/3^i,\, 2/3^i]$ such that $|f(a) - (g_1(a) + g_2(a) + \ldots + g_i(a)| \leqslant (2/3)^i$ for all $a \in A$, and disjoint closed subsets B_i, C_i of A such that $g_i(B_i) = -2/3^i$ and $g_i(C_i) = 2/3^i$. We define $f_n : A \to [-2/3^n,\, 2/3^n]$ by $f_n(a) = f(a) - (g_1(a) + g_2(a) + \ldots + g_n(a))$ for all $a \in A$. Again we divide $[-2/3^n,\, 2/3^n]$ into three equal subintervals. Let $B_{n+1} = f_n^{-1}([-2/3^n,\, -2/3^{n+1})$ and $C_{n+1} = f_n^{-1}([2/3^{n+1},\, 2/3^n])$. Then B_{n+1} and C_{n+1} are disjoint closed subsets of X, and by Corollary S.11.6 there exists a continuous function $g_{n+1} : X \to [-2/3^{n+1},\, 2/3^{n+1}]$ such that $g_{n+1}(B_{n+1}) = -2/3^{n+1}$, $g_{n+1}(C_{n+1}) = 2/3^{n+1}$. Then just as before we may prove that

$$|f(a) - g_1(a) - g_2(a) - \ldots - g_{n+1}(a)| = |f_n(a) - g_{n+1}(a)| \leqslant (2/3)^{n+1} \quad \text{for all } a \in A.$$

This completes the inductive step in the construction.

For each $n \in \mathbb{N}$ let $h_n = g_1 + g_2 + \ldots + g_n$. The sequence (h_n) converges uniformly on X, say to the function h, by Weierstrass M-test (see Chapter 16) since $|g_n(x)| \leqslant 2/3^n$ for all $x \in X$, and $\sum 1/3^n$ converges. Since each g_n is continuous, so is h by Theorem 16.10. Also, $h(a) = f(a)$ for all $a \in A$, since for any $a \in A$ and any $n \in \mathbb{N}$, $|f(a) - h_n(a)| \leqslant (2/3)^n$, so in the limit $f(a) = h(a)$. Now $h : X \to [-1,\, 1]$ is a continuous extension of f to X.

General case Since $\mathbb{R}$ is homeomorphic to $(-1,\, 1)$ we may suppose here that $f : A \to (-1,\, 1)$. We know from the above case that there is a continuous extension $h : X \to [-1,\, 1]$. We have to see that h may be replaced by a map into $(-1,\, 1)$. Define $D = h^{-1}(-1) \cup h^{-1}(1)$. Then D is a closed subset of X, and $D \cap A = \emptyset$ since h, which coincides with f on A, maps A into $(-1,\, 1)$. So by Urysohn's lemma there exists a continuous function $k : X \to [0,\, 1]$ such that $k(D) = 0$, $k(A) = 1$. Let $m : X \to (-1,\, 1)$ be the product $m(x) = h(x)k(x)$. Then m is continuous since h, k are. Also, for any $a \in A$ we have $k(a) = 1$ so $m(a) = h(a) = f(a)$. Thus m is a continuous extension of f. Finally, m maps all of X into $(-1,\, 1)$ since if $x \in D$ then $m(x) = 0$, while if $x \notin D$ then $h(x) \neq \pm 1$, so $|m(x)| = |h(x)k(x)| < |k(x)| \leqslant 1$, and $m(x) \in (-1,\, 1)$. □ ★

Hints for Exercise 11.9 Since f_A and f_B are continuous, so is g. Note that $f_A(a) = 0$ for any $a \in A$. Also $f_B(a) > 0$ for $a \in A$ since $A \cap B = \emptyset$ and $f_B(x) = 0$ iff $x \in B$. This gives $g(a) < 0$ for all $a \in A$, or equivalently $A \subseteq g^{-1}(-\infty,\, 0)$.

Supplementary material for Chapter 12

Here is a list of supplementary topics for Chapter 12.

Path-connectedness is a topological property (2) This follows in a standard fashion from Exercise 12.13, which says that if $f : X \to Y$ is a continuous map of topological spaces which is onto and X is path-connected then so is Y. If $f : X \to Y$ is a homeomorphism, it is certainly continuous and onto, so if X is path-connected then so is Y. But since f is a homeomorphism $f^{-1} : Y \to X$ is also continuous and onto, so if Y is path-connected then so is X.

More results on path-connectedness (2) Results analogous to 12.16, 12.17, 12.18 hold for path-connected spaces (these should really be exercises for Chapter 12). The *reverse* of a continuous path $f : [0, 1] \to X$ in a space X is the path $g : [0, 1] \to X$ defined by $g(t) = f(1-t)$. It is continuous since f is continuous and so is the map $t \mapsto 1 - t$ of $[0, 1]$.

Proposition S.12.1 Suppose that $\{A_i : i \in I\}$ is an indexed family of path-connected subsets of a topological space X with $A_i \cap A_j \neq \emptyset$ for each pair $i, j \in I$. Then $\bigcup_{i \in I} A_i$ is path-connected.

Proof Let $a_0 \in A_{i_0}$ for some particular $i_0 \in I$. It is enough to show that any point x in $\bigcup_{i \in I} A_i$ may be joined to a_0 by a continuous path in $\bigcup_{i \in I} A_i$, for then to join points $x, y \in \bigcup_{i \in I} A_i$ by a continuous path we juxtapose a continuous path from x to a_0 with the reverse of a continuous path from y to a_0, using Lemma 12.24. Now suppose $x \in A_i$. Since $A_i \cap A_{i_0} \neq \emptyset$ there exists a point $y \in A_i \cap A_{i_0}$. Since A_i, A_{i_0} are path-connected, there exist continuous paths from x to y and from y to a_0. The juxtaposition of these is a continuous path from x to a_0 by Lemma 12.24. □

Corollary S.12.2 Suppose that $\{C_i : i \in I\}$ and B are path-connected subsets of a space X such that for every $i \in I$ we have $C_i \cap B \neq \emptyset$. Then $B \cup \bigcup_{i \in I} C_i$ is path-connected.

This is deduced from Proposition S.12.1 as Corollary 12.17 is deduced from Proposition 12.16. □

Proposition S.12.3 The topological product $X \times Y$ of spaces X, Y is path-connected iff both X and Y are path-connected.

This follows using Corollary S.12.2 exactly as Theorem 12.18 follows using Corollary 12.17. $\square$

Components (3) Intuitively, these are the connected pieces that a space falls into. For example, if $X = [0, 1] \cup [2, 3] \subseteq \mathbb{R}$ then its components are $[0, 1]$ and $[2, 3]$. There are two versions, as for connectedness: components and path-components.

Definition S.12.4 A *component* of a space X is a maximal connected subset of X.

A maximal connected subset of X means a subset of X which is connected and is not properly contained in any connected subset of X.

Proposition S.12.5 The components of a space X partition X.

Proof Any two distinct components are disjoint for if distinct components C_1, C_2 satisfies $C_1 \cap C_2 \neq \emptyset$ then $C_1 \cup C_2$ would be connected by Proposition 12.16, so $C_1 \cup C_2 \subseteq C_1$ by maximality of C_1 as a connected set, so $C_2 \subseteq C_1$, and similarly $C_1 \subseteq C_2$, so $C_1 = C_2$, contradicting the choice of C_1 and C_2. Also, any point in X belongs to a maximal connected subset of X, namely the union C_x of all connected subsets of X containing x; for this is connected again by Proposition 12.16, and it is clearly maximal connected since if $C \supseteq C_x$ with C connected then $C \subseteq C_x$ since C is a connected set containing x. $\square$

Here is an alternative approach which is also popular.

Proposition S.12.6 The following defines an equivalence relation on a space X: $x \sim y$ iff there is a connected subset C of X with $x, y \in C$. The corresponding equivalence classes are the components of X.

Proof First we check that this is an equivalence relation. Reflexivity follows since any singleton set $\{x\}$ is connected. Symmetry follows from the definition of $\sim$, and transitivity follows from Proposition 12.16: for if $x \sim y$ and $y \sim z$ then there is a connected subset C of X with $x, y \in C$ and a connected subset D of X with $y, z \in D$. Then $y \in C \cap D$ so this intersection is non-empty, and hence $C \cup D$ is connected by Proposition 12.16. Since $x, z \in C \cup D$ this shows that $x \sim z$.

We now check that the equivalence classes are maximal connected sets. First, any equivalence class E is connected since we may choose a fixed $x_0 \in E$, and then for any $y \in E$ there is a connected subset C_y of X with $x_0, y \in C_y$. Then $C_y \subseteq E$ since for any $z \in C_y$ the set C_y is a connected subset of X containing x_0, z. Hence we can see that $E = \bigcup_{y \in E} C_y$: for any $y \in E$ is in C_y, while we have just seen that $C_y \subseteq E$ for each $y \in E$. Moreover, E is maximal connected, since if $E \subseteq F$ and F is a connected subset of X then we must have $F = E$ since if $y \in F \setminus E$

then F is a connected set containing x_0, y so $y \sim x_0$ and $y \in E$. Moreover every component occurs as an equivalence class by Proposition S.12.5. This completes the proof of Proposition S.12.6, which clearly provides an approach to components equivalent to Definition S.12.4. □

Example S.12.7 (a) As already mentioned, if X is the subspace of $\mathbb{R}$ consisting of $[0, 1] \cup [2, 3]$ then its components are $[0, 1]$ and $[2, 3]$. For these are connected sets by Theorem 12.10 and by Theorem 12.8 there is no connected subset of $\mathbb{R}$ containing points from both intervals.

(b) If $X = \mathbb{Q}$ with its usual topology, then the components are all singleton sets, since if r_1, r_2 are distinct rational numbers then there is no connected subset C of $\mathbb{Q}$ containing both r_1 and r_2 (for if x is any irrational number between r_1 and r_2 then a partition of C is given by $\{C \cap (-\infty, x), C \cap (x, \infty)\}$).

Proposition S.12.8 The components of a space X are closed in X.

Proof If C is a component of X, then C is connected hence $\overline{C}$ is also connected by Proposition 12.19. But a component is a maximal connected subset, so $\overline{C} = C$, which shows that C is closed in X. □

Example S.12.9 In Example S.12.7 (b) the components are not open in X.

Next we begin to see how components may be used in studying homeomorphisms.

Proposition S.12.10 A homeomorphism $f : X \to Y$ of topological spaces induces a bijection f_* of components where if C_x is the component containing a point $x \in X$ then $f_*(C_x) = C_{f(x)}$, the component of Y containing $f(x)$.

Proof We know that $f(C_x)$ is a connected subset of Y by Proposition 12.11. It is a maximal connected subset of Y: for suppose that $f(C_x) \subseteq B$ where B is a connected subset of Y. Then $C_x = f^{-1}(f(C_x)) \subseteq f^{-1}(B)$. Now the map f^{-1} is continuous so $f^{-1}(B)$ is connected by Proposition 12.11. Since C_x is maximal connected we have $C_x = f^{-1}(B)$, so $f(C_x) = B$, and $f(C_x)$ is also maximal connected hence a component. Thus f maps each component of X bijectively onto a component of Y. Since f is a bijection this induces a one-one correspondence f_* from the components of X to the components of Y as claimed. □

In particular if a space X has a finite number of components then this number is a topological invariant - it is the same for all spaces homeomorphic to X. We exploit this in the section giving a sample topological classification.

General open sets in $\mathbb{R}$ again (2) As promised in S.5 we now discuss general open sets in $\mathbb{R}$ in language which gives a smoother treatment.

Let $U \subseteq \mathbb{R}$ be an open subset of $\mathbb{R}$. Consider a component C of U. Since this is a connected set, it must be an interval by Theorem 12.8. It cannot be of the form $(-\infty, b]$, $(a, b]$, $[a, b)$, $[a, b]$ or $[a, \infty)$. For example if $C = [a, b)$, $[a, b]$ or $[a, \infty)$ then $a \in C \subseteq U$, so since U is open in

$\mathbb{R}$, for some $\varepsilon > 0$ we have $(a - \varepsilon,\, a + \varepsilon) \subseteq U$. Then $C \cup (a - \varepsilon,\, a + \varepsilon)$ is a connected subset of U which is strictly larger than C, contradicting the maximality of C in Definition S.12.4. A similar argument works for intervals of type $(-\infty,\, b]$ and $(a,\, b]$. Now the above list includes all the intervals which are not open, so C must be an open interval. The components of U are mutually disjoint sets, and the number of them is countable as before - for choosing a rational number in each component gives an injection from the set of components to $\mathbb{Q}$, which shows that the set of components is countable by Corollary S.2.7.

We have now proved that if a subset of $\mathbb{R}$ is open in $\mathbb{R}$ then it is a countable union of disjoint open intervals. The converse is as before: an open interval is an open set, hence so is any (disjoint) union of open intervals.

An example of topological classification (2) In this section we are going to use components to help illustrate a topological classification problem. The set of spaces to be classified are 'the letters of the alphabet': that is, certain subspaces of $\mathbb{R}^2$ which represent the letters of the alphabet. In order to be precise, we shall deal with capital letters, and suppose that a 'sans serif' font is used, which means for example that we use A rather than A. We shall use the letters (in their sans-serif form) as names for the subspaces of $\mathbb{R}^2$ which they are. This example was in the first edition of the book, as an exercise which experience shows was rather confusing. Our discussion of it here is semi-rigorous, in the sense that we shall rely on geometric intuition to see that for example the subspaces which are the letters I and C are homeomorphic, rather than describing these letters by formulae and constructing a specific homeomorphism. Our proof that letters on distinct lines below are not homeomorphic is incomplete, since we do not list all cases, but it will be rigorous, using components and Exercise 10.10, which we recall says that if $f : X \to Y$ is a homeomorphism of spaces and $A \subseteq X$ then $f|(X \setminus A) : X \setminus A \to Y \setminus f(A)$ is also a homeomorphism. We use also the remark at the end of the previous section, that if a space has a finite number of components then any homeomorphic space has this same number of components.

The answer we are aiming for is that the homeomorphism classes are:

A, R

B

C, G, I, J, L, M, N, S, U, V, W, Z

D, O

E, F, T, Y

H, K

P

Q

You will quickly spot that 'it's the number of loops and of junctions' that determines which homeomorphism class a letter is in. As mentioned, we rely on geometric intuition to see that the letters on any line are homeomorphic to each other.

You need to look closely at K to see that it should be on the same line as H rather than X.

Next here is a typical argument showing that spaces on distinct lines are not homeomorphic to each other: suppose that $f : \mathsf{T} \to \mathsf{I}$ were a homeomorphism. Let x be the junction point of the letter T and suppose that $f(x) = y$. We see that $\mathsf{T}\backslash\{x\}$ has precisely three components.

To be precise, think of the letter T as the subspace of $\mathbb{R}^2$ given by $(\{0\} \times [0, 1]) \cup ([-1, 1] \times \{1\})$, so that x is the point with coordinates $(0, 1)$, and $\mathsf{T}\backslash\{x\}$ consists of the disjoint union of three subspaces, $\{0\} \times [0, 1)$, $[-1, 0) \times \{1\}$, $(0, 1] \times \{1\}$, each of which is homeomorphic to a half-open interval. But no matter where y is in I, the space $\mathsf{I}\backslash\{y\}$ is either the union of two disjoint half-open intervals, or (if y is at either end of I) one half-open interval. Hence T and I are not homeomorphic.

Since there are nine topological equivalence classes listed above, we should really give thirty-six $(= 8.9/2)$ such arguments to show that no two letters on distinct lines are homeomorphic. We restrict to giving a few more examples.

If $f : \mathsf{X} \to \mathsf{P}$ were a homeomorphism, then letting x be the junction of X we see that $\mathsf{X}\backslash\{x\}$ consists of four half-open intervals, but wherever $f(x)$ is in P its complement has at most two components.

Next suppose that $f : \mathsf{B} \to \mathsf{O}$ were a homeomorphism. We remove both junction points $\{x, y\}$ from B, leaving three components. But wherever $f(x)$, $f(y)$ are in O the complement has just two components.

As a final example, suppose that $f : \mathsf{H} \to \mathsf{T}$ were a homeomorphism. If we remove both junction points $\{x, y\}$ from H the complement has five components, whereas no matter where $f(x)$ and $f(y)$ are in T their complement has at most four components.

★ **Connected spaces which are not path-connected (1)** In Proposition 12.25 we saw that any open connected subset of Euclidean space is path-connected. The traditional example to show that connected spaces are not in general path-connected is the 'topologist's sine curve' mentioned in the book, and we include it below. We begin with a slightly different example.

Example S.12.12 Let $X \subseteq \mathbb{R}^2$ be the following subset of the Euclidean plane:

$$X = C \cup \{x_0\}, \quad \text{where } C = ((0,\, 1] \times \{0\}) \cup \bigcup_{n=1}^{\infty} \{1/n\} \times [0,\, 1] \text{ and } x_0 \text{ has coordinates } (0,\, 1).$$

Thus X is a 'comb' C together with a single point x_0 at height 1 on the y-axis. Then it is easy to see that C is path-connected hence connected (any point on a 'tooth' $\{1/n\} \times [0,\, 1]$ of the comb is connected by a path in the tooth to the point $(1/n,\, 0)$, and all these points lie in the connected interval $(0,\, 1] \times \{0\} \subseteq C$). It is also easy to see that the point x_0 is in $\overline{C}$, since any open ball $B_\varepsilon(x_0)$ in the plane centred on $x_0 = (0,\, 1)$ contains the points $(1/n,\, 1) \in C$ whenever $1/n < \varepsilon$. So X is connected by Proposition 12.19.

However, X is not path-connected. We offer two proofs of this.

The first uses uniform continuity. Suppose $f : [0,\, 1] \to X$ is a continuous path with $f(0) = x_0$. We shall prove that $f(t) = x_0$ for all $t \in [0,\, 1]$, so that x_0 cannot be connected by a path to any other point in X. Since f is continuous on the compact space $[0,\, 1]$ it is uniformly continuous there by Proposition 13.24. So there exists $\delta > 0$ such that the euclidean distance $d_2(f(t),\, f(t')) < 1$ whenever $t,\, t' \in [0,\, 1]$ and $|t - t'| < \delta$. Choose points $t_1,\ t_2,\ \ldots,\ t_n$ such that $0 = t_1 < t_2 < \ldots < t_n = 1$ and $t_i - t_{i-1} < \delta$ for each $i = 2,\, 3,\, \ldots,\, n$. Suppose inductively that $f(t) = x_0$ for all $t \in [0,\, t_i]$ (this is true for $i = 1$). Suppose for a contradiction that there is some $t \in [t_i,\, t_{i+1}]$ such that $f(t) \neq x_0$. Let p_1 denote the projection of $\mathbb{R}^2$ onto the x-axis, and let $i : X \to \mathbb{R}^2$ be the inclusion. Then $p_1(i(f(t))) > 0$. Let x be an irrational number between 0 and $p_1(i(f(t)))$. Then by the intermediate value theorem applied to $p_1 \circ i \circ f$ on $[t_i,\, t]$, we have $p_1(i(f(s))) = x$ for some $s \in [t_i,\, t]$. This implies that $f(s) = (x,\, z)$ for some $z \in [0,\, 1]$. But the distance between the points $x_0 = f(t_i)$ and $f(s) = (x,\, z)$ is less than 1, so $z > 0$, and we see that $(x,\, z) \notin X$. This contradiction shows that $f([t_i,\, t_{i+1}]) = x_0$. This completes the inductive step, so $f(t) = x_0$ for all $t \in [0,\, t_n] = [0,\, 1]$. □

For the second proof, suppose for a contradiction that X is path-connected. Then there is a continuous function $f : [0,\, 1] \to X$ such that $f(0) = x_0$ and $f(1)$ has coordinates $(1,\, 0)$. Let $i : X \to \mathbb{R}^2$ be the inclusion function and let $p_1,\, p_2 : \mathbb{R}^2 \to \mathbb{R}$ be the projections onto the x-axis and the y-axis respectively. Write f_1 for $p_1 \circ i \circ f$ and f_2 for $p_2 \circ i \circ f$. Let (α_n) be a sequence of irrational numbers in $(0,\, 1)$ such that $\alpha_n \to 0$ as $n \to \infty$.

(a) Since $f_1(0) = 0$ and $f_1(1) = 1$ it follows by the intermediate value theorem that for each $n \in \mathbb{N}$ there exists $t_n \in [0,\, 1]$ such that $f_1(t_n) = \alpha_n$. Since $f(t_n) = (f_1(t_n),\, f_2(t_n)) = (\alpha_n,\, f_2(t_n))$ and this point is in X, we must have $f_2(t_n) = 0$ for every $n \in \mathbb{N}$.

(b) Since (t_n) is a bounded sequence, by the Bolzano-Weierstrass theorem it has a subsequence (t_{n_r}) converging to a point $a \in [0,\, 1]$.

By continuity, $f_1(t_{n_r}) \to f_1(a)$ and $f_2(t_{n_r}) \to f_2(a)$ as $r \to \infty$. Since $f_1(t_n) = \alpha_n$ for all

$n \in \mathbb{N}$, we get $f_1(a) = 0$, so $f_2(a) = 1$ since $f(a) \in X$. But $f_2(t_n) = 0$ for all $n \in \mathbb{N}$ and $f_2(t_{n_r}) \to f_2(a)$ as $r \to \infty$, so $f_2(a) = 0$. This contradiction shows that there is no continuous path in X from x_0 to the point $(1, 0)$. □

Now here is 'the topologist's sine curve'.

Example S.12.13 Let X be the following subset of the Euclidean plane:

$$X = G \cup \{(0, 0)\} \text{ where } G = \{(x, \sin(1/x)) : x \in (0, \infty)\} \text{ and } (0, 0) \text{ is the origin.}$$

Since G is the graph of a continuous function $x \mapsto \sin(1/x)$ on $(0, \infty)$, Proposition 10.18 together with connectedness of $(0, \infty)$ tells us that G is connected. Now the point $(0, 0)$ is in the closure of G since for any $\varepsilon > 0$ we may choose an integer n such that $x = 1/2n\pi < \varepsilon$ and then the point $(x, 0) = (x, \sin(1/x)) = (x, 0)$ is in $B_\varepsilon((0, 0)) \cap G$. Hence X is connected by Proposition 12.19.

But X is not path-connected. We follow a method similar to the second proof of Example S.12.12. Suppose for a contradiction that X is path-connected. Then there is a continuous function $f : [0, 1] \to X$ such that $f(0) = (0, 0)$ and $f(1) = (1, \sin 1)$. Let $i : X \to \mathbb{R}^2$ be the inclusion function and let $p_1, p_2 : \mathbb{R}^2 \to \mathbb{R}$ be the projections onto the x-axis and the y-axis respectively. Write f_1 for $p_1 \circ i \circ f$ and f_2 for $p_2 \circ i \circ f$. We get two results about these functions:

(a) Since $f_1(0) = 0$ and $f_1(1) = 1$ it follows by the intermediate value theorem that for each $n \in \mathbb{N}$ there exists $t_n \in [0, 1]$ such that $f_1(t_n) = 1/(2n + 1/2)\pi$. Since $f(t_n) = (f_1(t_n), f_2(t_n))$ is in X, we must have $f_2(t_n) = \sin(2n + 1/2)\pi = 1$ for every $n \in \mathbb{N}$.

(b) Since (t_n) is a bounded sequence, by the Bolzano-Weierstrass theorem it has a subsequence (t_{n_r}) converging to a point $a \in [0, 1]$.

By continuity, $f_1(t_{n_r}) \to f_1(a)$ and $f_2(t_{n_r}) \to f_2(a)$ as $r \to \infty$. Since $f_1(t_n) = 1/(2n + 1/2)\pi$ for all $n \in \mathbb{N}$, we get $f_1(a) = 0$, so $f_2(a) = 0$ since $f(a) \in X$. But $f_2(t_n) = 1$ for all $n \in \mathbb{N}$, so $f_2(a) = 1$. This contradiction shows that there is no continuous path in X from $(0, 0)$ to the point $(1, \sin 1)$. □ ★

Path-components There is a concept of path-component analogous to that of component, and the analogue of Proposition S.12.6 holds, where the equivalence relation is now $x \sim y$ iff x, y both belong to some path-connected subset of X. Also, the analogue of Proposition S.12.10 holds. However, the analogue of Proposition S.12.8 is not true in general, as we have seen in the previous section.

Hints for Exercise 12.10 This exercise needs some organization, or 'book-keeping'. We begin by supposing for a contradiction that there is a continuous function $f : \mathbb{R} \to \mathbb{R}$ such that for every $x \in \mathbb{R}$ the set $f^{-1}(x)$ contains exactly two points.

Choose some $x \in \mathbb{R}$ and let a, b be the two points in $f^{-1}(x)$. We may assume without loss of generality that $a < b$ (otherwise interchange their names). From standard properties of continuity of a real-valued function of a real variable (see Chapter 13 in the book) f is bounded on $[a, b]$ and attains its bounds there. Now the given condition means that f cannot be constant on the interval $[a, b]$, so either the maximum M of f on $[a, b]$ satisfies $M > f(a)$ or the minimum m satisfies $m < f(a)$.

Suppose that $M > f(a)$, and that f attains M at $c \in (a, b)$, i.e. $f(c) = M$. Choose some value d with $f(a) < d < M$. Since $f(b) = f(a)$, the intermediate value theorem tells us that there exist at least one value $x_1 \in (a, c)$ and at least one value $x_2 \in (c, b)$ such that $f(x_1) = d = f(x_2)$. Now choose some real number $\Delta > M$. We are given that $f(x) = \Delta$ for precisely two values of x. Since M is the maximum of f on $[a, b]$, neither of these values of x is in $[a, b]$. Suppose for example that $f(x) = \Delta$ for some $x < a$. Since d lies between $f(a)$ and Δ, the intermediate value theorem tells us that $f(x_3) = d$ for some $x_3 \in (x, a)$. This contradicts the assumption that $f^{-1}(d)$ contains exactly two points. A similar argument leads to the same contradiction if $f(x) = \Delta$ for some $x > b$.

If $m < f(a)$ a similar pair of arguments again lead to a contradiction.

Hints for Exercise 12.19 Think about an infinite ladder, and obtain V_n by removing the first n rungs, so that in the infinite intersection all the rungs are removed and we have a disconnected set (the two sides of the ladder) left.

Supplementary material for Chapter 13

Here is a list of supplementary topics for Chapter 13.

Uniform continuity and compactness (1) One promise in the book is to prove Proposition 13.24, which we recall as

Proposition S.13.1 If $f : X \to Y$ is a continuous map of metric spaces and X is compact, then f is uniformly continuous on X.

Proof This proof is not hard, but it is a little subtle in the choices made. Let d_X, d_Y be the metrics on X, Y. Let $\varepsilon > 0$ and let $a \in X$. By continuity of f at a there exists $\delta_a > 0$ such that $d_Y(f(x), f(a)) < \varepsilon/2$ whenever $d_X(x, a) < 2\delta_a$. (Slipping in the multiplier 2 on $2\delta_a$ here is rather important for the proof.) Since $\{B_{\delta_a}(a) : a \in X\}$ is an open cover of X and X is compact, there is a finite subcover, say $\{B_{\delta_{a_1}}(a_1), B_{\delta_{a_2}}(a_2), \ldots, B_{\delta_{a_n}}(a_n)\}$. Put $\delta = \min\{\delta_{a_1}, \delta_{a_2}, \ldots, \delta_{a_n}\}$. Suppose that x_1, x_2 are any points in X with $d_X(x_1, x_2) < \delta$. Then there exists $i \in \{1, 2, \ldots, n\}$ such that $d_X(x_1, a_i) < \delta_{a_i}$. From this we get also that $d_X(x_2, a_i) \leqslant d_X(x_2, x_1) + d_X(x_1, a_i) < \delta + \delta_{a_i} \leqslant 2\delta_{a_i}$. Hence

$$d_Y(f(x_1), f(x_2)) \leqslant d_Y(f(x_1), f(a_i)) + d_Y(f(a_i), f(x_2)) < \varepsilon/2 + \varepsilon/2 = \varepsilon,$$

and f is uniformly continuous on X as required. □

Example S.13.2 The function $x \mapsto x^2$ is uniformly continuous on any bounded interval but not on $\mathbb{R}$. Since $x \mapsto x^2$ is continuous, the first part of this assertion follows from Proposition S.13.1, since any bounded interval is contained in some closed bounded interval, which is compact by Proposition 13.9. But using $|x^2 - y^2| = |x+y||x-y|$ we shall see that uniform continuity fails on $\mathbb{R}$. For take $\varepsilon = 1$. Then no matter how small $\delta > 0$ is we may choose $x > 1/\delta$ and $y = x + \delta/2$ so that $|x - y| < \delta$ but $|x^2 - y^2| = |x - y||x + y| = (\delta/2)|x + y| > (\delta/2)(2/\delta) = 1$. So $x \mapsto x^2$ is not uniformly continuous on $\mathbb{R}$.

At various stages in the book when we prove continuity of some function, we actually prove uniform continuity. For example this is true for a contraction mapping as in Chapter 17. It is also true for the distance function $d : X \times X \to \mathbb{R}$ in a metric space (X, d) (see Exercise 5.17).

Uniform continuity has some general properties. For example

Proposition S.13.3 If $f : X \to Y$ and $g : Y \to Z$ are uniformly continuous functions of metric spaces then $g \circ f : X \to Z$ is also uniformly continuous.

Proof Suppose X, Y, Z have metrics d_X, d_Y, d_Z. Let $\varepsilon > 0$. Since g is uniformly continuous, there exists $\delta > 0$ such that $d_Z(g(y_1), g(y_2)) < \varepsilon$ whenever $d_Y(y_1, y_2) < \delta$. Since f is also uniformly continuous there exists $\gamma > 0$ such that $d_Y(f(x_1), f(x_2)) < \delta$ for all $x_1, x_2 \in X$ such that $d_X(x_1, x_2) < \gamma$. So for any $x_1, x_2 \in X$ with $d_X(x_1, x_2) < \gamma$ we have $d_Y(f(x_1), f(x_2)) < \delta$ and hence $d_Z(g(f(x_1)), g(f(x_2))) < \varepsilon$, i.e. $d_Z((g \circ f)(x_1), (g \circ f)(x_2)) < \varepsilon$. Hence $g \circ f$ is uniformly continuous on X. $\square$

Not everything works so well for uniform continuity.

Example S.13.4 Let $f : [1, \infty) \to [1, \infty)$ be the homeomorphism given by $f(x) = x^{1/2}$. Then f is uniformly continuous but f^{-1} is not.

To see that f is uniformly continuous note that

$$|f(x) - f(y)| = |x^{1/2} - y^{1/2}| = \left|\frac{x-y}{x^{1/2} + y^{1/2}}\right| \leqslant \frac{|x-y|}{2} \quad \text{since } x \geqslant 1,\ y \geqslant 1.$$

But $f^{-1}(x) = x^2$, and we have already seen that this is not uniformly continuous on $\mathbb{R}$, and the argument is the same on $[1, \infty)$.

★ **Compact subspaces which are not closed (2)** In the book, we try to emphasize the positive, so there are few examples that I would regard as 'pathological'. However, this section and the one on local compactness do contain some pathology. In view of Proposition 13.12, which says that in a Hausdorff space any compact subspace A is closed in X, we have to look for examples of compact subspaces which are not closed among the non-Hausdorff spaces described in Exercises 7.4, 7.5, 7.6. Recall that in Exercise 7.4 a topology is described for $\mathbb{N}$ in which the open sets are $\emptyset$, $\mathbb{N}$, and $\{1, 2, \ldots, n\}$ for each $n \in \mathbb{N}$. Recall also that in S.9 we saw that the closure in this space of the set $\{1, 2, \ldots, n\}$ is $\mathbb{N}$. So $\{1, 2, \ldots, n\}$ is not closed in this space. However, it is compact, since it is finite (see Example 13.8(a)).

A rather similar example is provided by Exercise 7.6. Any set of the form $(-\infty, b]$ has closure $\mathbb{R}$ so is not itself closed. But if $\mathcal{U}$ is any open cover of $(-\infty, b]$ there must be a set $U \in \mathcal{U}$ such that $b \in U$, and then since U is open, either $U = \mathbb{R}$ or $U = (-\infty, c)$ with $b < c$. In either case $(-\infty, b] \subseteq U$. So $(-\infty, b]$ is compact, but not closed in this space.

Finally we consider any infinite set X with the co-finite topology. Let A be any infinite subset of X which is not all of X. Then the only closed subset of X containing A is X itself, so A is not closed in X. But if $\mathcal{U}$ is any open cover of A, then there must exist some non-empty set $U_0 \in \mathcal{U}$, since $A \neq \emptyset$. Now $X \setminus U_0$ is finite, say $X \setminus U_0 = \{x_1, x_2, \ldots, x_n\}$. Let $\{x_1, x_2, \ldots, x_n\} \cap A = \{a_1, a_2, \ldots, a_r\}$. For each $i \in \{1, 2, \ldots, r\}$ there is some $U_i \in \mathcal{U}$ such that $a_i \in U_i$. Then $\{U_0, U_1, \ldots, U_r\}$ is a finite subcover of $\mathcal{U}$ for A. Hence A is compact. ★

★ **Local compactness (3)** Not all familiar spaces are compact (think of $\mathbb{R}$), but there are some weaker conditions which are almost as good; we mention one of them in this section, local compactness. There are various definitions of local compactness, not all of which are equivalent in general, but they all agree in a Hausdorff space. One can define local compactness 'at a point', but for simplicity and also because many of the locally compact spaces we are interested in are homogeneous - 'the same at every point' - we just consider local compactness at all points of a space. We begin by making a choice of definition in the general case.

Definition S.13.5 A space X is said to be *locally compact* if for any point $x \in X$ there exists an open subset U of X and a compact subset C of X such that $x \in U \subseteq C$.

Example S.13.6 Euclidean space $\mathbb{R}^n$ is locally compact. For any point is contained in some open set of the form $(a, b) \times (a, b) \times \ldots \times (a, b)$ (n copies) which is contained in the compact set $[a, b] \times [a, b] \times \ldots \times [a, b]$ (n copies).

Example S.13.7 Any compact space X is locally compact, for given $x \in X$ we may take $U = C = X$.

Example S.13.8 The space $\mathbb{Q}$ with its usual topology is not locally compact. For let $x \in \mathbb{Q}$, let U be an open subset of $\mathbb{Q}$ containing x, and let C be a subset of $\mathbb{Q}$ containing U. We want to show that C is not compact. First, U is of the form $V \cap \mathbb{Q}$ where V is an open subset of $\mathbb{R}$ containing x. Since V is open in $\mathbb{R}$, we may choose an irrational number $y \in V$, and there exists $\varepsilon > 0$ such that $(y - \varepsilon, y + \varepsilon) \subseteq V$. There exist rational numbers in $(y - \varepsilon, y + \varepsilon)$, which are therefore points of $V \cap \mathbb{Q} = U$. We may choose such rational points arbitrarily close to y. Hence y is a limit point of U and hence of C. But $y \notin C$ since $C \subseteq \mathbb{Q}$. Hence C is not closed in $\mathbb{R}$, so C is not compact (by Proposition 13.12).

There will be further examples of spaces which are not locally compact in S.14.

It is of interest to see which properties of compact spaces hold also for locally compact spaces. Some have no chance, such as Proposition 13.10, that any compact subspace of a metric space is bounded - the real line $\mathbb{R}$ is locally compact and not bounded. Likewise, the analogue of Proposition 13.12, that a compact subspace of a Hausdorff space X is closed in X is not true - for example $(0, 1)$ is a locally compact subspace of the Hausdorff space $\mathbb{R}$, but $(0, 1)$ is not closed in $\mathbb{R}$. However, the analogues of Proposition 13.20 and 13.21 are true and not hard.

Proposition S.13.9 Any closed subspace V of a locally compact space X is locally compact.

Proof Let $x \in V$. Since X is locally compact, there exist an open subset $U \subseteq X$ and a compact subset C such that $x \in U \subseteq C$. Then $U \cap V$ is open in V and $C \cap V$ is closed in C and hence compact, and we have $x \in U \cap V \subseteq C \cap V$ as required. □

Proposition S.13.10 The product of locally compact spaces X, Y is locally compact.

Proof Suppose $(x, y) \in X \times Y$. Since X is locally compact, there is an open subset U_1 of X and a compact subset C_1 of X such that $x \in U_1 \subseteq C_1$. Similarly there is an open subset U_2 of Y and a compact subset C_2 of Y such that $y \in U_2 \subseteq C_2$. Now $U_1 \times U_2$ is open in $X \times Y$ by definition of the product topology and $C_1 \times C_2$ is compact by Proposition 13.21, and $(x, y) \in U_1 \times U_2 \subseteq C_1 \times C_2$ as required. □

Remark This result does *not* extend to infinite products.

The next example prevents a possible error.

Example S.13.11 The continuous image of a locally compact space is not necessarily locally compact. Let $X = \mathbb{R} \setminus \mathbb{Z}$, the countable union of the open intervals between successive integers. Then X is locally compact - if $x \in (n, n+1)$ there exists $\varepsilon > 0$ such that $[x-\varepsilon, x+\varepsilon] \subseteq (n, n+1)$, and then $x \in (x-\varepsilon, x+\varepsilon) \subseteq [x-\varepsilon, x+\varepsilon] \subseteq X$ as required for local compactness of X. Now take the only space we yet know is not locally compact, namely $\mathbb{Q}$. This is countable, so we may let $f : X \to \mathbb{Q}$ map each open interval $(n, n+1)$ in X constantly to an element of $\mathbb{Q}$ in such a way that f is onto. Then $f|(n, n+1)$ is continuous for each $n \in \mathbb{Z}$ since it is a constant map, so f is continuous, for example by Exercise 10.7(a). But its image $\mathbb{Q}$ is not locally compact.

However, in spite of Example S.13.11, local compactness *is* a topological property.

Proposition S.13.12 If $f : X \to Y$ is a continuous, onto, open map and X is locally compact then so is Y.

Proof Let $y \in Y$ and suppose $x \in X$ is such that $f(x) = y$. Since X is locally compact, there is an open subset U of X and a compact subset C of X such that $x \in U \subseteq C$. Then $y \in f(U) \subseteq f(C)$, where $f(U)$ is open in Y because f is an open map and $f(C)$ is compact because f is continuous (Proposition 13.15). □

Corollary S.13.13 Local compactness is a topological property.

Proof A homeomorphism $f : X \to Y$ is continuous and onto; it is also open since if $U \subseteq X$ is open in X then $f(U) = (f^{-1})^{-1}(U)$ is open in Y by continuity of f^{-1}. Now if X is locally compact so is Y by Proposition S.13.12, and the converse is true since $f^{-1} : Y \to X$ is also a homeomorphism. □

The next result concerns the Alexandroff one-point compactification of Exercise 13.22. Recall from it that any topological space X may be regarded as a subspace of a compact space X' where X' contains just one more point than X, labelled ∞. We can now prove an addendum to that result, which also prepares the ground for an alternative definition of local compactness.

Proposition S.13.14 With the above notation, suppose that X is Hausdorff and locally compact. Then X' is Hausdorff.

Proof Let x, y be distinct points in X'. If x, y are both in X then there exist disjoint open subsets U, V of X such that $x \in U$, $y \in V$. Moreover U, V are also open in X'. Suppose now that one of x, y, say y, is ∞. By local compactness of X there is an open subset U of X and a compact subset C of X such that $x \in U \subseteq C$. Since X is Hausdorff, C is closed in X as well as compact. Let $V = X' \setminus C$. Then V is open in X' since its complement C is a compact closed subset of X. Now U, V are disjoint open sets in X' and $x \in U$, $y \in V$. □

The next proposition indicates an alternative definition which is not in general equivalent to Definition S.13.5 but is equivalent in any Hausdorff space.

Proposition S.13.15 Let X be a Hausdorff space. Then X is locally compact iff for each point $x \in X$ and each open set $U \ni x$ there is an open set V such that $x \in V$, $\overline{V} \subseteq U$, and $\overline{V}$ is compact.

Proof In one direction this is clear: if the latter condition holds, then for each point $x \in X$ there is an open set V with $x \in V$ and $\overline{V}$ compact, and $V \subseteq \overline{V}$, so X is locally compact.

For the converse we may use Alexandroff's one-point compactification X' of X. So suppose that X is Hausdorff and locally compact. By Exercise 13.22 and Proposition S.13.14, X' is a compact Hausdorff space. Hence X' is regular by Exercise 13.9. Suppose that $x \in X$ and U is an open subset of X with $x \in U$. Consider the closed subset $X' \setminus U$ of X'. By regularity of X', there exist disjoint open subsets V, W of X' such that $x \in V$ and $X' \setminus U \subseteq W$. Note that therefore $X \setminus (X \cap W) = X' \setminus W \subseteq U$. By definition, $X' \setminus W$ is compact and closed in X. Now $V \subseteq X' \setminus W = X \setminus (X \cap W)$, which is closed in X. So writing $\overline{V}$ for the closure of V in X, by Proposition 9.10 (f) we have $\overline{V} \subseteq X \setminus (X \cap W)$, so $\overline{V} \subseteq U$. Note that V is open in X. Also, $\overline{V}$ is a closed subset of the compact set $X \setminus (X \cap W)$ and hence is compact. □

Corollary S.13.16 An open subset U of a locally compact Hausdorff space X is locally compact.

Proof Let $x \in U$. Using Proposition S.13.15, choose an open subset V of X such that $x \in V$, $\overline{V}$ is compact, and $\overline{V} \subseteq U$. We see that the conditions are fulfilled for U to be locally compact.

□

We now pursue a little further the question of which subspaces of a locally compact Hausdorff space X are locally compact. We have already seen that a subspace of X is locally compact if it is either open or closed in X.

Proposition S.13.17 The intersection of two locally compact subspaces A, B of a Hausdorff space X is locally compact.

Proof Let $x \in A \cap B$. Since $x \in A$ and A is locally compact, $x \in U_1 \subseteq C_1$ for some open subset U_1 of A and compact subset C_1. By definition of the subspace topology, $U_1 = A \cap W_1$ for some W_1 open in X. Similarly $x \in U_2 \subseteq C_2$ where C_2 is compact and $U_2 = B \cap W_2$ for some W_2 open in X. Now $x \in U_1 \cap U_2 \subseteq C_1 \cap C_2$, where $U_1 \cap U_2 = (A \cap B) \cap (W_1 \cap W_2)$ is open in $A \cap B$ and $C_1 \cap C_2$ is compact by Exercise 13.10. So $A \cap B$ is locally compact. □

Corollary S.13.18 The intersection of an open subspace and a closed subspace of a locally compact Hausdorff space X is locally compact.

Proof This follows from Proposition S.13.9, Corollary S.13.16 and Proposition S.13.17. □

Remark Although we shall not prove it, conversely any locally compact subspace of a locally compact Hausdorff space is the intersection of an open subspace and a closed subspace. ★

Hints for Exercise 13.12 As the hint suggests, consider the sets: $W_n = V_n \cap (X \setminus U)$. Deduce from the fact that the V_n are nested that the same is true for the W_n. Prove that each W_n is closed in X. Now, using Exercise 13.11, get a contradiction from assuming that there is no integer n such that $V_n \subseteq U$.

Hints for Exercise 13.13 Show that the X_n form a nested sequence of closed subsets of X. Note that $X_1 = f(X_0) = f(X) \subseteq X_0$. Prove inductively that $X_n \subseteq X_{n-1}$ for all integers $n \geqslant 1$. Prove inductively that X_n is compact for all integers $n \geqslant 0$. But X is Hausdorff, so each X_n is closed in X. Also, each X_n is non-empty by inductive construction. Now use Exercise 13.11.

(b) The inclusion $f(A) \subseteq A$ is straightforward to check. To prove the opposite inclusion follow the hint, and for any $a \in A$ let $V_n = f^{-1}(a) \cap X_n$. Since X is Hausdorff, $\{a\}$ is closed in A, hence since f is continuous, $f^{-1}(a)$ is a closed subset of X by Proposition 9.5. Also, each X_n is compact hence closed in X since X is Hausdorff. So each V_n is closed in X. The V_n are nested since the X_n are. Moreover, for any integer $n \geqslant 0$, you know $a \in X_{n+1} = f(X_n)$ so there exists $x \in X_n$ such that $f(x) = a$. This says that $V_n = f^{-1}(a) \cap X_n$ is non-empty. Now by Exercise 13.11, $\bigcap_{n=0}^{\infty} V_n$ is non-empty. Let b be a point in this set. Then $b \in V_n = f^{-1}(a) \cap X_n$ for all integers $n \geqslant 0$. Show that $f(b) = a$, and that $b \in \bigcap_{n=0}^{\infty} X_n = A$.

Hints for Exercise 13.18 Suggest proof by contradiction. Suppose there is no point $x_0 \in X$ such that $f(x_0) = 0$ for all $f \in \mathcal{F}$. Then for any $x \in X$ there is a function $f_x \in \mathcal{F}$ such that $f_x(x) \neq 0$, so $f_x(x) > 0$ by (i). Now use continuity of f_x to get an open set $U(x) \ni x$ such that $f_x(y) > 0$ for all $y \in U(x)$. The open cover $\{U(x) : x \in X\}$ of compact X has a finite subcover, say $\{U(x_1), U(x_2), \ldots, U(x_r)\}$. Let $f = f_{x_1} + f_{x_2} + \ldots + f_{x_r}$. Iterate (ii) to see that $f \in \mathcal{F}$. Now prove that $f(x) > 0$ for all $x \in X$, contradicting (iii).

Hints for Exercise 13.19 Let W be a closed subset of a compact Hausdorff space X. Then W is compact (by Proposition 13.20). For any $w \in W$, by the Hausdorff condition there exist disjoint open subsets U_w, V_w of X such that $y \in U_w$, $w \in V_w$. Now the open cover $\{V_w : w \in W\}$ of compact W has a finite subcover $\{V_{w_1}, V_{w_2}, \ldots, V_{w_r}\}$. Put

$$U = \bigcap_{i=1}^{r} U_{w_i}, \quad V = \bigcup_{i=1}^{r} V_{w_i}.$$

Then U, V are open in X. Also, $W \subseteq V$ since $\{V_{w_1}, V_{w_2}, \ldots, V_{w_r}\}$ is a cover of W. Next, $y \in U$ since $y \in U_{w_i}$ for all $i \in \{1, 2, \ldots, r\}$. Finally check that U, V are disjoint.

The proof that X is normal is very similar. Suppose now that W, Y are disjoint closed subsets of X. Apply the first part, for each $y \in Y$, to get disjoint open subsets U_y, V_y of X such that $y \in U_y$, $W \subseteq V_y$. Now $\{U_y : y \in Y\}$ is an open cover of Y, and Y is compact (by Proposition 13.20) so there is a finite subcover $\{U_{y_1}, U_{y_2}, \ldots, U_{y_s}\}$. Put

$$U = \bigcup_{j=1}^{s} U_{y_j}, \quad V = \bigcap_{j=1}^{s} V_{y_j}$$

and check that these have the required properties (disjoint, open, containing Y, W respectively).

Hints for Exercise 13.20 You can show that $X \setminus p_X(W)$ is open in X by proving that if $x \in X \setminus p_X(W)$ then there is some open subset U of X such that $x \in U \subset X \setminus p_X(W)$.

If $x \in X \setminus p_X(W)$, there is no $y \in Y$ such that $(x, y) \in W$. So $(x, y) \notin W$ for any $y \in Y$. Now (x, y) is in the set $X \times (Y \setminus W)$ which is open in $X \times Y$. Use the definition of the product topology, to get open subsets U_y, V_y of X, Y respectively such that $(x, y) \in U_y \times V_y \subseteq X \times (Y \setminus W)$. Now $\{V_y : y \in Y\}$ is an open cover of Y, and Y is compact, so there exists a finite subcover $\{V_{y_1}, V_{y_2}, \ldots, V_{y_r}\}$. Put

$$U = \bigcap_{i=1}^{r} U_{y_i}.$$

Then U is open in X, and $x \in U$ since $x \in U_{y_i}$ for each $i \in \{1, 2, \ldots, r\}$. Show that $U \subseteq X \setminus p_X(W)$ (if $x' \in U$ then given any $y \in Y$ we know that $y \in V_{y_i}$ for some $i \in \{1, 2, \ldots, r\}$, so from $x' \in U_{y_i}$ and $(U_{y_i} \times V_{y_i}) \cap W = \emptyset$ we get $(x', y) \notin W$).

(b) Think about Exercise 10.15(b) to get an example.

Supplementary material for Chapter 14

Here is a list of supplementary topics for Chapter 14.

Limit point version of Bolzano-Weierstrass theorem (2)

Proposition S.14.1 Any bounded set X in $\mathbb{R}$ with an infinite number of members has at least one limit point.

Proof We use the bisection method. Since X is bounded, it is contained in some interval $[a, b]$ in $\mathbb{R}$. Since the number of members in X is infinite, at least one of the sub-intervals $[a, (a+b)/2]$, $[(a+b)/2, b]$ contains an infinite number of members of X. Write $[a_1, b_1]$ for such a subinterval (if both contain infinitely many members of X then let $[a_1, b_1]$ be the left-hand one for definiteness).

Suppose inductively that real numbers $a_1, a_2, \ldots a_n, b_1, b_2, \ldots, b_n$ have been chosen so that $a \leqslant a_1 \leqslant a_2 \leqslant \ldots \leqslant a_n < b_n \leqslant b_{n-1} \leqslant \ldots \leqslant b_2 \leqslant b_1$, for any $i \in \{1, 2, \ldots n\}$ the interval $[a_i, b_i]$ contains infinitely many points of X, and $b_i - a_i = (b-a)/2^i$. Then at least one of $[a_n, (a_n + b_n)/2]$, $[(a_n + b_n)/2, b_n]$ contains infinitely many points of X. Let $[a_{n+1}, b_{n+1}]$ be such an interval (or if both contain infinitely many points of X, choose $[a_{n+1}, b_{n+1}]$ to be the left-hand one for definiteness). Then the inductive hypotheses all hold up to stage $n+1$. Now the sequence (a_n) is monotonic increasing and bounded above by any b_m, so (a_n) converges by Proposition 4.16, and its limit x satisfies $x \leqslant b_m$ for all $m \in \mathbb{N}$. Also (b_m) is monotonic decreasing and bounded below by x, so it too converges, to a limit y such that $y \geqslant x$. But for any $n \in \mathbb{N}$ we have $a_n \leqslant x \leqslant y \leqslant b_n$, and $b_n - a_n = (b-a)/2^n$, so $y = x$. This point x is a limit point of X, for given any $\varepsilon > 0$ we know that for $b_n - a_n < \varepsilon$ for all sufficiently large $n \in \mathbb{N}$, so $(x-\varepsilon, x+\varepsilon)$ contains $[a_n, b_n]$, hence a point (indeed infinitely many points) of $X \setminus \{x\}$. This completes the proof. □

Remark It is possible that X has only one limit point in $\mathbb{R}$: e.g. the set $\{1, 1/2, \ldots, 1/n, \ldots\}$ has 0 as its only limit point in $\mathbb{R}$. On the other hand, if $X = [0, 1]$ then every point of X is a limit point of X in $\mathbb{R}$.

Another compact space (2) There will be further examples of compact spaces in the extra section on compactness in function spaces, C.1; here we give an example in Hilbert space $\mathbf{l^2}$, the space of real sequences (x_i) such that $\sum_{i=1}^{\infty} x_i^2$ converges, with the norm $||(x_i)||_2 = \sqrt{\sum_{i=1}^{\infty} x_i^2}$ (see S.5).

★**Example S.14.2** The Hilbert cube C is compact, where $C \subseteq \mathbf{l^2}$ consists of all sequences (x_i) of real numbers such that $|x_i| \leqslant 1/i$ for every $i \in \mathbb{N}$.

Note that $C \subseteq \mathbf{l_2}$ since $|x_i| \leqslant 1/i$ so $\sum_{i=1}^{\infty} x_i^2$ converges by comparison, since $\sum 1/i^2$ converges.

Proof We know from Chapter 14 that it is enough to show that C is sequentially compact. We shall use a 'diagonal trick' to prove this. As usual, there is a slight notational complication in dealing with sequences of points in sequence spaces, and even more in dealing with subsequences of those. We shall use the notation introduced in S.2 for subsequences of subsequences. Suppose that $\boldsymbol{x^{(1)}, x^{(2)}, \ldots, x^{(n)}, \ldots}$ is a sequence in C, so that writing $\boldsymbol{x^{(n)}} = (x_1^{(n)}, x_2^{(n)}, \ldots, x_i^{(n)}, \ldots)$ we have $|x_i^{(n)}| \leqslant 1/i$ for all $n, i \in \mathbb{N}$. The real sequence $(x_1^{(n)})$ is bounded ($|x_1^{(n)}| \leqslant 1$ for all $n \in \mathbb{N}$). So by the sequence form of the Bolzano-Weierstrass theorem, which says in particular that $[-1, 1]$ is sequentially compact, there is a subsequence $(x_1^{(n(r,1))})$ converging to a point x_1 say in $[-1, 1]$. Recall that $x_1^{(n(r,1))}$ is the rth term in this subsequence of $(x_1^{(n)})$. We consider next the corresponding subsequence $(x_2^{(n(r,1))})$ of the sequence $(x_2^{(n)})$ of second coordinates of the original sequence $(\boldsymbol{x^{(n)}})$. It again is bounded, so has a second level subsequence $(x_2^{(n(r,2))})$ converging to a point x_2 in $[-1/2, 1/2]$. Suppose inductively that for some positive integer s there is an sth level subsequence $(n(r, s))$ of (n) such that for each $i = 1, 2, \ldots s$ the subsequence $(x_i^{(n(r,i))})$ of $(x_i^{(n)})$ converges to a point x_i in $[-1/i, 1/i]$, and also for each $i = 2, \ldots s$ the sequence $(n(r, i))$ is a subsequence of the sequence $(n(r, i-1))$. Then by sequential compactness of $[-1/(s+1), 1/(s+1)]$ the sequence $(x_{s+1}^{(n(r,s))})$ has a subsequence $(x_{s+1}^{(n(r,s+1))})$ converging to a point x_{s+1} in $[-1/(s+1), 1/(s+1)]$. Recall that $x_{s+1}^{(n(r,s+1))}$ is the rth term in this $(s+1)$th level subsequence of the sequence $(x_{s+1}^{(n)})$ of $(s+1)$th coordinates of the original sequence $(\boldsymbol{x^{(n)}})$.

Since there are infinitely many values of n, stopping the above induction at any finite stage doesn't work. The trick is to consider the 'diagonal subsequence'. Let us write the successive subsequences of the integers which we have selected, on the rows of the following array, in which each row is a subsequence of any rows above it:

$$\begin{array}{llllll} n(1,1), & n(2,1), & n(3,1), & \ldots, & n(r,1) & \ldots \\ n(1,2), & n(2,2), & n(3,2), & \ldots, & n(r,2) & \ldots \\ \ldots & \ldots, & \ldots & & & \\ n(1,s), & n(2,s), & n(3,s), & \ldots, & n(r,s) & \ldots \end{array}$$

Now consider the diagonal sequence $(n(r,r))$ of this array; write $n_r = n(r,r)$. The claim is that the subsequence $(\boldsymbol{x}^{(n_r)})$ of $(\boldsymbol{x}^{(n)})$ converges to the point $(x_1, x_2, \ldots, x_i, \ldots)$ in $\mathbf{l^2}$. Note that $(x_1, x_2, \ldots, x_i, \ldots)$ *is* in $\mathbf{l^2}$ since $|x_i| \leqslant 1/i$ for each $i \in \mathbb{N}$. Note also that the sequence $(x_i^{(n_r)})$ of ith co-ordinates of the $(\boldsymbol{x}^{(n_r)})$ converges to x_i, since from stage $r = i$ onwards, $(\boldsymbol{x}^{(n_r)})$ is a subsequence of $(\boldsymbol{x}^{(n(r,i))})$, whose sequence of ith coordinates $(x_i^{n(r,i)})$ converges to x_i.

Let $\varepsilon > 0$. First choose $N \in \mathbb{N}$ such that

$$\sum_{i=N+1}^{\infty} 1/i^2 < \varepsilon^2/2.$$

For each $i = 1, 2, \ldots, N$ let $R_i \in \mathbb{N}$ be such that $|x_i^{(n(r,i))} - x_i| < \varepsilon/\sqrt{2N}$ for all $r \geqslant R_i$. We may assume also that $R_i \geqslant i$. Now let $R = \max\{R_1, R_2, \ldots, R_N\}$, and suppose that $r \geqslant R$. Then $r \geqslant R_i$ for each $i = 1, 2, \ldots, N$. Also, $(n_r) = (n(r,r))$ is a subsequence of $(n(r,i))$ since $r \geqslant R_i \geqslant i$. Hence $(x_i^{(n_r)} - x_i)^2 < \varepsilon^2/2N$. Now

$$\sum_{i=1}^{\infty}(x_i^{(n_r)} - x_i)^2 < \sum_{i=1}^{N}(x_i^{(n_r)} - x_i)^2 + \varepsilon^2/2 < \varepsilon^2/2 + \varepsilon^2/2 = \varepsilon^2,$$

so the subsequence $(\boldsymbol{x}^{(n_r)})$ of $(\boldsymbol{x}^{(n)})$ converges to the point $(x_1, x_2, \ldots, x_i, \ldots)$ in $\mathbf{l^2}$. □★

More examples of non-compact spaces (2) There are plenty examples of non-compact spaces among sequence spaces.

First, $\mathbf{l}^\infty$ is not compact. For consider the sequence $(\boldsymbol{x}^{(r)})$ in $\mathbf{l}^\infty$ where $(\boldsymbol{x}^{(r)})$ has a 1 in the rth place and zeros elsewhere. Then for any distinct pair of integers r, s we have $d_\infty(\boldsymbol{x}^{(r)}, \boldsymbol{x}^{(s)}) = 1$, so the sequence $(\boldsymbol{x}^{(r)})$ has no Cauchy subsequence and hence no convergent subsequence.

The same example shows that neither $\mathbf{l^1}$ nor $\mathbf{l^2}$ is compact.

Closed bounded sets in these sequence spaces are not generally compact either. The example we have used above to show that $\mathbf{l}^\infty$ is not compact is a sequence which lies in the 'unit cube' $\{\boldsymbol{x} \in \mathbf{l}^\infty : ||\boldsymbol{x}|| \leqslant 1\}$ and similarly in the unit cube in $\mathbf{l^1}$ or $\mathbf{l^2}$.

★ **More examples of spaces which are not locally compact (3)** Similar examples to those in the previous section show that sequence spaces are not in general locally compact either. For let U be a open set containing $\mathbf{0}$ in, say, $\mathbf{l^\infty}$. Then for some $\varepsilon > 0$ we must have that $B_\varepsilon(\mathbf{0}) \subseteq U$. So any compact set C containing U must contain $B_\varepsilon(\mathbf{0})$. But then C contains the sequence $(\boldsymbol{x^{(r)}}_\varepsilon)$ which has an ε in the rth place and zeros elsewhere, and just as in the previous section this has no Cauchy and hence no convergent subsequence. So C cannot be compact. Hence $\mathbf{l^\infty}$ is not locally compact. Similar arguments show that neither $\mathbf{l^1}$ nor $\mathbf{l^2}$ is locally compact.★

Hints for Exercise 14.14 If $U_i = X$ for some $i \in \{1, 2, \ldots, n\}$ then any $\varepsilon > 0$ is a Lebesgue number for $\mathcal{U}$, since for any $\varepsilon > 0$, any set of diameter at most ε is contained in X and hence in U_i.

(i) Suppose now that $C_i \neq \emptyset$ for every $i \in \{1, 2, \ldots, n\}$. Use Exercise 6.16 (c) to get continuity of the function $f_i : X \to \mathbb{R}$ defined by $f_i(x) = d(x, C_i)$. Check that from the definition all the values of $f_i(x)$ are non-negative.

(ii) Use continuity of each f_i and Proposition 5.17 to get continuity of f. Let $x \in X$. Since $\mathcal{U}$ is a cover for X, $x \in U_i$ for at least one $i \in \{1, 2, \ldots, n\}$ so x is not in $C_i = X \setminus U_i$. Now C_i is closed in X, so $f_i(x) = d(x, C_i) > 0$ by Exercise 6.16 (a). But also $f_j(x) \geqslant 0$ for all $j \in \{1, 2, \ldots, n\}$ so $f(x) > 0$ as required.

(iii) Exercise 14.8 applies, by sequential compactness of X: so there exists $\varepsilon > 0$ such that $f(x) \geqslant \varepsilon$ for all $x \in X$.

(iv) Since there are just n values $d(x, C_i)$, get

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} d(x, C_i) \leqslant \max\{d(x, C_i) : i \in \{1, 2, \ldots, n\}\}.$$

(v) For a given $x \in X$ let $\max\{d(x, C_i) : i \in \{1, 2, \ldots, n\}\} = d(x, C_{k(x)})$. Prove $B_\varepsilon(x) \subseteq U_{k(x)}$ where ε is as in (iii) above. Hints: suppose $d(y, x) < \varepsilon$. Then $\varepsilon \leqslant f(x) \leqslant d(x, C_{k(x)})$ so $d(y, x) < d(x, C_{k(x)})$. This says $d(y, x)$ is less than the distance from x to $C_{k(x)} = X \setminus U_{k(x)}$, so $y \in U_{k(x)}$. Hence $B_\varepsilon(x) \subseteq U_{k(x)}$. It follows that for any $x \in X$ there is a set $U \in \mathcal{U}$ such that $B_\varepsilon(x) \subseteq U$, so ε is a Lebesgue number for the cover $\mathcal{U}$.

Hints for Exercise 14.15 Check that if one of the V_n is empty then both sides of the equation are zero.

Suppose now that all the V_n are non-empty. (We already know from Exercise 14.11 that their intersection is non-empty.) Now $\bigcap_{n=1}^{\infty} V_n \subseteq V_m$ for any $m \in \mathbb{N}$, so diam $\bigcap_{n=1}^{\infty} V_n \leqslant$ diam V_m, and diam $\left(\bigcap_{n=1}^{\infty} V_n\right) \leqslant \inf\{\text{diam } V_n : n \in \mathbb{N}\} = m_0$ say.

Conversely, m_0 is a lower bound for the diameters of the V_n, so for any $\varepsilon > 0$ and any $n \in \mathbb{N}$ we know that diam $V_n > m_0 - \varepsilon$. Hence there exist points $x_n,\, y_n \in V_n$ such that $d(x_n,\, y_n) > m_0 - \varepsilon$. Since X is sequentially compact, (x_n) has a subsequence $(x_{n(r)})$ converging to a point $x \in X$, and then $(y_{n(r)})$ has a subsequence $(y_{m(r)})$ converging to a point $y \in X$. Check that then $(x_{m(r)})$ converges to x, and that $d(x_{m(r)},\, y_{m(r)}) \to d(x,\, y)$ as $r \to \infty$. This leads to $d(x,\, y) \geqslant m_o - \varepsilon$. Also, since V_n is closed in X, observe that $x,\, y \in V_n$, and since this is true for all $n \in \mathbb{N}$, that $x,\, y \in \bigcap_{n=1}^{\infty} V_n$. Deduce that diam $\bigcap_{n=1}^{\infty} V_n \geqslant m_0 - \varepsilon$ and from this that diam $\bigcap_{n=1}^{\infty} V_n \geqslant m_0$.

Hints for Exercise 14.17 Let $x \in X$. We want to show that $x \in f(X)$. Consider the sequence (x_n) in X defined by: $x_1 = x,\ x_{n+1} = f(x_n)$ for all integers $n \geqslant 1$. Since X is sequentially compact, there is a convergent subsequence, say (x_{n_r}). Any convergent sequence is Cauchy, so given $\varepsilon > 0$ there exists $R \in \mathbb{N}$ such that $d(x_{n_r},\, x_{n_s}) < \varepsilon$ whenever $s > r \geqslant R$, in particular $d(x_{n_R},\, x_{n_r})$ whenever $r > R$. Now we use the isometry condition, iterated $n_R - 1$ times, to see that $d(x_1,\, x_{n_r - n_R + 1}) < \varepsilon$ whenever $r \geqslant R$. But $x_1 = x$ and $x_{n_r - n_R + 1} \in f(X)$ whenever $r > R$. Hence $x \in \overline{f(X)}$. But X is compact and f is continuous, so $f(X)$ is compact. Also, X is metric hence Hausdorff, so $f(X)$ is closed in X. Hence $\overline{f(X)} = f(X)$. So $x \in f(X)$ for any $x \in X$, which says that f is onto. Hence f is an isometry.

(b) You can apply (a) to the compositions $g \circ f : X \to X$ and $f \circ g : Y \to Y$ to see that these are both onto. Since $g \circ f$ is onto, g is onto. Similarly since $f \circ g$ is onto, f is onto. Hence both f and g are isometries.

(c) Think of a translation, say $f(x) = x + 1$.

Supplementary material for Chapter 15

Here is a list of supplementary topics for Chapter 15.

Inevitability of the quotient topology (3) Here is another approach to seeing what the quotient topology should be. Given an equivalence relation $\sim$ on a topological space $(X, \mathcal{T})$, we want to give $X/\sim$ a topology $\tilde{\mathcal{T}}$ such that any map $\tilde{f} : X/\sim \to Y$ is continuous iff $f = \tilde{f} \circ p$ is continuous, where $p : X \to X/\sim$ is the natural map. It turns out that such a topology $\tilde{\mathcal{T}}$ is unique. First, let $\tilde{\mathcal{T}} = \{\tilde{U} \subseteq X/\sim : p^{-1}(\tilde{U}) \in \mathcal{T}\}$. We show in Chapter 15 that $\tilde{\mathcal{T}}$ is a topology on $X/\sim$ satisfying the given condition (that $\tilde{f}$ is continuous iff $\tilde{f} \circ p$ is continuous). Then if $\mathcal{T}'$ is any other topology on $X/\sim$ satisfying the given condition, we take $\tilde{f}_1$ to be the identity function of $(X/\sim, \mathcal{T}')$ which is continuous. Hence since $\mathcal{T}'$ satisfies the given condition, $p = \tilde{f}_1 \circ p : (X, \mathcal{T}) \to (X/\sim, \mathcal{T}')$ is continuous. So $p^{-1}(\tilde{U}) \in \mathcal{T}$ for any $\tilde{U} \in \mathcal{T}'$, which says that $\mathcal{T}' \subseteq \tilde{\mathcal{T}}$. Now we take $\tilde{f}_2$ to be the identity map from $(X/\sim, \mathcal{T}')$ to $(X/\sim, \tilde{\mathcal{T}})$, and we know that $p = \tilde{f}_2 \circ p : (X, \mathcal{T}) \to (X/\sim, \tilde{\mathcal{T}})$ is continuous, hence $\tilde{f}_2$ is continuous, which says that $\tilde{\mathcal{T}} \subseteq \mathcal{T}'$. Hence $\mathcal{T}' = \tilde{\mathcal{T}}$, and $\tilde{\mathcal{T}}$ is the unique topology on $X/\sim$ satisfying the given condition.

Equivalence relation for the torus (1) In Chapter 15 we define the torus by putting an equivalence relation on $X = [0, 1] \times [0, 1]$. We let $(s_1, t_1) \sim (s_2, t_2)$ iff one of the following holds:

(i) $s_1 = s_2$ and $t_1 = t_2$.

(ii) $\{s_1, s_2\} = \{0, 1\}$, $t_1 = t_2$.

(iii) $\{t_1, t_2\} = \{0, 1\}$, $s_1 = s_2$.

(iv) $\{s_1, s_2\} = \{0, 1\}$, $\{t_1, t_2\} = \{0, 1\}$.

Geometric contemplation may persuade you that this is an equivalence relation. Here are the rather lengthy details done analytically. The three such calculations in this section are slightly tedious, but are included to exhibit that they *can* be done. An alternative approach, suggested by Chris Norman, is given at the end of this file.

Reflexivity For any $(s_1,\ t_1) \in [0,\ 1] \times [0,\ 1]$ it follows from (i) that $(s_1,\ t_1) \sim (s_1,\ t_1)$.

Symmetry Suppose $(s_1,\ t_1) \sim (s_2,\ t_2)$.

Case (1) If $s_1 = s_2$ and $t_1 = t_2$ then $s_2 = s_1$ and $t_2 = t_1$ so $(s_2,\ t_2) \sim (s_1,\ t_1)$ by (i) above.

Case (2) If $\{s_1,\ s_2\} = \{0,\ 1\}$ and $t_1 = t_2$ then $\{s_2,\ s_1\} = \{0,\ 1\}$ and $t_2 = t_1$ so $(s_2,\ t_2) \sim (s_1,\ t_1)$ by (ii) above.

Case (3) If $\{t_1,\ t_2\} = \{0,\ 1\}$ and $s_1 = s_2$ then $(s_2,\ t_2) \sim (s_1,\ t_1)$ as in Case (2).

Case (4) If $\{s_1,\ s_2\} = \{0,\ 1\}$ and $\{t_1,\ t_2\} = \{0,\ 1\}$ then $\{s_2,\ s_1\} = \{0,\ 1\}$ and $\{t_2,\ t_1\} = \{0,\ 1\}$ so $(s_2,\ t_2) \sim (s_1,\ t_1)$ by (iv) above.

Transitivity Suppose that $(s_1,\ t_1) \sim (s_2,\ t_2)$ and $(s_2,\ t_2) \sim (s_3,\ t_3)$. We want to prove that $(s_1,\ t_1) \sim (s_3,\ t_3)$. We distinguish several cases:

Case (1) If $s_1 = s_2$, $t_1 = t_2$ and the analogue of any one of (i) - (iv) holds for $(s_2,\ t_2)$, $(s_3,\ t_3)$ then the analogue of one of (i) - (iv) holds for $(s_1,\ t_1)$, $(s_3,\ t_3)$ so $(s_1,\ t_1) \sim (s_3,\ t_3)$.

Case (2) Similarly to Case (1), if any of (i) - (iv) above holds and also $s_2 = s_3$, $t_2 = t_3$ then the analogue of one of (i) - (iv) holds for $(s_1,\ t_1)$ and $(s_3,\ t_3)$ so $(s_1,\ t_1) \sim (s_3,\ t_3)$.

Case (3a) If (ii) holds, i.e. $\{s_1,\ s_2\} = \{0,\ 1\}$, $t_1 = t_2$, and the analogue of (ii) holds for $(s_2,\ t_2)$, $(s_3,\ t_3)$, i.e. $\{s_2,\ s_3\} = \{0,\ 1\}$, $t_2 = t_3$, then $s_1 = s_3$ and $t_1 = t_3$ so $(s_1,\ t_1) \sim (s_3,\ t_3)$ by (i).

Case (3b) If (ii) is replaced by (iii) in Case (3a), then similarly $(s_1,\ t_1) \sim (s_3,\ t_3)$ by (i).

Case (3c) If (ii) is replaced by (iv) in Case (3a), then similarly $(s_1,\ t_1) \sim (s_3,\ t_3)$ by (i).

Case (4a) If (ii) holds, i.e. $\{s_1,\ s_2\} = \{0,\ 1\}$, $t_1 = t_2$ and the analogue of (iii) holds for $(s_2,\ t_2)$, $(s_3,\ t_3)$, i.e. $s_2 = s_3$, $\{t_2,\ t_3\} = \{0,\ 1\}$, then $\{s_1,\ s_3\} = \{0,\ 1\}$ and $\{t_1,\ t_3\} = \{0,\ 1\}$ so $(s_1,\ t_1) \sim (s_3,\ t_3)$ by (iv).

Case (4b) Similarly to Case (4a), if (iii) holds and $(s_2,\ t_2)$, $(s_3,\ t_s)$ satisfy the analogue of (ii) then $(s_1,\ t_1) \sim (s_3,\ t_3)$ by (iv).

Case (5a) If (ii) holds, i.e. $\{s_1,\ s_2\} = \{0,\ 1\}$, $t_1 = t_2$ and the analogue of (iv) holds for $(s_2,\ t_2)$, $(s_3,\ t_3)$, i.e. $\{s_2,\ s_3\} = \{0,\ 1\}$, $\{t_2,\ t_3\} = \{0,\ 1\}$, then $s_1 = s_3$ and $\{t_1,\ t_3\} = \{0,\ 1\}$ so $(s_1,\ t_1) \sim (s_3,\ t_3)$ by (iii).

Case (5b) Similarly to Case (5a), if (iii) holds and the analogue of (iv) holds for $(s_2,\ t_2)$, $(s_3,\ t_3)$ then $(s_1,\ t_1) \sim (s_3,\ t_3)$ by (ii).

Case (6a) If (iv) holds, i.e. $\{s_1,\ s_2\} = \{0,\ 1\}$, $\{t_1,\ t_2\} = \{0,\ 1\}$ and the analogue of (ii) holds for $(s_2,\ t_2)$, $(s_3,\ t_3)$, i.e. $\{s_2,\ s_3\} = \{0,\ 1\}$ and $t_2 = t_3$ then $s_1 = s_3$ and $\{t_1,\ t_3\} = \{0,\ 1\}$ so $(s_1,\ t_1) \sim (s_3,\ t_3)$ by (iii).

Case (6b) Similarly to Case (6a), if (iv) holds and the analogue of (iii) holds for $(s_2,\ t_2)$, $(s_3,\ t_3)$ then $(s_1,\ t_1) \sim (s_3,\ t_3)$ by (ii).

It is tempting to try to do without (iv), but then we'd need to assume transitivity in order to get for example $(0, 0) \sim (1, 1)$, which follows from (iv).

It is straightforward to check that the corresponding equivalence classes are:

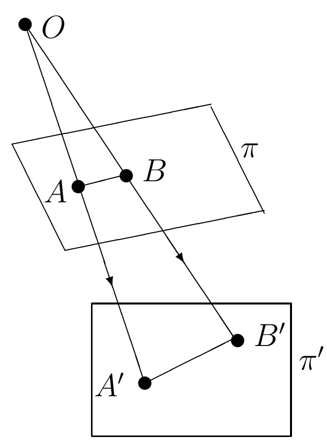
(a) $\{(s, t)\}$ where $0 < s < 1$ and $0 < t < 1$,

(b) $\{(0, t), (1, t)\}$ where $0 < t < 1$,

(c) $\{(s, 0), (s, 1)\}$ where $0 < s < 1$,

(d) $\{(0, 0), (1, 0), (0, 1), (1, 1)\}$.

Background on real projective plane (1) As the name suggests, projective geometry has to do with projections and properties which remain invariant under projections. For example we might project one plane π onto another plane π' from a point O in neither, as in the following diagram

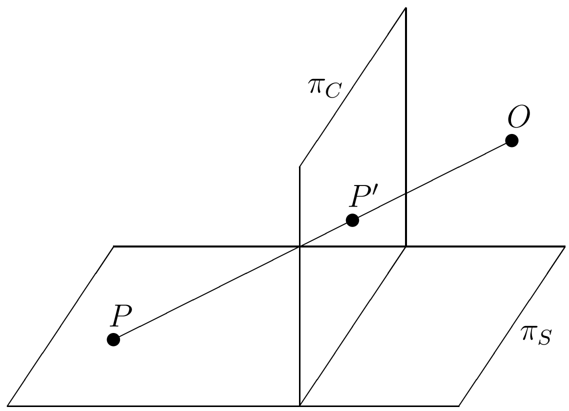


Neither Euclidean distances nor angles are preserved under such a map, but at least straight lines are projected into straight lines.

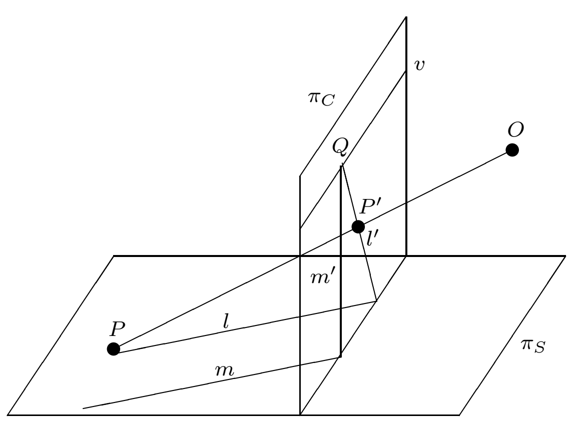
We shall mention two strands in the history of projective geometry, which are related to the versions of the real projective plane given in Proposition 15.10 (a) and (d) respectively. We discuss first the version in Proposition 15.10 (a), which we recall is $\mathbb{R}^3 \setminus \{0\}/\sim$ where $x \sim y$ iff $x = \lambda y$ for some non-zero real number λ. This is the space of lines through the origin in $\mathbb{R}^3$.

The rules of perspective in art influenced the beginnings of projective geometry. In Europe at least, these rules are usually attributed to Brunelleschi, the Renaissance architect who designed the great dome of the cathedral in Florence, and the first mathematical treatment is attributed to Desargues, a French architect. Here is a brief account of the rules of perspective, which it has to be admitted are designed for a one-eyed artist who keeps his or her head completely still while painting. One should not be too surprised to discover that not even the old masters stuck strictly to the rules.

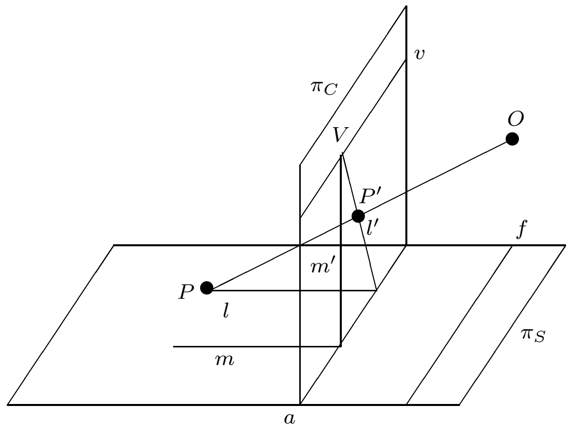
Suppose that an object is lying in a horizontal plane π_S (S for scenery) and that you want to paint a picture of it on a canvas in a vertical plane π_C (C for canvas). Suppose your eye is at the point O. The rules of perspective are very simple: they just say, think of π_C as a glass screen through which you look at the scenery π_S. Then a point P on π_S should be painted as the point P' where the line OP intersects the plane of the canvas π_C. This is illustrated in the next diagram.



Next, referring to the diagram below, let us call a the line of intersection of π_S and π_C, and v the horizontal line in π_C at the same height as O. Now let l be any line in π_S not parallel to a. The image of l in π_S is part of the line l' in which π_C meets the plane through O and l. As the point P in l moves further and further away from a on the line l, its image P' approaches but never quite reaches the line v. This line v is called 'the vanishing line'. It is the line in the painting corresponding to 'points at infinity' on the plane π_S. Next notice that if m is another line in π_S which is parallel to l, the lines l' and m' meet on v. (Proof: let h be the horizontal line through O parallel to l and m, and let it meet v at Q. The plane determined by O and l contains l' and h, so it contains their point of intersection. But the only point in which h meets π_C is Q, so l' passes through Q.)



In particular, the lines in π_C corresponding to lines in π_S which are perpendicular to a meet on v in a point V called 'the vanishing point', as in the diagram on the next page. It is interesting to look at some old paintings and see whether the artist has used the vanishing point to draw attention to a particular point in the painting. In fact there are several intriguing deductions which you may be able to make from a painting that obeys the rules of perspective. For example if the painting includes a square-tiled floor, as several Dutch interiors do, you should be able to deduce how far the artist's eye was from the canvas!



Next notice that, even when we extend these rectangles in the diagram to complete planes, the correspondence $P \leftrightarrow P'$ between π_S and π_C is not one-one, since points on v do not correspond to any points on π_S, and likewise points on the line f in π which is parallel to a and vertically below O do not correspond to any points in π_C. (This latter fact scarcely mattered to old masters, who didn't normally want to include their own feet in the painting.) But otherwise the correspondence is one-one, though this is not so obvious from the diagram - points on π_C above the line v correspond to points on π_S 'behind' the artist, and points on π_S between the artist and the canvas correspond to points on π_C below the line a.

This can be extended to give a one-one correspondence between the two planes, given by projecting through O, if we add a 'line at infinity' to each of π_S and π_C. Then points on v correspond to points on the line at infinity in π_S and points on f correspond to points on the line at infinity in π_C.

Now instead of thinking about the points P and P' let us shift attention to the whole *line* through O and P. To each point P in π_S there corresponds such a line, and the slightly mysterious 'points at infinity' on π_S just correspond to lines through O which go through points of v so are parallel to π_S.

This long story suggests that we shall get a good concrete model of the real projective plane (the extended version of π_S) by taking as 'points' all the lines in Euclidean 3-space through some fixed point O. This is what modern algebraic approaches to projective geometry do - points in an n-dimensional projective space are defined to be lines through the origin (i.e. 1-dimensional subspaces) in an $(n+1)$-dimensional vector space. This is the approach typified by (a) in Proposition 15.10.

The other historical strand we look at is due to Kepler. This again has to do with 'points at infinity'. In Euclidean geometry it can be a nuisance that one has to make special arguments to cover cases where two lines are parallel. Kepler suggested adding points at infinity to get around this. Consider all lines in the Euclidean plane, and let us say that two are *equivalent* if they are equal or parallel. Then add one new 'point at infinity' to the Euclidean plane for each such equivalence class, and deem that any two lines in the equivalence class 'meet' in that point. To tie things up, define the set of all the points at infinity to be a line. By doing this Kepler achieved a situation in which *any* two distinct lines meet in exactly one point, and *any* two points lie on exactly one line. This eliminates the need to deal with parallel lines separately, and it makes incidence properties now so symmetrical with respect to interchanging lines and points that we get a principle of duality, or 'two theorems for the price of one': any true statement about points and lines remains true when 'point' and 'line' are interchanged. To get a model of Kepler's real projective plane, we note that the whole plane is homeomorphic to the open unit disc D. If we add in the boundary that gives a line at infinity, and each equivalence class of parallel lines is represented by a line though the centre of D, or equivalently the two points in which it meets the boundary circle - but there should be just one point at infinity corresponding to each equivalence class of parallel lines, so we need to 'identify' pairs of antipodal points on the boundary circle. This leads to the model of the real projective plane given in (d) of Proposition 15.10.

Equivalence relation for P (1) We defined $(s_1, t_1) \sim (s_2, t_2)$ for points of $[0, 1] \times [0, 1]$ iff one of the following holds:

(i) $s_1 = s_2$ and $t_1 = t_2$;

(ii) $\{s_1, s_2\} = \{0, 1\}$ and $t_2 = 1 - t_1$;

(iii) $\{t_1, t_2\} = \{0, 1\}$ and $s_2 = 1 - s_1$.

Reflexivity Each point in $[0, 1] \times [0, 1]$ is equivalent to itself by (i).

Symmetry Suppose that $(s_1, t_1) \sim (s_2, t_2)$.

Case (1) If $s_1 = s_2$ and $t_1 = t_2$ then of course $s_2 = s_1$ and $t_2 = t_1$ so $(s_2, t_2) \sim (s_1, t_1)$.

Case (2) If $\{s_1, s_2\} = \{0, 1\}$ and $t_2 = 1 - t_1$ then $\{s_2, s_1\} = \{0, 1\}$ and $t_1 = 1 - t_2$, so $(s_2, t_2) \sim (s_1, t_1)$.

Case (3) The case when $\{t_1, t_2\} = \{0, 1\}$ and $s_2 = 1 - s_1$ is exactly like Case (2).

Transitivity Suppose that $(s_1, t_1) \sim (s_2, t_2)$ and that $(s_2, t_2) \sim (s_3, t_3)$. We want to prove that $(s_1, t_1) \sim (s_3, t_3)$.

Case (1) If $s_1 = s_2$ and $t_1 = t_2$ then $(s_1, t_1) = (s_2, t_2) \sim (s_3, t_3)$. Similarly if $s_2 = s_3$ and $t_2 = t_3$ then $(s_1, t_1) \sim (s_2, t_2) = (s_3, t_3)$.

Case (2) If $\{s_1, s_2\} = \{0, 1\}$ and $t_2 = 1 - t_1$ and also $\{s_2, s_3\} = \{0, 1\}$ and $t_3 = 1 - t_2$, then $s_3 = s_1$ and $t_3 = 1 - t_2 = 1 - (1 - t_1) = t_1$ so $(s_1, t_1) \sim (s_3, t_3)$ by (i).

Case (3) If $\{t_1, t_2\} = \{0, 1\}$ and $s_2 = 1 - s_1$ and also $\{t_2, t_3\} = \{0, 1\}$ and $s_3 = 1 - s_2$, then $(s_1, t_1) \sim (s_3, t_3)$ exactly as in Case (2).

Case (4) If $\{s_1, s_2\} = \{0, 1\}$ and $t_2 = 1 - t_1$ and also $\{t_2, t_3\} = \{0, 1\}$ and $s_3 = 1 - s_2$, then $s_1 = 1 - s_2 = s_3$ and $t_1 = 1 - t_2 = t_3$ so $(s_1, t_1) \sim (s_3, t_3)$ by (i).

Case (5) If $\{t_1, t_2\} = \{0, 1\}$ and $s_2 = 1 - s_1$ and also $\{s_2, s_3\} = \{0, 1\}$ and $t_3 = 1 - t_2$, then $(s_1, t_1) \sim (s_3, t_3)$ exactly as in Case (4).

Hence $\sim$ is an equivalence relation.

The corresponding equivalence classes are easily seen to be:

(a) $\{(s, t)\}$ where $0 < s < 1$, $0 < t < 1$,

(b) $\{(0, t), (1, 1 - t)\}$ where $0 \leqslant t \leqslant 1$,

(c) $\{(s, 0), (1 - s, 1)\}$ where $0 \leqslant s \leqslant 1$.

A homeomorphism between different versions of P (1) We prove that the versions of the real projective plane in (c) and (d) of Theorem 15.10 are homeomorphic. Recall that (c) is $D^+/\sim$ where D^+ is the upper hemisphere of the unit sphere in $\mathbb{R}^3$ and $\sim$ identifies each pair of antipodal points on the boundary circle of D^+. On the other hand (d) is $D/\sim$ where D is the unit disc in $\mathbb{R}^2$ and $\sim$ identifies each pair of antipodal points on the boundary circle of D. We define $f : D^+ \to D$ by $f(x, y, z) = (x, y)$. This is continuous since each coordinate function is continuous. If $(x_1, y_1, z_1) \sim (x_2, y_2, z_2)$ then either $x_1 = x_2$, $y_1 = y_2$, $z_1 = z_2$ or $z_1 = z_2 = 0$ and $(x_2, y_2) = (-x_1, -y_1)$. In either case $(x_1, y_1) \sim (x_2, y_2)$. So f respects the identifications and defines a continuous map $g : D^+/\sim \to D/\sim$. A suitable continuous inverse is induced by the map $h : D \to D^+$ defined by $h(x, y) = (x, y, \sqrt{1 - x^2 - y^2})$. It is straightforward to check that this respects identifications, and gives an inverse for g.

Equivalence relation for K (1) We defined $(x_1, y_1) \sim (x_2, y_2)$ for points in $[0, 2\pi] \times [0, \pi]$ iff one of the following holds:

(i) $x_1 = x_2$ and $y_1 = y_2$;

(ii) $\{y_1, y_2\} = \{0, \pi\}$ and $x_2 = 2\pi - x_1$;

(iii) $\{x_1, x_2\} = \{0, 2\pi\}$ and $y_2 = y_1$:

(iv) $x_1, x_2 \in \{0, 2\pi\}$ and $y_1, y_2 \in \{0, \pi\}$.

Reflexivity Each point in $[0, 2\pi] \times [0, \pi]$ is equivalent to itself by (i).

Symmetry Suppose that $(x_1, y_1) \sim (x_2, y_2)$.

Case (1) If $x_1 = x_2$ and $y_1 = y_2$ then $x_2 = x_1$ and $y_2 = y_1$ and $(x_2, y_2) \sim (x_1, y_1)$ by (i).

Case (2) If $\{y_1, y_2\} = \{0, \pi\}$ and $x_2 = 2\pi - x_1$ then $\{y_2, y_1\} = \{0, \pi\}$ and $x_1 = 2\pi - x_2$ so $(x_2, y_2) \sim (x_1, y_1)$ by (ii).

Case (3) If $\{x_1, x_2\} = \{0, 2\pi\}$, $y_2 = y_1$ then $\{x_2, x_1\} = \{0, 2\pi\}$, $y_1 = y_2$ so $(x_2, y_2) \sim (x_1, y_1)$ by (iii).

Case (4) If $x_1, x_2 \in \{0, 2\pi\}$ and $y_1, y_2 \in \{0, \pi\}$ then $x_2, x_1 \in \{0, 2\pi\}$ and $y_2, y_1 \in \{0, \pi\}$ so $(x_2, y_2) \sim (x_1, y_1)$ by (iv).

Transitivity Suppose that $(x_1, y_1) \sim (x_2, y_2)$ and that $(x_2, y_2) \sim (x_3, y_3)$. We want to prove that $(x_1, y_1) \sim (x_3, y_3)$

Case (1) If $x_1 = x_2$ and $y_1 = y_2$ then $(x_1, y_1) = (x_2, y_2) \sim (x_3, y_3)$. Similarly if $x_2 = x_3$ and $y_2 = y_3$ then $(x_1, y_1) \sim (x_2, y_2) = (x_3, y_3)$.

Case (2) If $\{y_1, y_2\} = \{0, \pi\}$ and $x_2 = 2\pi - x_1$ and also $\{y_2, y_3\} = \{0, \pi\}$ and $x_3 = 2\pi - x_2$ then $y_3 = y_1$ and $x_3 = 2\pi - x_2 = 2\pi - (2\pi - x_1) = x_1$ so $(x_1, y_)) \sim (x_3, y_3)$ by (i).

Case (3) If $\{x_1, x_2\} = \{0, 2\pi\}$ and $y_2 = y_1$, and also $\{x_2, x_3\} = \{0, 2\pi\}$ and $y_3 = y_2$ then $x_3 = x_1$ and $y_3 = y_1$ so $(x_1, y_1) \sim (x_3, y_3)$ by (i).

Case (4) If $x_1, x_2 \in \{0, 2\pi\}$ and $y_1, y_2 \in \{0, \pi\}$ and also $x_2, x_3 \in \{0, 2\pi\}$ and $y_1, y_3 \in \{0, \pi\}$ then $x_1, x_3 \in \{0, 2\pi\}$ and $y_1, y_3 \in \{0, \pi\}$ so $(x_1, y_1) \sim (x_3, y_3)$ by (iv).

Case (5a) If $\{y_1, y_2\} = \{0, \pi\}$ and $x_2 = 2\pi - x_1$ and also $\{x_2, x_3\} = \{0, 2\pi\}$ and $y_3 = y_2$ then $\{y_1, y_3\} = \{y_1, y_2\} = \{0, \pi\}$ so $y_1, y_3 \in \{0, \pi\}$ and $x_3 = 2\pi - x_2 = 2\pi - (2\pi - x_1) = x_1$ so $x_1, x_3 \in \{0, 2\pi\}$. Hence $(x_1, y_1) \sim (x_3, y_3)$ by (iv).

Case (5b) Similarly if $\{x_1, x_2\} = \{0, 2\pi\}$, $y_1 = y_2$ and also $\{y_2, y_3\} = \{0, \pi\}$, $x_3 = 2\pi - x_2$ then $(x_1, y_1) \sim (x_3, y_3)$ by (iv).

Case (6a) If $\{y_1, y_2\} = \{0, \pi\}$ and $x_2 = 2\pi - x_1$ and also $x_2, x_3 \in \{0, 2\pi\}$ and $y_2, y_3 \in \{0, \pi\}$ then $x_1, x_3 \in \{0, 2\pi\}$ and $y_1, y_3 \in \{0, \pi\}$ so $(x_1, y_1) \sim (x_3, y_3)$ by (iv).

Case (6b) Similarly if $x_1, x_2 \in \{0, 2\pi\}$ and $y_1, y_2 \in \{0, \pi\}$ and also $\{y_2, y_3\} = \{0, \pi\}$ and $x_3 = 2\pi - x_2$ then $(x_1, y_1) \sim (x_3, y_3)$ by (iv).

Case (7a) If $\{x_1, x_2\} = \{0, 2\pi\}$ and $y_1 = y_2$ and also $x_2, x_3 \in \{0, 2\pi\}$ and $y_2, y_3 \in \{0, \pi\}$ then $x_1, x_3 \in \{0, 2\pi\}$ and $y_1, y_3 \in \{0, \pi\}$ so $(x_1, y_1) \sim (x_3, y_3)$ by (iv).

Case (7b) Similarly if $x_1, x_2 \in \{0, 2\pi\}$ and $y_1, y_2 \in \{0, \pi\}$ and also $\{x_2, x_3\} = \{0, 2\pi\}$ and $y_2 = y_3$ then $(x_1, y_1) \sim (x_3, y_3)$ by (iv).

Again, the corresponding equivalence classes are easily seen to be:

(a) $\{(s, t)\}$ where $0 < s < 2\pi$, $0 < t < \pi$,

(b) $\{(s, 0), (2\pi - s, \pi)\}$ where $0 < s < 2\pi$,

(c) $\{(0, t), (2\pi, t)\}$ where $0 < t < \pi$,

(d) $\{(0, 0), (0, \pi), (2\pi, 0), (2\pi, \pi)\}$.

A non-Hausdorff locally Euclidean space (1) We give a 2-dimensional example, 'the plane with two origins', since the point is to illustrate why the Hausdorff condition does not come free with the locally Euclidean property of surfaces. An entirely similar example can be given of 'the line with two origins'.

To construct our example, we begin with the disjoint union $\pi_0 \sqcup \pi_1$ of two copies of $\mathbb{R}^2$. For convenience we take π_0 to be $\mathbb{R}^2 \times \{0\}$ and π_1 to be $\mathbb{R}^2 \times \{1\}$. We write points of π_0 as $(\boldsymbol{x}, 0)$ and points of π_1 as $(\boldsymbol{x}, 1)$ where $\boldsymbol{x} \in \mathbb{R}^2$. We define an equivalence relation $\sim$ on the disjoint union $\pi_0 \sqcup \pi_1$ by identifying corresponding points except the origins: in other words, let $(\boldsymbol{x}, 0) \sim (\boldsymbol{x}, 1)$ iff $\boldsymbol{x} \neq \mathbf{0}$ (where $\mathbf{0}$ denotes the origin in $\mathbb{R}^2$). It is clear that this *is* an equivalence relation: there is one equivalence class containing two points for each point $\boldsymbol{x} \in \mathbb{R}^2$ other than the origin and there are two singleton equivalence classes, $\{(\mathbf{0}, 0)\}$ and $\{(\mathbf{0}, 1)\}$. Let X denote the quotient space, with the quotient topology, and let $p : \pi_0 \sqcup \pi_1 \to X$ be the natural projection, which we recall is the map that sends any point to the equivalence class to which it belongs. We wish to prove two things about X:

(a) each point in X has a neighbourhood which is homeomorphic to an open disc in $\mathbb{R}^2$;

(b) X is not Hausdorff.

Proof of (a) Let $\tilde{x} \in X$. Choose a point $(\boldsymbol{x}, \varepsilon)$ in $p^{-1}(\tilde{x})$ (there is a choice of $\varepsilon = 0$ or 1 except when $\tilde{x}$ is $p(\mathbf{0}, 0)$ or $p(\mathbf{0}, 1)$ when there is no choice). Let D be any open disc in π_ε centred on $(\boldsymbol{x}, \varepsilon)$, and consider $p|D : D \to p(D)$. This is one-one onto, since no two points in D are equivalent under $\sim$. It is also continuous and open; continuity follows from the definition of the quotient topology on X; we check that it is open. Let $U \subseteq D$ be any open subset of D; to see that $p(U)$ is open in the quotient topology we just have to check that $p^{-1}(p(U))$ is open in $\pi_0 \sqcup \pi_1$. Now $p^{-1}(p(U)$ is the union of U with its mirror image in $\pi_{1-\varepsilon}$; (explicitly, if $U = V \times \{\varepsilon\}$ where V is open in $\mathbb{R}^2$ then by its mirror image we mean the open set $V \times \{1-\varepsilon\}$). Hence $p^{-1}(p(U))$ is open in $\pi_0 \sqcup \pi_1$ as required. Now $p|D$ is a homeomorphism of D onto $p(D)$ (the continuity of the inverse follows since $p|D$ is open).

Proof of (b) Suppose that $\tilde{U}$ and $\tilde{V}$ are any open subsets of X containing $p(\mathbf{0}, 0)$ and $p(\mathbf{0}, 1)$ respectively. Then $p^{-1}(\tilde{U})$ and $p^{-1}(\tilde{V})$ are open in $\pi_0 \sqcup \pi_1$. In particular $p^{-1}(\tilde{U})$ contains the open disc $B_\varepsilon((\mathbf{0}, 0))$ in the plane π_0 for some $\varepsilon > 0$. Now p is onto, so $p(p^{-1}(\tilde{U}))) = \tilde{U}$, hence $\tilde{U}$ contains $p(B_\varepsilon((\mathbf{0}, 0)))$. Similarly $p^{-1}(\tilde{V})$ contains the open disc $B_\delta((\mathbf{0}, 1))$ in the plane π_1 for some $\delta > 0$, and $\tilde{V}$ contains $p(B_\delta((\mathbf{0}, 1)))$. Let $\kappa = \min\{\varepsilon, \delta\}$. Then each point in $p(B_\kappa((\mathbf{0}, 0))) \setminus \{(\mathbf{0}, 0\}$ coincides with the corresponding point in $p(B_\kappa((\mathbf{0}, 1))) \setminus \{(\mathbf{0}, 1\}$. All such points are in $\tilde{U} \cap \tilde{V}$, and the latter is non-empty. So X is not Hausdorff. $\square$

P cannot be embedded in $\mathbb{R}^3$ (1) We give an argument which, though not completely rigorous, is reasonably persuasive. Recall that in Chapter 15 we saw by cutting and pasting that if you remove an open disc from P what you get is a Möbius band. Now suppose for a contradiction that P is embedded in $\mathbb{R}^3$ (meaning that it is homeomorphic to some subspace of $\mathbb{R}^3$). By deforming this embedding a little we assume that some disc D in P lies in some plane. (This is the heuristic part - how do we know we can do this for an arbitrary, possibly weird, embedding?) Now if D has radius r, remove some slightly smaller disc, say with radius $s < r$. As we saw in Chapter 15, what is left is a Möbius band M. Now the part of the disc D remaining is a planar annulus, in other words M has a 'collar' which is an annulus. But we also saw in Chapter 15, in a practical way, that cutting all round near the edge of a Möbius band produces not the disjoint union of an annulus and another Möbius band, but an interlinked pair consisting of a Möbius band and a twisted cylinder. This contradiction shows (at least if you are not too sceptical) that P does not embed in $\mathbb{R}^3$.

Remark I know of no reference for the above argument; the standard proof depends on more advanced algebraic topology. Reg Wood gave me the idea, but his is not the responsibility for errors.

Alternative approach to equivalence classes (1) In this approach we identify the equivalence classes first, and afterwards show that they arise from the given formulae, for example (i) - (iv) on p.1 in the case of the torus. The equivalence classes in the case of the torus are:

(a) one singleton class $\{(s, t)\}$ for each (s, t) with $0 < s < 1$, $0 < t < 1$.

(b) one class containing a pair of points $\{(0, t), (1, t)\}$ for each t such that $0 < t < 1$,

(c) one class containing a pair of points $\{(s, 0), (s, 1)\}$ for each s such that $0 < s < 1$, and

(d) a single class containing four points $\{(0, 0), (1, 0), (0, 1), (1, 1)\}$.

It is clear that this gives a partition of $[0, 1] \times [0, 1]$, and we know that there is a corresponding equivalence relation. To show that it is given by (i) - (iv), we note that an equivalence class of finite size k involves k^2 instances of $a \sim b$. An equivalence class as in (a) above can be traced back to just one instance of $a \sim b$ from (i). An equivalence class of type (b) above involves four instances of $a \sim b$, two from (i) and two from (ii). Similarly an equivalence class of type (c) above involves four instances of $a \sim b$, two from (i) and two from (iii). The equivalence class in (d) above involves sixteen instances of $a \sim b$, four from each of (i) - (iv). This uses up all instances of $a \sim b$, so no two of these equivalence classes coalesce.

The cases of P and K may be treated similarly, with some minor changes.

Supplementary material for Chapter 16

It is possible to give more complicated versions of material in Chapter 16, but to keep things simple we just give

Hints for Exercise 16.9 First we simplify the problem by showing it is enough to consider the case when the pointwise limit is the zero function. For if f_n and f are as in the question, we may set $g_n = f_n - f$. Then g_n is continuous since the f_n and f are. Also, $g_n(x) \geqslant g_{n+1}(x)$ for all $n \in \mathbb{N}$ and all $x \in X$ since this holds for the f_n. Also, (g_n) converges pointwise to the zero function on X, since (f_n) converges pointwise to f on X. Now if we prove that (g_n) converges to the zero function uniformly on X, it will follow that (f_n) converges to f uniformly on X.

So let (g_n) be a monotonic decreasing sequence of continuous functions on X converging pointwise to the zero function. Let $\varepsilon > 0$. For each $x \in X$ there exists $N_x \in \mathbb{N}$ such that $g_n(x) < \varepsilon/2$ whenever $n \geqslant N_x$. Now use continuity of g_{N_x} to see that there is some open subset U_x of X with $x \in U_x$ and such that $g_{N_x}(y) < \varepsilon$ for all $y \in U_x$. Next use monotonicity to see that $g_n(y) < \varepsilon$ whenever $n \geqslant N_x$ and $y \in U_x$. The open cover $\{U_x : x \in X\}$ for X has a finite subcover since X is compact. Take N to be the largest of the finite number of integers N_x corresponding to this finite subcover, and show that $0 \leqslant g_n(x) < \varepsilon$ whenever $n \geqslant N$ and for any $x \in X$.

Supplementary material for Chapter 17

Here is a list of supplementary topics for Chapter 17.

Although the first three topics arise from promises in the book, they are a little sophisticated.

★**Completeness in sequence spaces (1)** This is studied in functional analysis. Here we restrict to proving that $\mathbf{l}^\infty$, $\mathbf{l}^1$ and $\mathbf{l}^2$ are complete. Then as we know from Proposition 17.7 we may deduce completeness of any closed subspace, such as the subspace $\mathbf{c}$ of $\mathbf{l}^\infty$ considered in Example S.6.1. As usual there are slight notational complications in considering Cauchy sequences in sequence spaces. We stick to the notation used in S.6.

Before tackling any completeness proofs, we recall a result from the theory of real sequences:

Lemma S.17.1 If (s_m) is a sequence of real numbers which converges to s, and if $s_m < B$ for all $m \in \mathbb{N}$ then $s \leqslant B$.

The proof is easy, by contradiction. □

Completeness of $\mathbf{l}^\infty$ Let $(\boldsymbol{x}^{(n)})$ be a Cauchy sequence in $\mathbf{l}^\infty$, $\boldsymbol{x}^{(n)} = (x_1^{(n)}, x_2^{(n)}, \ldots, x_i^{(n)}, \ldots)$. For any $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that $\|\, \boldsymbol{x}^{(m)} - \boldsymbol{x}^{(n)} \,\|_\infty < \varepsilon$ whenever $m \geqslant n \geqslant N$. By the definition of the l^∞ norm, for each $i \in \mathbb{N}$ this gives $|x_i^{(m)} - x_i^{(n)}| < \varepsilon$ whenever $m \geqslant n \geqslant N$. So for each fixed $i \in \mathbb{N}$ the sequence $(x_i^{(n)})$ is a Cauchy sequence of real numbers, so converges to some number x_i by completeness of $\mathbb{R}$. We shall prove that $\boldsymbol{x} = (x_1, x_2, \ldots, x_i, \ldots)$ is in $\mathbf{l}^\infty$ and that $(\boldsymbol{x}^{(n)})$ converges to $\boldsymbol{x}$.

Keeping n fixed (and with $n \geqslant N$) and letting $m \to \infty$ in the inequality $|x_i^{(m)} - x_i^{(n)}| < \varepsilon$ whenever $m \geqslant n \geqslant N$, we get by Lemma S.17.1 that $|x_i - x_i^{(n)}| \leqslant \varepsilon$ whenever $n \geqslant N$, for each $i \in \mathbb{N}$. This shows in particular that $\boldsymbol{x} - \boldsymbol{x}^{(N)}$ is in $\mathbf{l}^\infty$. Since we showed in S.5 that $\mathbf{l}^\infty$ is a vector space, and also $\boldsymbol{x}^{(N)}$ is in $\mathbf{l}^\infty$, so is $\boldsymbol{x}$.

Since $|x_i - x_i^{(n)}| \leqslant \varepsilon$ whenever $n \geqslant N$ and for all $i \in \mathbb{N}$, it follows that $\|\, \boldsymbol{x} - \boldsymbol{x}^{(n)} \,\|_\infty \leqslant \varepsilon$ whenever $n \geqslant N$, so $(\boldsymbol{x}^{(n)})$ converges to $\boldsymbol{x}$ in $\mathbf{l}^\infty$, and the latter is complete. □

Completeness of $\mathbf{l^1}$ Let $(\boldsymbol{x^{(n)}})$ be a Cauchy sequence in $\mathbf{l^1}$, where $\boldsymbol{x^{(n)}} = (x_1^{(n)}, x_2^{(n)}, \ldots, x_i^{(n)}, \ldots)$. Then for any $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that $||\, \boldsymbol{x^{(m)}} - \boldsymbol{x^{(n)}} \,||_1 < \varepsilon$ whenever $m \geqslant n \geqslant N$, i.e.

$$\sum_{i=1}^{\infty} |x_i^{(m)} - x_i^{(n)}| < \varepsilon \text{ whenever } m \geqslant n \geqslant N.$$

Then for each fixed $i \in \mathbb{N}$ we certainly have $|x_i^{(m)} - x_i^{(n)}| < \varepsilon$ whenever $m \geqslant n \geqslant N$. So for fixed $i \in \mathbb{N}$ the sequence $(x_i^{(n)})$ is a Cauchy sequence of real numbers, so converges to say x_i in $\mathbb{R}$. Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_i, \ldots)$. We shall prove that $\boldsymbol{x} \in \mathbf{l^1}$ and that $(\boldsymbol{x^{(n)}})$ converges to $\boldsymbol{x}$ in $\mathbf{l^1}$.

First, from

$$\sum_{i=1}^{\infty} |x_i^{(m)} - x_i^{(n)}| < \varepsilon \text{ whenever } m \geqslant n \geqslant N,$$

we get

$$\sum_{i=1}^{k} |x_i^{(m)} - x_i^{(n)}| < \varepsilon \text{ whenever } m \geqslant n \geqslant N \text{ and for any } k \in \mathbb{N}.$$

Letting $m \to \infty$ in this we get for each $k \in \mathbb{N}$ that

$$\sum_{i=1}^{k} |x_i - x_i^{(n)}| \leqslant \varepsilon \text{ whenever } n \geqslant N.$$

Now letting $k \to \infty$ we get

$$\sum_{i=1}^{\infty} |x_i - x_i^{(n)}| \leqslant \varepsilon \text{ whenever } n \geqslant N. \qquad (*)$$

This shows in particular that $\boldsymbol{x} - \boldsymbol{x^{(N)}}$ is in $\mathbf{l^1}$, and since $\boldsymbol{x^{(N)}}$ is in $\mathbf{l^1}$ and since $\mathbf{l^1}$ is a vector space from S.5, we get that $\boldsymbol{x}$ is in $\mathbf{l^1}$.

Now $(\boldsymbol{x^{(n)}})$ converges to $\boldsymbol{x}$ in $\mathbf{l^1}$, since $(*)$ says that $||\, \boldsymbol{x} - \boldsymbol{x^{(n)}} \,||_1 \leqslant \varepsilon$ whenever $n \geqslant N$. □

Completeness of Hilbert space $\mathbf{l^2}$ This is very similar to the above. (In fact there is a similar argument showing that $\mathbf{l^p}$ is complete for all real $p \geqslant 1$). Let $(\boldsymbol{x^{(n)}})$ be a Cauchy sequence in $\mathbf{l^2}$, where $\boldsymbol{x^{(n)}} = (x_1^{(n)}, x_2^{(n)}, \ldots, x_i^{(n)}, \ldots)$. Then for any $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that $||\, \boldsymbol{x^{(m)}} - \boldsymbol{x^{(n)}} \,||_2 < \varepsilon$ whenever $m \geqslant n \geqslant N$, i.e.

$$\sum_{i=1}^{\infty} (x_i^{(m)} - x_i^{(n)})^2 < \varepsilon^2 \text{ whenever } m \geqslant n \geqslant N.$$

Then for each fixed $i \in \mathbb{N}$ we have $(x_i^{(m)} - x_i^{(n)})^2 < \varepsilon^2$ whenever $m \geqslant n \geqslant N$, and $(x_i^{(n)})$ is a Cauchy sequence of real numbers, so converges to say x_i in $\mathbb{R}$. Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_i, \ldots)$. We shall prove that $\boldsymbol{x} \in \mathbf{l^2}$ and that $(\boldsymbol{x^{(n)}})$ converges to $\boldsymbol{x}$ in $\mathbf{l^2}$.

First, from

$$\sum_{i=1}^{\infty}(x_i^{(m)} - x_i^{(n)})^2 < \varepsilon^2 \quad \text{whenever} \quad m \geqslant n \geqslant N,$$

we get

$$\sum_{i=1}^{k}(x_i^{(m)} - x_i^{(n)})^2 < \varepsilon^2 \quad \text{whenever} \quad m \geqslant n \geqslant N \quad \text{and for each } k \in \mathbb{N}.$$

Letting $m \to \infty$ in this we get for each $k \in \mathbb{N}$ that

$$\sum_{i=1}^{k}(x_i - x_i^{(n)})^2 \leqslant \varepsilon^2 \quad \text{whenever} \quad n \geqslant N.$$

Now letting $k \to \infty$ we get

$$\sum_{i=1}^{\infty}(x_i - x_i^{(n)})^2 \leqslant \varepsilon^2 \quad \text{whenever} \quad n \geqslant N.$$

This shows in particular that $\boldsymbol{x} - \boldsymbol{x^{(N)}}$ is in $\mathbf{l^2}$, and since $\boldsymbol{x^{(N)}}$ is also in $\mathbf{l^2}$ which is a vector space from S.5, we get that $\boldsymbol{x}$ is in $\mathbf{l^2}$.

We now show that $(\boldsymbol{x^{(n)}})$ converges to $\boldsymbol{x}$ in $\mathbf{l^2}$. For

$$\sum_{i=1}^{\infty}(x_i - x_i^{(n)})^2 \leqslant \varepsilon^2 \quad \text{whenever} \quad n \geqslant N$$

says that $||\,\boldsymbol{x} - \boldsymbol{x^{(n)}}\,||_2 \leqslant \varepsilon$ whenever $n \geqslant N$ as required. □★

★ **Cantor's and Baire's theorems (1)** The first thing to be said about Cantor's theorem is that it should probably be referred to as Cantor's intersection theorem, since 'Cantor's theorem' often means a result in set theory (roughly, that any set has got more subsets than elements).

Recall Exercise 13.11, that if X is a compact space and for each $n \in \mathbb{N}$ the set V_n is a non-empty closed subset of X such that the sequence (V_n) is 'nested', that is $V_n \supseteq V_{n+1}$ for every $n \in \mathbb{N}$, then the intersection of all the V_n is non-empty.

Cantor's theorem is a little similar.

Theorem S.17.2 If $(X,\, d)$ is a complete metric space and (V_n) is a nested sequence of non-empty closed subsets of X such that diam $V_n \to 0$ as $n \to \infty$, then the intersection of all the V_n is non-empty (and contains exactly one point).

Before proving this, here are a few remarks. The part in brackets is an easy consequence of the fact that the intersection is non-empty and that diam $V_n \to 0$ as $n \to \infty$. For if there were two distinct points $x,\, y$ in this intersection, then diam $V_n \geqslant d(x,\, y)$ for all $n \in \mathbb{N}$ so diam $V_n \to 0$ must be false.

It is perhaps surprising that one uses the hypothesis diam $V_n \to 0$ as $n \to \infty$ for the intersection to be non-empty - you might think that the bigger the sets V_n the more chance that the intersection of all of them would be non-empty. But think of $X = \mathbb{R}$ and $V_n = [n, \infty)$, a non-empty closed subset of the complete metric space $\mathbb{R}$. In this case the intersection of all the V_n is empty. The same can be true when the diameter is finite, though not in Euclidean space, since a closed bounded set there is compact, so the intersection of a nested sequence of non-empty bounded closed sets is non-empty by Exercise 13.11. But for example in $\mathbf{l}^\infty$ let $V_n = \{\mathbf{e}_m : m \geqslant n\}$ where $\mathbf{e_m}$ is the sequence with a 1 in the mth place and zeros elsewhere. Then diam$V_n = 1$ for each n, and each V_n is non-empty and closed in $\mathbf{l}^\infty$, and the V_n are nested, but the intersection of the V_n is empty. Another such example, in the function space $\mathcal{C}[0, 1]$, is given in Exercise 14.16.

We can't just drop completeness of X from the proof: if $X = (0, 1)$ we could take $V_n = (0, 1/n]$, to get a nested sequence of closed subsets of $(0, 1)$ with diam $V_n = 1/n \to 0$ as $n \to \infty$. Yet the intersection of all the V_n is empty.

The proof needs the V_n to be closed also: for suppose $X = \mathbb{R}$ and $V_n = (0, 1/n)$. Then again we have a nested sequence of sets with diam $V_n \to 0$ as $n \to \infty$, but the intersection of all the V_n is empty.

Proof of Cantor's theorem Let X, V_n be as in the statement of the theorem, and let $x_n \in V_n$. Then (x_n) is Cauchy: for given $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that diam $V_N < \varepsilon$. Then for all $m \geqslant n \geqslant N$ we have $x_m, x_n \in V_N$ so $d(x_m, x_n) < \varepsilon$. Since X is complete, (x_n) converges to a point $x \in X$. Now since $x_m \in V_n$ for all $m \geqslant n$ and V_n is closed in X, it follows by Corollary 6.30 that $x \in V_n$. This is true for all $n \in \mathbb{N}$, so x is in the intersection of all the V_n as required. □

The converse to Theorem S.17.2 is also true: if the intersection of any nested sequence (V_n) of closed subsets of a metric space X, with diam$V_n \to 0$ as $n \to \infty$, is non-empty then X is complete. For suppose that X is not complete. Let (x_n) be a Cauchy sequence in X which does not converge. Then (x_n) has no convergent subsequence either, by Lemma 17.10. Then by Corollary 14.13, for each $x \in X$ there exists $\varepsilon_x > 0$ such that $B_{\varepsilon_x}(x)$ contains x_n for only finitely many values of n. In particular the set $\{x_n : n \in \mathbb{N}\}$ has no limit points in X, for if x were a limit point then for every $\varepsilon > 0$ there would be infinitely many distinct points of $\{x_n : n \in \mathbb{N}\}$ in $B_\varepsilon(x)$, let alone infinitely many values of n for which $x_n \in B_\varepsilon(x)$. Now let $V_n = \{x_m : m \geqslant n\}$. Then by the above V_n has no limit points in X, so V_n is closed in X by Proposition 6.17. Also, (V_n) is clearly nested, and diam $V_n \to 0$ as $n \to \infty$ by the Cauchy condition. But the intersection of all the V_n is empty, since any point in the intersection would be a point to which (x_n) converged. (Let $\varepsilon > 0$. There exists $N \in \mathbb{N}$ such that $d(x_m, x_n) < \varepsilon$ for all $m \geqslant n \geqslant N$, so diam $V_n \leqslant \varepsilon$ for all $n \geqslant N$, and if $x \in V_n$ for all n then $d(x, x_n) \leqslant \varepsilon$ for all $n \geqslant N$.)

Baire's theorem comes in various degrees of generality and with various equivalent statements. To keep things simple, we discuss the following form.

Baire's theorem If (X, d) is a complete metric space and (D_n) is a sequence of dense open subsets of X then the intersection of all the D_n is also dense in X.

Note that this time we do not insist that the sequence (D_n) is nested. Also, we do not claim that the intersection of all the D_n is open in X. For example D_n might be $\mathbb{R} \setminus \{q_n\}$ where (q_n) is some listing of all the rational numbers. Then each D_n is open and dense in $\mathbb{R}$ but the intersection of all the D_n is the set of irrational numbers, which is dense but not open in $\mathbb{R}$.

Proof For any $x \in X$ and $r > 0$ we want to show that

$$B_r(x) \cap \bigcap_{n=1}^{\infty} D_n \neq \emptyset.$$

We shall find a sequence of points (y_n) in X and a sequence (r_n) of positive real numbers with $r_n \to 0$ as $n \to \infty$, such that the sequence of closed balls $V_n = \{y \in X : d(y, y_n) \leqslant r_n\}$ is nested and $V_n \subseteq D_n \cap B_r(x)$. We shall then be able to apply Cantor's theorem and find a point in all the V_n, which will be in the desired intersection. Note that $V_n \subseteq B_{2r_n}(y_n)$ for each $n \in \mathbb{N}$, since if $y \in V_n$ then $d(y, y_n) \leqslant r_n < 2r_n$.

First, since D_1 is dense in X there is a point $y_1 \in D_1 \cap B_r(x)$. Since $D_1 \cap B_r(x)$ is open, there exists $r_1 > 0$ such that $B_{2r_1}(y_1) \subseteq D_1 \cap B_r(x)$, so $V_1 \subseteq B_{2r_1}(y_1) \subseteq D_1 \cap B_r(x)$. We may choose the above $r_1 < 1$. Suppose inductively that we have positive real numbers $r_1, r_2, \ldots, r_n$ such that for each $i = 1, 2, \ldots, n$ both $r_i < 1/i$ and $V_i \subseteq D_i \cap B_r(x)$ hold (where V_i is as defined above). Then since D_{n+1} is dense in X there is some point $y_{n+1} \in D_{n+1} \cap B_r(x)$, and since $D_{n+1} \cap B_r(x)$ is open in X there is some r_{n+1}, which we may choose with $r_{n+1} < 1/(n+1)$, such that $B_{2r_{n+1}}(y_{n+1}) \subseteq D_{n+1} \cap B_r(x)$. Now $V_{n+1} \subseteq B_{2r_{n+1}}(y_{n+1}) \subseteq D_{n+1} \cap B_r(x)$. This inductive procedure gives a nested sequence (V_n) of non-empty closed sets with diam $V_n < 2/n$ (since radius $V_n = r_n < 1/n$), and by Cantor's theorem there is a unique point a in all of the V_n. In particular since $V_n \subseteq D_n$, and since $V_n \subseteq B_r(x)$ for all n, we have

$a \in B_r(x) \cap \bigcap_{n=1}^{\infty} D_n$ and the latter is non-empty as required. $\square$

Next we consider an equivalent statement of Baire's theorem, which focuses on the complementary situation.

Definition S.17.3 A subspace A of a topological space X is *nowhere dense* in X if the interior of $\overline{A}$ is empty.

Example S.17.4 The set of integers $\mathbb{Z}$ is nowhere dense in $\mathbb{R}$. For $\overline{\mathbb{Z}} = \mathbb{Z}$ and this has no interior points in $\mathbb{R}$. For a slightly less dull example, consider the Cantor set C defined in Exercise 6.5. We saw there that C is closed; but $\mathring{C} = \emptyset$ since no open interval is contained in C - some 'open middle third' has been removed from it. So C is nowhere dense in $\mathbb{R}$.

The next result explains the complementary relationship of 'dense' and 'nowhere dense'.

Proposition S.17.5 If A is a nowhere dense set in a topological space X then $X \setminus \overline{A}$ is dense in X. If U is a dense open subset of a topological space X then $X \setminus U$ is nowhere dense in X.

Proof Suppose that A is nowhere dense in X. Then no non-empty open set is contained in $\overline{A}$ so every non-empty open set has non-empty intersection with $X \setminus \overline{A}$, which says that $X \setminus \overline{A}$ is dense in X.

Suppose that U is dense and open in X. Then $X \setminus U$ is closed in X, so $\overline{X \setminus U} = X \setminus U$. Also, $X \setminus U$ has no interior points since any open subset of X has non-empty intersection with U and hence is not contained in $X \setminus U$. So $X \setminus U$ is nowhere dense in X. □

In view of the above we can re-state Baire's theorem to say that in a complete metric space the union of a sequence of nowhere dense closed sets has dense complement. For if A_n is closed and nowhere dense in X then $X \setminus A_n$ is dense and open, so the intersection of all of them is dense, and this is the complement of the union of all the A_n.

Conversely if we know that the union of a sequence of nowhere dense closed sets in X has dense complement, then by taking complements the intersection of a sequence of dense open sets in X is dense.

A more general approach is to define a *Baire space* to be a topological space such that the intersection of any sequence of dense open subsets is dense. Then 'our' Baire's theorem says that any complete metric space is a Baire space.

We conclude this section by describing one or two of the intriguing applications of Baire's theorem in analysis. We begin with two lemmas.

Lemma S.17.6 If A is a dense subset of a Hausdorff space X and $a \in A$ then $A \setminus \{a\}$ is dense in X.

Proof Since A is dense in X, given any open subset U of X we have $U \cap A \neq \emptyset$. Now X is Hausdorff, so $\{a\}$ is closed in X, and $X \setminus \{a\}$ is open in X. Hence for any open subset W of

X, $U = W \cap (X \setminus \{a\})$ is also open, so $U \cap A \neq \emptyset$. Hence

$$W \cap (A \setminus \{a\}) = (W \cap (X \setminus \{a\})) \cap A = U \cap A \neq \emptyset,$$

so $A \setminus \{a\}$ is dense in X. $\square$

This result extends by induction to removing any finite set of points from A. We do need some condition such as Hausdorff to ensure that a singleton set is closed in X - for example in the space X of Exercise 7.4 the singleton $\{1\}$ is dense in X, so we can't remove a point from it and still have a dense set!

Lemma S.17.7 The set $\mathbb{Q}$ of rational numbers is not the intersection of any sequence of open subsets U_n of $\mathbb{R}$.

Proof Suppose for a contradiction that $\mathbb{Q} = \bigcap_{n=1}^{\infty} U_n$ where U_n is open in $\mathbb{R}$. Then each U_n is also dense in $\mathbb{R}$, since $\mathbb{Q} \subseteq U_n$ and $\mathbb{Q}$ is dense in $\mathbb{R}$. Let (q_n) be an enumeration of $\mathbb{Q}$ and consider the sets $V_n = U_n \setminus \{q_n\}$. Then V_n is open in $\mathbb{R}$ and also dense in $\mathbb{R}$ by Lemma S.17.6. But consider the intersection of all the V_n. This is empty, so it is certainly not dense in $\mathbb{R}$, contradicting Baire's theorem. $\square$

Remark On the other hand, the set of irrational numbers is the countable intersection of the open sets $\mathbb{R} \setminus \{q_n\}$ where q_n ranges over $\mathbb{Q}$.

Proposition S.17.8 There does not exist a function $f : \mathbb{R} \to \mathbb{R}$ which is continuous at all rational numbers and discontinuous at all irrational numbers.

Proof Let $f : \mathbb{R} \to \mathbb{R}$ be any function. For each $n \in \mathbb{N}$, put

$U_n = \{a \in \mathbb{R} :$ there exists $\delta_a > 0$ with $|f(x) - f(y)| < 1/n$ whenever $x, y \in (a - \delta_a, a + \delta_a)\}$.

We shall prove two facts:

(a) U_n is open in $\mathbb{R}$.

(b) the set of points at which f is continuous is the intersection of all the U_n.

Proof of (a) Let $a \in U_n$. Then for some $\delta_a > 0$ we have $|f(x) - f(y)| < 1/n$ whenever $x, y \in (a - \delta_a, a + \delta_a)$. Let $b \in (a - \delta_a, a + \delta_a)$. We may take $\delta_b = \min\{a + \delta_a - b, b - (a - \delta_a)\}$ so that $(b - \delta_b, b + \delta_b) \subseteq (a - \delta, a + \delta_a)$. Then $|f(x) - f(y)| < 1/n$ whenever $x, y \in (b - \delta_b, b + \delta_b)$, which says $b \in U_n$. Hence $(a - \delta_a, a + \delta_a) \subseteq U_n$. This shows that U_n is open in $\mathbb{R}$.

Proof of (b) Any point a at which f is continuous is in U_n for all n, since given any $n \in \mathbb{N}$ there exists a $\delta_a > 0$ such that $|f(x) - f(a)| < 1/2n$ whenever $|x - a| < \delta_a$, and it follows that $|f(x) - f(y)| \leqslant |f(x) - f(a)| + |f(a) - f(y)| < 1/n$ whenever $x, y \in (a - \delta_a, a + \delta_a)$. On the other hand, if f is not continuous at a, then for some $\varepsilon > 0$ there is no $\delta > 0$ such that $|f(x) - f(a)| < \varepsilon$

whenever $|x - a| < \delta$. Choose $n \in \mathbb{N}$ such that $1/n < \varepsilon$. Then there is no $\delta > 0$ such that $|f(x) - f(a)| < 1/n$ whenever $x \in (a - \delta, a + \delta)$, which says that $a \notin U_n$. Hence the set of points of continuity of f is the intersection of all the U_n.

By (a), (b) and Lemma S.17.7, the set of points at which f is continuous is not $\mathbb{Q}$. □

Remark However, there does exist a function $f : \mathbb{R} \to \mathbb{R}$ such that f is continuous precisely at the irrational numbers. Suppose that $q_1, q_2, \ldots$ is a listing of all rational numbers, and define

$$f(x) = \begin{cases} 0 & \text{if } x \text{ is irrational} \\ 1/n & \text{if } x = q_n. \end{cases}$$

It is straightforward to show that f is continuous at a iff a is irrational.

One can also use Baire's theorem to establish the existence of a continuous real-valued function on [0, 1] which is nowhere differentiable - in fact the set of such functions is dense in the space $\mathcal{C}[0, 1]$ of continuous real-valued functions on [0, 1] with the sup metric. This is proved as follows: we know that $\mathcal{C}[0, 1]$ with the sup metric is complete, so Baire's theorem applies. We find that the set of continuous but nowhere differentiable functions in $\mathcal{C}[0, 1]$ is the intersection of a sequence of open dense sets U_n and hence is dense. The details are a little technical and we omit them.★

★ **Another application of Banach's fixed-point theorem (1)** The applications of Banach's theorem in Chapter 17 concern differential and integral equations. Here is an application to an inverse function theorem. This section requires some knowledge of calculus for several variables. We consider maps from one Euclidean space to another, where each has the Euclidean norm, which we denote here by just $\| \ \|$. Let $f : U \to \mathbb{R}^n$ be a map, where $U \subseteq \mathbb{R}^m$ is an open subset. Then we say that f is differentiable at a point $a \in U$ if there exists a linear map $Df_a : \mathbb{R}^m \to \mathbb{R}^n$ such that for $h \in \mathbb{R}^m$ with $h \neq 0$ and $a + h \in U$,

$$f(a + h) - f(a) = Df_a(h) + k, \quad \text{where } k \to 0 \ \text{ as } h \to 0.$$

In this situation, Df_a is often called 'the total derivative of f at a', to distinguish it from partial derivatives. The idea of the definition is that, near a, the linear map Df_a is a reasonably good approximation to the map f itself. We sometimes write $(Df)_a$ for clarity.

When f is differentiable at every point of U, $x \mapsto Df_x$ defines a map $D : U \to Lin(\mathbb{R}^m, \mathbb{R}^n)$, where the second space is the space of linear transformations of the vector space $\mathbb{R}^m$ to the vector space $\mathbb{R}^n$. This space is itself a vector space (of dimension mn) and may be given the sup norm $\| \ \|_\infty$. If the map D is continuous we say that f is *of class* C^1 *on* U. Here are a few standard results from calculus of several variables which we give without proofs. We shall also assume that functions of class C^1 on U are closed under addition and scalar multiplication.

Theorem S.17.9 (Chain rule) Let U, V be open sets in $\mathbb{R}^m$, $\mathbb{R}^n$ respectively. If $f : U \to \mathbb{R}^n$ is differentiable at $a \in U$ and $g : V \to \mathbb{R}^p$ is differentiable at $f(a) \in V$, then the composition $g \circ f$ is differentiable at a, and $D(g \circ f)_a = (Dg)_{f(a)} \circ Df_a$.

This says that the best linear approximation to a composition (near a point) is the composition of the best approximations, which is very plausible.

Proposition S.17.10 If $f : \mathbb{R}^m \to \mathbb{R}^n$ is given by $f(x) = L(x) + b$ where $L : \mathbb{R}^m \to \mathbb{R}^n$ is linear and $b \in \mathbb{R}^n$, then f is of class C^1 everywhere, and $Df_x = L$ for all $x \in \mathbb{R}^n$.

This roughly says that a linear map is its own best local linear approximation.

Theorem S.17.11 (Mean value inequality) Suppose that U is an open set in $\mathbb{R}^m$. If $f : U \to \mathbb{R}^n$ is of class C^1 on U and $||Df_x||_\infty \leqslant B$ for all $x \in U$, then for any x, $y \in U$,

$$||f(x) - f(y)|| \leqslant B||x - y||.$$

The inverse function theorem mentioned above concerns a map $f : U \to \mathbb{R}^n$ where U is an open subset of $\mathbb{R}^n$ and f is of class C^1 on U.

Theorem S.17.12 (Inverse function theorem) If $f : U \to \mathbb{R}^n$ is of class C^1 on $U \subset \mathbb{R}^n$, and for some $a \in U$ the linear transformation $Df_a : \mathbb{R}^n \to \mathbb{R}^n$ is invertible, then there is an open set U' such that $a \in U' \subseteq U$ and $f|U'$ is bijective from U' to $f(U')$ and f^{-1} is of class C^1 on $f(U')$. Moreover, $D(f^{-1})_{f(a)} = (Df_a)^{-1}$.

Remark Before getting into the technical details of the proof, here is one point of view on what the theorem says: what holds true infinitesimally at a point (i.e. for the derivative, or geometrically at the tangent space level) also holds locally (i.e. for a small enough neighbourhood of the point). This holds for other contexts, for example the implicit function theorem.

Proof We assume first that $a = 0$, $f(0) = 0$, and Df_0 is the identity transformation. (We later show how to deduce the general case.)

Let $g(x) = f(x) - x$. For a fixed $y \in \mathbb{R}^n$ we have $f(x) = y$ iff $g(x) + x = y$ iff $y - g(x) = x$ iff $F_y(x) = x$ where $F_y(x) = y - g(x)$. This converts the equation $f(x) = y$ to the fixed-point equation $F_y(x) = x$. We prove that if y is close enough to 0 then F_y is a contraction on a small enough closed ball around 0.

Now $D(F_y)_0 = -(Dg)_0 = -(Df)_0 + I = 0$. Since f and therefore F_y is of class C^1, there is some $r' > 0$ such that $||D(F_y)_x||_\infty < 1/2$ for all x in the open ball $B_{r'}(0)$. By Proposition S.17.11 (the mean value inequality) it follows that for all x, $x' \in B_{r'}(0)$ we have $||F_y(x) - F_y(x')|| \leqslant ||x - x'||/2$. In particular $||F_y(x) - F_y(0)|| \leqslant ||x||/2$. Now take r such that $0 < r < r'$. Write $V_r(0)$ for the closed ball of radius r around 0.

Suppose that $x \in V_r(0)$ and $y \in B_{r/2}(0)$. Then

$$||F_y(x)|| \leqslant ||F_y(x) - F_y(0)|| + ||F_y(0)|| \leqslant ||x||/2 + ||y|| < r/2 + r/2 = r.$$

So for $y \in B_{r/2}(0)$ the map F_y takes $V_r(0)$ into itself, and is a contraction on $V_r(0)$ because of the condition $||F_y(x) - F_y(x')|| \leqslant ||x - x'||/2$. Now $V_r(0)$ is a closed subspace of the complete metric space $\mathbb{R}^n$, so $V_r(0)$ is complete. Hence by Banach's fixed-point theorem, for $y \in B_{r/2}(0)$ there is a unique fixed point for F_y in $V_r(0)$ which we label $h(y)$. In fact $h(y) \in B_r(0)$. To see this, let $y \in B_{r/2}(0)$ and $x = h(y) \in V_r(0)$. Note that then $f(x) = y$ by the definition of F_y. Then $||x - f(x)|| = ||f(x) - x|| = ||g(x)|| = ||y - F_y(x)|| = ||F_y(0) - F_y(x)|| \leqslant ||x||/2$, hence

$$r/2 > ||y|| \geqslant ||x|| - ||x - y|| = ||x|| - ||x - f(x)|| \geqslant ||x|| - ||x||/2 = ||x||/2 = ||h(y)||/2,$$

so $||h(y)|| < r$. Now put $U' = f^{-1}(B_{r/2}(0)) \cap B_r(0)$. This is an open subset of $\mathbb{R}^n$ since f is continuous. Also, h maps $B_{r/2}(0)$ into U', and for $y \in B_{r/2}(0)$ we have $f(h(y)) = y$. Finally for $x \in U'$ we have $h(f(x)) = x$ since $f(x) = y \in B_{r/2}(0)$ so x is the unique point in $V_r(0)$ such that $f(x) = y$.

At this point we have established that $f|U' : U' \to B_{r/2}(0)$ is one-one onto, with set-theoretic inverse h.

We next show that h is continuous on $B_{r/2}(0)$. For if $y,\, y' \in B_{r/2}(0)$ then, bearing in mind that $h(y)$ is a fixed point for F_y and that $F_y(h(y)) = y - g(h(y))$,

$$||h(y) - h(y')|| = ||F_y(h(y)) - F_{y'}(h(y'))|| \leqslant ||g(h(y)) - g(h(y'))|| + ||y - y'||.$$

Now $F_y(x) = y - g(x)$ so $Dg_x = -D(F_y)_x$ and just as for F_y we get $||g(x) - g(x')|| \leqslant ||x - x'||/2$ for $x,\, x' \in B_r(0)$, and since $h(y),\, h(y') \in B_r(0)$ for $y,\, y' \in B_{r/2}(0)$ we get: for $y,\, y' \in B_{r/2}(0)$, $||h(y) - h(y')|| \leqslant ||h(y) - h(y')||/2 + ||y - y'||$, from which it follows that $||h(y) - h(y')|| \leqslant 2||y - y'||$, so h is continuous on $B_{r/2}(0)$.

Finally we show that h is differentiable on $B_{r/2}(0)$. For points $y,\, y' \in B_{r/2}(0)$ let $x = h(y)$ and $x' = h(y')$. Then

$$||h(y) - h(y') - (Df_x)^{-1}(y - y')|| = ||x - x' - (Df_x)^{-1}(f(x) - f(x')||$$

$$\leqslant ||Df_x||_\infty . ||(Df_x)^{-1}(x - x') - (f(x) - f(x'))||.$$

Since h is continuous, $||x - x'|| \to 0$ as $||y - y'|| \to 0$ so $(Df_x)^{-1}(x - x') \to 0$ as $||y - y'|| \to 0$, and also $||f(x) - f(x')|| \to 0$ as $||y - y'|| \to 0$ since f is differentiable and hence continuous at x. Now we have $||h(y) - h(y') - (Df_x)^{-1}(y - y')|| \to 0$ as $||y - y'|| \to 0$, showing that h is differentiable at y with derivative there $(Df_{h(y)})^{-1}$.

Now consider the general case when $a \in U$ and Df_a is just invertible rather than being the identity. Let k be the composition $T_{-f(a)} \circ f \circ T_a$ where T_b is the translation defined by $T_b(x) = x + b$. Then $k(0) = 0$ and k is differentiable with $Dk_0 = Df_a$, using the chain rule Proposition S.17.9 and also Proposition S.17.10. Finally consider $j = (Dk_0)^{-1} \circ k$. Then $j(0) = 0$ and again using Propositions S.17.9 and S.17.10 we get Dj_0 is the identity transformation. So the above special case of the inverse function theorem gives a local differentiable inverse i for j, and from this we get a differentiable inverse for f near a (explicitly, this inverse is the composition $T_a \circ i \circ (Dk_0)^{-1} \circ T_{-f(a)}$). □

Hints for Exercise 17.5 The answers to the question are Yes, No, No. For (a), suppose a sequence (x_n) is Cauchy in the sense of this new metric. Then (x_n^3) is Cauchy in the usual metric, so it converges to y say. Let $x = y^{1/3}$ and show that (x_n) converges to x in this new metric. For a negative result in (b), consider the sequence $(-n)$. Show that (e^{-n}) is Cauchy in the usual sense, so $(-n)$ is Cauchy in the new metric. Show that if $(-n)$ converged to a in the new metric then $\mathrm{e}^a = 0$, so get a contradiction. Similar argument for (c), using the sequence $(\tan^{-1}(n))$.

Hints for Exercise 17.6 The individual parts are not hard - it's mainly a matter of applying the definitions accurately. For example in (c) let (x_n) be a Cauchy sequence in X. Then $(f(x_n))$ is a Cauchy sequence in X' by (a). But X' is complete so $(f(x_n))$ is convergent in X', hence (x_n) is convergent in X by (b). Hence X is complete.

Hints for Exercise 17.15 Show that the map $x \mapsto d(x, f(x))$ is continuous on compact X so attains its inf, say l, at some point $x_0 \in X$. Now show that $l = 0$ by getting a contradiction from the given condition $d(f(x), f(y)) < d(x, y)$ if $l > 0$. Thus x_0 is a fixed point of f. Uniqueness follows as in the contraction map theorem.

Compact subspaces of function spaces

The aim of this file is to exhibit interesting examples of compact spaces which are not subspaces of Euclidean space. One natural place to look is in function spaces. We concentrate on function spaces with the sup metric. First we study a general criterion for compactness in metric spaces. This involves the introduction of the concept of 'total boundedness' which as the name suggests is stronger than boundedness. We use this to prove a hard theorem (the Arzelà-Ascoli theorem) about compactness in function spaces with the sup metric, and we end with some examples. So there are three sections, all of which count as (3), 'enhancement'.

A general criterion for compactness of a metric space

Definition C.1.1 A metric space X is said to be *totally bounded* (or *pre-compact*) if for any $\varepsilon > 0$ there exists a finite ε-net for X.

Recall the definition of an ε-net from Chapter 14 of the book: a subset $S \subseteq X$ is an ε-net for a metric space (X, d) if for every $x \in X$ there is some $p \in S$ such that $d(x, p) < \varepsilon$.

The eventual goal of this section is to show that a metric space is compact iff it is complete and totally bounded. But first we explore the concept of total boundedness.

Proposition C.1.2 Any totally bounded metric space X is bounded.

Proof Suppose that $\{x_1, x_2, \ldots, x_r\}$ is a finite 1-net for X. Then $X = \bigcup_{i=1}^{r} B_1(x_i)$. Now each $B_i(x_i)$ is bounded, and a finite union of bounded sets is bounded (Proposition 5.26). □

The converse of Proposition C.1.2 is not true in general, though it is in a Euclidean space. Before proving that, it is convenient to look at total boundedness for a subspace Y of a metric space X. It turns out that we don't need to insist that finite ε-nets live in Y, they can just be in X.

Lemma C.1.3 A subspace Y of a metric space (X, d) is totally bounded iff for any $\varepsilon > 0$ there exists a finite subset $F \subseteq X$ such that $Y \subseteq \bigcup_{x \in F} B_\varepsilon(x)$.

Proof One way around is easy: if Y is totally bounded then for each $\varepsilon > 0$ there is a finite set $F \subseteq Y$ such that $Y = \bigcup_{y \in F} B_\varepsilon^{d_Y}(y)$ where d_Y is the restriction of d to Y. Then $F \subseteq X$ and $B_\varepsilon^{d_Y}(y) \subseteq B_\varepsilon^{d}(y)$, and it follows that $Y \subseteq \bigcup_{y \in F} B_\varepsilon^{d}(y)$.

Conversely suppose that for each $\varepsilon > 0$ there exists a finite subset $F \subseteq X$ such that $Y \subseteq \bigcup_{x \in F} B_\varepsilon(x)$. For any $\varepsilon > 0$ let $F = \{x_1, x_2, \ldots, x_r\} \subseteq X$ be such that $Y \subseteq \bigcup_{i=1}^{r} B_{\varepsilon/2}(x_i)$.

We may assume that for each $x_j \in F$ the set $Y \cap B_{\varepsilon/2}(x_j)$ is non-empty, otherwise we may throw x_j out of F and still have $Y \subseteq \bigcup_{1 \leqslant i \leqslant r,\, i \neq j} B_{\varepsilon/2}(x_i)$. Now choose a point $y_i \in Y \cap B_{\varepsilon/2}(x_i)$ for each i. Then $\{y_1, y_2, \ldots, y_r\}$ is an ε-net for Y, since given any $y \in Y$ we have $y \in B_{\varepsilon/2}(x_i)$ for some i, and $d(y, y_i) \leqslant d(y, x_i) + d(x_i, y_i) < \varepsilon$. □

Corollary C.1.4 Any subspace of a totally bounded metric space is totally bounded.

Now we are ready to prove the promised result about Euclidean spaces.

Proposition C.1.5 A bounded subspace of $\mathbb{R}^n$ is totally bounded.

Proof By Corollary C.1.4 it is enough to show that $[a, b] \times [a, b] \times \ldots \times [a, b]$ (n copies) is totally bounded since any bounded subset X of $\mathbb{R}^n$ is contained in such a set. First, for an interval $[a, b]$, and for any $\varepsilon > 0$ we may let n be the greatest integer with $n\varepsilon < b - a$ and then the set $\{a, a + \varepsilon, a + 2\varepsilon, \ldots, a + n\varepsilon\}$ is a finite ε-net for $[a, b]$. It is now enough to apply inductively the 'if' part of the next result.

Lemma C.1.6 The product $X \times Y$ of two metric spaces (X, d_X), (Y, d_Y) is totally bounded iff both X and Y are totally bounded.

Proof Let us use d_1 as choice of product metric on $X \times Y$. Recall that

$$d_1((x, y), (x', y')) = d_X(x, x') + d_Y(y, y').$$

First suppose that $X \times Y$ is totally bounded. Then for any $\varepsilon > 0$ there exists a finite subset $\{(x_1, y_1), (x_2, y_2), \ldots, (x_r, y_r)\}$ of $X \times Y$ such that any point $(x, y) \in X \times Y$ is within distance ε of some (x_i, y_i). Then $\{x_1, x_2, \ldots x_r\}$ is a finite ε-net for X, since for any $x \in X$, take any $y \in Y$ and we know there exists $i \in \{1, 2, \ldots, r\}$ such that $d_1((x, y), (x_i, y_i)) < \varepsilon$, so $d_X(x, x_i) < \varepsilon$. Hence $\{x_1, x_2, \ldots, x_r\}$ is a finite ε-net for X. Similarly $\{y_1, y_2, \ldots, y_r\}$ is a finite ε-net for Y.

Conversely suppose that X and Y are both totally bounded. Given $\varepsilon > 0$ let $\{x_1, x_2, \ldots, x_r\}$, $\{y_1, y_2, \ldots, y_s\}$ be $\varepsilon/2$-nets for X, Y respectively. Then $\{(x_i, y_j) : 1 \leqslant i \leqslant r,\ 1 \leqslant j \leqslant s\}$ is an ε-net for $X \times Y$: for given any $(x, y) \in X \times Y$, we have $d_X(x, x_i) < \varepsilon/2$ and $d_Y(y, y_j) < \varepsilon/2$ for some i, j, and then
$d_1((x, y), (x_i, y_j)) = d_X(x, x_i) + d_Y(y, y_j) < \varepsilon$. □

To clarify the concept of total boundedness, we give several examples of metric spaces which are bounded but not totally bounded. The first example is a 'silly' one. The second illustrates that the situation depends on choice of metric, and the third and fourth are harder examples, from sequence spaces and function spaces respectively.

Example C.1.7 Any infinite discrete metric space (X, d) is bounded but not totally bounded. For choose any $x_0 \in X$. Then $d(x, x_0) \leqslant 1$ for all $x \in X$, so X is bounded. But there is no finite 1-net. For suppose $\{x_1, x_2, \ldots, x_r\}$ were a 1-net. Then for every $x \in X$ we would have $d(x, x_i) < 1$ for some $i \in \{1, 2, \ldots, r\}$. But the metric is discrete, so this would imply that $x = x_i$. Hence we should have at most r distinct points in X, contradicting its infiniteness.

Example C.1.8 Consider the set $\mathbb{R}$ with the metric d defined by $d(x, y) = \min\{1, |x - y|\}$ (cf. Exercise 5.12). The real line with its usual metric is not bounded and hence not totally bounded by Proposition C.1.2. But with the metric d it is bounded: for given any $x \in \mathbb{R}$ we have $d(x, 0) \leqslant 1$. But it is not totally bounded: we can show it has no finite 1-net. First, for any $x, y \in \mathbb{R}$, if $|x - y| < 1$ then $d(x, y) = |x - y|$. So $B_1^d(x) = (x - 1, x + 1)$. It follows that there cannot be any finite 1-net $\{x_1, x_2, \ldots, x_r\}$ since

$$\bigcup_{i=1}^{r} B_1^d(x_i) = \bigcup_{i=1}^{r} (x_i - 1, x_i + 1) \neq \mathbb{R}.$$

Example C.1.9 The unit cube C in $\mathbf{l}^\infty$ is bounded but not totally bounded.

By definition $C = \{\, \boldsymbol{x} = (x_1, x_2, \ldots, x_n, \ldots) \in \mathbf{l}^\infty : |x_n| \leqslant 1 \text{ for all } n \in \mathbb{N}\}$. Then C is certainly bounded, since $d_\infty(\boldsymbol{x}, \mathbf{0}) \leqslant 1$ for all $\boldsymbol{x} \in C$. But C is not totally bounded; there cannot exist a finite 1/2-net, for suppose that $\{\, \boldsymbol{x^{(1)}}, \boldsymbol{x^{(2)}}, \ldots, \boldsymbol{x^{(r)}} \,\}$ is a 1/2-net for C. For each $n \in \mathbb{N}$ let $\boldsymbol{e^{(n)}} \in \mathbf{l}^\infty$ be the sequence whose nth entry is 1 and all other entries 0. Then for some $s \in \{1, 2, \ldots, r\}$ we have $||\, \boldsymbol{x^{(s)}} - \boldsymbol{e_n} \,||_\infty < 1/2$. Hence the nth entry $x_n^{(s)}$ in $\boldsymbol{x^{(s)}}$ must satisfy $x_n^{(s)} > 1/2$ while the mth entry satisfies $x_m^{(s)} < 1/2$ for all $m \neq n$. But since there are infinitely many $\boldsymbol{e_n}$ and finitely many points in $\{\, \boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_r} \,\}$, for some $j \in \{1, 2, \ldots, r\}$ we must have $||\, \boldsymbol{x^{(j)}} - \boldsymbol{e^{(n)}} \,||_\infty < 1/2$ and $||\, \boldsymbol{x^{(j)}} - \boldsymbol{e^{(m)}} \,||_\infty < 1/2$ for distinct m and n. But this leads to the contradiction $x_n^{(j)} > 1/2$ and $x_n^{(j)} < 1/2$.

Example C.1.10 This is similar to Example C.1.9. Let B be the closed unit ball in the space $\mathcal{C}[0, 1]$ of continuous real-valued functions on $[0, 1]$, with the sup metric d_∞. Explicitly we have $B = \{f \in \mathcal{C}[0, 1] : d_\infty(f, 0) \leqslant 1\}$. This space is clearly bounded. But now let f_n be the functions used in Example 14.23. Each of these lies in B. Recall that for distinct m, n the functions f_m, f_n satisfy $d_\infty(f_m, f_n) = 1$. Suppose that there were a finite $1/2$-net $\{g_1, g_2, \ldots, g_r\}$ for B. Then for some $i \in \{1, 2, \ldots, r\}$ we should have $d_\infty(f_m, g_i) < 1/2$ and $d_\infty(f_n, g_i) < 1/2$ for some $m \neq n$. But then by the triangle inequality $d_\infty(f_m, f_n) \leqslant d(f_m, g_i) + d_\infty(g_i, f_n) < 1$. This contradiction shows there is no finite $1/2$-net for B.

The proof of the next result is similar to the proof of Lemma C.1.3.

Proposition C.1.11 If Y is a totally bounded subspace of a metric space X then $\overline{Y}$ is also totally bounded.

Proof Let $\varepsilon > 0$ and take $\{x_1, x_2, \ldots, x_r\}$ to be an $\varepsilon/2$-net for Y. Then for any $x \in \overline{Y}$, by definition of closure there is a point of Y within distance $\varepsilon/2$ of x and hence a point of $\{x_1, x_2, \ldots, x_r\}$ within distance ε of x, so $\{x_1, x_2, \ldots, x_r\}$ is an ε-net for $\overline{Y}$. □

The next result links total boundedness to completeness.

Proposition C.1.12 A metric space (X, d) is totally bounded iff every sequence (x_n) in X has a Cauchy subsequence.

Proof First suppose that X is totally bounded, and let (x_n) is a sequence in X. The argument to show that (x_n) has a Cauchy subsequence is a generalized bisection argument. First, there is a finite $1/2$-net for X, so X is covered by a finite number of open $1/2$-balls. At least one of these must contain x_n for infinitely many values of n. So there is a subsequence, say $(x_{n(r,1)})_{r=1}^\infty$ of (x_n), such that all the $x_{n(r,1)}$ belong to a single open $1/2$-ball say B_1, so $d(x_{n(r,1)}, x_{n(s,1)}) < 1$ for all $r, s \in \mathbb{N}$. Now B_1 is again totally bounded by Corollary C.1.10, so it may be covered by a finite number of open $1/4$-balls. At least one of these must contain $x_{n(r,1)}$ for infinitely many values of r, so $(x_{n(r,1)})$ has a subsequence, say $(x_{n(r,2)})$, all of whose members are in the same $1/4$-ball so are within distance $1/2$ of one another. Inductively suppose we have subsequences $(x_{n(r,1)}), (x_{n(r,2)}), \ldots (x_{n(r,i)})$ of (x_n) such that each is a subsequence of any preceding subsequences, and for $j = 1, 2, \ldots, i$ the members of $(x_{n(r,j)})$ are all within distance $1/2^{j-1}$ of one another. Consider the 'diagonal' subsequence $(x_{n(r,r)})$. This is a subsequence of $(x_{n(r,m)})$ for all $r \geqslant m$, so eventually all its members are within distance $1/2^{m-1}$ of one another. Hence $(x_{n(r,r)})$ is a Cauchy subsequence of (x_n).

Suppose on the other hand that X is not totally bounded. We may construct a sequence in X which has no Cauchy subsequence exactly as in Proposition 14.21. This completes the proof. □

We may now attain our goal of showing that a metric space is compact iff it is complete and totally bounded. One way around is easy.

Proposition C.1.13 If X is a compact metric space it is complete and totally bounded.

Proof Proposition 17.9 showed that if X is compact metric then it is complete. To show that X is totally bounded we can either use the 'open covering' approach to compactness or equivalently sequential compactness. For any $\varepsilon > 0$ the collection $\{B_\varepsilon(x) : x \in X\}$ is an open cover of X and the existence of a finite cover shows there is a finite ε-net. Alternatively by sequential compactness any sequence (x_n) in X has a convergent, hence Cauchy, subsequence and now total boundedness follows by Proposition C.1.12. □

We have also done the groundwork for proving the converse

Theorem C.1.14 Any complete totally bounded metric X space is compact.

Proof Let (x_n) be any sequence in X. By Proposition C.1.12 this has a Cauchy subsequence, and by completeness that subseqence converges to a point of X. Hence X is (sequentially) compact. □

Corollary C.1.15 The closure $\overline{Y}$ of a totally bounded subspace Y of a complete metric space is compact.

Proof It follows by Proposition 17.7 that $\overline{Y}$ is complete, and by Corollary C.1.10 that it is totally bounded. The result now follows from Theorem C.1.14. □

This result explains the alternative name 'precompactness' for total boundedness. If a totally bounded metric space X is not already compact, then its completion $\hat{X}$, in the sense of C.2, is compact. To see this we need to anticipate from C.2 the existence of $\hat{X}$ and the fact that X is dense in $\hat{X}$ so $\hat{X}$ is compact by Corollary C.1.15.

To complete this survey of total boundedness we show that it is invariant under uniform equivalence but not under homeomorphism.

Proposition C.1.16 If $f : X \to Y$ is a uniform equivalence of metric spaces (X, d_X), (Y, d_Y) then X is totally bounded iff Y is totally bounded.

Proof Recall that f is a bijection such that both f and f^{-1} are uniformly continuous. Suppose that X is totally bounded and let $\varepsilon > 0$. By uniform continuity of f there exists $\delta > 0$ such that $d_Y(f(x)\, f(x')) < \varepsilon$ whenever $d_X(x,\, x') < \delta$. Let $\{x_1,\, x_2,\, \ldots,\, x_n\}$ be a δ-net for X. Then $\{f(x_1),\, f(x_2),\, \ldots,\, f(x_n\}$ is an ε-net for Y: for given $y \in Y$ we know $y = f(x)$ for some $x \in X$. Now x is within distance δ of some x_i, and then $y = f(x)$ is within distance ε of $f(x_i)$. So Y is totally bounded.

Since f^{-1} is also uniformly continuous the converse is proved similarly. □

To see that total boundedness is not a topological invariant it is enough to consider the homeomorphic spaces $\mathbb{R}$ and the open interval $(0,\, 1)$, which is totally bounded (for example as a subspace of $[0,\, 1]$).

That is probably enough introduction to total boundedness. We mention without proof one more result.

Proposition C.1.17 A totally bounded metric space is separable. (See S.6.6 for the definition of separable.)

★ **The Arzelà-Ascoli theorem** Compact subspaces are harder to find in function spaces than in $\mathbb{R}^n$; being closed and bounded is often not enough. We have seen in Example C.1.8 that the closed unit ball in the space $\mathcal{C}[0,\, 1]$ with the sup metric is not totally bounded, so it is not compact. The Arzelà-Ascoli theorem will give a characterization of the compact subspaces of $\mathcal{C}[0,\, 1]$, but first we give some definitions and preliminary results.

Definition C.1.18 A subfamily $\mathcal{F}$ of $\mathcal{C}[0,\, 1]$ is *equicontinuous at a point* $a \in [0,\, 1]$ if given $\varepsilon > 0$ there exists $\delta > 0$ such that $|f(x) - f(a)| < \varepsilon$ whenever $x \in B_\delta(a)$ and $f \in \mathcal{F}$.

(In this definition $B_\delta(a)$ means the open ball with respect to the Euclidean metric restricted to $[0,\, 1]$: in other words $B_\delta(a) = (a - \delta,\, a + \delta) \cap [0,\, 1]$).

The force of the definition is that given an ε, the same δ has to do the business for *all* $f \in \mathcal{F}$.

Remark Definition C.1.18 can be generalized to a subfamily of the bounded continuous maps from any topological space X to any metric space $(Y,\, d)$: we just replace $B_\delta(a)$ by 'some open subset of X containing a' and $|f(x) - f(a)|$ by $d(f(x),\, f(a))$.

Proposition C.1.19 If for some $a \in [0,\, 1]$ the subfamilies $\mathcal{F}_1,\, \mathcal{F}_2,\, \ldots \mathcal{F}_r$ of $\mathcal{C}[0,\, 1]$ are all equicontinuous at a then so is their union.

Proof Let $\varepsilon > 0$. By Definition C.1.18, for each $i = 1,\, 2,\, \ldots,\, r$ there exists $\delta_i > 0$ such that $|f(x) - f(a)| < \varepsilon$ whenever $x \in B_{\delta_i}(a)$ and $f \in \mathcal{F}_i$. Let $\delta = \min\{\delta_i : 1 \leqslant i \leqslant r\}$. Then $|f(x) - f(a)| < \varepsilon$ whenever $x \in B_\delta(a)$ and $f \in \bigcup_{i=1}^{r} \mathcal{F}_i$ □

Example C.1.20 Consider the family of functions $\{f_n\}$ in $\mathcal{C}[0, 1]$ used in Example 14.23. We shall prove that it is equicontinuous at any $a \in (0, 1]$ and not at 0. First let $a \in (0, 1]$. Let $\varepsilon > 0$ and take $\delta = a/2$. Now choose N such that $2^{-N} < a/2$. Then for any $n > N$ the function f_n is zero throughout $[1/2^{n-1}, 1]$, so it is zero throughout $[a/2, 1]$. Hence the set $\{f_n : n > N\}$ is equicontinuous at a, since for any $x \in (a - \delta,\ a + \delta)$ and any $n > N$ we have $|f_n(x) - f_n(a)| = 0 < \varepsilon$. But any singleton $\{f_i\}$ containing just one continuous function is equicontinuous at any point, so $\{f_n : 1 \leqslant n \leqslant N\}$ is equicontinuous at a, and $\{f_n : n \in \mathbb{N}\}$, as a union of $N + 1$ sets all equicontinuous at a, is equicontinuous at a by Proposition C.1.19.

But consider $a = 0$. Take $\varepsilon = 1$ and let δ be any positive number. Then there exists $n \in \mathbb{N}$ such that $x = 2^{-n} + 2^{-(n+1)} < \delta$, and $f_n(x) = 1$ while $f_n(0) = 0$, so $|f_n(x) - f_n(0)| < 1$ fails for such n and x. So $\{f_n : n \in \mathbb{N}\}$ is not equicontinuous at 0. □

Just as for ordinary continuity we have the following definition.

Definition C.1.21 A subfamily $\mathcal{F} \subseteq \mathcal{C}[0, 1]$ is said to be *equicontinuous* if it is equicontinuous at every $a \in [0, 1]$.

An analogous definition may be made for subfamilies of bounded continuous functions from any topological space to any metric space.

Equicontinuity is sometimes taken to mean what we shall call uniform equicontinuity. We define it in the general context of the space $\mathcal{C}(X, Y)$ of bounded continuous maps from a metric space (X, d_X) to a metric space (Y, d_Y), with the sup metric.

Definition C.1.22 A subfamily $\mathcal{F}$ of $\mathcal{C}(X, Y)$ is said to be *uniformly equicontinous* if given any $\varepsilon > 0$ there exists $\delta > 0$ such that $d_Y(f(x), f(x')) < \varepsilon$ for any $x, x' \in X$ with $d_X(x, x') < \delta$ and any $f \in \mathcal{F}$.

The force of this definition is that for a given $\varepsilon > 0$ the same $\delta > 0$ has to do the business for all $f \in \mathcal{F}$ and also everywhere in X. As before, when X is compact we get the uniform bit for free.

Proposition C.1.23 If X, Y are metric spaces with X compact and $\mathcal{F} \subseteq \mathcal{C}(X, Y)$ is an equicontinuous family then it is uniformly equicontinuous.

Proof The proof is a matter of going through the corresponding proof of Proposition S.13.1 and inserting 'for all $f \in \mathcal{F}$' from time to time. □

Example C.1.24 Consider the family of functions $\mathcal{F} = \{f_n : n \in \mathbb{N}\}$ in the space of bounded continuous real-valued functions on $(0, 1)$ with the sup metric, where $f_n(x) = x^n$. We shall prove that this family is equicontinuous but not uniformly equicontinuous.

To prove that $\mathcal{F}$ is equicontinuous, let $a \in (0, 1)$ and let $\varepsilon > 0$. We shall use the identity $x^n - a^n = (x^{n-1} + x^{n-2}a + \ldots + xa^{n-2} + a^{n-1})(x - a)$. For all $x \in (0, (a + 1)/2)$ this gives

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

Now $(a + 1)/2 < 1$ so $n((a + 1)/2)^{n-1} \to 0$ as $n \to \infty$. We may therefore choose $N \in \mathbb{N}$ such that $n((a + 1)/2)^{n-1} < 1$ whenever $n \geqslant N$. Then $|x^n - a^n| < |x - a|$ for all $n \geqslant N$ and all $x \in (0, (a + 1)/2)$. Now take $\delta = \min\{\varepsilon, a, (1 - a)/2\}$ and $|f_n(x) - f_n(a)| = |x^n - a^n| < \varepsilon$ whenever $|x - a| < \delta$ and $n \geqslant N$. This says that $\{f_n : n \geqslant N\}$ is equicontinuous at a. Since each f_n for $n = 1, 2, \ldots, N$ is continuous at a, by Proposition C.1.20 the whole family $\{f_n : n \in \mathbb{N}\}$ is equicontinuous at a.

But $\mathcal{F}$ is not uniformly equicontinuous. For take $\varepsilon = 1/2$ and any $\delta > 0$. Take $n \in \mathbb{N}$ large enough that $n/4 > 1/\delta$. Now choose x, y in $(0, 1)$ close enough to 1 that $x^n > 1/2$, $y^n > 1/2$ and $|x - y| = \delta/2$. Then for $1 \leqslant r \leqslant n$ we have $x^{n-r}y^r > 1/4$ and

$$|f_n(x) - f_n(y)| = |x^n - y^n| = |x - y|(x^{n-1} + x^{n-2}y + \ldots + xy^{n-2} + y^{n-1}) \geqslant \frac{\delta}{2}\frac{n}{4} > 1/2.$$

This shows that $\mathcal{F}$ is not uniformly equicontinuous on $(0, 1)$. □

Actually, there is a simpler example but having less to do with *families* of functions than with a single function being continuous but not uniformly continuous.

Example C.1.25 Let $f(x) = \sin(1/x)$. Then f is continuous and bounded, but not uniformly continuous, on $(0, 1)$. So the 'family' $\{f\}$ is equicontinuous on $(0, 1)$ but not uniformly equicontinuous there.

The connection between equicontinuity and compactness begins to emerge in the next result.

Proposition C.1.26 If $\mathcal{F} \subseteq \mathcal{C}[0, 1]$ is totally bounded then it is equicontinuous.

Proof Let $a \in [0, 1]$ and let $\varepsilon > 0$. Let $\{f_1, f_2, \ldots, f_r\} \subseteq \mathcal{F}$ be an $\varepsilon/3$-net for $\mathcal{F}$. Since each f_i is continuous at a, there exists $\delta_i > 0$ such that $|f_i(x) - f_i(a)| < \varepsilon/3$ for all $x \in B_{\delta_i}(a)$. Let $\delta = \min\{\delta_1, \delta_2, \ldots, \delta_r\}$. Let $x \in B_\delta(a)$ and let $f \in \mathcal{F}$. There exists some $i \in \{1, 2, \ldots, r\}$ such that $d_\infty(f, f_i) < \varepsilon/3$. Then

$$|f(x) - f(a)| \leqslant |f(x) - f_i(x)| + |f_i(x) - f_i(a)| + |f_i(a) - f(a)| < \varepsilon, \quad \text{as required.}$$ □

Remark The same proof works for a totally bounded subfamily of the space $\mathcal{C}(X, Y)$ of bounded continuous functions from a topological space X to a metric space (Y, d): we have only to replace $B_{\delta_i}(a)$ by an open set U_i containing a and such that $f_i(U_i) \subseteq B_{\varepsilon/3}(f_i(a))$, the procedure of taking the minimum of the δ_i by taking the intersection of the U_i, and $|f(x) - f(a)|$ by $d(f(x), f(a))$.

Corollary C.1.27 Any compact subfamily of $\mathcal{C}[0, 1]$ is equicontinuous.

Proof Since any compact metric space is totally bounded this follows from Proposition C.1.26.
□

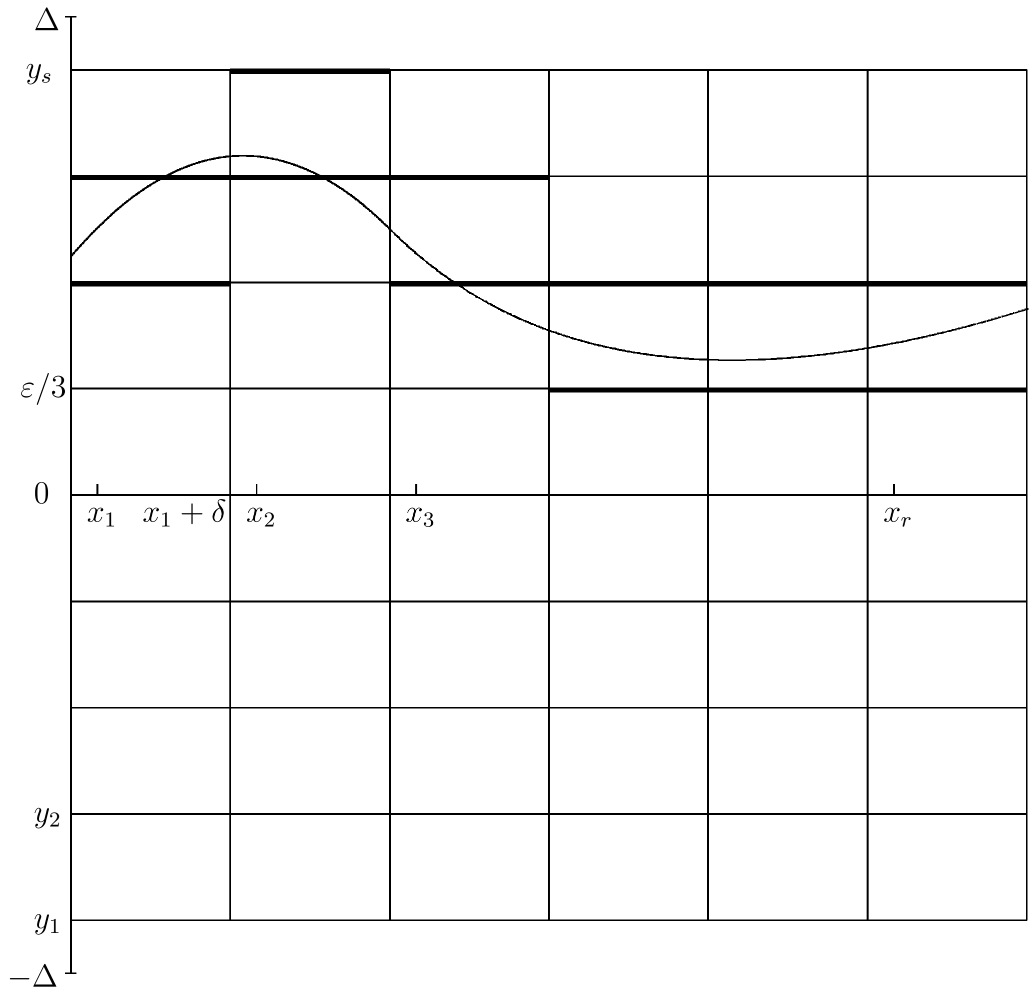
Theorem C.1.28 (Arzelà-Ascoli) A subset $\mathcal{F}$ of $\mathcal{C}[0, 1]$ is compact iff it is closed in $\mathcal{C}[0, 1]$, bounded and equicontinuous.

Proof We have already proved this in one direction, in Propositions 13.10, 13.12, C.1.13 and C.1.27. Suppose that $\mathcal{F}$ is closed, bounded and equicontinuous. Since $\mathcal{F}$ is a closed subspace of the complete space $\mathcal{C}[0, 1]$ it is complete. Compactness of $\mathcal{F}$ will follow from Proposition C.1.14 if we show that it is totally bounded.

Let $\varepsilon > 0$. We shall construct a finite ε-net for $\mathcal{F}$ in the space $\mathcal{B}[0, 1]$ of all bounded real-valued functions on $[0, 1]$ with the sup metric. Total boundedness of $\mathcal{F}$ then follows from Lemma C.1.9. Since $\mathcal{F}$ is equicontinuous and $[0, 1]$ is compact, $\mathcal{F}$ is uniformly equicontinuous by Proposition C.1.23. So there exists $\delta > 0$ such that $|f(x) - f(y)| < \varepsilon/3$ for all $f \in \mathcal{F}$ and all $x, y \in [0, 1]$ with $|x - y| < \delta$. Let $\{x_1, x_2, \ldots, x_r\}$ be a δ-net for $[0, 1]$. For the purposes of a diagram, we may assume that the x_i are equally spaced out and are labelled in increasing order. In order to define the functions that are going to provide an ε-net for $\mathcal{F}$ we replace the $B_\delta(x_i)$ by a *disjoint* union of sets S_i where $S_i \subseteq B_\delta(x_i)$ by setting $S_1 = B_\delta(x_1)$ (recall this means $(x_1 - \delta, x_1 + \delta) \cap [0, 1]$) and inductively letting

$$S_i = B_\delta(x_i) \setminus \bigcup_{j=1}^{i} B_\delta(x_j).$$

Assume each S_i is non-empty (otherwise just delete it). In the figure on the next page, the S_i are the intervals $[0, x_1 + \delta)$, $[x_1 + \delta, x_2 + \delta)$, $\ldots$, $[x_{r-1} + \delta, 1]$. Since $\mathcal{F}$ is bounded in the sup metric, there exists $\Delta \in \mathbb{R}$ such that $|f(x)| \leqslant \Delta$ for all $x \in [0, 1]$ and $f \in \mathcal{F}$. Choose an $\varepsilon/3$-net $\{y_1, y_2, \ldots, y_s\}$ for $[-\Delta, \Delta]$. In the figure the y_j are symmetric about 0 and evenly spaced at $\varepsilon/3$ apart. The figure shows the graph of one $f \in \mathcal{F}$. The idea is to approximate functions in $\mathcal{F}$ by 'step' functions, taking as constant value on each S_i one of the heights on the grid. The thick lines show the different possible choices of height to approximate f - the grid heights on either side of $f(x_i)$. (If $f(x_i) > y_s$ we use the top grid height and if $f(x_i) < y_1$ we use the bottom grid height.)



There are at most s^r 'step' functions on the grid, since for each i there are s possible constant values from $\{y_1, y_2, \ldots, y_s\}$. This finite number of (mostly discontinuous, but bounded) functions turns out to be an ε-net for $\mathcal{F}$.

Here are the details. In order to decide which step function should approximate a given $f \in \mathcal{F}$, consider the collection Ψ of all maps $\psi : \{1, 2, \ldots, r\} \to \{1, 2, \ldots, s\}$. For each $\psi \in \Psi$ we define g_ψ to be the step function taking the constant value $y_{\psi(i)}$ on the set S_i. There are s^r such maps g_ψ since $|\Psi| = s^r$.

Now for a given $f \in \mathcal{F}$ we decide on a step function of the form g_ψ to approximate f as follows: for each $i \in \{1, 2, \ldots, r\}$ we know that $f(x_i) \in B_{\varepsilon/3}(y_j)$ for at least one $j \in \{1, 2, \ldots, s\}$. (In general that is all one can say, though in the figure there are usually two adjacent values of j for which this holds.) If $f(x_i) \in B_{\varepsilon/3}(y_j)$ for more than one value of j, make a choice of one of them. This defines a map $\psi : \{1, 2, \ldots, r\} \to \{1, 2, \ldots, s\}$ and a corresponding g_ψ. (Explicitly, for each $i \in \{1, 2, \ldots, r\}$ we let $\psi(i)$ be a choice of $j \in \{1, 2, \ldots, s\}$ such that $f(x_i) \in B_{\varepsilon/3}(y_j)$. So ψ here depends on f.) We shall prove that $d_\infty(g_\psi, f) < \varepsilon$. Any $x \in [0, 1]$ is in a unique S_i, and $g_\psi(x) = y_{\psi(i)}$. Hence for $x \in S_i$, $|f(x) - g_\psi(x)| \leqslant |f(x) - f(x_i)| + |f(x_i) - g_\psi(x)| < 2\varepsilon/3$, so $d_\infty(f, g_\psi) = \sup_{x \in [0,1]} |f(x) - g_\psi(x)| \leqslant 2\varepsilon/3 < \varepsilon$. $\square$

Remark In general not *all* the $\psi \in \Psi$ are needed - for example in the figure on the previous page, the step function g_ψ determined by ψ for a given $f \in \mathcal{F}$ cannot jump more than two vertical steps in the grid as you move from one S_i to an adjacent one.

In some other versions of the proof of Theorem C.1.28, the ε-net for $\mathcal{F}$ consists of continuous functions. As you may guess, there are more general versions of the theorem, for example replacing $[0, 1]$ by a compact topological space and $\mathbb{R}$ by a complete metric space (and more general versions still).

Examples This section contains a few examples related to compactness in function spaces.

Example C.1.29 Suppose that $\mathcal{F}$ is a collection of real-valued functions on an interval $[a, b]$ all satisfying a Lipschitz condition $|f(x) - f(y)| \leqslant K|x - y|$ for all $x, y \in [a, b]$ and $f \in \mathcal{F}$. Then $\mathcal{F}$ is equicontinuous.

Let $\varepsilon > 0$. Choose $\delta = \varepsilon/K$. Then for any $f \in \mathcal{F}$ and any $x, y \in [, b]$ with $|x - y| < \delta$, we have $|f(x) - f(y)| \leqslant K|x - y| < K\delta = \varepsilon$.

Example C.1.30 Let $\mathcal{F}$ be a subset of real-valued functions which are differentiable on $[a, b]$ and whose derivatives are bounded on (a, b) by a constant K. Then $\mathcal{F}$ is equicontinuous.

Proof This follows from Example C.1.29 since functions f in $\mathcal{F}$ satisfy the Lipschitz condition $|f(x) - f(y)| \leqslant K|x - y|$ on $[a, b]$ by the mean value theorem. □

There are several applications of Theorem C.1.28 in the theory of differential and integral equations. These enable us to see that certain subsets of function spaces are compact, so that any sequence in such a subset will have a convergent subsequence.

Example C.1.31 Suppose that $K : [a, b] \times [c, d] \to \mathbb{R}$ is a continuous function, and we use it to define a map $\mathcal{K} : \mathcal{C}[c, d] \to \mathcal{C}[a, b]$ by

$$(\mathcal{K})(f)(x) = \int_c^d K(x, y)f(y)\mathrm{d}\,y \quad \text{for any } x \in [a, b].$$

If $\mathcal{B} \subseteq \mathcal{C}[c, d]$ is any subset which is bounded in the L^1 norm then $\mathcal{K}(\mathcal{B})$ is equicontinuous.

Proof Since K is continuous on the compact space $[a, b] \times [c, d]$, it is uniformly continous there, so given $\varepsilon > 0$ there exists $\delta > 0$ such that $|K(x_1, y) - K(x_2, y)| < \varepsilon$ whenever $|x_1 - x_2| < \delta$. Now

$$|(\mathcal{K})(f)(x_1) - (\mathcal{K}(f)(x_2)| = \left|\int_c^d (K(x_1, y) - K(x_2, y))f(y)\mathrm{d}\,y\right| \leqslant \varepsilon \left|\int_c^d f(y)\mathrm{d}\,y\right|,$$

whenever $|x_1 - x_2| < \delta$. This gives the result, since it shows that if the L_1-norm of f is bounded above for $f \in \mathcal{B}$ then $\mathcal{K}(\mathcal{B})$ is equicontinuous. In particular this holds if f belongs to a set which is bounded in the sup norm. □

Example C.1.32 Suppose that $K : [a, b] \times [c, d] \to \mathbb{R}$ is a continuous function, and as in Example C.1.31 we define $\mathcal{K} : \mathcal{C}[c, d] \to \mathcal{C}[a, b]$ by

$$(\mathcal{K})(f)(x) = \int_c^d K(x, y) f(y) \mathrm{d}\, y \quad \text{for any } x \in [a, b].$$

If $\mathcal{B} \subseteq \mathcal{C}[c, d]$ is any subset which is bounded in the L^2 norm then $\mathcal{K}(\mathcal{B})$ is equicontinuous.

Proof As in Example C.1.31 we get

$$\begin{aligned} |(\mathcal{K})(f)(x_1) - (\mathcal{K}(f)(x_2)| &= \left| \int_c^d (K(x_1, y) - K(x_2, y)) f(y) \mathrm{d}\, y \right| \\ &\leqslant \left\{ \int_c^d (K(x_1, y) - K(x_2, y))^2 \mathrm{d}\, y \right\}^{1/2} \left\{ \int_c^d (f(y))^2 \mathrm{d}\, y \right\}^{1/2}, \end{aligned}$$

where the inequality follows from the Cauchy-Schwarz-Bunyakovsky inequality. Now let $\varepsilon > 0$ and suppose that the L^2-norm of f is bounded above by B in $\mathcal{B}$. As in Example C.1.31, K is uniformly continuous on $[a, b] \times [c, d]$, and in particular there exists $\delta > 0$ such that $|K(x_1, y) - K(x_2, y)| < \varepsilon B/(\sqrt{d - c})$ whenever $x_1, x_2 \in [c, d]$ with $|x_1 - x_2| < \delta$. This shows that

$$|(\mathcal{K})(f)(x_1) - (\mathcal{K}(f)(x_2)| \leqslant \varepsilon B ||f||_2$$

whenever $|x_1 - x_2| < \delta$, so $\mathcal{K}(\mathcal{B})$ is (uniformly) equicontinuous. □

We may go on and show that $\mathcal{K}(\mathcal{B})$ is bounded in the sup metric, since for any $x \in [c, d]$,

$$|\mathcal{K} f(x)| = \left| \int_c^d K(x, y) f(y) \mathrm{d}\, y \right| \leqslant \left\{ \int_c^d (K(x, y))^2 \mathrm{d}\, y \right\}^{1/2} \left\{ \int_c^d (f(y))^2 \mathrm{d}\, y \right\}^{1/2} \leqslant MB$$

on $\mathcal{B}$, where M is an upper bound on $[c, d]$ for the continuous function

$$x \mapsto \left\{ \int_c^d (K(x, y))^2 \mathrm{d}\, y \right\}^{1/2}.$$

Thus $\mathcal{K}$ takes sets which are bounded in the L^2 metric into sets which are equicontinuous and bounded in the sup metric, and hence by Theorem C.1.28 are relatively compact (their closures are compact) in the sup metric. The way this is often used in classical analysis is the following: we have some sequence (f_n) of continuous functions on an interval $[a, b]$ which is bounded in the L^2 metric, so $(K f_n)$ lies in a set which is relatively compact in the sup metric and hence has a uniformly convergent subsequence. (If you are interested in seeing more about this, see any book on integral equations, for example G. Hoheisel 'Integral equations' Ungar 1968 is an attractive brief account, or F. Riesz and B Sz-Nagy 'Functional Analysis' (Dover 1991) if you wish to see how integral equations gave rise to the subject of functional analysis.)

Real numbers and completions

The aim of this section is to sketch a way of constructing $\mathbb{R}$ from $\mathbb{Q}$ and to look at completions of metric spaces. For the latter, we give the traditional approach via Cauchy sequences and also two other versions, using 'virtual points' and embedding in a function space.

Construction of $\mathbb{R}$ from $\mathbb{Q}$

In this section we sketch one way of constructing $\mathbb{R}$ from $\mathbb{Q}$. Logically this has no advantages over the approach in Chapter 4 of the book; for a rigorous approach we still need axioms (for $\mathbb{Q}$). But psychologically it may help for two reasons. First, it is close to the common-sense approach to real numbers. Secondly, the rational numbers may feel like more familiar objects than the real numbers.

We assume that $\mathbb{Q}$ is known. One could begin further back, with axioms for the integers or even for set theory, but $\mathbb{Q}$ is a familiar enough set which rarely gives rise to confusion or puzzlement. In practical calculations involving an irrational number such as $\sqrt{2}$ we use a rational number approximating it, perhaps a decimal expansion whose number of places is limited by the length of the display on our calculators. What is the number $\sqrt{2}$ itself? We might try defining it as the *sequence* (q_n) where q_n is the decimal expansion correct to n places: more precisely the decimal $q_n = 1.a_1a_2\dots a_n$ making $|q_n^2 - 2|$ minimal among the decimals of the form $x = 1.x_1x_2\dots x_n$. Then (q_n) is a Cauchy sequence, since $|q_m - q_n| < 10^{-n}$ whenever $m > n$. (We have to re-define Cauchy sequence: given any *rational number* $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that $|q_m - q_n| < \varepsilon$ whenever $m \geqslant n \geqslant N$.) However, there are disadvantages in defining a real number to be a single Cauchy sequence in $\mathbb{Q}$. There are many different sequences of rational numbers approximating a given real number: as well as ambiguities in decimal expansions, we could for example use a number other than 10 as the base.

It is better to define a real number as a whole equivalence class of Cauchy sequences (intuitively, *all* the Cauchy sequences which approximate it). We define Cauchy sequences (q_n) and (q'_n) to be equivalent iff intuitively they are approximating the same number: $(q_n) \sim (q'_n)$ iff

$(q_n - q'_n)$ converges to 0. It turns out that we may then define addition and multiplication on the resulting set of equivalence classes, and this construction gives us all we want for our system $\mathbb{R}$. Moreover, this construction is just a formalization of the practical viewpoint on real numbers, where we use different approximations to meet the needs of different situations. If you want only to feel less mystery about the real numbers you may have now read far enough in this section - in fact you may now feel that real numbers are a bit of a confidence trick.

► To get the algebraic and order properties of $\mathbb{R}$ we assume known the definitions of field, commutative ring, ideal, ring homomorphism, and the construction of the quotient ring of a commutative ring by an ideal; these are contained in most textbooks of modern algebra. But we define orderings since they are particularly central to what follows.

Definition C.2.1 An *order* for a field F is a subset P of F called the (strictly) positive elements, such that

(P1) If $x,\, y \in P$ then $x+y,\, xy \in P$.

(P2) For any $x \in F$ exactly one of the following holds: $x \in P$, $x = 0$, $-x \in P$.

Given an order P for a field F, we write $x > y$ (or equivalently $y < x$) if $x - y \in P$; we write $x \geqslant y$ if $x > y$ or $x = y$.

We need some results about Cauchy sequences in $\mathbb{Q}$.

Definition C.2.2 A sequence (q_n) in $\mathbb{Q}$ is called a *null sequence* if it converges to 0; otherwise it is called *non-null*.

We note that the definition of 'convergent', like the definition of 'Cauchy sequence', is modified to concern only *rational* $\varepsilon > 0$. It is clear that a null sequence is a Cauchy sequence.

Lemma C.2.3 Any Cauchy sequence in $\mathbb{Q}$ is bounded.

This follows as in the proof of Theorem 4.18. □

Lemma C.2.4 If (q_n) is a non-null Cauchy sequence in $\mathbb{Q}$ then there exists a rational number $\delta > 0$ and an integer N such that either

(a) $q_n > \delta$ for all $n \geqslant N$, or

(b) $q_n < -\delta$ for all $n \geqslant N$.

Proof Since (q_n) does not converge to 0, using the technique explained in S.2 we get that there exists a rational number $\varepsilon_0 > 0$ such that for any integer N there exists an integer $M \geqslant N$ such that $|q_M| \geqslant \varepsilon_0$. Put $\delta = \varepsilon_0/2$. Since (q_n) is Cauchy, there exists an integer N such that $|q_m - q_n| < \delta$ whenever $m \geqslant n \geqslant N$. By choice of ε_0 there exists some integer $M \geqslant N$ such that $|q_M| \geqslant 2\delta$. We now distinguish two cases.

(a) If $q_M \geqslant 2\delta$ then for any $n \geqslant N$ we have $q_M - q_n < \delta$, so $q_n > q_M - \delta \geqslant \delta$.

(b) If $q_M \leqslant -2\delta$ then for any $n \geqslant N$ we have $q_n - q_M < \delta$ so $q_n < q_M + \delta \leqslant -\delta$.

This completes the proof. In either case $|q_n| > \delta$ for $n \geqslant N$. □

Addition and multiplication of Cauchy sequences in $\mathbb{Q}$ are defined by:

$$(q_n) + (q'_n) = (q_n + q'_n), \quad (q_n).(q'_n) = (q_n q'_n).$$

If (q_n) and (q'_n) are Cauchy sequences, so are $(-q_n)$, $(q_n + q'_n)$, $(q_n q'_n)$: the proofs are easy, analogous to those for convergent sequences in S.4. Lemma C.2.3 is needed for the product, and one should remember to talk about *rational* ε. It also follows easily that the set $\mathcal{C}$ of all Cauchy sequences of rational numbers forms a commutative ring, with multiplicative identity $e = (1, 1, 1, \ldots)$.

Lemma C.2.5 The set $\mathcal{N}$ of all null sequences forms an ideal in $\mathcal{C}$.

One has to prove that if (q_n) and (q'_n) are in $\mathcal{N}$ and (x_n) is in $\mathcal{C}$ then $(q_n - q'_n)$ and $(q_n x_n)$ are in $\mathcal{N}$. The proofs are easy, using Lemma C.2.3 for the product. □

Now $\mathbb{R}$ is defined as the quotient ring $\mathcal{C}/\mathcal{N}$. So a real number is an equivalence class of Cauchy sequences of rational numbers, where $(q_n) \sim (q'_n)$ iff $(q_n - q'_n)$ is null. We say that the real number is 'represented' by any Cauchy sequence in the equivalence class. It follows by algebra that $\mathbb{R}$ is a commutative ring, with multiplicative identity $\hat{1}$, the equivalence class of e. (For clarity we denote elements of $\mathbb{R}$ by $\hat{x}$, $\hat{y}$ etc. in this section.)

Proposition C.2.6 $\mathbb{R}$ is a field.

Proof Given that $\mathbb{R}$ is a commutative ring with multiplicative identity, what remains to be proved is the existence of multiplicative inverses, that if $\hat{x} \neq \hat{0}$ in $\mathbb{R}$ then there exists $\hat{y}$ in $\mathbb{R}$ with $\hat{x}\hat{y} = \hat{1}$. Let $(q_n) \in \mathcal{C}$ represent $\hat{x}$. Then we have to find $(q'_n) \in \mathcal{C}$ (to represent $\hat{y}$) such that $(q_n q'_n) \sim e$, in other words such that $(q_n q'_n - 1) \in \mathcal{N}$.

Since $\hat{x} \neq \hat{0}$, (q_n) is non-null so by Lemma C.2.4 there exists (a rational number) $\delta > 0$ and an integer N such that $|q_n| > \delta$ for all $n \geqslant N$. Let $q'_n = 1/q_n$ for $n \geqslant N$ and let $q'_n = 0$ (say) for $n < N$. Since then $q_n q'_n - 1 = 0$ for all $n \geqslant N$, $(q_n q'_n - 1)$ is certainly a null sequence. We have only to prove that (q'_n) is Cauchy. Given (a rational number) $\varepsilon > 0$ let $N_1 \geqslant N$ be such that $|q_n - q_m| < \varepsilon\delta^2$ for all $m, n \geqslant N_1$. Then for $m, n \geqslant N_1$,

$|q'_m - q'_n| = |q_m - q_n|/|q_m q_n| < |q_m - q_n|/\delta^2 < \varepsilon$, as required. □

We want to see that $\mathbb{Q}$ is (or may be identified with) a subset of $\mathbb{R}$. Define $i : \mathbb{Q} \to \mathbb{R}$ by letting $i(q)$ be the equivalence class of $(q, q, q, \ldots)$. Formally $i = \pi \circ j$ where $j : \mathbb{Q} \to \mathcal{C}$ is given by

$j(q) = (q, q, q, \ldots)$ and $\pi : \mathcal{C} \to \mathcal{C}/\mathcal{N} = \mathcal{R}$ is the natural map assigning to each element of $\mathbb{C}$ its equivalence class in $\mathcal{C}/\mathcal{N}$. Since j and π are ring homomorphisms so is i. The kernel of i is $\{0\}$ since if $q \neq 0$ the sequence $(q, q, q, \ldots)$ is non-null. So i determines a ring isomorphism of $\mathbb{Q}$ onto $i(\mathbb{Q})$. In practice we normally identify $\mathbb{Q}$ with $i(\mathbb{Q})$ via this isomorphism, but for clarity in most of this section we preserve the distinction between them.

Next we define the order $P \subseteq \mathbb{R}$ to be the set of elements represented by Cauchy sequences satisfying condition (a) of Lemma C.2.4. To see that this is well-defined we need to show that if $(q_n) \in \mathcal{C}$ satisfies the condition (a) then so does $(q_n + q_n')$ for any (q_n') in $\mathcal{N}$. This is easy (since $q_n' \to 0$ as $n \to \infty$). The conditions (P1) and (P2) for an order are easy to check for P ((P2) follows from Lemma C.2.4). Notice that for $q \in \mathbb{Q}$, $i(q) \in P$ iff $q > 0$. This says that i is order-preserving: for if $q_1, q_2 \in \mathbb{Q}$ and $q_2 > q_1$ then $q_2 - q_1 > 0$ so $i(q_2) - i(q_1) = i(q_2 - q_1) \in P$ so $i(q_2) > i(q_1)$. The definitions of convergent and Cauchy now make sense in $\mathbb{R}$ since we have defined an order for $\mathbb{R}$.

For convenience later, we discuss the order in $\mathbb{R}$ and its relation with the order in $\mathbb{Q}$ a little further.

Proposition C.2.7 Suppose that (q_n) in $\mathcal{C}$ represents $\hat{x}$ in $\mathbb{R}$. If $q_n \geqslant 0$ for all sufficiently large n then $\hat{x} \geqslant 0$.

Proof If $\hat{x} < 0$ then by definition of P there would exist a rational number $\delta > 0$ such that $q_n < -\delta$ for all sufficiently large n, contradicting the hypothesis. □

The converse of Proposition C.2.7 is false: if $\hat{x} = \hat{0}$ and (q_n) in $\mathcal{C}$ represents $\hat{x}$ then (q_n) is null, but this does not force $q_n \geqslant 0$ for all sufficiently large n; for example we might have $q_n = (-1)^n/n$.

Corollary C.2.8 Suppose that (q_n) represents $\hat{x}$ and that $q_n \leqslant q$ for some rational number q and all sufficiently large n. Then $\hat{x} \leqslant i(q)$. The same holds with $\leqslant$ replaced by $\geqslant$.

Proof Suppose that $q_n \leqslant q$ for large n. Then, recalling the definitions of j, π from the top of the page, $i(q) - \hat{x} = \pi(j(q)) - \pi((q_n)) = \pi((q - q_n))$. So $(q - q_n)$ represents $i(q) - \hat{x}$, and by Proposition C.2.7, $i(q) - \hat{x} \geqslant 0$, so $\hat{x} \leqslant i(q)$. The case when $q_n \geqslant q$ for large n is similar. □

We also need to handle absolute values in $\mathbb{R}$.

Proposition C.2.9 If $(q_n) \in \mathcal{C}$ represents $\hat{x}$ then $(|q_n|)$ represents $|\hat{x}|$.

Proof If $\hat{x} = \hat{0}$ then (q_n) is a null sequence hence so is $(|q_n|)$ and so it represents $\hat{0} = |\hat{x}|$.

Suppose that $\hat{x} > \hat{0}$, so $|\hat{x}| = \hat{x}$. By definition of P there exists $\delta > 0$ and an integer N such that $q_n > \delta$ for all $n \geqslant N$. So $|q_n| - q_n = 0$ for $n \geqslant N$ and $(|q_n| - q_n)$ is null. Hence $(|q_n|)$

represents the same number as (q_n) represents, namely $\hat{x}$ which is $|\hat{x}|$.

Finally suppose that $\hat{x} < \hat{0}$, so $|\hat{x}| = -\hat{x}$. By definition of P there exist $\delta > 0$ and an integer N such that $q_n < -\delta$ for all $n \geqslant N$. So $|q_n| = -q_n$ for all $n \geqslant N$ and $(|q_n| + q_n)$ is null. Hence $(|q_n|)$ represents the same element of $\mathbb{R}$ as $(-q_n)$ represents, namely $-\hat{x}$ which is $|\hat{x}|$. □

Corollary C.2.10 If $(q_n) \in \mathcal{C}$ represents $\hat{x}$ and $|q_n| \leqslant q$ for all sufficiently large n then $|\hat{x}| \leqslant i(q)$.

This follows from Proposition C.2.9 and Corollary C.2.8. □

Corollary C.2.11 For any $\hat{x},\, \hat{y} \in \mathbb{R}$, $|\hat{x} + \hat{y}| \leqslant |\hat{x}| + |\hat{y}|$.

Proof Let (q_n), $(r_n) \in \mathcal{C}$ represent $\hat{x}$, $\hat{y}$. Then $(q_n + r_n)$ represents $\hat{x} + \hat{y}$, so by Proposition C.2.9 $(|q_n + r_n|)$ represents $|\hat{x} + \hat{y}|$. Similarly $(|q_n|)$, $(|r_n|)$ represent $|\hat{x}|$, $|\hat{y}|$. Thus $(|q_n| + |r_n| - |q_n + r_n|)$ represents $|\hat{x}| + |\hat{y}| - |\hat{x} + \hat{y}|$. The result now follows from the triangle inequality in $\mathbb{Q}$ and Proposition C.2.7. □

Finally we need a form of Archimedes' axiom.

Proposition C.2.12 Given any $\hat{\varepsilon} > \hat{0}$ in $\mathbb{R}$, there exists $\delta \in \mathbb{Q}$ such that $\hat{0} < i(\delta) \leqslant \hat{\varepsilon}$. In particular there exists an integer $n \in \mathbb{N}$ such that $i(1/n) \leqslant \hat{\varepsilon}$.

Proof Let (ε_n) represent $\hat{\varepsilon}$. Since $\hat{\varepsilon} > \hat{0}$, there exist a rational number $\delta > 0$ and an integer N such $\varepsilon_n > \delta$ for all $n \geqslant N$. Then $\hat{\varepsilon} \geqslant i(\delta)$ by Corollary C.2.8, and $i(\delta) > \hat{0}$ since $\delta > 0$ (i is order-preserving). The last part of the statement follows since there is an integer n with $n > 1/\delta$ and i is order-preserving. □

In particular it follows from Proposition C.2.12 that in defining convergent or Cauchy sequences in $\mathbb{R}$ it does not matter whether we use a 'rational or real ε' - precisely, whether $\hat{\varepsilon}$ is in $i(\mathbb{Q})$ or not.

The secret of why $\mathbb{R}$ is complete is divulged in the next result.

Proposition C.2.13 If $(q_n) \in \mathcal{C}$ represents $\hat{x}$ then $(i(q_n))$ converges to $\hat{x}$ in $\mathbb{R}$.

Proof Given $\hat{\varepsilon} > \hat{0}$, by Proposition C.2.12 there exists a rational $\delta > 0$ with $\hat{0} < i(\delta) \leqslant \hat{\varepsilon}$. Since $(q_n) \in \mathcal{C}$ there exists an integer N such that $|q_m - q_n| < \delta$ whenever $m \geqslant n \geqslant N$. Now think of a fixed $n \geqslant N$ and let m vary: the sequence $(|q_m - q_n|)$ represents $|\hat{x} - i(q_n)|$ by Proposition C.2.9. Hence $|\hat{x} - i(q_n)| \leqslant i(\delta) \leqslant \hat{\varepsilon}$ by Corollary C.2.10. This holds for all $n \geqslant N$, so $(i(q_n))$ converges to $\hat{x}$. □

Corollary C.2.14 $i(\mathbb{Q})$ is dense in $\mathbb{R}$.

This follows from the Proposition C.2.13 and Proposition 6.29, which tells us that the limit $\hat{x}$ of the convergent sequence $(i(q_n))$ is in the closure of $i(\mathbb{Q})$. □

We now prove completeness of $\mathbb{R}$.

Theorem C.2.15 Any Cauchy sequence $(\hat{x}_n)$ in $\mathbb{R}$ converges to a point $\hat{x}$ in $\mathbb{R}$.

Proof Since $i(\mathbb{Q})$ is dense in $\mathbb{R}$, for each $n \in \mathbb{N}$ there exists $q_n \in \mathbb{Q}$ with $|\hat{x}_n - i(q_n)| < i(1/n)$. We shall prove that (q_n) is a Cauchy sequence. Given a rational number $\delta > 0$, since $(\hat{x}_n)$ is Cauchy there exists an integer N such that $|\hat{x}_m - \hat{x}_n| < i(\delta/3)$ whenever $m \geqslant n \geqslant N$. Choose N large enough so that $N > 3/\delta$. Using the triangle inequality proved in Corollary C.2.11, for $m \geqslant n \geqslant N$,

$$|i(q_m) - i(q_n)| \leqslant |i(q_m) - \hat{x}_m| + |\hat{x}_m - \hat{x}_n| + |\hat{x}_n - i(q_n)| < i(\delta).$$

Since i is order-preserving, $|q_m - q_n| < \delta$ whenever $m \geqslant n \geqslant N$, so $(q_n) \in \mathcal{C}$. Thus (q_n) represents some real number $\hat{x}$.

We now prove that $(\hat{x}_n)$ converges to $\hat{x}$. Let $\hat{\varepsilon} > 0$. Since (q_n) represents $\hat{x}$, by Proposition C.2.13, $(i(q_n))$ converges to $\hat{x}$. So we may choose an integer N such that $|i(q_n) - \hat{x}| < i(1/2)\hat{\varepsilon}$ for all $n \geqslant N$. Also, using Proposition C.2.12 we may choose N large enough that $i(1/N) < i(1/2)\hat{\varepsilon}$ and then for $n \geqslant N$ we have $|\hat{x}_n - i(q_n)| < i(1/n) \leqslant i(1/N) < i(1/2)\hat{\varepsilon}$. So for $n \geqslant N$,

$$|\hat{x}_n - \hat{x}| \leqslant |\hat{x}_n - i(q_n)| + |i(q_n) - \hat{x}| < i(1)\hat{\varepsilon} = \hat{\varepsilon},$$

which shows that $(\hat{x}_n)$ converges to $\hat{x}$ as required. □

Finally we prove the form of completeness for $\mathbb{R}$ that is stated in Proposition 4.4: any non-empty subset of $\mathbb{R}$ which is bounded above has a least upper bound. Given the work we have already done, we may now without risk drop the hats from the names of real numbers and identify $\mathbb{Q}$ with $i(\mathbb{Q})$ via i.

Proof Let $S \subseteq \mathbb{R}$ be a non-empty set which is bounded above. Let $a_1 \in S$ and let v_1 be an upper bound for S. We are going to use a bisection procedure to construct sequences which will converge to a least upper bound for S.

Consider $(a_1 + v_1)/2$. If this is an upper bound for S let $a_2 = a_1$ and $v_2 = (a_1 + v_1)/2$. Otherwise let $a_2 \in S \cap ((a_1 + v_1)/2,\, v_1)$ and $v_2 = v_1$. Suppose inductively there exist real numbers $a_1,\, a_2,\, \ldots,\, a_n,\; v_1,\, v_2,\, \ldots,\, v_n$ with $a_1 \leqslant a_2 \leqslant \ldots \leqslant a_n < v_n \leqslant v_{n-1} \leqslant \ldots \leqslant v_2 \leqslant v_1$ such that for each $i \in \{1, 2, \ldots, n\}$, $a_i \in S$ and v_i is an upper bound for S, and $|v_i - a_i| \leqslant (v - a)/2^i$. If

$(a_n + v_n)/2$ is an upper bound for S we let $a_{n+1} = a_n$ and $v_{n+1} = (a_n + v_n)/2$. Otherwise we let $v_{n+1} = v_n$ and choose $a_{n+1} \in S \cap ((a_n + v_n)/2, v)$. Then the inductive statement holds good up to $n+1$. For any $m \geqslant n$ we have $a_m, v_m \in [a_n, v_n]$, and from Proposition C.2.12 we may deduce that $(v-a)/2^n$ can be made smaller than any prescribed positive number by choosing n sufficiently large. Hence both (a_n) and (v_n) are Cauchy sequences. So they converge by Theorem C.2.15. Also, since $|v_n - a_n| \leqslant (v-a)/2^n$, these two sequences converge to the same limit u say. By construction, each v_n is an upper boumd for S. If $x > u$ for some $x \in S$ we would get the contradiction that $x > v_n$ for all n sufficiently large. So u is an upper bound for S. Also, no $x < u$ is an upper bound for S since $a_n > x$ for such an x and sufficiently large n, and $a_n \in S$. So u is a least upper bound for S as required. □

There are other ways of constructing $\mathbb{R}$ from $\mathbb{Q}$, for example the method of Dedekind cuts. The method described above is generally ascribed to Cantor. ◀

★ **Completion of metric spaces via Cauchy sequences** The book indicates the advantages of having a complete metric space. In this section and the next two we'll give three distinct methods of constructing a completion of a metric space. Then in the final section we'll show that completions are unique so we know that we don't get different completions via these three different methods. We first describe the traditional process of completing a metric space via Cauchy sequences. This is similar to the construction of $\mathbb{R}$ from $\mathbb{Q}$ using Cauchy sequences, but the latter is *not* a special case of the completion process we are about to describe. For one thing, the existence of $\mathbb{R}$ is needed before we can even define what 'metric space' means. Also, we shall use the completeness of $\mathbb{R}$ in this section.

Definition C.2.16 Given a metric space (X, d), a *completion* of (X, d) consists of a metric space $(\hat{X}, \hat{d})$ together with a map $i : X \to \hat{X}$ such that

(C1) $(\hat{X}, \hat{d})$ is complete,

(C2) i is an isometry into; in other words $\hat{d}(i(x), i(y)) = d(x, y)$ for all $x, y \in X$,

(C3) $i(X)$ is dense in $\hat{X}$.

The third condition, (C3), is there to ensure some kind of uniqueness via minimality - without it we could take larger and larger metric spaces containing a given completion and still have a completion.

Remark If (X, d) is already complete, we would expect the completion process not to do much, and indeed we can immediately see that a completion for (X, d) in this case is just (X, d) with i taken to be the identity map of X. This clearly satisfies (C1), (C2) and (C3).

Our goal in this section is to prove

Theorem C.2.17 Any metric space (X, d) has a completion.

Proof Let $\hat{X}$ be the set of equivalence classes of Cauchy sequences in X, where $(x_n) \sim (x'_n)$ iff $d(x_n, x'_n) \to 0$ as $n \to \infty$. It is clear that this is an equivalence relation. Again we say that an element $\hat{x}$ of $\hat{X}$ is 'represented' by the Cauchy sequence (x_n) in X if (x_n) is one of the Cauchy sequences in the equivalence class $\hat{x}$. We define $\hat{d}$ as follows. If $\hat{x}, \hat{y} \in \hat{X}$ are represented by $(x_n), (y_n)$, let $\hat{d}(\hat{x}, \hat{y}) = \lim_{n\to\infty} d(x_n, y_n)$. There are several things to check. First we show that $(d(x_n, y_n))$ is a Cauchy sequence in $\mathbb{R}$, so does have a limit as $n \to \infty$. For

$$|d(x_m, y_m) - d(x_n, y_n)| \leqslant d(x_m, x_n) + d(y_m, y_n)$$

by Exercise 5.2, so $(d(x_n, y_n))$ is a Cauchy sequence in $\mathbb{R}$ since (x_n) and (y_n) are Cauchy sequences in X. Hence $\lim_{n\to\infty} d(x_n, y_n)$ exists by completeness of $\mathbb{R}$.

We need to check that the definition of $\hat{d}(\hat{x}, \hat{y})$ is independent of choice of Cauchy sequences in X representing $\hat{x}$ and $\hat{y}$. Suppose that (x'_n) is another Cauchy sequence in X representing $\hat{x}$, so we know $d(x_n, x'_n) \to 0$ as $n \to \infty$. Then $0 \leqslant |d(x'_n, y_n) - d(x_n, y_n)| \leqslant d(x'_n, x_n)$ so $\lim_{n\to\infty} d(x'_n, y_n) = \lim_{n\to\infty} d(x_n, y_n)$. Similarly if we change the representative Cauchy sequence for $\hat{y}$ the value of $\hat{d}(\hat{x}, \hat{y})$ remains unchanged. So $\hat{d}$ is well defined on $\hat{X} \times \hat{X}$.

Next we check that $\hat{d}$ is a metric on $\hat{X}$.

(M1) Suppose that $\hat{x}, \hat{y}$ in $\hat{X}$ are represented by Cauchy sequences $(x_n), (y_n)$ in X. Then

$$\hat{d}(\hat{x}, \hat{y}) = 0 \text{ iff } \lim_{n\to\infty} (d(x_n, y_n)) = 0 \text{ iff } (x_n) \sim (y_n) \text{ iff } \hat{x} = \hat{y}.$$

Also, clearly $\hat{d}(\hat{x}, \hat{y}) \geqslant 0$, so (M1) holds.

(M2) and (M3) follow by taking limits from the corresponding conditions for d.

Now we define $i : X \to \hat{X}$ by $i(x) = \hat{x}$ where $\hat{x}$ is the element of $\hat{X}$ represented by the Cauchy sequence $(x, x, x, \ldots)$. If $x, y \in X$ then $\hat{d}(\hat{x}, \hat{y}) = \lim_{n\to\infty} d(x, y) = d(x, y)$ so i is an isometry of X into $\hat{X}$.

The next two results complete the proof of Theorem C.2.17. The results, and their proofs, are very close to Corollary C.2.14 and Theorem C.2.15.

Proposition C.2.18 $i(X)$ is dense in $\hat{X}$.

Proof For any $\hat{x}$ and any real number $\varepsilon > 0$, let (x_n) be a Cauchy sequence in X representing $\hat{x}$, and let $N \in \mathbb{N}$ be such that $d(x_m, x_n) < \varepsilon/2$ whenever $m \geqslant n \geqslant N$. Consider $i(x_N)$. By definition $\hat{d}(i(x_N), \hat{x}) = \lim_{n\to\infty} d(x_N, x_n) \leqslant \varepsilon/2 < \varepsilon$, as required. □

Theorem C.2.19 $(\hat{X}, \hat{d})$ is complete.

Proof Suppose that $(\hat{x}_n)$ is a Cauchy sequence in $\hat{X}$. Since $i(X)$ is dense in $\hat{X}$, for each

$n \in N$ there exists $x_n \in X$ such that $\hat{d}(i(x_n), \hat{x}_n) < 1/n$. We shall prove that (x_n) is a Cauchy sequence and that $(\hat{x}_n)$ converges to the point of $\hat{X}$ represented by (x_n).

Let $\delta > 0$. Since $(\hat{x}_n)$ is a Cauchy sequence there exists $N \in \mathbb{N}$ such that $\hat{d}(\hat{x}_m, \hat{x}_n) < \delta/3$ for all $m \geqslant n \geqslant N$. We may choose N such that also $N > 3/\delta$ (so that $\hat{d}(i(x_n), \hat{x}_n) < \delta/3$ whenever $n \geqslant N$). Then for $m \geqslant n \geqslant N$

$$\hat{d}(i(x_m), i(x_n)) \leqslant \hat{d}(i(x_m), \hat{x}_m) + d(\hat{x}_m, \hat{x}_n) + \hat{d}(\hat{x}_n, i(x_n)) < \delta,$$

so since i is an isometry, (x_n) is a Cauchy sequence in X. Let $\hat{x}$ be the point of $\hat{X}$ represented by (x_n).

As in Proposition C.2.13 we can prove that $(i(x_n))$ converges to $\hat{x}$ in $\hat{X}$. For given $\delta > 0$ there exists $N \in \mathbb{N}$ such that $d(x_n, x_m) < \delta/2$ whenever $m \geqslant n \geqslant N$. Now think of a fixed $n \geqslant N$ and let m vary. We get $\hat{d}(i(x_n), \hat{x}) = \lim_{m\to\infty} d(x_n, x_m) \leqslant \delta/2 < \delta$.

Finally we prove that $(\hat{x}_n)$ converges to $\hat{x}$. Let $\varepsilon > 0$. Since $(i(x_n))$ converges to $\hat{x}$, we may choose $N \in \mathbb{N}$ such that $\hat{d}(i(x_n), \hat{x}) < \varepsilon/2$ whenever $n \geqslant N$. We may also choose N large enough that $1/N < \varepsilon/2$, so $\hat{d}(\hat{x}_n, i(x_n)) < 1/n < \varepsilon/2$ for $n \geqslant N$. So for all $n \geqslant N$,

$$\hat{d}(\hat{x}_n, \hat{x}) \leqslant \hat{d}(\hat{x}_n, i(x_n)) + \hat{d}(i(x_n), \hat{x}) < 1/n + \varepsilon/2 < \varepsilon$$

as required. □★

Completions of metric spaces via virtual points Our second approach to completion of a metric space is via virtual points. If x is a point in a metric space (X, d) then there is an associated function $f_x : X \to [0, \infty)$ defined by $f_x(y) = d(x, y)$ for any $y \in X$. This function has the following properties:

(1) $f_x(y) \leqslant f_x(z) + d(y, z)$ for any $y, z \in X$,

(2) $d(y, z) \leqslant f_x(y) + f_x(z)$ for any $y, z \in X$,

(3) $\inf f_x = 0$, where $\inf f_x$ means $\inf f_x(X)$.

The first two properties follow from the triangle inequality and symmetry, and the third one is clear since $f_x(x) = 0$ while $f_x(y) = d(x, y) \geqslant 0$ for any $y \in X$.

Taking our cue from the above, we make the

Definition C.2.20 Given a metric space (X, d) we say that a function $f : X \to [0, \infty)$ is a *virtual point of X* if

(V1) $f(y) \leqslant f(z) + d(y, z)$ for any $y, z \in X$;

(V2) $d(y, z) \leqslant f(y) + f(z)$ for any $y, z \in X$;

(V3) $\inf f = 0$.

Sometimes the term 'virtual point' is reserved for functions f satisfying (V1), (V2), (V3) for which there is no point x in X such that $f = f_x$.

Here are a few elementary deductions.

Proposition C.2.21 If f is a virtual point in a metric space (X, d) then $|f(y)-f(z)| \leqslant d(y, z)$ for all $y, z \in X$.

Proof From (V1), $f(y) - f(z) \leqslant d(y, z)$. By symmetry, also $f(z) - f(y) \leqslant d(z, y) = d(y, z)$. The result follows. □

Corollary C.2.22 Any virtual point in a metric space is continuous.

This follows from Proposition C.2.21 (given $\varepsilon > 0$ take $\delta = \varepsilon$, and then $|f(y) - f(z)| < \varepsilon$ whenever $d(y, z) < \delta$). □

Proposition C.2.23 If f and g are virtual points on a metric space (X, d), and $x, a \in X$ then $|f(x) - g(x)| \leqslant f(a) + g(a)$.

Proof From (V1) applied to f we have $f(x) - f(a) \leqslant d(x, a)$ and from (V2) applied to g we have $d(x, a) \leqslant g(x)+g(a)$. So $f(x)-f(a) \leqslant g(x)+g(a)$, which gives $f(x)-g(x) \leqslant f(a)+g(a)$. By symmetry, also $g(x) - f(x) \leqslant f(a) + g(a)$, and the result follows. □

This result shows that although in general a virtual point is an unbounded function, nevertheless the difference between any two virtual points f and g is bounded; for we may think of a as some fixed 'base point' in X, and then $|f(x) - g(x)| \leqslant f(a) + g(a)$ for all $x \in X$. This will enable us to define a metric on the set of all virtual points on X by $d_\infty(f, g) = \sup_{x\in X} |f(x) - g(x)|$. The proof that d_∞ is a metric is entirely similar to the proof in Example 5.13.

Proposition C.2.24 With the above notation, given any virtual point f in a metric space (X, d) and any point $x \in X$, $d_\infty(f, f_x) = f(x)$.

Proof For any $y \in X$, from Proposition C.2.23 applied with x, a replaced by y, x we get $|f(y)-f_x(y)| \leqslant f(x)+f_x(x) = f(x)$. But also when $y = x$ we have $|f(y)-f_x(y)| = f(x)$ (recall that $f(x) \geqslant 0$). Hence $d_\infty(f, f_x) = \sup_{y\in X} |f(y) - f_x(y)| = f(x)$. □

Next here is a result which explains how virtual points are related to completeness.

Proposition C.2.25 A metric space (X, d) is complete iff for every virtual point f in X there is some $x \in X$ such that $f = f_x$. ('All virtual points are actual points'.)

Proof Suppose first that (X, d) is complete and let f be a virtual point in X. Using condition (V3), let (z_n) be a sequence of points in X such that $f(z_n) \to 0$ as $n \to \infty$. By (V2), $d(z_m, z_n) \leqslant f(z_m) + f(z_n)$ so also $d(z_m, z_n) \to 0$ as $n \to \infty$. This shows that (z_n) is a Cauchy sequence in X, and since X is complete, (z_n) converges to some point $z \in X$. Since f is continuous then $f(z_n) \to f(z)$ as $n \to \infty$, so $f(z) = 0$. From Proposition C.2.21, for any $y \in X$ we have $|f(y) - d(z, y)| \leqslant f(z) = 0$, so $f(y) = d(z, y) = f_z(y)$ for all $y \in X$ and we get $f = f_z$. Conversely suppose that for every virtual point f there is a $z \in X$ such that $f = f_z$. Let (y_n) be a Cauchy sequence in X. Then for each $x \in X$ we get that $d(x, y_n)$ is a Cauchy sequence in $\mathbb{R}$, since $|d(x, y_m) - d(x, y_n)| \leqslant d(y_m, y_n)$. By completeness of $\mathbb{R}$, $\lim_{n\to\infty} d(x, y_n)$ exists. Define $f : X \to [0, \infty)$ by $f(x) = \lim_{n\to\infty} d(x, y_n)$. We check that f is a virtual point in X:

(V1) For any $y, z \in X$, $|d(y, y_n) - d(z, y_n)| \leqslant d(y, z)$, so in the limit as $n \to \infty$ we get $|f(y) - f(z)| \leqslant d(y, z)$.

(V2) For any $y, z \in X$ and any integer n, $d(y, z) \leqslant d(y, y_n) + d(z, y_n)$, so in the limit as $n \to \infty$ we get $d(y, z) \leqslant f(y) + f(z)$.

(V3) Since (y_n) is a Cauchy sequence, given any $\varepsilon > 0$ there exists an integer N such that $d(y_m, y_n) < \varepsilon$ whenever $m \geqslant n \geqslant N$. So $d(y_N, y_n) < \varepsilon/2$ for all $n \geqslant N$, and $f(y_N) = \lim_{n\to\infty} d(y_N, y_n) \leqslant \varepsilon/2 < \varepsilon$. Hence $\inf f = 0$.

By the given condition, there exists a point $z \in X$ such that $f = f_z$. So $f(x) = f_z(x) = d(x, z)$ for all $x \in X$. But by definition of f, $\lim_{n\to\infty} d(z, y_n) = f(z) = d(z, z) = 0$. Hence (y_n) converges to z. This shows that X is complete. □

We are now ready to establish that the set of all virtual points gives a completion.

Theorem C.2.26 Let (X, d) be a metric space, and let X^* denote the set of all virtual points in X, with the metric d_∞ introduced after Proposition C.2.23 above. Then the map $i : X \to X^*$ defined by $i(y) = f_y$ is a completion of X.

Proof Given any $y, z \in X$, by Proposition C.2.24 $d_\infty(f_y, f_z) = f_y(z) = d(y, z)$, which says that i is an isometry into.

Next we prove that $i(X)$ is dense in X^*. For let $f \in X^*$ and $\varepsilon > 0$. By (V3) there is some point $x \in X$ such that $f(x) < \varepsilon$. Then by Proposition C.2.24, $d_\infty(f, f_x) = f(x) < \varepsilon$.

Finally we prove that X^* is complete. By Proposition C.2.25 it is enough to show that any virtual point F on X^* is F_f for some $f \in X^*$. So let F be any virtual point on X^*. Then $F|i(X)$ satisfies the virtual point condition on $i(X)$, and since i is an isometry this gives a virtual point f say on X, in other words $f \in X^*$. We shall prove that $F = F_f$. Since $i(X)$ is dense in X^*, it is enough to show that $F|i(X) = F_f|i(X)$, and this is almost tautologous: we want to check that for any $x \in X$ we have $f(x) = d_\infty(f, f_x)$. But this follows from Proposition C.2.24. □

Completions of metric spaces via embedding in a function space This is the shortest of our three methods of establishing the existence of completions. We need a generalization mentioned in Chapter 16: the space $\mathcal{B}C(X, \mathbb{R})$ of bounded continuous real-valued functions on a metric space X, with the sup metric d_∞, is a complete metric space just as in the special case dealt with in Theorem 16.9. To show that a metric space (X, d) has a completion it is enough to prove that there is an isometry i from X into $\mathcal{B}C(X, \mathbb{R})$, since then we may take $\hat{X}$ to be the closure of $i(X)$ in $\mathcal{B}C(X, \mathbb{R})$ and $\hat{d}$ to be the restriction of d_∞ to $i(X)$. Density of $i(X)$ in its own closure is automatic, and $\hat{X}$ is complete since it is a closed subspace of the complete space $\mathcal{B}C(X, \mathbb{R})$.

We choose a fixed 'base point' $a \in X$, and for any $x \in X$ we define $i(x) = f_x$ where $f_x : X \to \mathbb{R}$ has the formula $f_x(y) = d(x, y) - d(a, y)$. The point of including the $d(a, y)$ term is to ensure that f_x is bounded.

We check first that each f_x is continuous and bounded. For any $y, z \in X$ we have

$$|f_x(y) - f_x(z)| = |d(x, y) - d(a, y) - d(x, z) + d(a, z)|$$

$$\leqslant |d(x, y) - d(x, z)| + |d(a, y) - d(a, z)| \leqslant 2d(y, z), \quad \text{using Exercise 5.1.}$$

(Uniform) continuity of f_x follows (by a 'given $\varepsilon > 0$ take $\delta = \varepsilon/2$' argument). Similarly for any $y \in X$, by Exercise 5.1 $|f_x(y)| = |d(x, y) - d(a, y)| \leqslant d(x, a)$, and for a fixed $x \in X$ this means f_x is bounded.

It remains to show that i is an isometry into, so we have to show that $d_\infty(f_x, f_{x'}) = d(x, x')$ for any $x, x' \in X$. But for any $y \in X$,

$$|f_x(y) - f_{x'}(y)| = |d(x, y) - d(a, y) - d(x', y) + d(a, y| = |d(x, y) - d(x', y)| \leqslant d(x, x'),$$

while putting $y = x'$ gives $|f_x(y) - f_{x'}(y)| = |d(x, y) - d(x', y)| = |d(x, x') - 0| = d(x, x')$. These together show that $\sup_{y \in X} |f_x(y) - f_{x'}(y)| = d(x, x')$, so $x \mapsto f_x$ is an isometry into as required. $\square$

Uniqueness of completions This section is independent of how our completions are constructed, so it is relevant to all three of the preceding sections. We show first that completions are unique up to isometry.

Lemma C.2.27 Suppose that Y is a dense subspace of a metric space (X, d_X) and that $f : Y \to Z$ is an isometry from Y into a complete metric space (Z, d_Z). Then f extends uniquely to an isometry g of X into Z.

Proof For any $x \in X$ there is a sequence (y_n) in Y converging to x, by denseness. Now f is an isometry, so $(f(y_n))$ is a Cauchy sequence in Z since (y_n) is a Cauchy sequence in Y. By completeness of Z, $(f(y_n))$ converges to a point, say $g(x)$, in Z. We now show that $g(x)$ is independent of the choice of (y_n). For if (y'_n) is another sequence in Y converging to x in X, then $d_X(y'_n, y_n) \leqslant d_X(y'_n, x) + d_X(x, y_n)$ so $d_X(y'_n, y_n) \to 0$ as $n \to \infty$, and since f is an isometry also $d_Z(f(y'_n), f(y_n)) \to 0$ as $n \to \infty$, so $\lim_{n\to\infty} f(y'_n) = \lim_{n\to\infty} f(y_n)$. Thus $g(x)$ is well-defined.

In particular if $x \in Y$ we may take (y_n) to be $(x, x, x, \ldots)$ so $g(x) = f(x)$, which shows that g is an extension of f (in other words, $g|Y = f$).

We now show that g is an isometry. Given $x, x' \in X$, suppose that (y_n), (y'_n) are sequences in Y converging to x, x' respectively. Then $d_X(x, x') = \lim_{n\to\infty} d_X(y_n, y'_n)$ by continuity of d_X (Exercise 5.17). Since f is an isometry, $d_X(y_n, y'_n) = d_Z(f(y_n), f(y'_n))$. Now

$$d_Z(g(x), g(x')) = \lim_{n\to\infty} d_Z(f(y_n), f(y'_n)) = \lim_{n\to\infty} d_X(y_n, y'_n) = d_X(x, x').$$

This shows that g is an isometry.

Finally we prove uniqueness of g. This uses the continuity of any isometry. Given any isometry $h : X \to Z$ extending f, and any point $x \in X$, we know there is a sequence (y_n) in Y converging to x, so

$$h(x) = h(\lim_{n\to\infty} y_n) = \lim_{n\to\infty} h(y_n) = \lim_{n\to\infty} f(y_n) = g(x),$$

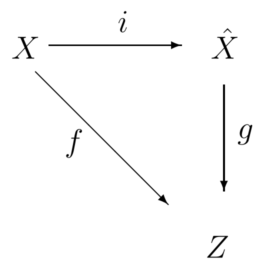
where the second equality follows from continuity of h. Alternatively we could observe that g, h are continuous, $h|Y = g|Y$ and Y is dense in X and apply Exercise 11.8. □

From Lemma C.2.27 we get the following 'universal property' of completions.

Proposition C.2.28 Let $(\hat{X}, i)$ be a completion of a metric space X and let $f : X \to Z$ be an isometry of X into a complete metric space Z. Then there is a unique isometry g of $\hat{X}$ into Z such that $g \circ i = f$.

Proof This follows from Lemma C.2.27 applied with Y taken to be the dense subset $i(X)$ of $\hat{X}$. □

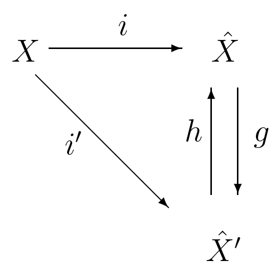
The conclusion of Proposition C.2.28 may be summed up in the following commutative diagram of isometries into:



We may now prove uniqueness of completions up to isometry.

Theorem C.2.29 Suppose that $(\hat{X}, i)$ and $(\hat{X}', i')$ are completions of the same metric space X. Then there is a unique isometry g of $\hat{X}$ onto $\hat{X}'$ such that $g \circ i = i'$.

Proof This may be followed using the following diagram:



By Proposition C.2.28 there is a unique isometry g of $\hat{X}$ into $\hat{X}'$ such that $g \circ i = i'$. Interchanging the roles of $\hat{X}$ and $\hat{X}'$, there is also a unique isometry h from $\hat{X}'$ into $\hat{X}$ such that $h \circ i' = i$. The next bit of the argument is typical trickery used in proofs about universal properties. We apply Proposition C.2.28 to the apparently trivial case when $\hat{X}' = \hat{X}$ and $i' = i$. We get that there is a unique isometry k from $\hat{X}$ into itself such that $k \circ i = i$. But on the one hand it is clear that the identity map of $\hat{X}$ is such an isometry, and on the other hand it follows from our previous arguments that $h \circ g$ is also such an isometry. The upshot is that $h \circ g$ is the identity map of $\hat{X}$. By the same trick applied to $\hat{X}'$, we get that $g \circ h$ is the identity isometry of $\hat{X}'$. Hence g and h are mutually inverse isometries onto, and each carries the 'copy' of X in the completion onto itself by the identity function. This is as much uniqueness as you could hope for. □

★ There is another universal property which completions have. To prove it we need a lemma.

Lemma C.2.30 Suppose that Y is a dense subspace of a metric space $(X,\, d_X)$ and that $f : Y \to Z$ is a uniformly continuous map to a complete metric space $(Z,\, d_Z)$. Then there is a unique uniformly continuous map $g : X \to Z$ such that $g|Y = f$.

Proof This is similar to the proof of Lemma C.2.27, but we give the details briefly . For any $x \in X$ there is a sequence (y_n) in Y converging to x (by Exercise 6.26). Since (y_n) converges in X, it is a Cauchy sequence. Then $(f(y_n))$ is also a Cauchy sequence by uniform continuity of f. But Z is complete, so $(f(y_n))$ converges to a point we call $g(x)$.

We need to check that $g(x)$ is well-defined (independent of the choice of (y_n)). But if also (y'_n) converges to x, then $d_Y(y_n,\, y'_n) \to 0$ as $n \to \infty$ and by uniform continuity of f then $d_Z(f(y_n),\, f(y'_n)) \to 0$ as $n \to \infty$, so the values for $g(x)$ arising from (y_n) and (y'_n) are equal.

The equality $g|Y = f$ follows since for $y \in Y$ a suitable sequence (y_n) is $(y,\, y,\, y,\, \ldots)$ and then $(f(y_n))$ is $(f(y),\, f(y),\, f(y),\, \ldots)$ which converges to $f(y)$.

To prove uniform continuity of g, let $\varepsilon > 0$. By uniform continuity of f, there exists $\delta > 0$ such that $d_Z(f(y),\, f(y')) < \varepsilon/3$ whenever $d_Y(y,\, y') < \delta$. Now let $x,\, x'$ be any points in X with $d_X(x,\, x') < \delta/3$, and let $(y_n),\, (y'_n)$ be sequences in Y converging to $x,\, x'$ respectively. Then there exist integers $N,\, N'$ with $d_X(x,\, y_n) < \delta/3$ whenever $n \geqslant N$ and $d_X(x',\, y'_n) < \delta/3$ whenever $n \geqslant N'$. Take $N_1 = \max\{N,\, N'\}$. Then $d_Y(y_n,\, y'_n) \leqslant d_X(y_n,\, x) + d_X(x,\, x') + d_X(x',\, y'_n) < \delta$ whenever $n \geqslant N_1$. So $d_Z(f(y_n),\, f(y'_n)) < \varepsilon/3$ whenever $n \geqslant N_1$. But $(f(y_n))$ converges to $g(x)$, so there exists an integer N_2 such that $d_Z(f(y_n),\, g(x)) < \varepsilon/3$ whenever $n \geqslant N_2$ and an integer N_3 such that $d_Z(f(y'_n),\, g(x')) < \varepsilon/3$ whenever $n \geqslant N_3$. Now using an $n \geqslant \max\{N_1,\, N_2,\, N_3\}$ we get $d_Z(g(x),\, g(x')) \leqslant d_Z(g(x),\, f(y_n)) + d_Z(f(y_n),\, f(y'_n)) + d_Z(f(y'_n),\, f(x')) < \varepsilon$. Thus g is uniformly continuous on X.

Finally we prove that g is unique. Suppose that g' is another uniformly continuous extension of f to X. Then $g|Y = g'|Y$, and since $g,\, g'$ are continuous and Y is dense in X it follows by Exercise 11.8 that $g = g'$. □

Proposition C.2.31 Suppose that $(\hat{X},\, i)$ is a completion of the metric space X, and that $f : X \to Z$ is a uniformly continuous map to a complete metric space Z. Then there is a unique uniformly continuous map $g : \hat{X} \to Z$ such that $g \circ i = f$.

Proof This follows from Lemma C.2.30 by taking Y to be the dense subset $i(X)$ of $\hat{X}$. □

Theorem C.2.32 Given two completions $(\hat{X},\, i)$ and $(\hat{X}',\, i')$ of the same metric space X, there is a unique uniform equivalence $g : \hat{X} \to \hat{X}'$ such that $g \circ i = i'$.

Proof This follows from Proposition C.2.31 exactly as Theorem C.2.29 followed from Proposition C.2.28. (Isometries are replaced by uniformly continuous maps, and isometries onto by uniform equivalences.) □

In Theorem C.2.32 g is in fact an isometry. The force of the theorem is that g is unique even within the wider class of uniform equivalences.

Lest the two different universal properties of completions contained in Propositions C.2.28 and C.2.32 seem mysterious, it should be mentioned that the book suppresses a whole class of spaces, called uniform spaces, which are more general that metric spaces but less general then topological spaces. There is a completion process for uniform spaces as well as metric spaces, and the universal property expressed in Proposition C.2.32 is really that possessed by the uniform space 'underlying' the metric space. For further information about uniform spaces, see for example 'Topologies and Uniformities' by I. M. James, Springer SUMS 1999. ★

We conclude with a simple example of completion.

Example C.2.33 Consider the open interval $(0, 1)$ in $\mathbb{R}$, with its usual metric. The general machinery for completing this would build up an elaborate space in which the points are equivalence classes of Cauchy sequences in $(0, 1)$. However, we know that $[0, 1]$ is a complete space and $(0, 1)$ is dense in $[0, 1]$ (more elaborately, if $i : (0, 1) \to [0, 1]$ is the inclusion then i is an isometry into $[0, 1]$ and $i((0, 1))$ is dense in $[0, 1]$). So by uniqueness, $[0, 1]$ *is* the completion of $(0, 1)$ 'up to isometry' (which means that any other completion is isometric to $[0, 1]$, and in such a way that $(0, 1)$ is carried identically onto itself.