Chapter 1

Metric and Normed Spaces

We are all familiar with the geometrical properties of ordinary, three-dimensional Euclidean space. A persistent theme in mathematics is the grouping of various kinds of objects into abstract spaces. This grouping enables us to extend our intuition of the relationship between points in Euclidean space to the relationship between more general kinds of objects, leading to a clearer and deeper understanding of those objects.

The simplest setting for the study of many problems in analysis is that of a metric space. A metric space is a set of points with a suitable notion of the distance between points. We can use the metric, or distance function, to define the fundamental concepts of analysis, such as convergence, continuity, and compactness.

A metric space need not have any kind of algebraic structure defined on it. In many applications, however, the metric space is a linear space with a metric derived from a norm that gives the "length" of a vector. Such spaces are called normed linear spaces. For example, n-dimensional Euclidean space is a normed linear space (after the choice of an arbitrary point as the origin). A central topic of this book is the study of infinite-dimensional normed linear spaces, including function spaces in which a single point represents a function. As we will see, the geometrical intuition derived from finite-dimensional Euclidean space remains essential, although completely new features arise in the case of infinite-dimensional spaces.

In this chapter, we define and study metric spaces and normed linear spaces. Along the way, we review a number of definitions and results from real analysis.

1.1 Metrics and norms

Let X be an arbitrary nonempty set.

Definition 1.1 A *metric*, or *distance function*, on X is a function

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

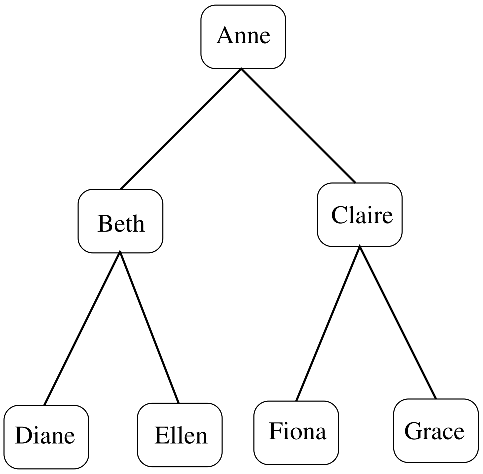

Fig. 1.1 A family tree used in the definition of the ultrametric in Example 1.3.

with the following properties:

(a) $d(x, y) \geq 0$ for all $x, y \in X$, and $d(x, y) = 0$ if and only if $x = y$;
(b) $d(x, y) = d(y, x)$, for all $x, y \in X$;
(c) $d(x, y) \leq d(x, z) + d(z, y)$, for all $x, y, z \in X$.

A *metric space* (X, d) is a set X equipped with a metric d.

When the metric d is understood from the context, we denote a metric space simply by the set X. In words, the definition states that:

(a) distances are nonnegative, and the only point at zero distance from x is x itself;
(b) the distance is a symmetric function;
(c) distances satisfy the *triangle inequality*.

For points in the Euclidean plane, the triangle inequality states that the length of one side of a triangle is less than the sum of the lengths of the other two sides.

Example 1.2 The set of real numbers $\mathbb{R}$ with the distance function $d(x, y) = |x - y|$ is a metric space. The set of complex numbers $\mathbb{C}$ with the distance function $d(z, w) = |z - w|$ is also a metric space.

Example 1.3 Let X be a set of people of the same generation with a common ancestor, for example, all the grandchildren of a grandmother (see Figure 1.1). We define the distance $d(x, y)$ between any two individuals x and y as the number of generations one has to go back along female lines to find the first common ancestor.

For example, the distance between two sisters is one. It is easy to check that d is a metric. In fact, d satisfies a stronger condition than the triangle inequality, namely

$$d(x, y) \leq \max\{d(x, z), d(z, y)\} \qquad \text{for all } x, y, z \in X. \tag{1.1}$$

A metric d which satisfies (1.1) is called an *ultrametric.* Ultrametrics have been used in taxonomy to characterize the genetic proximity of species.

Example 1.4 Let X be the set of n-letter words in a k-character alphabet $A = \{a_1, a_2, \ldots, a_k\}$, meaning that $X = \{(x_1, x_2, \ldots, x_n) \mid x_i \in A\}$. We define the distance $d(x, y)$ between two words $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ to be the number of places in which the words have different letters. That is,

$$d(x, y) = \#\{i \mid x_i \neq y_i\}.$$

Then (X, d) is a metric space.

Example 1.5 Suppose (X, d) is any metric space and Y is a subset of X. We define the distance between points of Y by restricting the metric d to Y. The resulting metric space $(Y, d|_Y)$, or (Y, d) for short, is called a *metric subspace* of (X, d), or simply a subspace when it is clear that we are talking about metric spaces. For example, $(\mathbb{R}, |\cdot|)$ is a metric subspace of $(\mathbb{C}, |\cdot|)$, and the space of rational numbers $(\mathbb{Q}, |\cdot|)$ is a metric subspace of $(\mathbb{R}, |\cdot|)$.

Example 1.6 If X and Y are sets, then the *Cartesian product* $X \times Y$ is the set of ordered pairs (x, y) with $x \in X$ and $y \in Y$. If d_X and d_Y are metrics on X and Y, respectively, then we may define a metric $d_{X\times Y}$ on the product space by

$$d_{X\times Y}\left((x_1, y_1), (x_2, y_2)\right) = d_X\left(x_1, x_2\right) + d_Y\left(y_1, y_2\right)$$

for all $x_1, x_2 \in X$ and $y_1, y_2 \in Y$.

We recall the definition of a linear, or vector, space. We consider only real or complex linear spaces.

Definition 1.7 A *linear space* X over the scalar field $\mathbb{R}$ (or $\mathbb{C}$) is a set of points, or vectors, on which are defined operations of vector addition and scalar multiplication with the following properties:

(a) the set X is a commutative group with respect to the operation $+$ of vector addition, meaning that for all $x, y, z \in X$, we have $x + y = y + x$ and $x + (y + z) = (x + y) + z$, there is a zero vector 0 such that $x + 0 = x$ for all $x \in X$, and for each $x \in X$ there is a unique vector $-x$ such that $x + (-x) = 0$;

(b) for all $x, y \in X$ and $\lambda, \mu \in \mathbb{R}$ (or $\mathbb{C}$), we have $1x = x$, $(\lambda + \mu)x = \lambda x + \mu x$, $\lambda(\mu x) = (\lambda\mu)x$, and $\lambda(x + y) = \lambda x + \lambda y$.

We assume that the reader is familiar with the elementary theory of linear spaces. Some references are given in Section 1.9.

A norm on a linear space is a function that gives a notion of the "length" of a vector.

Definition 1.8 A *norm* on a linear space X is a function $\|\cdot\| : X \to \mathbb{R}$ with the following properties:

(a) $\|x\| \geq 0$, for all $x \in X$ (nonnegative);
(b) $\|\lambda x\| = |\lambda| \|x\|$, for all $x \in X$ and $\lambda \in \mathbb{R}$ (or $\mathbb{C}$) (homogeneous);
(c) $\|x + y\| \leq \|x\| + \|y\|$, for all $x, y \in X$ (triangle inequality) ;
(d) $\|x\| = 0$ implies that $x = 0$ (strictly positive).

A *normed linear space* $(X, \|\cdot\|)$ is a linear space X equipped with a norm $\|\cdot\|$.

A normed linear space is a metric space with the metric

$$d(x, y) = \|x - y\|. \tag{1.2}$$

All the concepts we define for metric spaces therefore apply, in particular, to normed linear spaces. The metric associated with a norm in this way has the special properties of translation invariance, meaning that for all $z \in X$, $d(x+z, y+z) = d(x, y)$, and homogeneity, meaning that for all $\lambda \in \mathbb{R}$ (or $\mathbb{C}$), $d(\lambda x, \lambda y) = |\lambda| d(x, y)$.

The *closed unit ball* $\overline{B}$ of a normed linear space X is the set

$$\overline{B} = \{x \in X : \|x\| \leq 1\}.$$

A subset C of a linear space is *convex* if

$$tx + (1 - t)y \in C \tag{1.3}$$

for all $x, y \in C$ and all real numbers $0 \leq t \leq 1$, meaning that the line segment joining any two points in the set lies in the set. The triangle inequality implies that the unit ball is convex, and its shape gives a good picture of the norm's geometry.

Example 1.9 The set of real numbers $\mathbb{R}$ with the absolute value norm $\|x\| = |x|$ is a one-dimensional real normed linear space. More generally, $\mathbb{R}^n$, where $n = 1, 2, 3, \ldots$, is an n-dimensional linear space. We define the *Euclidean norm* of a point $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ by

$$\|x\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2},$$

and call $\mathbb{R}^n$ equipped with the Euclidean norm n-dimensional *Euclidean space*. We can also define other norms on $\mathbb{R}^n$. For example, the *sum* or 1-norm is given by

$$\|x\|_1 = |x_1| + |x_2| + \cdots + |x_n|.$$

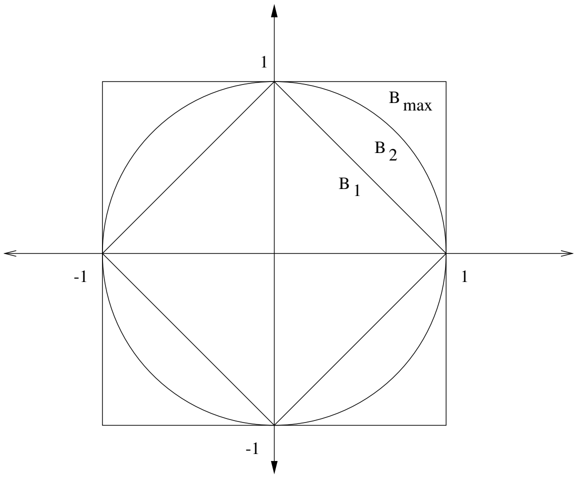

Fig. 1.2 The unit balls in $\mathbb{R}^2$ for the Euclidean norm (B_2), the sum norm (B_1), and the maximum norm ($B_{\max}$).

The *maximum norm* is given by

$$\|x\|_{\max} = \max\{|x_1|, |x_2|, \ldots, |x_n|\}.$$

We also call the maximum norm the ∞-*norm*, and denote it by $\|x\|_\infty$. The unit balls in $\mathbb{R}^2$ for each of these norms are shown in Figure 1.2. We will equip $\mathbb{R}^n$ with the Euclidean norm, unless stated otherwise.

Example 1.10 A *linear subspace* of a linear space, or simply a *subspace* when it is clear we are talking about linear spaces, is a subset that is itself a linear space. A subset M of a linear space X is a subspace if and only if $\lambda x + \mu y \in M$ for all $\lambda, \mu \in \mathbb{R}$ (or $\mathbb{C}$) and all $x, y \in M$. A subspace of a normed linear space is a normed linear space with norm given by the restriction of the norm on X to M.

We will see later on that all norms on a finite-dimensional linear space lead to exactly the same notion of convergence, so often it is not important which norm we use. Different norms on an infinite-dimensional linear space, such as a function space, may lead to completely different notions of convergence, so the specification of a norm is crucial in this case.

We will always regard a normed linear space as a metric space with the metric defined in equation (1.2), unless we explicitly state otherwise. Nevertheless, this equation is not the only way to define a metric on a normed linear space.

Example 1.11 If $(X, \|\cdot\|)$ is a normed linear space, then

$$d(x,y) = \frac{\|x-y\|}{1+\|x-y\|} \tag{1.4}$$

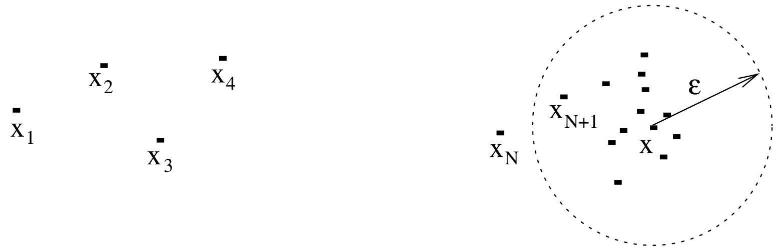

Fig. 1.3 A sequence (x_n) converging to x.

defines a nonhomogeneous, translation invariant metric on X. In this metric, the distance between two points is always less than one.

1.2 Convergence

We first consider the convergence of sequences of real numbers. A *sequence* of real numbers is a map from the natural numbers $\mathbb{N} = \{1, 2, 3, \dots\}$ to $\mathbb{R}$. That is, with each $n \in \mathbb{N}$, we associate a real number $x_n \in \mathbb{R}$. We denote a sequence by (x_n), or $(x_n)_{n=1}^{\infty}$ when we want to indicate the range of the index n. The index n is a "dummy" index, and we may also write the sequence as (x_k) or $(x_k)_{k=1}^{\infty}$.

Another common notation for a sequence is $\{x_n\}$. This notation is a little ambiguous because a sequence is not the same thing as a set. For example,

$$(0, 1, 0, 1, 0, \dots) \text{ and } (1, 0, 0, 0, 0, \dots)$$

are different sequences, but the set of terms is $\{0, 1\}$ in each case.

A *subsequence* of a sequence (x_n) is a sequence of the form (x_{n_k}), where for each $k \in \mathbb{N}$ we have $n_k \in \mathbb{N}$, and $n_k < n_{k+1}$ for all k. That is, $k \mapsto n_k$ is a strictly increasing function from the set of natural numbers to itself. For example, $(1/k^2)_{k=1}^{\infty}$ is a subsequence of $(1/n)_{n=1}^{\infty}$.

The most important concept concerning sequences is convergence.

Definition 1.12 A sequence (x_n) of real numbers *converges* to $x \in \mathbb{R}$ if for every $\epsilon > 0$ there is an $N \in \mathbb{N}$ such that $|x_n - x| < \epsilon$ for all $n \geq N$. The point x is called the *limit* of (x_n).

In this definition, the integer N depends on ϵ, since smaller ϵ's usually require larger N's, and we could write $N(\epsilon)$ to make the dependence explicit. Common ways to write the convergence of (x_n) to x are

$$x_n \to x \text{ as } n \to \infty, \qquad \lim_{n \to \infty} x_n = x.$$

A sequence that does not converge is said to *diverge*. If a sequence diverges because its terms eventually become larger than any number, it is often convenient to regard the sequence as converging to ∞. That is, we say $x_n \to \infty$ if for every $M \in \mathbb{R}$ there is an $N \in \mathbb{N}$ such that $x_n > M$ for all $n \geq N$. Similarly, we say $x_n \to -\infty$ if for every $M \in \mathbb{R}$ there is an $N \in \mathbb{N}$ such that $x_n < M$ for all $n \geq N$.

Example 1.13 Here are a few examples of the limits of convergent sequences:

$$\lim_{n\to\infty} \frac{1}{n} = 0, \qquad \lim_{n\to\infty} n \sin\left(\frac{1}{n}\right) = 1, \qquad \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n = e.$$

The sequence $(\log n)$ diverges because $\log n \to \infty$ as $n \to \infty$. The sequence $((-1)^n)$ diverges because its terms oscillate between -1 and 1, and it does not converge to either ∞ or $-\infty$.

A sequence is said to be *Cauchy* if its terms eventually get arbitrarily close together.

Definition 1.14 A sequence (x_n) is a *Cauchy sequence* if for every $\epsilon > 0$ there is an $N \in \mathbb{N}$ such that $|x_m - x_n| < \epsilon$ for all $m, n \geq N$.

Suppose that (x_n) converges to x. Given $\epsilon > 0$, there is an integer N such that $|x_n - x| < \epsilon/2$ when $n \geq N$. If $m, n \geq N$, then use of the triangle inequality implies that

$$|x_m - x_n| \leq |x_m - x| + |x - x_n| < \epsilon,$$

so (x_n) is Cauchy. Thus, every convergent sequence is a Cauchy sequence. For the real numbers, the converse is also true, and every Cauchy sequence is convergent. The convergence of Cauchy sequences is a fundamental defining property of the real numbers, called *completeness*. We will discuss completeness for general metric spaces in greater detail below.

Example 1.15 The sequence (x_n) with $x_n = \log n$ is not a Cauchy sequence, since $\log n \to \infty$. Nevertheless, we have

$$|x_{n+1} - x_n| = \log\left(1 + \frac{1}{n}\right) \to 0$$

as $n \to \infty$. This example shows that it is not sufficient for successive terms in a sequence to get arbitrarily close together to ensure that the sequence is Cauchy.

We can use the definition of the convergence of a sequence to define the sum of an infinite series as the limit of its sequence of partial sums. Let (x_n) be a sequence in $\mathbb{R}$. The sequence of *partial sums* (s_n) of the *series* $\sum x_n$ is defined by

$$s_n = \sum_{k=1}^{n} x_k. \tag{1.5}$$

If (s_n) converges to a limit s, then we say that the series $\sum x_n$ *converges* to s, and write

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

If the sequence of partial sums does not converge, or converges to infinity, then we say that the series *diverges*. The series $\sum x_n$ is said to be *absolutely convergent* if the series of absolute values $\sum |x_n|$ converges. Absolute convergence implies convergence, but not conversely. A useful property of an absolutely convergent series of real (or complex) numbers is that any series obtained from it by a permutation of its terms converges to the same sum as the original series.

The definitions of convergent and Cauchy sequences generalize to metric spaces in an obvious way. A sequence (x_n) in a metric space (X, d) is a map $n \mapsto x_n$ which associates a point $x_n \in X$ with each natural number $n \in \mathbb{N}$.

Definition 1.16 A sequence (x_n) in X *converges* to $x \in X$ if for every $\epsilon > 0$ there is an $N \in \mathbb{N}$ such that $d(x_n, x) < \epsilon$ for all $n \geq N$. The sequence is *Cauchy* if for every $\epsilon > 0$ there is an $N \in \mathbb{N}$ such that $d(x_m, x_n) < \epsilon$ for all $m, n \geq N$.

Figure 1.3 shows a convergent sequence in the Euclidean plane. Property (a) of the metric in Definition 1.1 implies that if a sequence converges, then its limit is unique. That is, if $x_n \to x$ and $x_n \to y$, then $x = y$. The fact that convergent sequences are Cauchy is an immediate consequence of the triangle inequality, as before. The property that every Cauchy sequence converges singles out a particularly useful class of metric spaces, called complete metric spaces.

Definition 1.17 A metric space (X, d) is *complete* if every Cauchy sequence in X converges to a limit in X. A subset Y of X is *complete* if the metric subspace $(Y, d|_Y)$ is complete. A normed linear space that is complete with respect to the metric (1.2) is called a *Banach space.*

Example 1.18 The space of rational numbers $\mathbb{Q}$ is not complete, since a sequence of rational numbers which converges in $\mathbb{R}$ to an irrational number (such as $\sqrt{2}$ or π) is a Cauchy sequence in $\mathbb{Q}$, but does not have a limit in $\mathbb{Q}$.

Example 1.19 The finite-dimensional linear space $\mathbb{R}^n$ is a Banach space with respect to the sum, maximum, and Euclidean norms defined in Example 1.9. (See Exercise 1.6.)

Series do not make sense in a general metric space, because we cannot add points together. We can, however, consider series in a normed linear space X. Just as for real numbers, if (x_n) is a sequence in X, then the series $\sum_{n=1}^{\infty} x_n$ *converges* to $s \in X$ if the sequence (s_n) of partial sums, defined in (1.5), converges to s.

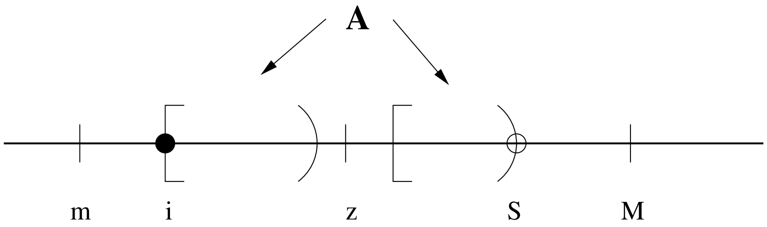

Fig. 1.4 The number M is an upper bound of A and m is a lower bound of A. The number z is neither an upper bound nor a lower bound. The number S is the supremum of A, but does not belong to A. The number i is the infimum of A, and since $i \in A$ it is also the minimum of A.

1.3 Upper and lower bounds

The real numbers have a natural ordering which we can use to define the supremum and infimum of a set of real numbers, and the lim sup and lim inf of a real sequence. Even a metric space as simple as the Euclidean plane cannot be ordered in a way that is compatible with its metric structure. Thus, the definitions in this section are restricted to real sets and sequences. We begin with the definitions of *upper bound* and *lower bound.*

Definition 1.20 Let A be a subset of $\mathbb{R}$. We say that $M \in \mathbb{R}$ is an *upper bound* of A if $x \leq M$ for all $x \in A$, and $m \in \mathbb{R}$ is a *lower bound* of A if $m \leq x$ for all $x \in A$. The set A is *bounded from above* if it has an upper bound, *bounded from below* if it has a lower bound, and *bounded* if it has both an upper and a lower bound.

If A has an upper bound M, then A has many upper bounds. For example, any number $M' \geq M$ is an upper bound.

Definition 1.21 A number M is the *supremum*, or *least upper bound*, of a set $A \subset \mathbb{R}$ if M is an upper bound of A and $M \leq M'$ for all upper bounds M' of A. A number m is the *infimum*, or *greatest lower bound*, of A if m is a lower bound of A and $m \geq m'$ for all lower bounds m' of A. We denote the supremum of A by $\sup A$, and the infimum of A by $\inf A$.

If A is given in the form $A = \{x_\alpha \mid \alpha \in \mathcal{A}\}$, where $\mathcal{A}$ is an indexing set, we also denote the supremum of A by $\sup_{\alpha \in \mathcal{A}} x_\alpha$, or $\sup x_\alpha$ for short.

The supremum and infimum are unique if they exist. For example, if M_1 and M_2 are both least upper bounds of a set A, then the definition implies that $M_1 \leq M_2$ and $M_2 \leq M_1$, so $M_1 = M_2$. The existence of the supremum of every set bounded from above, or the existence of the infimum of every set bounded from below, is a consequence of the completeness of $\mathbb{R}$, and is in fact equivalent to it.

Example 1.22 The subset $A = \{x \in \mathbb{Q} \mid x < \sqrt{2}\}$ of the rational numbers $\mathbb{Q}$ is bounded from above by $\sqrt{2}$, but has no supremum in $\mathbb{Q}$. The supremum in $\mathbb{R}$ is the irrational number $\sqrt{2}$. In this example, the supremum of A does not belong to A.

If A does not have an upper bound, we define $\sup A = \infty$, and if A does not have a lower bound, we define $\inf A = -\infty$. The convention that every number is both an upper and a lower bound of the empty set $\emptyset$ is sometimes convenient, so that $\sup \emptyset = -\infty$ and $\inf \emptyset = \infty$.

The supremum of a set A may, or may not, belong to A itself. If it does, then $\sup A$ is called the *maximum* of A, and is also denoted by $\max A$. Similarly, if the infimum belongs to A, then $\inf A$ is called the *minimum* of A, and is also denoted by $\min A$. The illustration in Figure 1.4 shows an example.

Thus, provided we allow the values $\pm\infty$, every set of real numbers has a supremum and an infimum, but it does not necessarily have a maximum or a minimum.

Next, we define the $\liminf$ and $\limsup$ of a real sequence. First, we consider monotone sequences. A sequence (x_n) is said to be *monotone increasing* if $x_n \leq x_{n+1}$, for every n, and *monotone decreasing* if $x_n \geq x_{n+1}$, for every n. A *monotone* sequence is a sequence that is monotone increasing or monotone decreasing. A monotone increasing sequence converges to its supremum (which could be ∞), and a monotone decreasing sequence converges to its infimum (which could be $-\infty$). Thus, provided that we allow for convergence to $\pm\infty$, all monotone sequences converge.

Now suppose that (x_n) is an arbitrary sequence of real numbers. We construct a new sequence (y_n) by taking the supremum of successively truncated "tails" of the original sequence, $y_n = \sup \{x_k \mid k \geq n\}$. The sequence (y_n) is monotone decreasing because the supremum is taken over smaller sets for larger n's. Therefore, the sequence (y_n) has a limit, which we call the $\limsup$ of the sequence (x_n), and denote by $\limsup x_n$. Similarly, taking the infimum of the successively truncated "tails" of (x_n), we get a monotone increasing sequence. We call the limit of that sequence, the $\liminf$ of (x_n), and denote it by $\liminf x_n$. Thus, we have the following definition.

Definition 1.23 Let (x_n) be a sequence of real numbers. Then

$$\limsup_{n\to\infty} x_n = \lim_{n\to\infty} \left[\sup \{x_k \mid k \geq n\}\right],$$
$$\liminf_{n\to\infty} x_n = \lim_{n\to\infty} \left[\inf \{x_k \mid k \geq n\}\right].$$

Another common notation for the $\limsup$ and $\liminf$ is

$$\limsup x_n = \overline{\lim}\, x_n, \qquad \liminf x_n = \underline{\lim}\, x_n.$$

We make the natural convention that if

$$\sup \{x_k \mid k \geq n\} = \infty, \quad \text{or} \quad \inf \{x_k \mid k \geq n\} = -\infty,$$

for every n, then $\limsup x_n = \infty$, or $\liminf x_n = -\infty$, respectively. In contrast to the limit, the $\liminf$ and $\limsup$ of a sequence of real numbers always exist, provided that we allow the values $\pm\infty$. The $\limsup$ of a sequence whose terms are bounded from above is finite or $-\infty$, and the $\liminf$ of a sequence whose terms are bounded from below is finite or ∞.

It follows from the definition that

$$\liminf_{n\to\infty} x_n \leq \limsup_{n\to\infty} x_n.$$

Moreover, a sequence (x_n) converges if and only if

$$\liminf_{n\to\infty} x_n = \limsup_{n\to\infty} x_n,$$

and, in that case, the limit is the common value of $\liminf x_n$ and $\limsup x_n$.

Example 1.24 If $x_n = (-1)^n$, then

$$\liminf_{n\to\infty} x_n = -1, \qquad \limsup_{n\to\infty} x_n = 1.$$

The lim inf and lim sup have different values and the sequence does not have a limit.

Example 1.25 If $\{x_{n,\alpha} \in \mathbb{R} \mid n \in \mathbb{N}, \alpha \in \mathcal{A}\}$ is a set of real numbers indexed by the natural numbers $\mathbb{N}$ and an arbitrary set $\mathcal{A}$, then

$$\sup_{\alpha\in\mathcal{A}} \left[\liminf_{n\to\infty} x_{n,\alpha}\right] \leq \liminf_{n\to\infty} \left[\sup_{\alpha\in\mathcal{A}} x_{n,\alpha}\right].$$

See Exercise 1.10 for the proof, and the analogous inequality with inf and lim sup.

Suppose that A is a nonempty subset of a general metric space X. The *diameter* of A is

$$\operatorname{diam} A = \sup\{d(x,y) \mid x,y \in A\}.$$

The set A is *bounded* if its diameter is finite. It follows that A is bounded if and only if there is an $M \in \mathbb{R}$ and an $x_0 \in X$ such that $d(x_0, x) \leq M$ for all $x \in A$. The *distance* $d(x, A)$ of a point $x \in X$ from the set A is defined by

$$d(x,A) = \inf\{d(x,y) \mid y \in A\}.$$

The statement $d(x, A) = 0$ does not imply that $x \in A$.

We say that a function $f : X \to Y$ is *bounded* if its range $f(X)$ is bounded. For example, a real-valued function $f : X \to \mathbb{R}$ is bounded if there is a finite number M such that $|f(x)| \leq M$ for all $x \in X$. We say that $f : X \to \mathbb{R}$ is *bounded from above* if there is an $M \in \mathbb{R}$ such that $f(x) \leq M$ for all $x \in X$, and *bounded from below* if there is an $M \in \mathbb{R}$ such that $f(x) \geq M$ for all $x \in X$.

1.4 Continuity

A real function $f : \mathbb{R} \to \mathbb{R}$ is *continuous* at a point $x_0 \in \mathbb{R}$ if for every $\epsilon > 0$ there is a $\delta > 0$ such that $|x - x_0| < \delta$ implies $|f(x) - f(x_0)| < \epsilon$. Thus, continuity of f at x_0 is the property that the value of f at a point close to x_0 is close to the value of f at x_0. The definition of continuity for functions between metric spaces is an obvious generalization of the definition for real functions. Let (X, d_X) and (Y, d_Y) be two metric spaces.

Definition 1.26 A function $f : X \to Y$ is *continuous* at $x_0 \in X$ if for every $\epsilon > 0$ there is a $\delta > 0$ such that $d_X(x, x_0) < \delta$ implies $d_Y(f(x), f(x_0)) < \epsilon$. The function f is *continuous on* X if it is continuous at every point in X.

If f is not continuous at x, then we say that f is *discontinuous* at x. There are continuous functions on any metric space. For example, every constant function is continuous.

Example 1.27 Let $a \in X$, and define $f : X \to \mathbb{R}$ by $f(x) = d(x, a)$. Then f is continuous on X.

We can also define continuity in terms of limits. If $f : X \to Y$, we say that $f(x) \to y_0$ as $x \to x_0$, or

$$\lim_{x \to x_0} f(x) = y_0,$$

if for every $\epsilon > 0$ there is a $\delta > 0$ such that $0 < d_X(x, x_0) < \delta$ implies that $d_Y(f(x), y_0) < \epsilon$. More generally, if $f : D \subset X \to Y$ has domain D, and x_0 is a limit of points in D, then we say $f(x) \to y_0$ as $x \to x_0$ in D if for every $\epsilon > 0$ there is a $\delta > 0$ such that $0 < d_X(x, x_0) < \delta$ and $x \in D$ implies that $d_Y(f(x), y_0) < \epsilon$. A function $f : X \to Y$ is continuous at $x_0 \in X$ if

$$\lim_{x \to x_0} f(x) = f(x_0),$$

meaning that the limit of $f(x)$ as $x \to x_0$ exists and is equal to the value of f at x_0.

Example 1.28 If $f : (0, a) \to Y$ for some $a > 0$, and $f(x) \to L$ as $x \to 0$, then we write

$$\lim_{x \to 0^+} f(x) = L.$$

Similarly, if $f : (-a, 0) \to Y$, and $f(x) \to L$ as $x \to 0$, then we write

$$\lim_{x \to 0^-} f(x) = L.$$

If $f : X \to Y$ and E is a subset of X, then we say that f is *continuous on* E if it is continuous at every point $x \in E$. This property is, in general, not equivalent to the continuity of the restriction $f|_E$ of f on E.

Example 1.29 Let $f : \mathbb{R} \to \mathbb{R}$ be the characteristic function of the rationals, which is defined by

$$f(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q}, \\ 0 & \text{if } x \notin \mathbb{Q}. \end{cases}$$

The function f is discontinuous at every point of $\mathbb{R}$, but $f|_{\mathbb{Q}} : \mathbb{Q} \to \mathbb{R}$ is the constant function $f|_{\mathbb{Q}}(x) = 1$, so $f|_{\mathbb{Q}}$ is continuous on $\mathbb{Q}$.

A subtle, but important, strengthening of continuity is *uniform continuity.*

Definition 1.30 A function $f : X \to Y$ is *uniformly continuous* on X if for every $\epsilon > 0$ there is a $\delta > 0$ such that $d_X(x, y) < \delta$ implies $d_Y(f(x), f(y)) < \epsilon$ for all $x, y \in X$.

The crucial difference between Definition 1.30 and Definition 1.26 is that the value of δ does not depend on the point $x \in X$, so that $f(y)$ gets closer to $f(x)$ at a uniform rate as y gets closer to x.

In the following, we will denote all metrics by d when it is clear from the context which metric is meant.

Example 1.31 The function $r : (0, 1) \to \mathbb{R}$ defined by $r(x) = 1/x$ is continuous on $(0, 1)$ but not uniformly continuous. The function $s : \mathbb{R} \to \mathbb{R}$ defined by $s(x) = x^2$ is continuous on $\mathbb{R}$ but not uniformly continuous. If $[a, b]$ is any bounded interval, then $s|_{[a,b]}$ is uniformly continuous on $[a, b]$.

Example 1.32 A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is *affine* if

$$f(tx + (1-t)y) = tf(x) + (1-t)f(y) \qquad \text{for all } x, y \in \mathbb{R}^n \text{ and } t \in [0, 1].$$

Every affine function is uniformly continuous. An affine function f can be written in the form $f(x) = Ax + b$, where A is a constant $m \times n$ matrix and b is a constant m-vector. Affine functions are more general than linear functions, for which $b = 0$.

There is a useful equivalent way to characterize continuous functions on metric spaces in terms of sequences.

Definition 1.33 A function $f : X \to Y$ is *sequentially continuous* at $x \in X$ if for every sequence (x_n) that converges to x in X, the sequence $(f(x_n))$ converges to $f(x)$ in Y.

Proposition 1.34 Let X, Y be metric spaces. A function $f : X \to Y$ is continuous at x if and only if it is sequentially continuous at x.

Proof. First, we show that if f is continuous, then it is sequentially continuous. Suppose that f is continuous at x, and $x_n \to x$. Let $\epsilon > 0$ be given. By the continuity of f, we can choose $\delta > 0$ so that $d(x, x_n) < \delta$ implies $d(f(x), f(x_n)) < \epsilon$. By the convergence of (x_n), we can choose N so that $n \geq N$ implies $d(x, x_n) < \delta$. Therefore, $n \geq N$ implies $d(f(x), f(x_n)) < \epsilon$, and $f(x_n) \to f(x)$.

To prove the converse, we show that if f is discontinuous, then it is not sequentially continuous. If f is discontinuous at x, then there is an $\epsilon > 0$ such that for every $n \in \mathbb{N}$ there exists $x_n \in X$ with $d(x, x_n) < 1/n$ and $d(f(x), f(x_n)) \geq \epsilon$. The sequence (x_n) converges to x but $(f(x_n))$ does not converge to $f(x)$. □

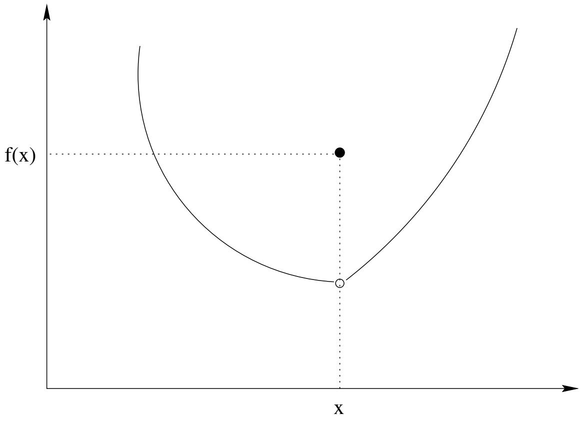

Fig. 1.5 The function f is upper semicontinuous, but not continuous, at the point x.

There are two kinds of "half-continuous" real-valued functions, defined as follows.

Definition 1.35 A function $f : X \to \mathbb{R}$ is *upper semicontinuous* on X if for all $x \in X$ and every sequence $x_n \to x$, we have

$$\limsup_{n\to\infty} f(x_n) \leq f(x).$$

A function f is *lower semicontinuous* on X if for all $x \in X$ and every sequence $x_n \to x$, we have

$$\liminf_{n\to\infty} f(x_n) \geq f(x).$$

The definition is illustrated in Figure 1.5. A function $f : X \to \mathbb{R}$ is continuous if and only if it is upper and lower semicontinuous.

1.5 Open and closed sets

Open sets provide another way to formulate the concepts of convergence and continuity. In this section, we define open sets in a metric space. We will discuss open sets in the more general context of topological spaces in Chapter 4.

Let (X, d) be a metric space. The *open ball*, $B_r(a)$, with radius $r > 0$ and center $a \in X$ is the set

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

The *closed ball*, $\overline{B}_r(a)$, is the set

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

Definition 1.36 A subset G of a metric space X is *open* if for every $x \in G$ there is an $r > 0$ such that $B_r(x)$ is contained in G. A subset F of X is *closed* if its complement $F^c = X \setminus F$ is open.

For example, an open ball is an open set, and a closed ball is a closed set. The following properties of open and closed sets are easy to prove from the definition.

Proposition 1.37 Let X be a metric space.

(a) The empty set $\emptyset$ and the whole set X are open and closed.
(b) A finite intersection of open sets is open.
(c) An arbitrary union of open sets is open.
(d) A finite union of closed sets is closed.
(e) An arbitrary intersection of closed sets is closed.

Example 1.38 The interval $I_n = (-1/n, 1)$ is open in $\mathbb{R}$ for every $n \in \mathbb{N}$, but the intersection

$$\bigcap_{n=1}^{\infty} I_n = [0, 1)$$

is not open. Thus, an infinite intersection of open sets need not be open.

Example 1.39 Let $\{q_n \mid n \in \mathbb{N}\}$ be an enumeration of the rational numbers $\mathbb{Q}$, and $\epsilon > 0$. We define the open interval I_n in $\mathbb{R}$ by

$$I_n = \left(q_n - \frac{\epsilon}{2^n}, q_n + \frac{\epsilon}{2^n}\right).$$

Then $G = \bigcup_{n=1}^{\infty} I_n$ is an open set which contains $\mathbb{Q}$. The sum of the lengths of the intervals I_n is 2ϵ, which can be made as small as we wish. Nevertheless, every interval in $\mathbb{R}$ contains infinitely many rational numbers, and therefore infinitely many intervals I_n.

A subset of $\mathbb{R}$ has *Lebesgue measure zero* if for every $\epsilon > 0$ there is a countable collection of open intervals whose union contains the subset such that the sum of the lengths of the intervals is less than ϵ. Thus, the previous example shows that the set of rational numbers $\mathbb{Q}$, or any other countable subset of $\mathbb{R}$, has measure zero. A property which holds everywhere except on a set of measure zero is said to hold *almost everywhere*, abbreviated *a.e.* For example, the function $\chi_{\mathbb{Q}} : \mathbb{R} \to \mathbb{R}$ that is one on the rational numbers and zero on the irrational numbers is zero almost everywhere.

Every open set in $\mathbb{R}$ is a countable union of disjoint open intervals. The structure of open sets in $\mathbb{R}^n$ for $n \geq 2$ may be much more complicated.

Example 1.40 We define a closed set F_1 in $\mathbb{R}$ by removing the "middle third" $(1/3, 2/3)$ of the interval $[0, 1]$. That is,

$$F_1 = [0, 1/3] \cup [2/3, 1].$$

We define F_2 by removing the middle thirds of the intervals in F_1, so that

$$F_2 = [0, 1/9] \cup [2/9, 1/3] \cup [2/3, 7/9] \cup [8/9, 1].$$

Continuing this removal of middle thirds, we obtain a nested sequence of closed sets (F_n). The intersection $F = \bigcap_{n=1}^{\infty} F_n$ is a closed set called the *Cantor set*. A number $x \in [0, 1]$ belongs to the Cantor set if and only if it has a base three expansion that contains no 1's. The endpoints of the closed intervals in the F_n's do not have a unique expansion. For example, we can write $1/3 \in F$ in base three as $0.1000\ldots$ and as $0.0222\ldots$. The Cantor set is an uncountable set of Lebesgue measure zero which contains no open intervals, and is a simple example of a *fractal*. Heuristically, any part of the set — for example, the left part contained in the interval $[0, 1/3]$ — is a scaled version of the whole set. The name fractal refers to the fact that, with a suitable definition of the Hausdorff dimension of a set, the Cantor set has a fractional dimension of $\log 2/\log 3 \approx 0.631$. The Hausdorff dimension of the Cantor set lies between that of a point, which has dimension 0, and an interval, which has dimension 1.

Closed sets in a metric space can be given an alternative, sequential characterization as sets that contain their limit points.

Proposition 1.41 A subset F of a metric space is closed if and only if every convergent sequence of elements in F converges to a limit in F. That is, if $x_n \to x$ and $x_n \in F$ for all n, then $x \in F$.

Example 1.42 A subset of a complete metric space is complete if and only if it is closed.

The *closure* $\overline{A}$ of a set $A \subset X$ is the smallest closed set containing A. From property (e) of Proposition 1.37, the closure $\overline{A}$ is the intersection of all closed sets that contain A. In a metric space, the closure of a set A can also be obtained by adding to A all limits of convergent sequences of elements of A. That is,

$$\overline{A} = \{x \in X \mid \text{there exist } a_n \in A \text{ such that } a_n \to x\}. \tag{1.6}$$

The closure of the set of rational numbers $\mathbb{Q}$ in the space of real numbers $\mathbb{R}$ is the whole space $\mathbb{R}$. Sets with this property are said to be dense.

Definition 1.43 A subset A of a metric space X is *dense* in X if $\overline{A} = X$.

It follows from (1.6) that A is a dense subset of the metric space X if and only if for every $x \in X$ there is a sequence (a_n) in A such that $a_n \to x$. Thus, every point in X can be approximated arbitrarily closely by points in the dense set A. We will encounter many dense sets later on. Theorem 2.9, the Weierstrass approximation theorem, gives one example.

Definition 1.44 A metric space is *separable* if it has a countable dense subset.

For example, $\mathbb{R}$ with its usual metric is separable because $\mathbb{Q}$ is a countable dense subset. On the other hand, $\mathbb{R}$ with the discrete metric $d(x, y) = 1$ when $x \neq y$ is not separable.

Definition 1.45 Let x be a point in a metric space X. A set $U \subset X$ is a *neighborhood* of x if there is an open set $G \subset U$ with $x \in G$.

Equivalently, a set U is a neighborhood of x if U contains a ball $B_r(x)$ centered at x for some $r > 0$. Definition 1.16 for the convergence of a sequence can therefore be rephrased in the following way. A sequence (x_n) *converges* to x if for every neighborhood U of x there is an $N \in \mathbb{N}$ such that $x_n \in U$ for all $n \geq N$.

The following proposition characterizes continuous functions as functions that "pull back" open sets to open sets.

Proposition 1.46 Let X, Y be metric spaces and $f : X \to Y$. The function f is continuous on X if and only if $f^{-1}(G)$ is open in X for every open set G in Y.

Proof. Suppose that f is continuous and $G \subset Y$ is open. If $a \in f^{-1}(G)$, then there is a $b \in G$ with $b = f(a)$. Since G is open, there is an $\epsilon > 0$ with $B_\epsilon(b) \subset G$. Since f is continuous, there is a $\delta > 0$ such that $d(x, a) < \delta$ implies $d(f(x), b) < \epsilon$. It follows that $B_\delta(a) \subset f^{-1}(G)$, so $f^{-1}(G)$ is open.

Conversely, suppose that f is discontinuous at some point a in X. Then there is an $\epsilon > 0$ such that for every $\delta > 0$, there is an $x \in X$ with $d(x, a) < \delta$ and $d(f(x), f(a)) \geq \epsilon$. It follows that, although a belongs to the inverse image of the open set $B_\epsilon(f(a))$ under f, the inverse image does not contain $B_\delta(a)$ for any $\delta > 0$, so it is not open. □

Example 1.47 If $s : \mathbb{R} \to \mathbb{R}$ is the function $s(x) = x^2$, then $s^{-1}((-4, 4)) = (-2, 2)$ is open, as required by continuity. On the other hand, $s((-2, 2)) = [0, 4)$ is not open. Thus, continuous functions need not map open sets to open sets.

1.6 The completion of a metric space

Working with incomplete metric spaces is very inconvenient. For example, suppose we wish to solve an equation for which we cannot write an explicit expression for the solution. We may instead construct a sequence (x_n) of approximate solutions, for example, by use of an iterative method or some kind of numerical scheme. If the approximate solutions get closer and closer together with increasing n, meaning that they form a Cauchy sequence in a metric space, then we would like to conclude that the approximate solutions have a limit, and then try to show that the limit is a solution. We cannot do this unless the metric space in which the approximations lie is complete.

In this section we explain how to extend an incomplete metric space X to a larger, complete metric space, called the completion of X. We construct the completion of X as a set of equivalence classes of Cauchy sequences in X which "ought" to converge to the same point. For a brief review of equivalence relations and equivalence classes, see Exercise 1.22. A point $x \in X$ is naturally identified with the class of Cauchy sequences in X that converge to x, while classes of Cauchy sequences that do not converge in X correspond to new points in the completion. In effect, we construct the completion by filling the "holes" in X that are detected by its Cauchy sequences.

Example 1.48 The completion of the set of rational numbers $\mathbb{Q}$ is the set of real numbers $\mathbb{R}$. A real number x is identified with the equivalence class of rational Cauchy sequences that converge to x. When we write a real number in decimal notation, we give a Cauchy sequence of rational numbers that converges to it.

In order to give a formal definition of the completion, we require the notion of an isometry between two metric spaces (X, d_X) and (Y, d_Y).

Definition 1.49 A map $\imath : X \to Y$ which satisfies

$$d_Y(\imath(x_1), \imath(x_2)) = d_X(x_1, x_2) \tag{1.7}$$

for all $x_1, x_2 \in X$ is called an *isometry* or an *isometric embedding* of X into Y. An isometry which is onto is called a *metric space isomorphism*, or an *isomorphism* when it is clear from the context that we are dealing with metric spaces. Two metric spaces X and Y are *isomorphic* if there is an isomorphism $\imath : X \to Y$.

Equation (1.7) implies that an isometry $\imath$ is one-to-one and continuous. We think of $\imath$ as "identifying" a point $x \in X$ with its image $\imath(x) \in Y$, so that $\imath(X)$ is a "copy" of X embedded in Y. Two isomorphic metric spaces are indistinguishable as metric spaces, although they may differ in other ways.

Example 1.50 The map $\imath : \mathbb{C} \to \mathbb{R}^2$ defined by $\imath(x + iy) = (x, y)$ is a metric space isomorphism between the complex numbers $(\mathbb{C}, |\cdot|)$ and the Euclidean plane $(\mathbb{R}^2, \|\cdot\|)$. In fact, since $\imath$ is linear, the spaces $\mathbb{C}$ and $\mathbb{R}^2$ are isomorphic as real normed linear spaces.

We can now define the completion of a metric space. The example of the real and rational numbers is helpful to keep in mind while reading this definition.

Definition 1.51 A metric space $(\widetilde{X}, \widetilde{d})$ is called the *completion* of (X, d) if the following conditions are satisfied:

(a) there is an isometric embedding $\imath : X \to \widetilde{X}$;
(b) the image space $\imath(X)$ is dense in $\widetilde{X}$;
(c) the space $(\widetilde{X}, \widetilde{d})$ is complete.

The main theorem about the completion of metric spaces is the following.

Theorem 1.52 Every metric space has a completion. The completion is unique up to isomorphism.

Proof. First, we prove that the completion is unique up to isomorphism, if it exists. Suppose that $(\widetilde{X}_1, \widetilde{d}_1)$ and $(\widetilde{X}_2, \widetilde{d}_2)$ are two completions of (X, d), with corresponding isometric embeddings $\imath_1 : X \to \widetilde{X}_1$ and $\imath_2 : X \to \widetilde{X}_2$. We will use $\imath_1$ to extend $\imath_2$ from X to the completion $\widetilde{X}_1$ and obtain an isomorphism $\widetilde{\imath} : \widetilde{X}_1 \to \widetilde{X}_2$.

To define $\widetilde{\imath}$ on $\widetilde{x} \in \widetilde{X}_1$, we pick a sequence (x_n) in X such that $(\imath_1(x_n))$ converges to $\widetilde{x}$ in $\widetilde{X}_1$. Such a sequence exists because $\imath_1(X)$ is dense in $\widetilde{X}_1$. The sequence $(\imath_1(x_n))$ is Cauchy because it converges. Since $\imath_1$ and $\imath_2$ are isometries, it follows that (x_n) and $(\imath_2(x_n))$ are also Cauchy. The space $\widetilde{X}_2$ is complete, hence $(\imath_2(x_n))$ converges in $\widetilde{X}_2$. We define

$$\widetilde{\imath}(\widetilde{x}) = \lim_{n\to\infty} \imath_2(x_n). \tag{1.8}$$

If (x'_n) is another sequence in X such that $(\imath_1(x'_n))$ converges to $\widetilde{x}$ in $\widetilde{X}_1$, then

$$\widetilde{d}_2\left(\imath_2(x'_n), \imath_2(x_n)\right) = d\left(x'_n, x_n\right) = \widetilde{d}_1\left(\imath_1(x'_n), \imath_1(x_n)\right) \to 0$$

as $n \to \infty$. Thus, $(\imath_2(x'_n))$ and $(\imath_2(x_n))$ converge to the same limit, and $\widetilde{\imath}(\widetilde{x})$ is well-defined.

If $\widetilde{x}$, $\widetilde{y}$ belong to $\widetilde{X}_2$, and

$$\widetilde{\imath}(\widetilde{x}) = \lim_{n\to\infty} \imath_2(x_n), \qquad \widetilde{\imath}(\widetilde{y}) = \lim_{n\to\infty} \imath_2(y_n),$$

then

$$\widetilde{d}_2\left(\widetilde{\imath}(\widetilde{x}), \widetilde{\imath}(\widetilde{y})\right) = \lim_{n\to\infty} \widetilde{d}_2(\imath_2(x_n), \imath_2(y_n)) = \lim_{n\to\infty} d(x_n, y_n) = \widetilde{d}_1(\widetilde{x}, \widetilde{y}).$$

Therefore $\widetilde{\imath}$ is an isometry of $\widetilde{X}_1$ into $\widetilde{X}_2$. By using constant sequences in X, we see that $\widetilde{\imath} \circ \imath_1(x) = \imath_2(x)$ for all $x \in X$, so that $\widetilde{\imath}$ identifies the image of X in $\widetilde{X}_1$ under $\imath_1$ with the image of X in $\widetilde{X}_2$ under $\imath_2$.

To show that $\widetilde{\imath}$ is onto, we observe that $\widetilde{X}_1$ contains the limit of all Cauchy sequences in $\imath_1(X)$, so the isomorphic space $\widetilde{\imath}(\widetilde{X}_1)$ contains the limit of all Cauchy sequences in $\imath_2(X)$. Therefore $\overline{\imath_2(X)} \subset \widetilde{\imath}(\widetilde{X}_1)$. By assumption, $\imath_2(X)$ is dense in $\widetilde{X}_2$, so $\overline{\imath_2(X)} = \widetilde{X}_2$, and $\widetilde{\imath}(\widetilde{X}_1) = \widetilde{X}_2$. This shows that any two completions are isomorphic.

Second, we prove the completion exists. To do this, we construct a completion from Cauchy sequences in X. We define a relation $\sim$ between Cauchy sequences $x = (x_n)$ and $y = (y_n)$ in X by

$$x \sim y \quad \text{if and only if} \quad \lim_{n\to\infty} d(x_n, y_n) = 0.$$

Two convergent Cauchy sequences x, y satisfy $x \sim y$ if and only if they have the same limit. It is straightforward to check that $\sim$ is an equivalence relation on the set $\mathcal{C}$ of Cauchy sequences in X. Let $\widetilde{X}$ be the set of equivalence classes of $\sim$ in $\mathcal{C}$. We call an element $(x_n) \in \widetilde{x}$ of an equivalence class $\widetilde{x} \in \widetilde{X}$, a *representative* of $\widetilde{x}$.

We define $\widetilde{d} : \widetilde{X} \times \widetilde{X} \to \mathbb{R}$ by

$$\widetilde{d}(\widetilde{x}, \widetilde{y}) = \lim_{n\to\infty} d(x_n, y_n), \tag{1.9}$$

where (x_n) and (y_n) are any two representatives of $\widetilde{x}$ and $\widetilde{y}$, respectively. The limit in (1.9) exists because $(d(x_n, y_n))_{n=1}^{\infty}$ is a Cauchy sequence of real numbers. For this definition to make sense, it is essential that the limit is independent of which representatives of $\widetilde{x}$ and $\widetilde{y}$ are chosen. Suppose that (x_n), (x'_n) represent $\widetilde{x}$ and (y_n), (y'_n) represent $\widetilde{y}$. Then, by the triangle inequality, we have

$$\begin{aligned} d(x_n, y_n) &\leq d(x_n, x'_n) + d(x'_n, y'_n) + d(y'_n, y_n), \\ d(x_n, y_n) &\geq d(x'_n, y'_n) - d(x_n, x'_n) - d(y'_n, y_n). \end{aligned}$$

Taking the limit as $n \to \infty$ of these inequalities, and using the assumption that $(x_n) \sim (x'_n)$ and $(y_n) \sim (y'_n)$, we find that

$$\lim_{n\to\infty} d(x_n, y_n) = \lim_{n\to\infty} d(x'_n, y'_n).$$

Thus, the limit in (1.9) is independent of the representatives, and $\widetilde{d}$ is well-defined. It is straightforward to check that $\widetilde{d}$ is a metric on $\widetilde{X}$.

To show that the metric space $(\widetilde{X}, \widetilde{d})$ is a completion of (X, d), we define an embedding $\imath : X \to \widetilde{X}$ as the map that takes a point $x \in X$ to the equivalence class of Cauchy sequences that contains the constant sequence (x_n) with $x_n = x$ for all n. This map is an isometric embedding, since if (x_n) and (y_n) are the constant sequences with $x_n = x$ and $y_n = y$, we have

$$\widetilde{d}(\imath(x), \imath(y)) = \lim_{n\to\infty} d(x_n, y_n) = d(x, y).$$

The image $\imath(X)$ consists of the equivalence classes in $\widetilde{X}$ which have a constant representative Cauchy sequence. To show the density of $\imath(X)$ in $\widetilde{X}$, let (x_n) be a representative of an arbitrary point $\widetilde{x} \in \widetilde{X}$. We define a sequence $(\widetilde{y}_n)$ of constant sequences by $\widetilde{y}_n = (y_{n,k})_{k=1}^{\infty}$ where $y_{n,k} = x_n$ for all $n, k \in \mathbb{N}$. From the definition of $(\widetilde{y}_n)$ and the fact that (x_n) is a Cauchy sequence, we have

$$\lim_{n\to\infty} \widetilde{d}(\widetilde{y}_n, \widetilde{x}) = \lim_{n\to\infty} \lim_{k\to\infty} d(x_n, x_k) = 0.$$

Thus, $\imath(X)$ is dense in $\widetilde{X}$.

Finally, we prove that $(\widetilde{X}, \widetilde{d})$ is complete. We will use Cantor's "diagonal" argument, which is useful in many other contexts as well. Let $(\widetilde{x}_n)$ be a Cauchy sequence in $\widetilde{X}$. In order to prove that a Cauchy sequence is convergent, it is enough to prove that it has a convergent subsequence, because the whole sequence converges to the limit of any subsequence. Picking a subsequence, if necessary, we can assume that $(\widetilde{x}_n)$ satisfies

$$\widetilde{d}\,(\widetilde{x}_m, \widetilde{x}_n) \le \frac{1}{N} \quad \text{for all } m, n \ge N. \tag{1.10}$$

For each term $\widetilde{x}_n$, we choose a representative Cauchy sequence in X, denoted by $(x_{n,k})_{k=1}^{\infty}$. Any subsequence of a representative Cauchy sequence of $\widetilde{x}_n$ is also a representative of $\widetilde{x}_n$. We can therefore choose the representative so that

$$d\,(x_{n,k}, x_{n,l}) < \frac{1}{n} \quad \text{for all } k, l \ge n. \tag{1.11}$$

We claim that the "diagonal" sequence $(x_{k,k})_{k=1}^{\infty}$ is a Cauchy sequence, and that the equivalence class $\widetilde{x}$ to which it belongs is the limit of $(\widetilde{x}_n)$ in $\widetilde{X}$. The fact that we can obtain the limit of a Cauchy sequence of sequences by taking a diagonal sequence is the key point in proving the existence of the completion.

To prove that the diagonal sequence is Cauchy, we observe that for any $i \in \mathbb{N}$,

$$d\,(x_{k,k}, x_{l,l}) \le d\,(x_{k,k}, x_{k,i}) + d\,(x_{k,i}, x_{l,i}) + d\,(x_{l,i}, x_{l,l})\,. \tag{1.12}$$

The definition of $\widetilde{d}$ and (1.10) imply that for all $k, l \ge N$,

$$\widetilde{d}(\widetilde{x}_k, \widetilde{x}) = \lim_{i \to \infty} d(x_{k,i}, x_{l,i}) \le \frac{1}{N}. \tag{1.13}$$

Taking the lim sup of (1.12) as $i \to \infty$, and using (1.11) and (1.13) in the result, we find that for all $k, l \ge N$,

$$d\,(x_{k,k}, x_{l,l}) \le \frac{3}{N}.$$

Therefore $(x_{k,k})$ is Cauchy.

By a similar argument, we find that for all $k, n \ge N$,

$$d\,(x_{n,k}, x_{k,k}) \le \limsup_{i \to \infty} \left\{d\,(x_{n,k}, x_{n,i}) + d\,(x_{n,i}, x_{k,i}) + d\,(x_{k,i}, x_{k,k})\right\} \le \frac{3}{N}.$$

Therefore, for $n \ge N$, we have

$$\widetilde{d}\,(\widetilde{x}_n, \widetilde{x}) = \lim_{k \to \infty} d(x_{n,k}, x_{k,k}) \le \frac{3}{N}.$$

Hence, the Cauchy sequence $(\widetilde{x}_n)$ converges to $\widetilde{x}$ as $n \to \infty$, and $\widetilde{X}$ is complete. □

It is slightly annoying that the completion $\widetilde{X}$ is constructed as a space of equivalence classes of sequences in X, rather than as a more direct extension of X. For example, if X is a space of functions, then there is no guarantee that its completion can be identified with a space of functions that is obtained by adding more functions to the original space.

Example 1.53 Let $C([0,1])$ be the set of continuous functions $f : [0,1] \to \mathbb{R}$. We define the L^2-norm of f by

$$\|f\| = \left(\int_0^1 |f(x)|^2 \, dx \right)^{1/2}.$$

The associated metric $d(f,g) = \|f - g\|$ is a very useful one, analogous to the Euclidean metric on $\mathbb{R}^n$, but the space $C([0,1])$ is not complete with respect to it. The completion is denoted by $L^2([0,1])$, and it can nearly be identified with the space of Lebesgue measurable, square-integrable functions. More precisely, a point in $L^2([0,1])$ can be identified with an equivalence class of square-integrable functions, in which two functions that differ on a set of Lebesgue measure zero are equivalent. According to the Riesz-Fisher theorem, if (f_n) is a Cauchy sequence with respect to the L^2-norm, then there is a subsequence (f_{n_k}) that converges pointwise-a.e. to a square-integrable function, and this fact provides one way to identify an element of the completion with an equivalence class of functions. Many of the usual operations on functions can be defined on equivalence classes, independently of which representative function is chosen, but the pointwise value of an element $f \in L^2([0,1])$ cannot be defined unambiguously.

In a similar way, the space $L^2(\mathbb{R})$ of equivalence classes of Lebesgue measurable, square integrable functions on $\mathbb{R}$ is the completion of the space $C_c(\mathbb{R})$ of continuous functions on $\mathbb{R}$ with compact support (see Definition 2.6) with respect to the L^2-norm

$$\|f\| = \left(\int_{\mathbb{R}} |f(x)|^2 \, dx \right)^{1/2}.$$

We will see later on that these L^2 spaces are fundamental examples of infinite-dimensional Hilbert spaces. We discuss measure theory in greater detail in Chapter 12. We will use facts from that chapter as needed throughout the book, including Fubini's theorem for the exchange in the order of integration, and the dominated convergence theorem for passage to the limit under an integral sign.

1.7 Compactness

Compactness is one the most important concepts in analysis. A simple and useful way to define compact sets in a metric space is by means of sequences.

Definition 1.54 A subset K of a metric space X is *sequentially compact* if every sequence in K has a convergent subsequence whose limit belongs to K.

We can take $K = X$ in this definition, so that X is sequentially compact if every sequence in X has a convergent subsequence. A subset K of (X, d) is sequentially compact if and only if the metric subspace $(K, d|_K)$ is sequentially compact.

Example 1.55 The space of real numbers $\mathbb{R}$ is not sequentially compact. For example, the sequence (x_n) with $x_n = n$ has no convergent subsequence because $|x_m - x_n| \geq 1$ for all $m \neq n$. The closed, bounded interval $[0, 1]$ is a sequentially compact subset of $\mathbb{R}$, as we prove below. The half-open interval $(0, 1]$ is not a sequentially compact subset of $\mathbb{R}$, because the sequence $(1/n)$ converges to 0, and therefore has no subsequence with limit in $(0, 1]$. The limit does, however, belong to $[0, 1]$.

The full importance of compact sets will become clear only in the setting of infinite-dimensional normed spaces. It is nevertheless interesting to start with the finite-dimensional case. Compact subsets of $\mathbb{R}^n$ have a simple, explicit characterization.

Theorem 1.56 (Heine-Borel) A subset of $\mathbb{R}^n$ is sequentially compact if and only if it is closed and bounded.

The fact that closed, bounded subsets of $\mathbb{R}^n$ are sequentially compact is a consequence of the following theorem, called the Bolzano-Weierstrass theorem, even though Bolzano had little to do with its proof. We leave it to the reader to use this theorem to complete the proof of the Heine-Borel theorem.

Theorem 1.57 (Bolzano-Weierstrass) Every bounded sequence in $\mathbb{R}^n$ has a convergent subsequence.

Proof. We will construct a Cauchy subsequence from an arbitrary bounded sequence. Since $\mathbb{R}^n$ is complete, the subsequence converges.

Let (x_k) be a bounded sequence in $\mathbb{R}^n$. There is an $M > 0$ such that $x_k \in [-M, M]^n$ for all k. The set $[-M, M]^n$ is an n-dimensional cube of side $2M$. We denote this cube by C_0. We partition C_0 into 2^n cubes of side M. We denote by C_1 one of the smaller cubes that contains infinitely many terms of the sequence (x_k), meaning that $x_k \in C_1$ for infinitely many $k \in \mathbb{N}$. Such a cube exists because there is a finite number of cubes and an infinite number of terms in the sequence. Let k_1 be the smallest index such that $x_{k_1} \in C_1$. We pick x_{k_1} as the first term of the subsequence.

To choose the second term, we form a new sequence (y_k) by deleting from (x_k) the term x_{k_1} and all terms which do not belong to C_1. We repeat the procedure described in the previous paragraph, but with (x_k) replaced by (y_k), and C_0 replaced

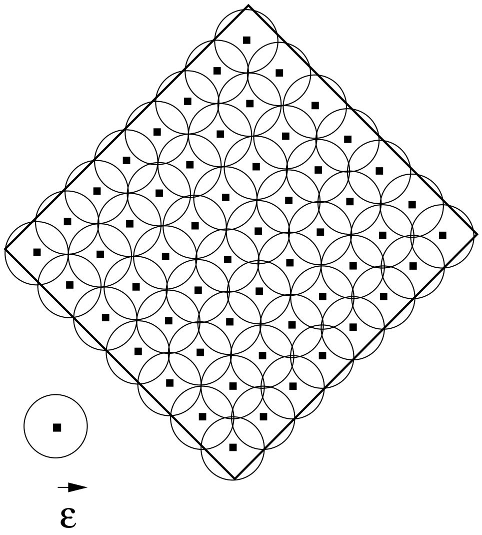

Fig. 1.6 A set with a finite ϵ-net for it.

by C_1. This procedure gives a subcube C_2 of C_1 of side $M/2$, which contains infinitely many terms of the original sequence, and an element y_{k_1}. We pick $x_{k_2} = y_{k_1}$ as the second element of the subsequence.

By repeating this procedure, we obtain a subsequence $(x_{k_i})_{i=1}^{\infty}$. We never "exhaust" the original sequence, because every cube in the construction contains infinitely many terms. We have $x_{k_i} \in C_j$ for all $i \geq j$ where C_j is a cube of side $M/2^{j-1}$. Therefore (x_{k_i}) is a Cauchy sequence, and hence it converges. □

The following criterion for the sequential compactness of a metric space is often easier to verify than the definition. Let A be a subset of a metric space X. We say that a collection $\{G_\alpha \mid \alpha \in \mathcal{A}\}$ of subsets of X is a *cover* of A if its union contains A, meaning that

$$A \subset \bigcup_{\alpha \in \mathcal{A}} G_\alpha.$$

The number of sets in the cover is not required to be countable. If every G_α in the cover is open, then we say that $\{G_\alpha\}$ is an *open cover* of A.

Let $\epsilon > 0$. A subset $\{x_\alpha \mid \alpha \in \mathcal{A}\}$ of X is called an *ϵ-net* of the subset A if the family of open balls $\{B_\epsilon(x_\alpha) \mid \alpha \in \mathcal{A}\}$ is an open cover of A. If the set $\{x_\alpha\}$ is finite, then we say that $\{x_\alpha\}$ is a *finite ϵ-net* of A (see Figure 1.6).

Definition 1.58 A subset of a metric space is *totally bounded* if it has a finite ϵ-net for every $\epsilon > 0$.

That is, a subset A of a metric space X is totally bounded if for every $\epsilon > 0$ there is a finite set of points $\{x_1, x_2, \dots, x_n\}$ in X such that $A \subset \bigcup_{i=1}^{n} B_\epsilon(x_i)$.

Theorem 1.59 A subset of a metric space is sequentially compact if and only if it is complete and totally bounded.

Proof. The proof that a complete, totally bounded set K is sequentially compact is the same as the proof of the Bolzano-Weierstrass theorem 1.57. Suppose that (x_n) is a sequence in K. Then, since K is totally bounded, there is a sequence of balls (B_k) such that B_k has radius $1/2^k$ and every intersection $A_k = \bigcap_{i=1}^k B_i$ contains infinitely many terms of the sequence. We can therefore choose a subsequence (x_{n_k}) such that $x_{n_k} \in A_k$ for every k. This subsequence is Cauchy, and, since K is complete, it converges.

To prove the converse, we show that a sequentially compact space is complete, and that a space which is not totally bounded is not sequentially compact.

If (x_n) is a Cauchy sequence in a sequentially compact space K, then it has a convergent subsequence. The whole Cauchy sequence converges to the limit of any convergent subsequence. Hence K is complete.

Now suppose that K is not totally bounded. Then there is an $\epsilon > 0$ such that K has no finite ϵ-net. For every finite subset $\{x_1, \dots, x_n\}$ of K, there is a point $x_{n+1} \in K$ such that $x_{n+1} \notin \bigcup_{i=1}^n B_\epsilon(x_i)$. Consequently, we can find an infinite sequence (x_n) in K such that $d(x_m, x_n) \geq \epsilon$ for all $m \neq n$. This sequence does not contain a Cauchy subsequence, and hence has no convergent subsequence. Therefore K is not sequentially compact. □

Another way to define compactness is in terms of open sets. We say that a cover $\{G_\alpha\}$ of A has a finite subcover if there is a finite subcollection of sets $\{G_{\alpha_1}, \dots, G_{\alpha_n}\}$ such that $A \subset \bigcup_{i=1}^n G_{\alpha_i}$.

Definition 1.60 A subset K of a metric space X is *compact* if every open cover of K has a finite subcover.

Example 1.61 The space of real numbers $\mathbb{R}$ is not compact, since the open cover $\{(n-1, n+1) \mid n \in \mathbb{Z}\}$ of $\mathbb{R}$ has no finite subcover. The half-open interval $(0, 1]$ is not compact, since the open cover $\{(1/2n, 2/n) \mid n \in \mathbb{N}\}$ has no finite subcover. If this open cover is extended to an open cover of $[0, 1]$, then the extension must contain an open neighborhood of 0. This open neighborhood, together with a finite number of sets from the cover of $(0, 1]$, is a finite subcover of $[0, 1]$.

For metric spaces, compactness and sequential compactness are equivalent.

Theorem 1.62 A subset of a metric space is compact if and only if it is sequentially compact.

Proof. First, we prove that sequential compactness implies compactness. We will show that an arbitrary open cover of a sequentially compact set has a countable subcover, and that a countable cover has a finite subcover.

Lemma 1.63 A sequentially compact metric space is separable.

Proof. By Theorem 1.59, there is a finite $(1/n)$-net A_n of a sequentially compact space K for every $n \in \mathbb{N}$. Let $A = \bigcup_{n=1}^{\infty} A_n$. Then A is countable, because it is a countable union of finite sets, and A is dense in K by construction. □

Suppose that $\{G_\alpha \mid \alpha \in \mathcal{A}\}$ is an arbitrary open cover of a sequentially compact space K. From Lemma 1.63, the space K has a countable dense subset A. Let $\mathcal{B}$ be the collection of open balls with rational radius and center in A, and let $\mathcal{C}$ be the subcollection of balls in $\mathcal{B}$ that are contained in at least one of the open sets G_α. The collection $\mathcal{B}$ is countable because it is a countable union of countable sets. Hence, the subcollection $\mathcal{C}$ is also countable.

For every $x \in K$, there is a set G_α in the open cover of K with $x \in G_\alpha$. Since G_α is open, there is an $\epsilon > 0$ such that $B_\epsilon(x) \subset G_\alpha$. Since A is dense in K, there is a point $y \in A$ such that $d(x, y) < \epsilon/3$. Then $x \in B_{\epsilon/3}(y)$, and $B_{2\epsilon/3}(y) \subset G_\alpha$. (It may help to draw a picture!) Thus, if q is a rational number with $\epsilon/3 < q < 2\epsilon/3$, then $x \in B_q(y)$ and $B_q(y) \subset G_\alpha$. It follows that $B_q(y) \in \mathcal{C}$, so any point x in K belongs to a ball in $\mathcal{C}$. Hence $\mathcal{C}$ is an open cover of K. For every $B \in \mathcal{C}$, we pick an $\alpha_B \in \mathcal{A}$ such that $B \subset G_{\alpha_B}$. Then $\{G_{\alpha_B} \mid B \in \mathcal{C}\}$ is a countable subcover of K, because $\bigcup_{B \in \mathcal{C}} G_{\alpha_B}$ contains $\bigcup_{B \in \mathcal{C}} B$, which contains K.

We will show by contradiction that a countable open cover has a finite subcover. Suppose that $\{G_n \mid n \in \mathbb{N}\}$ is a countable open cover of a sequentially compact space K that does not have a finite subcover. Then the finite union $\bigcup_{n=1}^{N} G_n$ does not contain K for any N. We can therefore construct a sequence (x_k) in K as follows. We pick a point $x_1 \in K$. Since $\{G_n\}$ covers K, there is an N_1 such that $x_1 \in G_{N_1}$. We pick $x_2 \in K$ such that $x_2 \notin \bigcup_{n=1}^{N_1} G_n$, and choose N_2 such that $x_2 \in G_{N_2}$. Then we pick $x_3 \in K$ such that $x_3 \notin \bigcup_{n=1}^{N_2} G_n$, and so on. Since

$$x_k \in G_{N_k}, \quad \text{and} \quad x_k \notin \bigcup_{n=1}^{N_{k-1}} G_n,$$

the open set G_{N_k} is not equal to G_n for any $n \leq N_{k-1}$. Thus, the sequence (N_k) is strictly increasing, and $N_k \to \infty$ as $k \to \infty$. It follows that, for any n, there is an integer K_n such that $x_k \notin G_n$ when $k \geq K_n$. If $x \in G_n$, then all points of the sequence eventually leave the open neighborhood G_n of x, so no subsequence of (x_k) can converge to x. Since the collection $\{G_n\}$ covers K, the sequence (x_n) has no subsequence that converges to a point of K. This contradicts the sequential compactness of K, and proves that sequential compactness implies compactness.

To prove the converse, we show that if a space is not sequentially compact, then it is not compact. Suppose that K has a sequence (x_n) with no convergent subsequence. Such a sequence must contain an infinite number of distinct points, so we can assume without loss of generality that $x_m \neq x_n$ for $m \neq n$.

Let $x \in K$. If the open ball $B_\epsilon(x)$ contains a point in the sequence that is distinct from x for every $\epsilon > 0$, then x is the limit of a subsequence, which contradicts the assumption that the sequence has no convergent subsequence in K. Hence, there is an $\epsilon_x > 0$ such that the open ball $B_{\epsilon_x}(x)$ contains either no points in the sequence, if x itself does not belong to the sequence, or one point, if x belongs to the sequence.

The collection of open balls $\{B_{\epsilon_x}(x) \mid x \in K\}$ is an open cover of K. Every finite subcollection of n open balls contains at most n terms of the sequence. Since the terms of the sequence are distinct, no finite subcollection covers K. Thus, K has an open cover with no finite subcover, and K is not compact. □

In future, we will abbreviate "sequentially compact" to "compact" when referring to metric spaces. The following terminology is often convenient.

Definition 1.64 A subset A of a metric space X is *precompact* if its closure in X is compact.

The term "relatively compact" is frequently used instead of "precompact." This definition means that A is precompact if every sequence in A has a convergent subsequence. The limit of the subsequence can be any point in X, and is not required to belong to A. Since compact sets are closed, a set is compact if and only if it is closed and precompact. A subset of a complete metric space is precompact if and only if it is totally bounded.

Example 1.65 A subset of $\mathbb{R}^n$ is precompact if and only if it is bounded.

Continuous functions on compact sets have several nice properties. From Proposition 1.34, continuous functions preserve the convergence of sequences. It follows immediately from Definition 1.54 that continuous functions preserve compactness.

Theorem 1.66 Let $f : K \to Y$ be continuous on K, where K is a compact metric space and Y is any metric space. Then $f(K)$ is compact.

Since compact sets are bounded, continuous functions on a compact set are bounded. Moreover, continuous functions on compact sets are uniformly continuous.

Theorem 1.67 Let $f : K \to Y$ be a continuous function on a compact set K. Then f is uniformly continuous.

Proof. Suppose that f is not uniformly continuous. Then there is an $\epsilon > 0$ such that for all $\delta > 0$, there are $x, y \in X$ with $d(x, y) < \delta$ and $d(f(x), f(y)) \geq \epsilon$. Taking $\delta = 1/n$ for $n \in \mathbb{N}$, we find that there are sequences (x_n) and (y_n) in X such that

$$d(x_n, y_n) < \frac{1}{n}, \qquad d(f(x_n), f(y_n)) \geq \epsilon. \tag{1.14}$$

Since K is compact there are convergent subsequences of (x_n) and (y_n) which, for simplicity, we again denote by (x_n) and (y_n). From (1.14), the subsequences converge to the same limit, but the sequences $(f(x_n))$ and $(f(y_n))$ either diverge or converge to different limits. This contradicts the continuity of f. □

1.8 Maxima and minima

Maximum and minimum problems are of central importance in applications. For example, in many physical systems, the equilibrium state is one which minimizes energy or maximizes entropy, and in optimization problems, the desirable state of a system is one which minimizes an appropriate cost function. The mathematical formulation of these problems is the maximization or minimization of a real-valued function f on a state space X. Each point of the state space, which is often a metric space, represents a possible state of the system. The existence of a maximizing, or minimizing, point of f in X may not be at all clear; indeed, such a point may not exist. The following theorem gives sufficient conditions for the existence of maximizing or minimizing points — namely, that the function f is continuous and the state space X is compact. Although these conditions are fundamental, they are too strong to be useful in many applications. We will return to these issues later on.

Theorem 1.68 Let K be a compact metric space and $f : K \to \mathbb{R}$ a continuous, real-valued function. Then f is bounded on K and attains its maximum and minimum. That is, there are points $x, y \in K$ such that

$$f(x) = \inf_{z \in K} f(z), \qquad f(y) = \sup_{z \in K} f(z). \tag{1.15}$$

Proof. From Theorem 1.66, the image $f(K)$ is a compact subset of $\mathbb{R}$, and therefore f is bounded by the Heine-Borel theorem in Theorem 1.56.

It is enough to prove that f attains its infimum, because the application of this result to $-f$ implies that f attains its supremum. Since f is bounded, it is bounded from below, and the infimum m of f on K is finite. By the definition of the infimum, for each $n \in \mathbb{N}$ there is an $x_n \in K$ such that

$$m \leq f(x_n) < m + \frac{1}{n}.$$

This inequality implies that

$$\lim_{n \to \infty} f(x_n) = m. \tag{1.16}$$

The sequence $(x_n)_{n=1}^{\infty}$ need not converge, but since K is compact the sequence has a convergent subsequence, which we denote by $(x_{n_k})_{k=1}^{\infty}$. We denote the limit of the subsequence by x. Then, since f is continuous, we have from (1.16) that

$$f(x) = \lim_{k \to \infty} f\left(x_{n_k}\right) = m.$$

Therefore, f attains its infimum m at x. □

The strategy of this proof is typical of many compactness arguments. We construct a sequence of approximate solutions of our problem, in this case a minimizing sequence (x_n) that satisfies (1.16). We use compactness to extract a convergent subsequence, and show that the limit of the convergent subsequence is a solution of our problem, in this case a point where f attains its infimum. The following examples illustrate Theorem 1.68 and some possible behaviors of minimizing sequences.

Example 1.69 The function $f(x) = x^4/4 - x^2/2$ is continuous and bounded on $[-2, 2]$. It attains its minimum at $x = \pm 1$. An example of a minimizing sequence (x_n) is given by $x_n = (-1)^n$. In fact, $f(x_n) = \inf f(x)$ for all n. This minimizing sequence does not converge because its terms jump back and forth between $x = -1$ and $x = 1$. The subsequences (x_{2k+1}) and (x_{2k}) converge, to $x = -1$ and $x = 1$, respectively.

As this example shows, the compactness argument does not imply that a point where f attains its minimum is unique. There are many possible minimizing sequences, and there may be subsequences of a given minimizing sequence that converge to different limits. If, however, the function f attains its minimum at a unique point, then it follows from Exercise 1.27 that every minimizing sequence must converge to that point.

Example 1.70 The function $f(x) = e^{-x}$ is continuous and bounded from below on the noncompact set $\mathbb{R}$. The infimum of f on $\mathbb{R}$ is zero, but f does not attain its infimum. An example of a minimizing sequence (x_n) is given by $x_n = n$. The terms of the minimizing sequence "escape" to infinity, and it has no convergent subsequence.

Example 1.71 The discontinuous function f on the compact set $[0, 1]$ defined by

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

is not bounded from below. A sequence (x_n) is a minimizing sequence if $x_n \to 0$ as $n \to \infty$. In that case, $f(x_n) \to -\infty$ as $n \to \infty$, but f is discontinuous at the limit point $x = 0$.

Some of the conclusions of Theorem 1.68 still hold for semicontinuous functions. An almost identical proof shows the following result.

Theorem 1.72 Let K be a compact metric space. If $f : K \to \mathbb{R}$ is upper semicontinuous, then f is bounded from above and attains its supremum. If $f : K \to \mathbb{R}$ is lower semicontinuous, then f is bounded from below and attains its infimum.

Example 1.73 We define $f, g : [0, 1] \to \mathbb{R}$ by

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

The function f is upper semicontinuous, and does not attain its infimum, while g is lower semicontinuous and attains its minimum at $x = 0$.

1.9 References

For introductions to basic real analysis, see Marsden and Hoffman [37] or Rudin [47]. Simmons [50] gives a clear and accessible discussion of metric, normed, and topological spaces. For linear algebra, see Halmos [19] and Lax [30]. Two other books with a similar purpose to this one are Naylor and Sell [40] and Stakgold [52].

1.10 Exercises

Exercise 1.1 A set A is *countably infinite* if there is a one-to-one, onto map from A to $\mathbb{N}$. A set is *countable* if it is finite or countably infinite, otherwise it is *uncountable*.

(a) Prove that the set $\mathbb{Q}$ of rational numbers is countably infinite.
(b) Prove that the set $\mathbb{R}$ of real numbers is uncountable.

Exercise 1.2 Give an ϵ-δ proof that

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x},$$

when $|x| < 1$.

Exercise 1.3 If x, y, z are points in a metric space (X, d), show that

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

Exercise 1.4 Suppose that (X, d_X) and (Y, d_Y) are metric spaces. Prove that the Cartesian product $Z = X \times Y$ is a metric space with metric d defined by

$$d(z_1, z_2) = d_X(x_1, x_2) + d_Y(y_1, y_2),$$

where $z_1 = (x_1, y_1)$ and $z_2 = (x_2, y_2)$.

Exercise 1.5 Suppose that $(X, \|\cdot\|)$ is a normed linear space. Prove that (1.2) and (1.4) define metrics on X.

Exercise 1.6 Starting from the fact that $\mathbb{R}$ equipped with its usual distance function is complete, prove that $\mathbb{R}^n$ equipped with the sum, maximum, or Euclidean norm is complete.

Exercise 1.7 Show that the series

$$\sum_{n=1}^{\infty} \frac{(-1)^n}{n}$$

is not absolutely convergent. Show that by permuting the terms of this series one can obtain series with different limits.

Exercise 1.8 Let (x_n) be a sequence of real numbers. A point $c \in \mathbb{R} \cup \{\pm\infty\}$ is called a *cluster point* of (x_n) if there is a convergent subsequence of (x_n) with limit c. Let C denote the set of cluster points of (x_n). Prove that C is closed and

$$\limsup x_n = \max C \qquad \text{and} \qquad \liminf x_n = \min C.$$

Exercise 1.9 Let (x_n) be a bounded sequence of real numbers.

(a) Prove that for every $\epsilon > 0$ and every $N \in \mathbb{N}$ there are $n_1, n_2 \geq N$, such that

$$\limsup_{n\to\infty} x_n \leq x_{n_1} + \epsilon, \quad x_{n_2} - \epsilon \leq \liminf_{n\to\infty} x_n.$$

(b) Prove that for every $\epsilon > 0$ there is an $N \in \mathbb{N}$ such that

$$x_m \leq \limsup_{n\to\infty} x_n + \epsilon, \quad x_m \geq \liminf_{n\to\infty} x_n - \epsilon$$

for all $m \geq N$.

(c) Prove that (x_n) converges if and only if

$$\liminf_{n\to\infty} x_n = \limsup_{n\to\infty} x_n.$$

Exercise 1.10 Consider a family $\{x_{n,\alpha}\}$ of real numbers indexed by $n \in \mathbb{N}$ and $\alpha \in A$. Prove the following relations:

$$\begin{aligned} \limsup_{n\to\infty} \left(\inf_{\alpha} x_{n,\alpha}\right) &\leq \inf_{\alpha} \left(\limsup_{n\to\infty} x_{n,\alpha}\right); \\ \sup_{\alpha} \left(\liminf_{n\to\infty} x_{n,\alpha}\right) &\leq \liminf_{n\to\infty} \left(\sup_{\alpha} x_{n,\alpha}\right). \end{aligned}$$

Exercise 1.11 If (x_n) is a sequence of real numbers such that

$$\lim_{n\to\infty} x_n = x,$$

and $a_n \leq x_n \leq b_n$, prove that

$$\limsup_{n\to\infty} a_n \leq x \leq \liminf_{n\to\infty} b_n.$$

Exercise 1.12 Let $(X, d_X), (Y, d_Y)$, and (Z, d_Z) be metric spaces and let $f : X \to Y$, and $g : Y \to Z$ be continuous functions. Show that the composition

$$h = g \circ f : X \to Z,$$

defined by $h(x) = g(f(x))$, is also continuous.

Exercise 1.13 A function $f : \mathbb{R} \to \mathbb{R}$ is said to be *differentiable* at x if the following limit exists and is finite:

$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h}.$$

(a) Prove that if f is differentiable at x, then f is continuous at x.
(b) Show that the function

$$f(x) = \begin{cases} x^2 \sin\left(1/x^2\right) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

is differentiable at $x = 0$ but the derivative is not continuous at $x = 0$.
(c) Prove that if f is differentiable at x and has a local maximum or minimum at x, then $f'(x) = 0$.

Exercise 1.14 If $f : [a, b] \to \mathbb{R}$ is continuous on $[a, b]$ and differentiable in (a, b), then prove that there is a $a < \xi < b$ such that

$$f(b) - f(a) = f'(\xi)\,(b - a)\,.$$

This result is called the *mean value theorem.* Deduce that if $f'(x) = 0$ for all $a < x < b$ then f is a constant function.

Exercise 1.15 Prove that every compact subset of a metric space is closed and bounded. Prove that a closed subset of a compact space is compact.

Exercise 1.16 Suppose that F and G are closed and open subsets of $\mathbb{R}^n$, respectively, such that $F \subset G$. Show that there is a continuous function $f : \mathbb{R}^n \to \mathbb{R}$ such that:

(a) $0 \leq f(x) \leq 1$;
(b) $f(x) = 1$ for $x \in F$;
(c) $f(x) = 0$ for $x \in G^c$.

HINT. Consider the function

$$f(x) = \frac{d(x, G^c)}{d(x, G^c) + d(x, F)}.$$

This result is called *Urysohn's lemma.*

Exercise 1.17 Let (X, d) be a complete metric space, and $Y \subset X$. Prove that (Y, d) is complete if and only if Y is a closed subset of X.

Exercise 1.18 Let (X, d) be a metric space, and let (x_n) be a sequence in X. Prove that if (x_n) has a Cauchy subsequence, then, for any decreasing sequence of positive $\epsilon_k \to 0$, there is a subsequence (x_{n_k}) of (x_n) such that

$$d(x_{n_k}, x_{n_l}) \leq \epsilon_k \qquad \text{for all } k \leq l.$$

Exercise 1.19 Following the construction of the Cantor set C by the removal of middle thirds, we define a function F on the complement of the Cantor set $[0, 1] \setminus C$ as follows. First, we define $F(x) = 1/2$ for $1/3 < x < 2/3$. Then $F(x) = 1/4$ for $1/9 < x < 2/9$ and $F(x) = 3/4$ for $7/9 < x < 8/9$, and so on. Prove that F extends to a unique continuous function $F : [0,1] \to \mathbb{R}$. Prove that F is differentiable at every $x \in \mathbb{R} \setminus C$ and $F'(x) = 0$. This function is called the *Cantor function.* Its graph is sometimes called the *devil's staircase.*

Exercise 1.20 Let X be a normed linear space. A series $\sum x_n$ in X is *absolutely convergent* if $\sum \|x_n\|$ converges to a finite value in $\mathbb{R}$. Prove that X is a Banach space if and only if every absolutely convergent series converges.

Exercise 1.21 Suppose that X is a Banach space, and (x_{mn}) is a doubly indexed sequence in X such that

$$\sum_{m=1}^{\infty}\sum_{n=1}^{\infty} \|x_{mn}\| < \infty.$$

Prove that

$$\sum_{m=1}^{\infty}\left(\sum_{n=1}^{\infty} x_{mn}\right) = \sum_{n=1}^{\infty}\left(\sum_{m=1}^{\infty} x_{mn}\right).$$

Exercise 1.22 Let S be a set. A relation $\sim$ between points of S is called an *equivalence relation* if, for all $a, b, c \in S$, we have:

(a) $a \sim a$;
(b) $a \sim b$ implies $b \sim a$;
(c) $a \sim b$ and $b \sim c$ implies $a \sim c$.

Define the equivalence class C_a associated with $a \in S$ by

$$C_a = \{b \in S \mid a \sim b\}.$$

Prove that two equivalence classes are either disjoint or equal, so $\sim$ partitions S into a union of disjoint equivalence classes. Show that the relation $\sim$ between Cauchy sequences defined in the proof of Theorem 1.52 is an equivalence relation.

Exercise 1.23 Suppose that $f : X \to \mathbb{R}$ is lower semicontinuous and M is a real number. Define $f_M : X \to \mathbb{R}$ by

$$f_M(x) = \min\left(f(x), M\right).$$

Prove that f_M is lower semicontinuous.

Exercise 1.24 Let $f : X \to \mathbb{R}$ be a real-valued function on a set X. The *epigraph* $\operatorname{epi} f$ of f is the subset of $X \times \mathbb{R}$ consisting of points that lie above the graph of f:

$$\operatorname{epi} f = \{(x,t) \in X \times \mathbb{R} \mid t \geq f(x)\}.$$

Prove that a function is lower semicontinuous if and only if its epigraph is a closed set.

Exercise 1.25 A function $f : \mathbb{R}^n \to \mathbb{R}$ is *coercive* if

$$\lim_{\|x\| \to \infty} f(x) = \infty. \tag{1.17}$$

Explicitly, this condition means that for any $M > 0$ there is an $R > 0$ such that $\|x\| \geq R$ implies $f(x) \geq M$. Prove that if $f : \mathbb{R}^n \to \mathbb{R}$ is lower semicontinuous and coercive, then f is bounded from below and attains its infimum.

Exercise 1.26 Let $p : \mathbb{R}^2 \to \mathbb{R}$ be a polynomial function of two real variables. Suppose that $p(x, y) \geq 0$ for all $x, y \in \mathbb{R}$. Does every such function attain its infimum? Prove or disprove.

Exercise 1.27 Suppose that (x_n) is a sequence in a compact metric space with the property that every convergent subsequence has the same limit x. Prove that $x_n \to x$ as $n \to \infty$.