Introduction to Metric and Topological Spaces

Second Edition

WILSON A SUTHERLAND
Emeritus Fellow of New College, Oxford

Companion web site: www.oup.com/uk/companion/metric

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford ox2 6DP
Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in
Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto
With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam
Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries
Published in the United States
by Oxford University Press Inc., New York

First published 2009

British Library Cataloguing in Publication Data
Data available
Library of Congress Cataloging in Publication Data
Typeset by SPI Publisher Services, Pondicherry, India
Printed in Great Britain
on acid-free paper by
Clays, St Ives plc
ISBN 978-0-19-956307-4 (Hbk.)
978-0-19-956308-1 (Pbk.)
1 3 5 7 9 10 8 6 4 2

Preface

Preface to the second edition

One technical advance since the first edition is the possibility of having a companion web site, and I have tried to use this to the full. The address of the companion web site is www.oup.com/uk/companion/metric. Parts of the first edition have been moved there. This makes room for new material on standard surfaces, intended both to give a brief introduction to geometric topology and also to amplify the section on quotient spaces, hopefully without losing the advantage of brevity. Also, more explanations and examples have been added both to the book and on the companion web site. Accordingly the numbering in the preface to the first edition no longer applies, although the progression of ideas described there is still roughly followed. To help convey familiarity, concepts such as closure and interior are introduced first for metric spaces, and then repeated for topological spaces; I have tried to vary the accompanying examples and exercises to suit the context.

I am grateful for the opportunity to update notation and references.

A colleague who liked other aspects of the first edition complained that his students too readily looked at the answers; now as before I have a concern about students working from this book on their own, but I have moved the answers to a restricted web page.

It is a pleasure to thank anonymous referees for their thoughtful suggestions for improvements; and equally to thank two distinguished ex-students of New College for the comforting advice to change as little as possible. I hope to have steered a middle course in response to all this advice. It is also a pleasure to thank several ex-students and other friends for corrections and improvements to this edition.

It is also more than a pleasure to thank Ruth for many things, in particular her encouragement for writing a second edition.

Oxford, 2008 W.A.S.

Preface to the first edition

One of the ways in which topology has influenced other branches of mathematics in the past few decades is by putting the study of continuity and convergence into a general setting. This book introduces metric and topological spaces by describing some of that influence. The aim is to move gradually from familiar real analysis to abstract topological spaces; the main topics in the abstract setting are related back to familiar ground as far as possible. Apart from the language of metric and topological spaces, the topics discussed are compactness, connectedness, and completeness. These form part of the central core of general topology which is now used in several branches of mathematics. The emphasis is on *introduction*; the book is not comprehensive even within this central core, and algebraic and geometric topology are not mentioned at all. Since the approach is via analysis, it is hoped to add to the reader's insight on some basic theorems there (for example, it can be helpful to some students to see the Heine–Borel theorem and its implications for continuous functions placed in a more general context).

The stage at which a student of mathematics should see this process of generalization, and the degree of generality he should see, are both controversial. I have tried to write a book which students can read quite soon after they have had a course on analysis of real-valued functions of one real variable, not necessarily including uniform convergence.

The first chapter reviews real numbers, sequences, and continuity for real-valued functions of one real variable. Most readers will find nothing new there, but we shall continually refer back to it. With continuity as the motivating concept, the setting is generalized to metric spaces in Chapter 2 and to topological spaces in Chapter 3. The pay-off begins in Chapter 5 with the study of compactness, and continues in later chapters on connectedness and completeness. In order to introduce uniform convergence, Chapter 8 reverts to the traditional approach for real-valued functions of a real variable before interpreting this as convergence in the sup metric.

Most of the methods of presentation used are the common property of many mathematicians, but I wish to acknowledge that the way of introducing compactness is influenced by Hewitt (1960). It is also a pleasure to acknowledge the influence of many teachers, colleagues, and ex-students on this book, and to thank Peter Strain of the Open University for helpful comments and the staff of the Clarendon Press for their encouragement during the writing.

Oxford, 1974 W.A.S.

Preface to reprinted edition

I am grateful to all who have pointed out errors in the first printing (even to those who pointed out that the proof of Corollary 1.1.7 purported to establish the existence of a *positive* rational number between any two real numbers). In particular, it is a pleasure to thank Roy Dyckhoff, Ioan James, and Richard Woolfson for valuable comments and corrections.

Oxford, 1981 W.A.S.

Contents

1 Introduction

In this book we are going to generalize theorems about convergence and continuity which are probably familiar to the reader in the case of sequences of real numbers and real-valued functions of one real variable. The kind of result we shall be trying to generalize is the following: *if a real-valued function f is defined and continuous on the closed interval $[a, b]$ in the real line, then f is bounded on $[a, b]$, i.e. there exists a real number K such that $|f(x)| \leqslant K$ for all x in $[a, b]$.* Several such theorems about real-valued functions of a real variable are true and useful in a more general framework, after suitable minor changes of wording. For example, if we suppose that a real-valued function f of *two* real variables is defined and continuous on a rectangle $[a, b] \times [c, d]$, then f is bounded on this rectangle. Once we have seen that the result generalizes from one to two real variables, it is natural to suspect that it is true for any finite number of real variables, and then to go a step further by asking: how general a situation can the theorem be formulated for, and how generally is it true? These questions lead us first to metric spaces and eventually to topological spaces.

Before going on to study such questions, it is fair to ask: what is the point of generalization? One answer is that it saves time, or at least avoids tedious repetition. If we can show by a single proof that a certain result holds for functions of n real variables, where n is any positive integer, this is better than proving it separately for one real variable, two real variables, three real variables, etc. In the same vein, generalization often gives a unified mental grasp of several results which otherwise might just seem vaguely similar, and in addition to the satisfaction involved, this more efficient organization of material helps some people's understanding. Another gain is that generalization often illuminates the proof of a theorem, because to see how generally a given result can be proved, one has to notice exactly which properties or hypotheses are used at each stage in the proof.

Against this, we should be aware of some dangers in generalization. Most mathematicians would agree that it can be carried to an excessive extent. Just when this stage is reached is a matter of controversy, but the potential reader is warned that some mathematicians would say 'Enough,

no more (at least as far as analysis is concerned)' when we get into metric spaces. Also, there is an initial barrier of unfamiliarity to be overcome in moving to a more general framework, with its new language; the extent to which the pay-off is worthwhile is likely to vary from one student to another.

Our successive generalizations lead to the subject called topology. Applications of topology range from analysis, geometry, and number theory to mathematical physics and computer science. Topology is a language for many mathematical topics, just as mathematics is a language for many sciences. But it also has attractive results of its own. We have mentioned that some of these generalize theorems the reader has already met for real-valued functions of a real variable. Moreover, topology has a geometric aspect which is familiar in popular expositions as 'rubber-sheet geometry', with pictures of doughnuts, Möbius bands, Klein bottles, and the like; we touch on this in the chapter on quotients, trying to indicate how such topics are part of the same story as the more analytic aspects. From the point of view of analysis, topology is the study of continuity, while from the point of view of geometry, it is the study of those properties of geometric objects which are preserved when the objects are stretched, compressed, bent, and otherwise mistreated—everything is legitimate except tearing apart and sticking together. This is what gives rise to the old joke that a topologist is a person who cannot tell the difference between a coffee cup and a doughnut—the point being that each of these is a solid object with just one hole through it.

As a consequence of introducing abstractions gradually, the theorem density in this book is low. The title of theorem is reserved for substantial results, which have significance in a broad range of mathematics.

Some exercises are marked $\star$ or even $\star\star$ and some passages are enclosed between $\bigstar$ signs to denote that they are tentatively thought to be more challenging than the rest. A few paragraphs are enclosed between ► and ◄ signs to denote that they require some knowledge of abstract algebra.

We shall try to illustrate the exposition with suitable diagrams; in addition readers are urged to draw their own diagrams wherever possible.

A word about the exercises: there are lots. Rather than being daunted, try a sample at a first reading, some more on revision, and so on. Hints are given with some of the exercises, and there are further hints on the web site. When you have done most of the exercises you will have an excellent understanding of the subject.

A previous course in real analysis is a prerequisite for reading this book. This means an introduction (including rigorous proofs) to continuity, differential and preferably also integral calculus for real-valued functions of one real variable, and convergence of real number sequences. This material is included, for example, in Hart (2001) or, in a slightly more sophisticated but very complete way, in Spivak (2006) (names followed by dates in parentheses refer to the bibliography at the end of the book). The experience of abstraction gained from a previous course, in say, linear algebra, would help the reader in a general way to follow the abstraction of metric and topological spaces. However, the student is likely to be the best judge of whether he/she is ready, or wants, to read this book.

2 Notation and terminology

We use the logical symbols $\Rightarrow$ and $\Leftrightarrow$ meaning *implies* and *if and only if*. We also use iff to mean 'if and only if'; although not pretty, it is short and we use it frequently. Most introductions to algebra and analysis survey many parts of the language of sets and maps, and for these we just list notation.

If an object a belongs to a set A we write $a \in A$, or occasionally $A \ni a$, and if not we write $a \notin A$. If A is a subset of B (perhaps equal to B) we write $A \subseteq B$, or occasionally $B \supseteq A$. The subset of elements of A possessing some property P is written $\{a \in A : P(a)\}$. A finite set is sometimes specified by listing its elements, say $\{a_1, a_2, \ldots, a_n\}$. A set containing just one element is called a *singleton* set. Intersection and union of sets are denoted by $\cap$, $\cup$, or $\bigcap$, $\bigcup$. The empty set is written $\emptyset$. If $A \cap B = \emptyset$ we say that A and B are *disjoint*. Given two sets A and B, the set of elements which are in B but not in A is written $B \setminus A$. Thus in particular if $A \subseteq B$ then $B \setminus A$ is the complement of A in B. If S is a set and for each i in some set I we are given a subset A_i of S, then we denote by $\bigcup_{i \in I} A_i$, $\bigcap_{i \in I} A_i$ (or just $\bigcup_I A_i$, $\bigcap_I A_i$) the union and intersection of the A_i over all $i \in I$; for example, in the case of union what this means is

$$s \in \bigcup_{i \in I} A_i, \ \Leftrightarrow \ \text{there exists } i \in I \text{ such that } s \in A_i.$$

In this situation I is called an *indexing* set. We use De Morgan's laws, which with the above notation assert

$$S \setminus \bigcup_I A_i = \bigcap_I (S \setminus A_i), \qquad S \setminus \bigcap_I A_i = \bigcup_I (S \setminus A_i).$$

In particular, if the indexing set is the positive integers $\mathbb{N}$ we usually write

$$\bigcup_{i=1}^{\infty} A_i, \ \bigcap_{i=1}^{\infty} A_i \ \text{ for } \ \bigcup_{i \in \mathbb{N}} A_i, \ \bigcap_{i \in \mathbb{N}} A_i.$$

The Cartesian product $A \times B$ of sets A, B is the set of all ordered pairs (a, b) where $a \in A$, $b \in B$. This generalizes easily to the product of any

finite number of sets; in particular we use A^n to denote the set of ordered n-tuples of elements from A.

A map or function f (we use the terms interchangably) between sets X, Y is written $f : X \to Y$. We call X the *domain* of f, and we avoid calling Y anything. We think of f as assigning to each x in X an element $f(x)$ in Y, although logically it is preferable to define a map as a pair of sets X, Y together with a certain type of subset of $X \times Y$ (intuitively the graph of f). Persisting with our way of thinking about f, we define *the graph of* f to be the subset $G_f = \{(x, y) \in X \times Y : f(x) = y\}$ of $X \times Y$.

We call $f : X \to Y$ *injective* if $f(x) = f(x') \Rightarrow x = x'$ (we prefer this to 'one–one' since the latter is a little ambiguous). We should therefore call $f : X \to Y$ surjective if for every $y \in Y$ there is an $x \in X$ with $f(x) = y$, but we usually call such an f *onto*. If $f : X \to Y$ is both injective and onto we call it *bijective* or *a one–one correspondence*.

If $f : X \to Y$ is a map and $A \subseteq X$ then the *restriction of* f *to* A, written $f|A$, is the map $f|A : A \to Y$ defined by $(f|A)(a) = f(a)$ for every $a \in A$. In traditional calculus the function $f|A$ would not be distinguished from f itself, but when we are being fussy about the precise domains of our functions it is important to make the distinction: f has domain X while $f|A$ has domain A.

If $f : X \to Y$ and $g : Y \to Z$ are maps then their *composition* $g \circ f$ is the map $g \circ f : X \to Z$ defined by $(g \circ f)(x) = g(f(x))$ for each $x \in X$. This is the abstract version of 'function of a function' that features, for example, in the chain rule in calculus.

There are some more concepts relating to sets and functions which we shall focus on in the next chapter.

We shall occasionally assume that the terms *equivalence relation* and *countable set* are understood.

We use $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ to denote the sets of positive integers, integers, rational numbers, real numbers, and complex numbers, respectively. We often refer to $\mathbb{R}$ as the *real line* and we call the following subsets of $\mathbb{R}$ *intervals*:

$$\begin{aligned}
&\text{(i)} && [a, b] = \{x \in \mathbb{R} : a \leqslant x \leqslant b\},\\
&\text{(ii)} && (a, b) = \{x \in \mathbb{R} : a < x < b\},\\
&\text{(iii)} && (a, b] = \{x \in \mathbb{R} : a < x \leqslant b\},\\
&\text{(iv)} && [a, b) = \{x \in \mathbb{R} : a \leqslant x < b\},\\
&\text{(v)} && (-\infty, b] = \{x \in \mathbb{R} : x \leqslant b\},\\
&\text{(vi)} && (-\infty, b) = \{x \in \mathbb{R} : x < b\},
\end{aligned}$$

$$\begin{aligned}
&\text{(vii)} && [a, \infty) = \{x \in \mathbb{R} : x \geqslant a\},\\
&\text{(viii)} && (a, \infty) = \{x \in \mathbb{R} : x > a\},\\
&\text{(ix)} && (-\infty, \infty) = \mathbb{R}.
\end{aligned}$$

This is our definition of *interval*—a subset of $\mathbb{R}$ is an interval iff it is on the above list. The intervals in (i), (v), (vii) (and (ix)) are called *closed* intervals; those in (ii), (vi), (viii) (and (ix)) are called *open* intervals; and (iii), (iv) are called *half-open* intervals. When we refer to an interval of types (i)–(iv), it is always to be understood that $b > a$, except for type (i), when we also allow $a = b$. We shall try to avoid the occasional risk of confusing an interval (a, b) in $\mathbb{R}$ with a point (a, b) in $\mathbb{R}^2$ by stating which of these is meant when there might be any doubt.

The reader has probably already had practice working with sets; here as revision exercises are a few facts which appear later in the book. The last two exercises, involving equivalence relations, are relevant to the chapter on quotient spaces (and only there). They look more complicated than they really are.

Exercise 2.1 Suppose that C, D are subsets of a set X. Prove that

$$(X \setminus C) \cap D = D \setminus C.$$

Exercise 2.2 Suppose that A, V are subsets of a set X. Prove that

$$A \setminus (V \cap A) = A \cap (X \setminus V).$$

Exercise 2.3 Suppose that V, X, Y are sets with $V \subseteq X \subseteq Y$ and suppose that U is a subset of Y such that $X \setminus V = X \cap U$. Prove that

$$V = X \cap (Y \setminus U).$$

Exercise 2.4 Suppose that U, V are subsets of sets X, Y, respectively. Prove that

$$U \times V = (X \times V) \cap (U \times Y).$$

Exercise 2.5 Suppose that U_1, U_2 are subsets of a set X and that V_1, V_2 are subsets of a set Y. Prove that

$$(U_1 \times V_1) \cap (U_2 \times V_2) = (U_1 \cap U_2) \times (V_1 \cap V_2).$$

Exercise 2.6 Suppose that for some set X and some indexing sets I, J we have $U = \bigcup_{i\in I} B_{i1}$ and $V = \bigcup_{j\in J} B_{j2}$ where each B_{i1}, B_{j2} is a subset of X. Prove that

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

Exercise 2.7 (a) Let $\sim$ be an equivalence relation on a set X. Show that the corresponding equivalence classes partition X into a union of pairwise disjoint non-empty subsets $\{A_i : i \in I\}$ for some indexing set I. (This means that for all $i, j \in I$, we have $A_i \subseteq X$, $A_i \neq \emptyset$, $A_i \cap A_j = \emptyset$ for $i \neq j$, and $\bigcup_{i\in I} A_i = X$).

(b) Conversely show that a partition of X into pairwise disjoint non-empty subsets, say $\mathcal{P} = \{A_i : i \in I\}$, determines an equivalence relation $\sim$ on X where $x_1 \sim x_2$ iff x_1 and x_2 belong to the same set A_i in $\mathcal{P}$.

Exercise 2.8 Continuing with the notation of Exercise 2.7, let the partition determined by an equivalence relation $\sim$ on X be denoted by $\mathcal{P}(\sim)$ and the equivalence relation determined by a partition $\mathcal{P}$ be denoted by $\sim(\mathcal{P})$. Show that $\sim(\mathcal{P}(\sim)) = \sim$ and $\mathcal{P}(\sim(\mathcal{P})) = \mathcal{P}$. This shows that there is a one–one correspondence between equivalence relations on X and partitions of X.

3 More on sets and functions

In the previous chapter we assumed familiarity with a certain amount of notation and terminology about sets and functions; but some readers may not yet be as much at ease with the concepts in the present chapter. In topology the idea of the *inverse image* of a set under a map is much used, so it is good to be familiar with it. If you are at ease with Definitions 3.1 and 3.2 below, then you could safely skip the rest of this chapter. (If in doubt, skip it now but come back to it later if necessary.)

Direct and inverse images

Let $f : X \to Y$ be any map, and let A, C be subsets of X, Y respectively.

Definition 3.1 *The* (direct *or* forwards) image $f(A)$ *of* A *under* f *is the subset of* Y *given by* $\{y \in Y : y = f(a) \text{ for some } a \in A\}$.

Definition 3.2 *The* inverse image $f^{-1}(C)$ *of* C *under* f *is the subset of* X *given by* $\{x \in X : f(x) \in C\}$.

We note immediately that in order to make sense Definition 3.2 does *not* require the existence of an 'inverse function' f^{-1}. *Pre-image* is possibly a safer name, but *inverse image* is more common so we shall stick to it. For the same reason, to avoid confusion with inverse functions, at least one textbook has very reasonably tried to popularize the notation $f^{\dashv}(C)$ in place of $f^{-1}(C)$, but this has not caught on, so we shall grasp the nettle and use $f^{-1}(C)$.

A particularly confusing case is $f^{-1}(y)$ for $y \in Y$. The confusion is enhanced by the notation: $f^{-1}(y)$ should really be written $f^{-1}(\{y\})$. It is the special case of $f^{-1}(C)$ when C is the singleton set $\{y\}$. We shall see examples below in which $f^{-1}(y)$ contains more than one element. We follow common usage by writing $f^{-1}(y)$ for $f^{-1}(\{y\})$ except in the next example.

Example 3.3 Let $X = \{x, y, z\}$, $Y = \{1, 2, 3\}$ and define $f : X \to Y$ by $f(x) = 1$, $f(y) = 2$, $f(z) = 1$. Then we have $f(\{x, y\}) = \{1, 2\}$, $f(\{x, z\}) = \{1\}$, $f^{-1}(\{1\}) = \{x, z\}$, and $f^{-1}(\{2, 3\}) = \{y\}$.

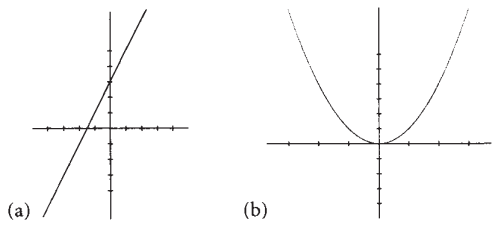

Figure 3.1. (a) Graph of f and (b) graph of g

As mentioned, we henceforth write $f^{-1}(\{1\})$ as $f^{-1}(1)$. Note $f^{-1}(1)$ here is *not* a singleton set.

Example 3.4 Let $X = Y = \mathbb{R}$ and define $f : X \to Y$ by $f(x) = 2x + 3$. The graph of this function is a straight line (see Figure 3.1(a)):

Then for example,

$$f([0, 1]) = [3, 5],\ f((1, \infty)) = (5, \infty),\ \ f^{-1}([0, 1]) = [-3/2, -1].$$

Example 3.5 Again let $X = Y = \mathbb{R}$. Define g by $g(x) = x^2$. The graph of this function has the familiar parabolic shape as in Figure 3.1(b). Then for example,

$$g([0, 1]) = [0, 1],\ \ g([1, 2]) = [1, 4],\ \ g(\{-1, 1\}) = \{1\},$$
$$g^{-1}([0, 1]) = [-1, 1],\ g^{-1}([1, 2]) = [-\sqrt{2}, -1] \cup [1, \sqrt{2}],\ g^{-1}([0, \infty)) = \mathbb{R}.$$

The special case of direct image and inverse image of the empty set are worth noting: for any map $f : X \to Y$ we have $f(\emptyset) = \emptyset$ and $f^{-1}(\emptyset) = \emptyset$: for example, $f^{-1}(\emptyset)$ consists of all elements of X which are mapped by f into the empty set, and there are no such elements so $f^{-1}(\emptyset) = \emptyset$.

We now come to some important formulae involving direct and inverse images. We state those about unions and intersections first in the case of just two subsets.

Proposition 3.6 *Suppose that $f : X \to Y$ is a map, that A, B are subsets of X and that C, D are subsets of Y. Then:*

$$f(A \cup B) = f(A) \cup f(B),\ \ f(A \cap B) \subseteq f(A) \cap f(B),$$
$$f^{-1}(C \cup D) = f^{-1}(C) \cup f^{-1}(D),\ \ f^{-1}(C \cap D) = f^{-1}(C) \cap f^{-1}(D).$$

Equality does not necessarily hold in the second formula, as we shall see shortly. There is a more general form of Proposition 3.6.

Proposition 3.7 *Suppose that $f : X \to Y$ is a map, and that 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*

$$f\left(\bigcup_{i\in I} A_i\right) = \bigcup_{i\in I} f(A_i), \qquad f\left(\bigcap_{i\in I} A_i\right) \subseteq \bigcap_{i\in I} f(A_i),$$

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

As a sample of the proof we show that

$$f^{-1}\left(\bigcap_{i\in I} C_i\right) = \bigcap_{i\in I} f^{-1}(C_i).$$

(Proofs of the other parts of Proposition 3.7 are on the web site.) First let $x \in f^{-1}\left(\bigcap_{i\in I} C_i\right)$. Then $f(x) \in \bigcap_{i\in I} C_i$, so $f(x) \in C_i$ for every $i \in I$. This tells us that $x \in f^{-1}(C_i)$ for every $i \in I$, so $x \in \bigcap_{i\in I} f^{-1}(C_i)$. Hence,

$$f^{-1}\left(\bigcap_{i\in I} C_i\right) \subseteq \bigcap_{i\in I} f^{-1}(C_i).$$

The reverse inclusion is proved by running the argument backwards. Explicitly, if $x \in \bigcap_{i\in I} f^{-1}(C_i)$ then for every $i \in I$ we have $x \in f^{-1}(C_i)$, so $f(x) \in C_i$. This tells us that $f(x) \in \bigcap_{i\in I} C_i$, so $x \in f^{-1}\left(\bigcap_{i\in I} C_i\right)$ as required.

Next we give results about complements, again preceded by a special case.

Proposition 3.8 *Suppose that $f : X \to Y$ is a map and $B \subseteq X, D \subseteq Y$. Then*

$$f(X \setminus B) \supseteq f(X) \setminus f(B), \quad f^{-1}(Y \setminus D) = X \setminus f^{-1}(D).$$

This follows by taking $A = X$, $C = Y$ in the next proposition (for the second part of Proposition 3.8 we use also $f^{-1}(Y) = X$).

Proposition 3.9 *With the notation of Proposition 3.6,*

$$f(A \setminus B) \supseteq f(A) \setminus f(B) \text{ and } f^{-1}(C \setminus D) = f^{-1}(C) \setminus f^{-1}(D).$$

The proof is on the web site.

We now explore Propositions 3.6 and 3.8 further, in order to gain familiarity. Here are two examples in which $f(A \cap B) = f(A) \cap f(B)$ fails and one in which $f(A \setminus B) = f(A) \setminus f(B)$ fails.

Example 3.10 Let $X = \{a, b\}$, $Y = \{1, 2\}$ and $f(a) = 1$, $f(b) = 1$. Put $A = \{a\}$, $B = \{b\}$. Then $A \cap B = \emptyset$, so $f(A \cap B) = \emptyset$. But on the other hand $f(A) \cap f(B) = \{1\} \neq \emptyset$.

Example 3.11 Let $X = Y = \mathbb{R}$, define $g(x) = x^2$, and let $A = [0, 1)$, $B = (-1, 0]$ so that $A \cap B = \{0\}$. Then $g(A \cap B) = \{0\}$ but on the other hand $g(A) \cap g(B) = [0, 1)$.

Example 3.12 Let $X = \{x, y, z\}$, $Y = \{1, 2, 3\}$ and as in Example 3.3 let $f(x) = 1 = f(z)$, $f(y) = 2$. Put $B = \{z\}$. Then $f(X \setminus B) = \{1, 2\}$, but on the other hand $f(X) \setminus f(B) = \{2\}$

The next result is useful later.

Proposition 3.13 *Suppose that $f : X \to Y$ is a map, $B \subseteq Y$ and for some indexing set I there is a family $\{A_i : i \in I\}$ of subsets of X with $X = \bigcup_I A_i$. Then*

$$f^{-1}(B) = \bigcup_I (f|A_i)^{-1}(B).$$

Proof First suppose $x \in f^{-1}(B)$. Since $X = \bigcup_I A_i$ we have $x \in A_{i_0}$ for some $i_0 \in I$. Then $(f|A_{i_0})(x) = f(x) \in B$, so $x \in (f|A_{i_0})^{-1}(B)$, which is contained in $\bigcup_I (f|A_i)^{-1}(B)$.

Conversely suppose that $x \in \bigcup_I (f|A_i)^{-1}(B)$. Then $x \in (f|A_{i_0})^{-1}(B)$ for some $i_0 \in I$. This says $(f|A_{i_0})(x) \in B$. But $(f|A_{i_0})(x) = f(x)$, so $f(x) \in B$ which gives $x \in f^{-1}(B)$. □

★ We occasionally want to look at sets such as $f^{-1}(f(A))$ or $f(f^{-1}(C))$; we look at a few basic facts about these, and explore them further in the exercises.

Proposition 3.14 *Let X, Y be sets and $f : X \to Y$ a map. For any subset $C \subseteq Y$ we have $f(f^{-1}(C)) = C \cap f(X)$. In particular, $f(f^{-1}(C)) = C$ if f is onto. For any subset $A \subseteq X$ we have $A \subseteq f^{-1}(f(A))$.*

Proof First let $y \in f(f^{-1}(C))$. Then $y = f(x)$ for some $x \in f^{-1}(C)$. But for such an x we have $f(x) \in C$, so $y \in C$. But also $y = f(x)$ so $y \in f(X)$. Hence $y \in C \cap f(X)$ and we have proved $f(f^{-1}(C)) \subseteq C \cap f(X)$. Suppose conversely that $y \in C \cap f(X)$. Then $y \in C$, and also $y = f(x)$ for some $x \in X$. Now for this x we have $f(x) = y \in C$, so $x \in f^{-1}(C)$. So $y = f(x) \in f(f^{-1}(C))$ as required, and we have proved the reverse inclusion $C \cap f(X) \subseteq f(f^{-1}(C))$. Thus $f(f^{-1}(C)) = C \cap f(X)$. When f is onto, $f(X) = Y$ so $f(f^{-1}(C)) = C$.

Secondly, for any $a \in A$ we have $f(a) \in f(A)$ so $a \in f^{-1}(f(A))$ as required. □

It is easy to find examples where the inclusion in the last part is strict.

Example 3.15 Following Example 3.10 let $X = \{a, b\}$, $Y = \{1, 2\}$, and $f(a) = 1 = f(b)$, $A = \{a\}$. Then $f^{-1}(f(A)) = f^{-1}(1) = \{a, b\} \neq A$.

Example 3.16 Let $X = Y = \mathbb{R}$ and let $g(x) = x^2$. Put $A = [0, 1]$. Then $g^{-1}(g(A)) = g^{-1}([0, 1]) = [-1, 1] \neq A$. ★

Inverse functions

We have emphasized that in order for the inverse image $f^{-1}(C)$ to be defined, there need not exist any inverse function f^{-1}. We now look at the case when such an inverse does exist.

Definition 3.17 *A map $f : X \to Y$ is said to be* invertible *if there exists a map $g : Y \to X$ such that the composition $g \circ f$ is the identity map of X and the composition $f \circ g$ is the identity map of Y.*

We immediately get a criterion on f for it to be invertible:

Proposition 3.18 *A map $f : X \to Y$ is invertible if and only if it is bijective.*

Proof Suppose first that f is invertible and let g be as in Definition 3.17. Then

$$f(x) = f(x') \Rightarrow g(f(x)) = g(f(x')) \Rightarrow x = x'$$

so f is injective. Also, given any $y \in Y$ we have $y = f(g(y))$ so $y \in f(X)$, which says that f is onto. Hence f is bijective.

Secondly suppose that f is bijective. We may define $g : Y \to X$ as follows: for any $y \in Y$ we know f is onto, so $y = f(x)$ for some $x \in X$. Moreover this x is unique for a given y since f is injective. Put $g(y) = x$, and we can see that f and g satisfy Definition 3.17, so f is invertible as required. □

The last part of the above proof also proves

Proposition 3.19 *When f is invertible, there is a* unique *g satisfying Definition 3.17. This unique g is called* the inverse of f, *written f^{-1}.*

For given $y \in Y$, in order to satisfy Definition 3.17 we have to choose $g(y)$ to be the unique $x \in X$ such that $f(x) = y$.

The final result in this chapter is slightly tricky, but it is very useful for one important theorem later (Theorem 13.26).

Proposition 3.20 *Suppose that $f : X \to Y$ is a one–one correspondence of sets X and Y and that $V \subseteq X$. Then the inverse image of V under the inverse map $f^{-1} : Y \to X$ equals the image set $f(V)$.*

Proof Let us write $g : Y \to X$ for the inverse function f^{-1} of $f : X \to Y$. We want to show for any $V \subseteq X$ that $g^{-1}(V) = f(V)$.

First suppose y is in $f(V)$. Then $y = f(x)$ for some $x \in V$, and this x is unique since f is injective. By definition of inverse function $x = g(y)$. But since $x \in V$ this gives $y \in g^{-1}(V)$. We have now proved $f(V) \subseteq g^{-1}(V)$.

Secondly suppose $y \in g^{-1}(V)$. Then $g(y) \in V$. So $f(g(y)) \in f(V)$. But g is the inverse function to f, so $f(g(y)) = y$, and we have $y \in f(V)$. This shows that $g^{-1}(V) \subseteq f(V)$. So we have proved $g^{-1}(V) = f(V)$ as required.

We may write the conclusion in the following rather mind-boggling way: $(f^{-1})^{-1}(V) = f(V)$. The inner superscript -1 indicates the function f^{-1} is inverse to f, and the outer one indicates the inverse image of the set V under that inverse function. □

Although some textbooks write f^{-1} only when f is invertible, others take the more relaxed view that if $f : X \to Y$ is injective, then it defines a bijective function $f_1 : X \to f(X)$, and they write $f^{-1} : f(X) \to X$ for the inverse of f_1 in the sense of Definition 3.17 and Proposition 3.19 above. This is a useful alternative, although we shall stick to the narrower interpretation.

Of the exercises, 3.5, 3.6, and 3.9 involve the starred section above.

Exercise 3.1 Let $f : X \to Y$ be a map and suppose that $A \subseteq B \subseteq X$ and that $C \subseteq D \subseteq Y$. Prove that $f(A) \subseteq f(B) \subseteq Y$ and that $f^{-1}(C) \subseteq f^{-1}(D) \subseteq X$.

Exercise 3.2 Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = \sin x$. Describe the sets:

$$f([0, \pi/2]),\ f([0, \infty)),\ f^{-1}([0,\ 1]),\ f^{-1}([0,\ 1/2]),\ f^{-1}([-1,\ 1]).$$

Exercise 3.3 Suppose that $f : X \to Y$ and $g : Y \to Z$ are maps and $U \subseteq Z$. Prove that $(g \circ f)^{-1}(U) = f^{-1}(g^{-1}(U))$.

Exercise 3.4 Let $f : \mathbb{R} \to \mathbb{R}^2$ be defined by $f(x) = (x,\ 2x)$. Describe the sets:

$$f([0,\ 1]),\ f^{-1}([0,\ 1] \times [0,\ 1]),\ f^{-1}(D) \text{ where } D = \{(x,\ y) \in \mathbb{R}^2 : x^2 + y^2 \leqslant 1\}.$$

Exercise 3.5 Show that a map $f : X \to Y$ is onto iff $f(f^{-1}(C)) = C$ for all subsets $C \subseteq Y$.

Exercise 3.6 Show that a map $f : X \to Y$ is injective iff $A = f^{-1}(f(A))$ for all subsets $A \subseteq X$.

Exercise 3.7 Let $f : X \to Y$ be a map. For each of the following determine whether it is true in general or whether it is sometimes false. (Give a proof or a counterexample for each.)

(i) If $y,\ y' \in Y$ with $y \neq y'$ then $f^{-1}(y) \neq f^{-1}(y')$.

(ii) If $y,\ y' \in Y$ with $y \neq y'$ and f is onto then $f^{-1}(y) \neq f^{-1}(y')$.

Exercise 3.8 Let $f : X \to Y$ be a map and let A, B be subsets of X. Prove that $f(A \setminus B) = f(A) \setminus f(B)$ if and only if $f(A \setminus B) \cap f(B) = \emptyset$. Deduce that if f is injective then $f(A \setminus B) = f(A) \setminus f(B)$.

Exercise 3.9 Let $f : X \to Y$ be a map and $A \subseteq X$, $C \subseteq Y$. Prove that

(a) $f(A) \cap C = f(A \cap f^{-1}(C))$.

(b) if also $B \subseteq X$ and $f^{-1}(f(B)) = B$ then $f(A) \cap f(B) = f(A \cap B)$.

Exercise 3.10 Suppose that $f : X \to Y$ is a map from a set X onto a set Y. Show that the family of subsets $\{f^{-1}(y) : y \in Y\}$ forms a partition of X in the sense of Exercise 2.7.

4 Review of some real analysis

The point of this chapter is to review a few basic ideas in real analysis which will be generalized in later chapters. It is not intended to be an introduction to these concepts for those who have never seen them before.

Real numbers

Two popular ways of thinking about the real number system are:

(1) geometrically, as corresponding to all the points on a straight line;

(2) in terms of decimal expansions, where if a number is irrational we think of longer and longer decimal expansions approximating it more and more closely.

Neither of these intuitive ideas is precise enough for our purposes, although each leads to a way of constructing the real numbers from the rational numbers. The second of these ways is described on the web site.

One approach to real numbers is axiomatic. This means we write down a list of properties and define the real numbers to be any system satisfying these properties. The properties are called *axioms* when they are used in this way. Another approach is constructive: we construct the real numbers from the rationals. The rational numbers may in turn be constructed from the integers, and so on—we can follow the trail backwards through the positive integers and back to set theory. (One has to begin with axioms at some stage, however.) In either approach the set of real numbers has certain properties; depending on the approach we have in mind, we call these properties either axioms or propositions. We shall assume that the construction of $\mathbb{R}$ has already been carried out for us, and we are interested in its properties.

Many introductions to analysis contain a list of properties of real numbers (see, for example, Hart (2001) or Spivak (2006)). A large number of these may be summed up technically by saying that the real numbers form an ordered field. Roughly this means that addition, subtraction, multiplication, and division of real numbers all work in the way we expect them to, and that the same is true of the way in which inequalities $x < y$ work and interact with addition and multiplication. We shall not review these properties, but concentrate on the so-called completeness property. The reasons for this strange behaviour are, first, that this is the property

which distinguishes the real numbers from the rational numbers (and in a sense analysis and topology from algebra) and secondly that our intuition is unlikely to let us down on properties deducible from those of an ordered field, whereas arguments using completeness tend to be more subtle.

To state the completeness property we need some terminology. Let S be a non-empty set of real numbers. An *upper bound for* S is a number x such that $y \leqslant x$ for all y in S. If an upper bound for S exists we say that S is bounded above. Lower bounds are defined similarly.

Example 4.1 (a) The set $\mathbb{R}$ of all real numbers has no upper or lower bound.

(b) The set $\mathbb{R}_-$ of all strictly negative real numbers has no lower bound, but for example 0 is an upper bound (as is any positive real number).

(c) The half-open interval (0, 1] is bounded above and below.

If S has an upper bound u, then S has (infinitely) many upper bounds, since any $x \in \mathbb{R}$ satisfying $x \geqslant u$ is also an upper bound. This gives the next definition some point.

Definition 4.2 *Given a non-empty subset S of $\mathbb{R}$ which is bounded above, we call u a* least upper bound for S *if*

(a) u is an upper bound for S,

(b) $x \geqslant u$ for any upper bound x for S.

Example 4.3 In Example 4.1 (b), 0 is a least upper bound for $\mathbb{R}_-$. For 0 is an upper bound, and it is a *least* upper bound because any $x < 0$ is not an upper bound for $\mathbb{R}_-$ (since any such x satisfies $x/2 > x$ and $x/2 \in \mathbb{R}_-$). Examples 4.1 (c) and (b) show that a least upper bound of a set S may or may not be in S.

It follows from Definition 4.2 that least upper bounds are unique when they exist. For if u, u' are both least upper bounds for a set S, then since u' is an upper bound for S it follows that $u \leqslant u'$ by leastness of u ('leastness' means the property in Definition 4.2 (b)). Interchanging the roles of u and u' in this argument shows that also $u' \leqslant u$, so $u' = u$.

Greatest lower bounds are defined similarly to least upper bounds.

We can now state one form of the completeness property for $\mathbb{R}$.

Proposition 4.4 *Any non-empty subset of $\mathbb{R}$ which is bounded above has a least upper bound.*

Since our interest is in generalizing real analysis rather than studying its foundations, we offer no proof of Proposition 4.4. The completeness property is quite subtle, and it is difficult to grasp its full significance until it has been used several times. It corresponds to the intuitive idea that there are no gaps in the real numbers, thought of as the points on a straight line; but the transition from the intuitive idea to the formal statement is not immediately obvious. For some sets of real numbers, such as Examples 4.1 (b) and (c), it is 'obvious' that a least upper bound exists (strictly speaking, this means that it follows from the properties of an ordered field). But this is not the case for all bounded non-empty sets of real numbers—for example, consider $S = \{x \in \mathbb{Q} : x^2 < 2\}$: the least upper bound turns out to be $\sqrt{2}$, and we need Proposition 4.4 to establish its existence—indeed, the existence of $\sqrt{2}$ cannot follow from the ordered field properties alone, since $\mathbb{Q}$ is an ordered field, but there is no rational number whose square is 2 (see Exercise 4.5).

For any non-empty subset S of $\mathbb{R}$ which is bounded above we call its unique least upper bound $\sup S$ (sup is short for supremum). Other notation sometimes used is l.u.b. S.

Although the completeness property was stated in terms of sets bounded above, it is equivalent to the corresponding property for sets bounded below. The next proposition formally states half of this equivalence.

Proposition 4.5 *If a non-empy subset S of $\mathbb{R}$ is bounded below then it has a greatest lower bound.*

Proof Let $T = \{x \in \mathbb{R} : -x \in S\}$. The idea of the proof is simply that l is a lower bound for S if and only if $-l$ is an upper bound for T. The details are left as Exercise 4.7. □

Just as in the case of least upper bounds, a non-empty subset S of $\mathbb{R}$ which is bounded below has a unique greatest lower bound called $\inf S$ (short for infimum) or g.l.b. S.

The next proposition and its corollary are applications of the completeness property.

Proposition 4.6 *The set $\mathbb{N}$ of positive integers is not bounded above.*

Proof Suppose for a contradiction that $\mathbb{N}$ is bounded above. Then by the completeness property there is a real number $u = \sup \mathbb{N}$. For any $n \in \mathbb{N}$, $n + 1$ is also in $\mathbb{N}$, so $n + 1 \leqslant u$. But then $n \leqslant u - 1$. Hence $n \leqslant u - 1$ for any $n \in \mathbb{N}$, so $u - 1$ is an upper bound for $\mathbb{N}$, contradicting the leastness of u. This contradiction shows that $\mathbb{N}$ cannot be bounded above. □

Corollary 4.7 *Between any two distinct real numbers x and y there is a rational number.*

Proof Suppose first that $0 \leqslant x < y$. Since $y - x > 0$, by Proposition 4.6 there is an n in $\mathbb{N}$ such that $n > 1/(y-x)$ and hence $1/n < y - x$. Let $M = \{m \in \mathbb{N} : m/n > x\}$. By Proposition 4.6 $M \neq \emptyset$, otherwise nx would be an upper bound for $\mathbb{N}$. Hence, since $M \subseteq \mathbb{N}$, M contains a least integer m_0. So $m_0/n > x$ and $(m_0 - 1)/n \leqslant x$, from which $m_0/n \leqslant x + 1/n$. Hence $x < m_0/n \leqslant x + 1/n < x + (y - x) = y$, and m_0/n is a suitable rational number, between x and y. Now suppose that $x < 0$. If $y > 0$ then 0 is a rational number between x and y, while if $y \leqslant 0$ then the first case supplies a rational number r such that $-y < r < -x$, so $x < -r < y$ which says that the rational number $-r$ is between x and y. □

The above proofs of Proposition 4.6 and Corollary 4.7 assume several 'obvious' facts about $\mathbb{R}$ which we should really prove beforehand. For example, we deduced $n \leqslant u - 1$ from $n + 1 \leqslant u$, a consequence of the property often stated as follows: if a, b, $c \in \mathbb{R}$ and $a \leqslant b$ then $a+c \leqslant b+c$. Also, we assumed that any non-empty subset of $\mathbb{N}$ has a least element. We leave the reader to spot other such assumptions.

Remark 4.8 *Between any two distinct real numbers there is also an irrational number (see Exercise 4.8).*

We conclude this brief review of real numbers by recalling two useful inequalities, often called the triangle inequality and the reverse triangle inequality. There are proofs on the web site.

Proposition 4.9 $|x + y| \leqslant |x| + |y|$ *for any x, y in $\mathbb{R}$.*

Corollary 4.10 $|x - y| \geqslant ||x| - |y||$ *for any x, y in $\mathbb{R}$.*

Real sequences

Formally an infinite sequence of real numbers is a map $s : \mathbb{N} \to \mathbb{R}$. This definition is useful for discussing topics such as subsequences and rearrangements without being vague. In practice, however, given such a map s we denote $s(n)$ by s_n and think of the sequence in the traditional way as an infinite ordered string of numbers, using the notation (s_n) or $s_1, s_2, s_3, \ldots$ for the whole sequence.

It is important to distinguish between a sequence (s_n) and the set of its members $\{s_n : n \in \mathbb{N}\}$. The latter can easily be finite. For example if (s_n) is 1, 0, 1, 0, ... then its set of members is $\{0, 1\}$. Formally, this is a matter of distinguishing between a map $s : \mathbb{N} \to \mathbb{R}$ and its image set $s(\mathbb{N})$.

Sequences can arise, for example, in solving algebraic or differential equations. On the theoretical side, convergent sequences might be used to prove the existence of solutions to equations. On the practical side, s_n might be the answer at the nth stage in some method of successive approximations for finding a root of an equation. The only difference between theory and practice here is that in practice one is interested in how quickly the sequence gives a good approximation to the answer. Also, in applications we might be dealing with a sequence of vectors or of functions instead of real numbers.

We now review real number sequences, emphasizing those definitions and results whose analogues we shall later study for more general sequences.

Example 4.11 (a) $\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \ldots,$ (b) $1, -\frac{1}{2}, \frac{1}{4}, -\frac{1}{8}, \ldots,$

(c) $\frac{1}{2}, 1, -\frac{1}{2}, -1, \frac{1}{4}, \frac{1}{2}, -\frac{1}{4}, -\frac{1}{2}, \ldots,$ (d) $1, 2, 3, \ldots,$ (e) $1, 0, 1, 0, \ldots,$

(f) $s_1 = 1$, $s_2 = 0$, $s_n = \frac{1}{2}(s_{n-2} + s_{n-1})$ for $n > 2$,

(g) s_n is the nth stage in some specified algorithm for approximating $\sqrt{2}$.

In examples (a)–(e), there is a simple formula for s_n in terms of n, which the reader will spot. This is convenient for illustrating the basic theory of sequences, but in practice a sequence might be generated by an iterative process, as in examples (f) and (g), or by the results of a probabilistic experiment repeated more and more often, or by some other means, and in such cases there may not be any simple formula for s_n in terms of n.

In Examples 4.11 (a), (b), (c) the sequence seems intuitively to be heading towards a definite number, whether steadily, or by alternately overshooting and undershooting the target, or irregularly, whereas in Examples 4.11 (d) (e) this is not the case. The mathematical term for 'heading towards' is 'converging', and the precise definition, as the reader probably knows, is as follows.

Definition 4.12 *The sequence (s_n)* converges to *(the real number) l if given (any real number) $\varepsilon > 0$, there exists (an integer) N_ε such that $|s_n - l| < \varepsilon$ for all $n \geqslant N_\varepsilon$.*

This is usually shortened by omitting the phrases in parentheses, and we often write just N in place of N_ε, although we need to remember that the value of N needed will usually vary with ε—intuitively, the smaller ε

is, the larger N will need to be. When Definition 4.12 holds, the number l is called the *limit* of the sequence. Other ways of writing '(s_n) converges to l' are '$s_n \to l$ as $n \to \infty$' and '$\lim_{n\to\infty} s_n = l$'. Here are two ways of thinking about the definition.

(1) (s_n) converges to l if, given any required degree of accuracy, then by going far enough along the sequence we can be sure that the terms beyond that stage all approximate l to within the required degree of accuracy.

(2) Let us take coordinate axes in the plane and mark the points with coordinates (n, s_n). Let us also draw a horizontal line L at height l. Then (s_n) converges to l if given any horizontal band of positive width centred on L, there exists a vertical line such that all marked points to the right of this vertical line lie within the prescribed horizontal band. Figure 4.1 is the kind of picture this suggests. The sequence promises to stay out of the shaded territory.

Two points are easy to get wrong when one is first trying to wield the formal definition. First, the order in which ε, N occur is crucial: given any $\varepsilon > 0$ *first*, there must *then* be an N_ε such that ... etc. Secondly, to prove convergence it is not enough to show that given $\varepsilon > 0$ there exists an N such that $|s_n - l| < \varepsilon$ for *some* $n \geqslant N$: this would be true of the sequence $1, 0, 1, 0, \ldots$, with $l = 0$, any $\varepsilon > 0$, and $N = 1$, yet the sequence does not converge.

The first deduction from the formal definition is an obvious part of the intuitive idea of convergence.

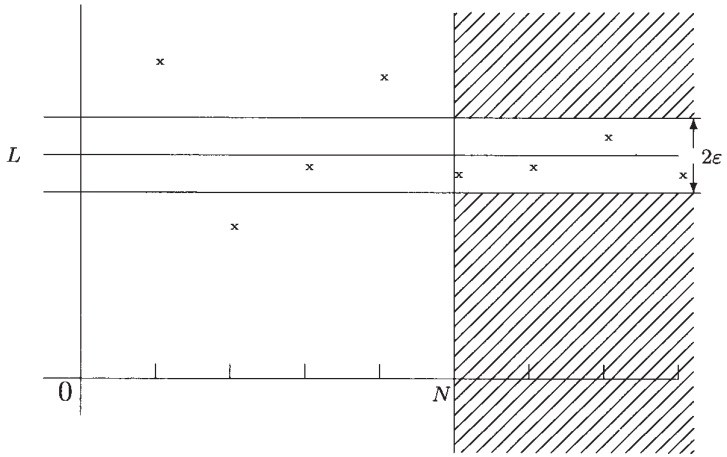

Figure 4.1. 'Graph' of a convergent sequence

Proposition 4.13 *A convergent sequence has a unique limit.*

Proof Suppose that (s_n) converges to l and also to l' where $l' \neq l$. Put $\varepsilon = \frac{1}{2}|l - l'|$. Since (s_n) converges to l, there is an integer N_1 such that $|s_n - l| < \varepsilon$ for all $n \geqslant N_1$. Similarly, since (s_n) converges to l', there is an integer N_2 such that $|s_n - l'| < \varepsilon$ for all $n \geqslant N_2$. Put $N = \max\{N_1, N_2\}$. Then, using the triangle inequality (Proposition 4.9),

$$|l - l'| = |l - s_N + s_N - l'| \leqslant |l - s_N| + |s_N - l'| < 2\varepsilon = |l - l'|.$$

This contradiction shows that $l' = l$. □

Before going further it is convenient to state explicitly a technical detail which is often used in convergence proofs.

Lemma 4.14 *Suppose there is a positive real number K such that given $\varepsilon > 0$ there exists N with $|s_n - l| < K\varepsilon$ for all $n \geqslant N$. Then (s_n) converges to l.*

Proof Let $\varepsilon > 0$. Then $\varepsilon/K > 0$, and if the stated condition holds, then there exists N such that $|s_n - l| < K(\varepsilon/K) = \varepsilon$ for all $n \geqslant N$, as required. In practice K is often an integer such as 2 or 3; we note that it needs to be independent of the choice of ε. □

In simple cases such as Example 4.11 (a) we can guess the limit and prove convergence directly. In general, however, it may be hard to guess the limit, and more importantly there may be no more convenient way to name a real number than as the limit of a given sequence. As an example consider:

$$s_n = 1 + \frac{1}{1!} + \frac{1}{2!} + \ldots + \frac{1}{n!}.$$

The reader may be able to think of a way to define the number e other than as the limit of the sequence (s_n), but it will also directly or indirectly involve taking the limit of this or some other sequence such as (t_n) where $t_n = (1 + 1/n)^n$.

We shall consider two theorems which provide ways of proving convergence without using a known value of the limit. As the above discussion indicates, both will depend heavily on the completeness property for $\mathbb{R}$.

Definition 4.15 *A sequence (s_n) is said to be* monotonic increasing (decreasing) *if $s_{n+1} \geqslant s_n$ ($s_{n+1} \leqslant s_n$) for all n in $\mathbb{N}$. It is* monotonic *if it has either of these properties.*

Theorem 4.16 *Every bounded monotonic sequence of real numbers converges.*

The proof is on the companion web site. As well as being useful on its own, Theorem 4.16 helps to prove the next convergence criterion. First we give a name to sequences in which the terms get closer and closer together as we get further along in the sequence.

Definition 4.17 *A sequence (s_n) is a* Cauchy *sequence if given $\varepsilon > 0$ there exists N such that if $m, n \geqslant N$ (i.e. if $m \geqslant N$ and $n \geqslant N$) then $|s_m - s_n| < \varepsilon$.*

Theorem 4.18 (Cauchy's convergence criterion) *A sequence (s_n) of real numbers converges if and only if it is a Cauchy sequence.*

Proof Suppose that (s_n) converges to l. Then given $\varepsilon > 0$, there exists N such that $|s_n - l| < \varepsilon$ for all $n \geqslant N$, so for $m, n \geqslant N$ the triangle inequality gives

$$|s_m - s_n| = |s_m - l + l - s_n| \leqslant |s_m - l| + |l - s_n| < 2\varepsilon.$$

Hence (s_n) is a Cauchy sequence (cf. Lemma 4.14).

Suppose conversely that (s_n) is a Cauchy sequence in $\mathbb{R}$. We show first that (s_n) is bounded. Take $\varepsilon = 1$, say, in the Cauchy condition. Thus there exists an N such that $m, n \geqslant N$ imply $|s_m - s_n| < 1$, so for any $m \geqslant N$ we have $|s_m - s_N| < 1$, and hence, using the triangle inequality,

$$|s_m| = |s_m - s_N + s_N| \leqslant |s_m - s_N| + |s_N| < 1 + |s_N|.$$

From this we get $|s_n| \leqslant \max\{|s_1|, |s_2|, \ldots |s_{N-1}|, 1 + |s_N|\}$ for all n, so (s_n) is bounded. (We could have used any fixed positive choice of ε in place of 1 in this part of the proof—for example, 10^{10} or 10^{-10}.)

Next, in order to use Theorem 4.16, we manufacture a monotonic sequence out of (s_n) in the following subtle fashion. For each $m \in \mathbb{N}$ we let S_m be the set of members of the sequence from the mth stage onwards, $S_m = \{s_n : n \geqslant m\}$. Since the whole set of members $S = S_1$ of the sequence is bounded, so is S_m. Hence by the completeness property $\sup S_m$ exists. Let $t_m = \sup S_m$. Since $S_{m+1} \subseteq S_m$, we have $\sup S_{m+1} \leqslant \sup S_m$ (see Exercise 4.1). Thus the sequence (t_m) is monotonic decreasing. Also, $t_m \geqslant s_m$ by definition of t_m, and S is bounded below, so (t_m) is bounded below. So by Theorem 4.16, (t_m) converges, say to l.

Finally we prove, by a 3ε-argument, that (s_n) also converges to l. Given $\varepsilon > 0$ there exists N_1 such that $|s_m - s_n| < \varepsilon$ for $m, n \geqslant N_1$ and there exists N_2 such that $|l - t_m| < \varepsilon$ for $m \geqslant N_2$. Put $N = \max\{N_1, N_2\}$. Since t_N is $\sup S_N$, we know that $t_N - \varepsilon$ is not an upper bound of S_N, so there exists $M \geqslant N$ such that $s_M > t_N - \varepsilon$; also, $s_M \leqslant t_N$ since $s_M \in S_N$ and t_N is an upper bound for S_N. Hence $|s_M - t_N| < \varepsilon$. Now for any $n \geqslant N$, using the triangle inequality twice,

$$|s_n - l| = |s_n - s_M + s_M - t_N + t_N - l| \leqslant |s_n - s_M| + |s_M - t_N| + |t_N - l| < 3\varepsilon.$$

Hence (s_n) converges to l (using Lemma 4.14). □

There is a further result about sequences which we record here for later reference: it is a version of the Bolzano–Weierstrass theorem.

Theorem 4.19 *Every bounded sequence of real numbers has at least one convergent subsequence.*

There is a proof on the web site.

Before leaving sequences we recall that their limits behave well under algebraic operations in the following sense.

Proposition 4.20 *Suppose that (s_n), (t_n) converge to s, t. Then*
(a) $(s_n + t_n)$ converges to $s + t$,
(b) $(s_n t_n)$ converges to st,
(c) $(1/t_n)$ converges to $1/t$ provided $t \neq 0$.

A few particular limits which we need are included in the exercises below.

Limits of functions

Limits of functions are used in the theoretical study of continuity, differentiability, and integration, and in practical estimates of the behaviour of particular functions.

Suppose first for simplicity that we have a function $f : \mathbb{R} \to \mathbb{R}$. (In general the domain could be smaller.) Let $a \in \mathbb{R}$.

Definition 4.21 *We say that $f(x)$* tends to the limit *l as x* tends to *a, and write $\lim_{x \to a} f(x) = l$, if given (any real number) $\varepsilon > 0$ there exists (a real number) $\delta > 0$ such that $|f(x) - l| < \varepsilon$ for all real numbers x which satisfy $0 < |x - a| < \delta$.*

This is similar to the definition of convergence of a sequence (s_n), but instead of looking at s_n for large values of n, we look at $f(x)$ for x close to, **but not equal to**, a. Again the phrases in parentheses are usually omitted, and we note that the size of δ needed will in general depend

on ε. **The value $f(a)$ is irrelevant to the existence of $\lim_{x\to a} f(x)$, and the limit, if it exists, may or may not equal $f(a)$.** Exercise 4.12 is a good test of whether this important point has been fully absorbed.

Example 4.22 Let $f : \mathbb{R} \to \mathbb{R}$ be given by

$$f(x) = x \quad \text{for } x \neq 0, \qquad f(0) = 1.$$

Then $\lim_{x\to 0} f(x) = 0$. For given $\varepsilon > 0$, put $\delta = \varepsilon$. If $0 < |x - 0| < \delta$, then $|f(x) - 0| = |x| < \varepsilon$, as required.

To emphasize further that $f(a)$ is irrelevant to the existence or value of $\lim_{x\to a} f(x)$, we note that Definition 4.21 makes sense even if $f(a)$ is not defined—it is enough to assume that f is defined on some subset $A \subseteq \mathbb{R}$, where A contains numbers arbitrarily close to (but not equal to) a. We shall not study this general case, but we note two especially useful ways of generalizing Definition 4.21. Suppose first that the domain A of f contains the open interval (a, d) for some $d > a$.

Definition 4.23 *The* right-hand limit $\lim_{x\to a+} f(x)$ *is equal to l if given $\varepsilon > 0$ there exists $\delta > 0$ such that $|f(x) - l| < \varepsilon$ for all x in $(a, a + \delta)$.*

(Note that δ may be chosen small enough so that $(a, a + \delta) \subseteq (a, d)$, and therefore $f(x)$ is defined for all x in $(a, a + \delta)$.) Left-hand limits are defined similarly.

Next, here are two examples much used in illustrating theoretical points.

Example 4.24 Let $f, g : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ be given by

$$f(x) = x \sin 1/x, \qquad g(x) = \sin 1/x.$$

Then $\lim_{x\to 0} f(x) = 0$, while $\lim_{x\to 0} g(x)$ does not exist.

The proofs are left as Exercise 4.14.

Results about limits of functions may be proved by analogy with the proofs about sequences or we may deduce them from the latter using the following conversion lemma.

Lemma 4.25 *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 we have $x_n \neq a$, then $(f(x_n))$ converges to l.

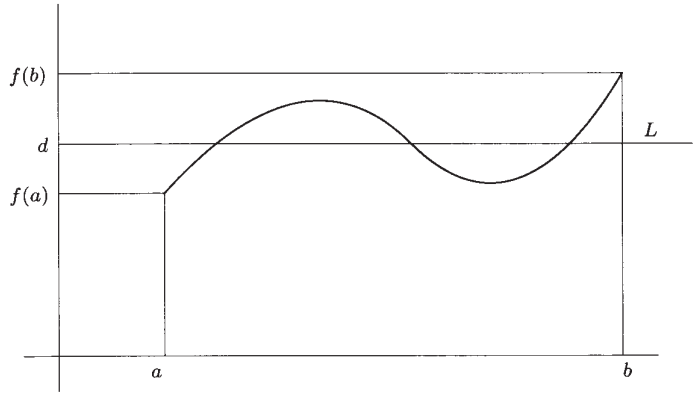

Figure 4.2. Intermediate value property

The proof is on the web site. One may also prove analogues of Theorem 4.18 and Proposition 4.20 for limits of functions, and for left- and right-hand limits.

Continuity

In this section we review the way in which a precise definition of continuity is derived from the intuitive notion. We first make a false start.

One statement containing something of the intuitive idea of continuity is that a function is continuous if its graph can be drawn without lifting pencil from paper. To formulate this more mathematically, let $f : \mathbb{R} \to \mathbb{R}$ be a function and let $(a, f(a))$, $(b, f(b))$ be two points on its graph (see Figure 4.2).

Let L be the horizontal line at some height d between $f(a)$ and $f(b)$. Then to satisfy our intuition about continuity, the graph of f has to cross the line L at least once on its way from $(a, f(a))$ to $(b, f(b))$. In other words, there exists at least one point c in $[a, b]$ such that $f(c) = d$. Formally, we make the following definition.

Definition 4.26 *A function $f : \mathbb{R} \to \mathbb{R}$ has the* intermediate value property (IVP) *if given any a, b, d in $\mathbb{R}$ with $a < b$ and d between $f(a)$ and $f(b)$, there exists at least one c satisfying $a \leqslant c \leqslant b$ and $f(c) = d$.*

This definition also applies when the domain $\mathbb{R}$ in Definition 4.26 is replaced by an interval in $\mathbb{R}$.

A tentative definition of continuity would be that f is continuous if it has the IVP. However, this fails to capture completely the intuitive idea of continuity, as the next example shows.

Example 4.27 Let f be given by

$$f(x) = \begin{cases} 0 & \text{for } x \leqslant 0, \\ \sin 1/x & \text{for } x > 0. \end{cases}$$

Part of the graph of f is shown in Figure 4.3.

Although we shall not prove it now, it is easy to believe by inspection that f does have the IVP. But f does not satisfy our intuitive requirements for a continuous function—something is wrong near $x = 0$. On closer scrutiny, we realize that our intuition includes the requirement that for all values of x near 0, $f(x)$ should be reasonably close to $f(0)$, not oscillating with amplitude 1 as it does in this example. More precisely, the reason we are dissatisfied with f is that $\lim_{x\to 0} f(x)$ does not exist. Considerations such as these lead to the accepted definition.

Definition 4.28 *A function $f : \mathbb{R} \to \mathbb{R}$ is* continuous at a *if $\lim_{x\to a} f(x)$ exists and is $f(a)$.*

Using Definition 4.21 this translates into $\varepsilon - \delta$ form.

Definition 4.29 *A function $f : \mathbb{R} \to \mathbb{R}$ is* continuous at a *if given any $\varepsilon > 0$, there exists $\delta > 0$ such that $|f(x) - f(a)| < \varepsilon$ for any x such that $|x - a| < \delta$.*

As usual, the size of δ needed in general depends on ε, though we do not exhibit that in the notation.

A third way of expressing continuity of a function f is to say that $\lim_{x\to a-} f(x)$ and $\lim_{x\to a+} f(x)$ both exist and both equal $f(a)$. This has the advantage of identifying the ways in which continuity at a might fail: the left-hand limit, or the right-hand limit of f at a (or both of these) might fail to exist; or both left- and right-hand limits exist, but at least one of them fails to equal $f(a)$. (In this last case we say that f has a *simple jump discontinuity at a*.)

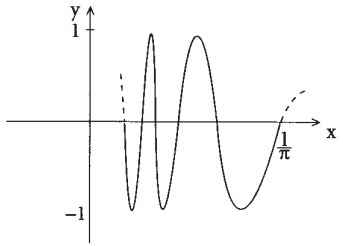

Figure 4.3. Graph of $\sin 1/x$

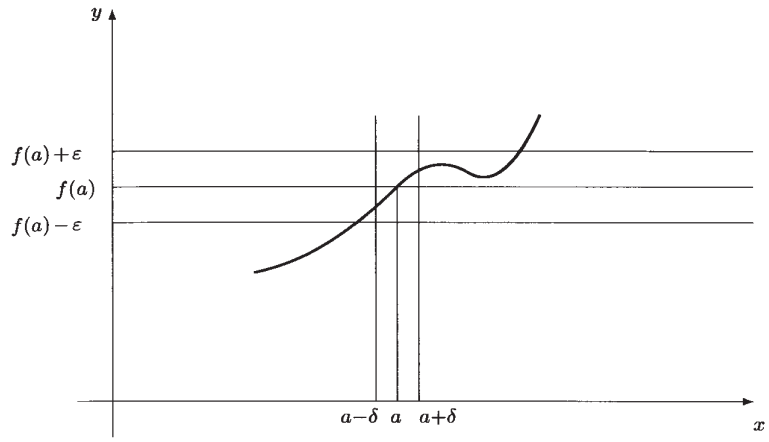

Figure 4.4. Continuity at a

Here are two ways of thinking about continuity of f at a.

(1) In terms of approximations: we can ensure that $f(x)$ approximates $f(a)$ within any prescribed degree of accuracy by choosing x to approximate a sufficiently accurately.

(2) Geometrically: given a horizontal band of any positive width 2ε centred on height $f(a)$, we can choose a vertical band of some suitable width 2δ centred on $x = a$ such that the part of the graph of f in this vertical band is also in the horizontal band (see Figure 4.4): if an aeroplane is flying at 10000 ft at time $t = a$ then it is between 9000 ft and 11000 ft for a non-zero time interval around $t = a$, unless it is capable of discontinuous flight.

The same idea motivates the next result.

Proposition 4.30 *Suppose that $f : \mathbb{R} \to \mathbb{R}$ is continous at $a \in \mathbb{R}$ and that $f(a) \neq 0$. Then there exists $\delta > 0$ such that $f(x) \neq 0$ whenever $|x - a| < \delta$.*

Proof Take $\varepsilon = |f(a)|/2$ in Definition 4.29. Then there exists $\delta > 0$ such that $|f(x) - f(a)| < |f(a)|/2$ whenever $|x - a| < \delta$. For such x, using the reverse triangle inequality (Corollary 4.10) we get

$$|f(x)| = |f(a)-(f(a)-f(x))| \geqslant |f(a)|-|f(a)-f(x)| > |f(a)|-|f(a)|/2 > 0,$$

so $f(x) \neq 0$. □

In view of such results, continuity is sometimes called 'the principle of inertia'.

As in the definition of $\lim_{x\to a} f(x)$, it is not necessary for f to be defined on all of $\mathbb{R}$ for the definition of continuity of f at a to make sense. It is certainly enough for f to be defined on some open interval I containing a, since then in Definition 4.29 we can take δ small enough so that $x \in I$ whenever $|x-a| < \delta$. Also, we say that f is *continuous at a from the right (or the left)* if $\lim_{x\to a+} f(x)$ (or $\lim_{x\to a-} f(x)$) exists and equals $f(a)$.

Examples of continuous functions

In this section we review how to build up many examples of continuous functions. If $f, g : \mathbb{R} \to \mathbb{R}$ are functions then we can define functions $|f|$, $f+g$, $f.g : \mathbb{R} \to \mathbb{R}$ by the formulae

$|f|(x) = |f(x)|$, $(f+g)(x) = f(x)+g(x)$, $(f.g)(x) = f(x)g(x)$ for all $x \in \mathbb{R}$.

Also, if $\mathcal{Z} = \{x \in \mathbb{R} : g(x) = 0\}$, 'the zero set of g', then we may define $1/g : \mathbb{R} \setminus \mathcal{Z} \to \mathbb{R}$ by $(1/g)(x) = 1/g(x)$ for all $x \in \mathbb{R} \setminus \mathcal{Z}$.

Proposition 4.31 *Suppose that $f, g : \mathbb{R} \to \mathbb{R}$ are continuous at $a \in \mathbb{R}$. Then so are (a) $|f|$, (b) $f+g$ and (c) $f.g$. (d) If $g(a) \neq 0$ then $1/g$ is continuous at a.*

Proof (a) Let $\varepsilon > 0$. We know there exists $\delta > 0$ with $|f(x) - f(a)| < \varepsilon$ whenever $|x-a| < \delta$. Then using the reverse triangle inequality (Corollary 4.10), whenever $|x-a| < \delta$ we have

$$||f|(x) - |f|(a)| = ||f(x)| - |f(a)|| \leqslant |f(x) - f(a)| < \varepsilon$$

so $|f|$ is continuous at a.

(b) For $\varepsilon > 0$ there exists $\delta_1 > 0$ such that $|f(x) - f(a)| < \varepsilon/2$ whenever $|x-a| < \delta_1$, and $\delta_2 > 0$ such that $|g(x) - g(a)| < \varepsilon/2$ whenever $|x-a| < \delta_2$. Let $\delta = \min\{\delta_1, \delta_2\}$. Then whenever $|x-a| < \delta$, we have

$$\begin{aligned}|(f+g)(x) - (f+g)(a)| &= |f(x) - f(a) + g(x) - g(a)| \\ &\leqslant |f(x) - f(a)| + |g(x) - g(a)| < \varepsilon/2 + \varepsilon/2 = \varepsilon.\end{aligned}$$

so $f+g$ is continuous at a.

(c) For the proof that $f.g$ is continuous at $a \in X$ when f and g are, it makes sense to 'begin at the end'. We are going to use a trick way of writing $f(x)g(x) - f(a)g(a)$, as $f(x)(g(x)-g(a)) + (f(x)-f(a))g(a)$ (the roles of f and g could be exchanged). From this we see

$$\begin{aligned}|f(x)g(x) - f(a)g(a)| &= |f(x)(g(x) - g(a)) + (f(x) - f(a))g(a)| \\ &\leqslant |f(x)||g(x) - g(a)| + |f(x) - f(a)||g(a)|.\end{aligned}$$

We know that $|f(x) - f(a)|$ is small when $|x-a|$ is sufficiently small, and $|g(a)|$ is a constant so gives no trouble—given $\varepsilon > 0$ we may choose

$\delta_1 > 0$ such that $|f(x) - f(a)| < \varepsilon/2(|g(a)| + 1)$ whenever $|x-a| < \delta_1$. [The extra 1 is added on the denominator just to avoid making a special case when $g(a) = 0$.] So $|f(x) - f(a)||g(a)| < \varepsilon/2$ whenever $|x-a| < \delta_1$. But $|f(x)||g(x) - g(a)|$ is slightly more awkward to deal with since $|f(x)|$ varies. However, it does not vary too wildly near a since f is continuous at a: there exists $\delta_2 > 0$ such that $|f(x) - f(a)| < 1$ whenever $|x-a| < \delta_2$, so for all such x, we have $|f(x)| = |f(x) - f(a) + f(a)| \leqslant 1 + |f(a)|$ by the triangle inequality. Finally, by continuity of g at a there exists $\delta_3 > 0$ such that $|g(x) - g(a)| < \varepsilon/2(1 + |f(a)|)$ whenever $|x-a| < \delta_3$. Put $\delta = \min\{\delta_1, \delta_2, \delta_3\}$. Then for any x with $|x-a| < \delta$ we have

$$\begin{aligned}|f(x)g(x) - f(a)g(a)| &\leqslant |f(x)||g(x) - g(a)| + |f(x) - f(a)||g(a)| \\ &< \frac{(1+|f(a)|)\varepsilon}{2(1+|f(a)|)} + \frac{\varepsilon|g(a)|}{2(|g(a)|+1)} \\ &< \varepsilon/2 + \varepsilon/2 = \varepsilon.\end{aligned}$$

So $f.g$ is continuous at $a \in X$.

(d) First we note that by continuity of g at a, there is an open interval containing a on which $1/g$ is defined because g is never zero (see Proposition 4.30). Now beginning at the end again, we are going to use

$$\left|\frac{1}{g(x)} - \frac{1}{g(a)}\right| = \frac{|g(a) - g(x)|}{|g(x)||g(a)|}. \qquad (\ddagger)$$

We know that $|g(x) - g(a)|$ is small when $|x-a|$ is small, and $|g(a)|$ is a non-zero constant, so it is easy to handle. But $|g(x)|$ varies, and might 'come dangerously close to 0', so that $\ddagger$ might become large. We get around that as follows. By continuity of g at a, there exists $\delta_1 > 0$ such that $|g(x) - g(a)| < |g(a)|/2$ whenever $|x-a| < \delta_1$. For all such x we have, using the reverse triangle inequality (Corollary 4.10.),

$$|g(x)| = |(g(a) - (g(a) - g(x))| \geqslant |g(a)| - |g(a) - g(x)| \geqslant |g(a)|/2.$$

Continuity of g at a gives $\delta_2 > 0$ such that $|g(x) - g(a)| < \varepsilon|g(a)|^2/2$ whenever $|x-a| < \delta_2$. Put $\delta = \min\{\delta_1, \delta_2\}$. Then using $(\ddagger)$ above, for any x with $|x-a| < \delta$,

$$\left|\frac{1}{g(x)} - \frac{1}{g(a)}\right| = \frac{|g(a) - g(x)|}{|g(x)||g(a)|} \leqslant \frac{|g(x) - g(a)|}{|g(a)|^2/2} < \varepsilon.$$

So $1/g$ is continuous at a. □

The above proofs can be shortened by hiding the secrets of how they are constructed. To illustrate the shorter version, and to see how to reassemble the proof forwards, here is a rabbit-out-of-a-hat proof of (c):

Proof of (c) Let $\varepsilon > 0$. By continuity of f at a, there exists $\delta_1 > 0$ such that $|f(x) - f(a)| < \varepsilon/2(|g(a)| + 1)$ whenever $|x - a| < \delta_1$. Also by continuity of f at a, there exists $\delta_2 > 0$ such that $|f(x) - f(a)| < 1$, and hence $|f(x)| < 1 + |f(a)|$, whenever $|x - a| < \delta_2$. Finally, by continuity of g at a there exists $\delta_3 > 0$ such that $|g(x) - g(a)| < \varepsilon/2(1 + |f(a)|)$ whenever $|x - a| < \delta_3$. Put $\delta = \min\{\delta_1, \delta_2, \delta_3\}$. Then for any x with $|x - a| < \delta$ we have

$$|f(x)g(x) - f(a)g(a)| \leqslant |f(x)||g(x) - g(a)| + |f(x) - f(a)||g(a)| < \varepsilon.$$

□

We can use Proposition 4.31 and induction to see that other real-valued functions are continuous.

Proposition 4.32 *(i) Let $p : \mathbb{R} \to \mathbb{R}$ be the 'polynomial function' defined by $p(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0$ where the a_i are constants. Then p is continuous.*

(ii) Let $r : \mathbb{R} \setminus \mathcal{Z} \to \mathbb{R}$ be the rational function $x \mapsto p(x)/q(x)$ where p and q are polynomial functions and $\mathcal{Z}$ is the zero set of q. Then r is continuous on $\mathbb{R} \setminus \mathcal{Z}$.

Proof For (i), it is easy to check that the map $x \mapsto x$ is continuous on $\mathbb{R}$, and so too is any constant function. Next we show inductively that $x \mapsto x^n$ is continuous on $\mathbb{R}$ for any $n \in \mathbb{N}$. The case $n = 1$ is continuity of $x \mapsto x$. Suppose inductively that $x \mapsto x^n$ is continuous on $\mathbb{R}$. Then using (c) of Proposition 4.31, $x \mapsto x.x^n = x^{n+1}$ is continuous on $\mathbb{R}$. Hence by induction, $x \mapsto x^n$ is continuous for any positive integer n. Since the constant map $x \mapsto a_n$ is also continuous on $\mathbb{R}$, another application of (c) shows that $x \mapsto a_n x^n$ is continuous on $\mathbb{R}$. Now an easy induction on (b) shows that any polynomial function is continuous on $\mathbb{R}$. For (ii), an application of (d) shows that $x \mapsto 1/q(x)$ is continuous on $\mathbb{R} \setminus \mathcal{Z}$, and then (c) shows that $x \mapsto p(x)/q(x)$ is continuous on $\mathbb{R} \setminus \mathcal{Z}$. □

Proposition 4.33 *Suppose that $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ are such that f is continuous at $a \in \mathbb{R}$, and g is continuous at $f(a)$. Then $g \circ f$ is continuous at a.*

Proof Let $\varepsilon > 0$. By continuity of g at $f(a)$ there exists $\delta_1 > 0$ such that $|g(y) - g(f(a))| < \varepsilon$ whenever $|y - f(a)| < \delta_1$. By continuity of f at a, there exists $\delta_2 > 0$ such that $|f(x) - f(a)| < \delta_1$ whenever $|x - a| < \delta_2$. Now for any x with $|x - a| < \delta_2$ we have $|f(x) - f(a)| < \delta_1$ so $|g(f(x)) - g(f(a))| < \varepsilon$. This shows that $g \circ f$ is continuous at a. □

Like Proposition 4.31, this result helps build up a store of continuous functions, especially when used in conjunction with continuity of specific functions such as the exponential and log functions, cosine and sine functions, and the like, whose continuity properties we know from analysis (see for example 7.4 of Hart (2001) or Part III of Spivak (2006)). So functions such as $x \mapsto \sin(x^2 + 3x + 1)$, $x \mapsto \mathrm{e}^{-x^2}$, $x \mapsto \mathrm{e}^{x^3 + \cos x}$ are continuous on $\mathbb{R}$.

★ Here is a more general approach to continuity for real-valued functions of a real variable.

Definition 4.34 *Let $f : X \to \mathbb{R}$ be a function defined on a subset $X \subseteq \mathbb{R}$ and let $a \in X$. We say f* is continuous at a *if given $\varepsilon > 0$ there exists $\delta > 0$ such that $|f(x) - f(a)| < \varepsilon$ whenever $|x - a| < \delta$ and $x \in X$.*

The more general analogue of Proposition 4.31 can be proved similarly; after each occurrence of the phrase 'whenever $|x - a| < \delta$' we just insert 'and $x \in X$'. This is a special case of the later Proposition 5.17. ★

In connection with the false start we made on defining continuity, the following theorem, usually called the intermediate value theorem, is true.

Theorem 4.35 *Any continuous function $f : \mathbb{R} \to \mathbb{R}$ has the IVP. The same is true for a continuous function $f : I \to \mathbb{R}$ for any interval I in $\mathbb{R}$.*

We could give the proof now, using the completeness property, but before proving this and other basic results about continuity we raise the stakes by generalizing to functions between more general 'spaces' than subsets of $\mathbb{R}$. The motives for this were mentioned in the introduction.

Exercise 4.1 Show that if $\emptyset \neq A \subseteq B \subseteq \mathbb{R}$ and B is bounded above then A is bounded above and $\sup A \leqslant \sup B$.

Exercise 4.2 Show that if A and B are non-empty subsets of $\mathbb{R}$ which are bounded above then $A \cup B$ is bounded above and

$$\sup A \cup B = \max\{\sup A, \sup B\}.$$

Exercise 4.3 Formulate and prove analogues of Exercises 4.1 and 4.2 for inf.

Exercise 4.4 For each of the following subsets of $\mathbb{R}$ find the sup if it exists, and decide whether it is in the set:

$$\{x : x^2 \leqslant 2x - 1\}, \qquad \{x : x^2 + 2x \leqslant 1\},$$
$$\{x : x^3 < 8\}, \qquad \{x : x \sin x < 1\}.$$

Exercise 4.5 Show that there is no rational number q such that $q^2 = 2$.
[Hint: express q as a quotient of integers m/n where m, n are mutually prime, and show that $m^2 = 2n^2$ leads to a contradiction.]

Exercise 4.6* Show that if m and n are positive integers with highest common factor 1, then m/n is the square of a rational number if and only if m and n are both squares of integers.

Exercise 4.7 Deduce from the completeness property Proposition 4.4 that a non-empty set of real numbers which is bounded below has a greatest lower bound.

Exercise 4.8 Prove that between any two distinct real numbers there is an irrational number.

Exercise 4.9 Prove that if y, α are real numbers with $y > 1$ then $n^\alpha / y^n \to 0$ as $n \to \infty$.
[Hint: use the binomial expansion of $(1 + x)^n$ where $y = 1 + x$.]

Exercise 4.10 Prove that $\lim_{n\to\infty} n^{1/n} = 1$.
[Hint: Put $n^{1/n} = 1 + a_n$ and note that $a_n > 0$ for $n > 1$. Using $(1 + a_n)^n = n$, deduce that $n - 1 \geqslant n(n-1)a_n^2/2$ for $n > 1$ and hence $0 \leqslant a_n^2 \leqslant 2/n$.]

Exercise 4.11 Given a set of r non-negative real numbers $\{a_1, a_2, \ldots, a_r\}$, let $a = \max\{a_1, a_2, \ldots, a_r\}$. Prove that for any positive integer n,

$$a^n \leqslant a_1^n + a_2^n + \ldots + a_r^n \leqslant r a^n.$$

By taking nth roots throughout, deduce that

$$a \leqslant (a_1^n + a_2^n + \ldots + a_r^n)^{1/n} \leqslant r^{1/n} a,$$

and hence that $\lim_{n\to\infty} (a_1^n + a_2^n + \ldots + a_r^n)^{1/n} = a$.

Exercise 4.12 Give an example where $f(x) \to b$ as $x \to a$ and $g(y) \to c$ as $y \to b$ but $g(f(x)) \not\to c$ as $x \to a$.

Exercise 4.13 (a) Prove that for any y, z in $\mathbb{R}$,

$$\max\{y, z\} = \frac{1}{2}(y + z + |y - z|), \quad \min\{y, z\} = \frac{1}{2}(y + z - |y - z|).$$

(b) Given that $f, g : \mathbb{R} \to \mathbb{R}$ are continuous at a, prove that h and k are continuous at $a \in \mathbb{R}$, where for any x in $\mathbb{R}$

$$h(x) = \max\{f(x), g(x)\}, \quad k(x) = \min\{f(x), g(x)\}.$$

Exercise 4.14 Let $f, g : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ be given by

$$f(x) = x \sin 1/x, \qquad g(x) = \sin 1/x.$$

Prove that $\lim_{x\to 0} f(x) = 0$, while $\lim_{x\to 0} g(x)$ does not exist.

Exercise 4.15 Let $f : \mathbb{R} \to \mathbb{R}$ be given by

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

Show that f is not continuous at any point in $\mathbb{R}$.

Exercise 4.16* Let $f : \mathbb{R} \to \mathbb{R}$ be given by

$$f(x) = \begin{cases} 0 & \text{if } x = 0 \text{ or } x \notin \mathbb{Q}, \\ 1/q & \text{if } x = p/q,\ p, q \text{ integers with highest common factor 1},\ q > 0. \end{cases}$$

Prove that f is discontinuous at any non-zero a in $\mathbb{Q}$, but continuous at 0 and at any irrational a in $\mathbb{R}$.

Exercise 4.17* A function $f : \mathbb{R} \to \mathbb{R}$ is said to be *convex* if

$$f(\lambda x + (1 - \lambda)y) \leqslant \lambda f(x) + (1 - \lambda) f(y)$$

for all x, y in $\mathbb{R}$ and all λ in $[0, 1]$. Prove that any convex function is continuous.

Exercise 4.18** For any function $f : \mathbb{R} \to \mathbb{R}$ show that the set of points $a \in \mathbb{R}$ at which f has a simple jump discontinuity is countable.

5 Metric spaces

In this chapter we begin to study metric spaces. These are a bit more concrete than the topological spaces that we shall study later, but they give valuable pointers for the more abstract material. They are also related to analysis and geometry in intuitively appealing ways.

Motivation and definition

The motivation for metric spaces comes from studying continuity. We begin by rephrasing Definition 4.29 using more English and no Greek: a real-valued function of a real variable is continuous at a if we can make the distance $|f(x) - f(a)|$ between $f(x)$ and $f(a)$ as small as we please by choosing x so that the distance $|x - a|$ between x and a is sufficiently small. (The reader is reminded that the terms function and map are interchangeable. We tend to use the former when dealing with functions of real variables, in which case this terminology is long established, and the latter when dealing with maps between more general sets.) Next let us consider a real-valued function f of two real variables. We again get a definition corresponding to our intuitive idea of continuity by changing the above wording very slightly: f is continuous at a point (a, b) in $\mathbb{R}^2$ if we can make the distance between $f(x, y)$ and $f(a, b)$ as small as we please by choosing (x, y) so that its distance from (a, b) is sufficiently small. We may recover an $\varepsilon - \delta$ form of this definition by using the formulae for the distances involved. Since $f(x, y)$ and $f(a, b)$ are real numbers (f is real-valued) the distance between them is $|f(x, y) - f(a, b)|$. Since (x, y) and (a, b) are points in the plane, the distance between them is $\sqrt{[(x-a)^2 + (y-b)^2]}$, where as always this means the non-negative square root. Thus $f : \mathbb{R}^2 \to \mathbb{R}$ is continuous at (a, b) if given $\varepsilon > 0$ there exists $\delta > 0$ such that $|f(x, y) - f(a, b)| < \varepsilon$ for all (x, y) in $\mathbb{R}^2$ satisfying $\sqrt{[(x-a)^2 + (y-b)^2]} < \delta$.

Now given any positive integer n let us try to define continuity for a real-valued function f of n real variables, $f : \mathbb{R}^n \to \mathbb{R}$. We shall denote a point in $\mathbb{R}^n$ by $x = (x_1, x_2, \ldots, x_n)$. By analogy with our previous definitions we may try writing: f is continuous at $a = (a_1, a_2, \ldots, a_n)$ if the distance between $f(x)$ and $f(a)$ can be made as small as we please by choosing x so that the distance between x and a is sufficiently small. Let

us see if this means anything. Since $f(x)$ and $f(a)$ are real numbers, the distance between them is $|f(x)-f(a)|$. However, continuity for $f : \mathbb{R}^n \to \mathbb{R}$ at a does not make sense (for $n > 3$) until we find a suitable meaning for 'the distance between x and a in $\mathbb{R}^n$' for general n.

Such a meaning is not hard to guess when we look at the particular cases $n = 1, 2, 3$. When $n = 3$ and x and a are points in Euclidean 3-space, the distance between them is $\sqrt{[(x_1 - a_1)^2 + (x_2 - a_2)^2 + (x_3 - a_3)^2]}$. We have already used similar formulae for $n = 1$ and $n = 2$. (For the case $n = 1$ note that $|x_1 - a_1| = \sqrt{(x_1 - a_1)^2}$.) It is therefore plausible to define 'the Euclidean distance between $(x_1, x_2, \ldots, x_n)$ and $(a_1, a_2, \ldots, a_n)$' to be

$$\sqrt{\left[\sum_{i=1}^{n}(x_i - a_i)^2\right]}.$$

Definition 5.1 *A function $f : \mathbb{R}^n \to \mathbb{R}$ is* continuous at a point $a \in \mathbb{R}^n$, say $a = (a_1, a_2, \ldots, a_n)$, *if given $\varepsilon > 0$ there exists $\delta > 0$ such that $|f(x) - f(a)| < \varepsilon$ for every $x = (x_1, x_2, \ldots, x_n)$ satisfying*

$$\sqrt{\left[\sum_{i=1}^{n}(x_i - a_i)^2\right]} < \delta.$$

For general n this does not have any graphical interpretation except by analogy with the cases $n = 1$ and $n = 2$. However, it still has familiar physical interpretations. It often happens that some physical quantity depends on several variables. For example, the energy of a given solid body in a gravitational field depends on its height, its linear velocity, and its angular velocity, and these may be described by seven real variables. Continuity in the sense of Definition 5.1 for the function giving the energy in terms of these seven variables means that if the variables are altered slightly the energy changes only slightly.

Now let us try generalizing one step further. The way in which we arrived at Definition 5.1 suggests a plausible definition of continuity for any map $f : X \to Y$ provided that we can give an adequate meaning to 'the distance between' any two elements or points in X and likewise for any two points in Y. Formally, a function giving the distance between any two points of a set X will be a map $d : X \times X \to \mathbb{R}$, since for any two points x, y of X it should give a real number (the distance between x and y). What properties should this distance function, or *metric* d have? The

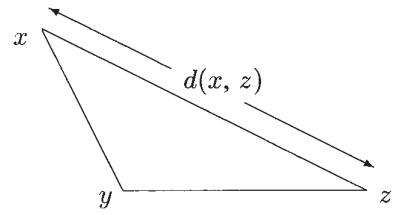

Figure 5.1. Triangle inequality in Euclidean space

choice of these was probably historically a matter of trial and error, but we shall go straight to the historical winners: we pick out some properties of Euclidean distances in the line, plane, and 3-space, and then use them as the *axioms* for a general metric space.

First, the distance between two points is greater than zero except when the points coincide:

(M1) for all $x, y \in X$, $d(x, y) \geqslant 0$; and $d(x, y) = 0$ iff $x = y$.

Secondly, the distance from y to x is the same as the distance from x to y:

(M2) (Symmetry) for all $x, y \in X$, $d(y, x) = d(x, y)$.

Finally we use the *triangle inequality*

(M3) for all $x, y, z \in X$, $d(x, z) \leqslant d(x, y) + d(y, z)$.

Geometrically this says: in any triangle, the length of a side is less than or equal to the sum of the lengths of the other two sides. This is familiar in the plane or in 3-space (see Figure 5.1); in the line, the 'triangle' collapses and we have Proposition 4.9.

It turns out that any function d satisfying just these three properties is similar enough to Euclidean distance for a lot of our geometric intuition about distances to work, so we formalize this into the definition of a metric space.

Definition 5.2 *A* metric space *consists of a non-empty set X together with a function $d : X \times X \to \mathbb{R}$ such that* (M1), (M2), *and* (M3) *above hold.*

We often just talk about 'the metric space X' for short, but there is always a metric d attached to it, which we name only when necessary. The elements of X are called 'points' of the space, and d is called the metric or the distance function. Note that the set X is assumed to be non-empty. The choice between this and allowing the empty set is a matter of swings

and roundabouts—there are advantages and disadvantages in each. We have chosen the non-empty option because if we allow the empty set, then some later results would need certain spaces to be non-empty in a context where it would be easy to forget to say so.

► At this point readers familiar with vector spaces may refer to the web site for the definition and examples of a concept which lies between Euclidean spaces and metric spaces in degree of generality, that of a *normed vector space*. Any norm on a vector space gives rise to a metric on it. Many of the metric spaces below are actually normed vector spaces. ◄

Before giving examples of metric spaces, we follow the train of thought that led us to them by defining continuity in this context.

Definition 5.3 *Suppose that* (X, d_X) *and* (Y, d_Y) *are metric spaces and let* $f : X \to Y$ *be a map.*

(*a*) *We say* f *is* continuous at $x_0 \in X$ *if given* $\varepsilon > 0$, *there exists* $\delta > 0$ *such that* $d_Y(f(x), f(x_0)) < \varepsilon$ *whenever* $d_X(x, x_0) < \delta$.

(*b*) *We say* f *is* continuous *if* f *is continuous at every* $x_0 \in X$.

When there are other metrics around we say 'f is (d_X, d_Y)-continuous'.

Examples of metric spaces

We shall look at several examples of metric spaces, to get familiar with the definition and to explore its scope. The extent to which metric spaces are a fruitful generalization of Euclidean spaces depends largely on how many interesting examples there are of metric spaces. Some of our examples are designed to illustrate phenomena internal to metric space theory, but others are of interest in analysis. The first example is the one from which we abstracted the definition.

Example 5.4 Euclidean n-space $(\mathbb{R}^n, d_2)$ where for

$$x = (x_1, x_2, \ldots, x_n) \text{ and } y = (y_1, y_2, \ldots, y_n),\ d_2(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

It is easy to see that (M1) and (M2) are satisfied. In order to check (M3), let $z = (z_1, z_2, \ldots, z_n)$, let all summations be over $i = 1, 2, \ldots, n$ and write $x_i - y_i = r_i$, $y_i - z_i = s_i$. Then we have to prove

$$\left(\sum r_i^2\right)^{\frac{1}{2}} + \left(\sum s_i^2\right)^{\frac{1}{2}} \geqslant \left(\sum (r_i + s_i)^2\right)^{\frac{1}{2}}.$$

Since both sides are non-negative, it is equivalent (squaring both sides) to prove

$$\sum r_i^2 + \sum s_i^2 + 2\left(\sum r_i^2\right)^{\frac{1}{2}} \left(\sum s_i^2\right)^{\frac{1}{2}} \geqslant \sum r_i^2 + \sum s_i^2 + 2\sum r_i s_i.$$

This in turn is equivalent to proving Cauchy's inequality,

$$\left(\sum r_i^2\right)\left(\sum s_i^2\right) \geqslant \left(\sum r_i s_i\right)^2.$$

The reader may have seen a proof of this inequality before. One proof is on the web site.

When we consider $\mathbb{R}^n$ as a metric space, this Euclidean metric will always be understood unless some other is specified.

Next we see that from the metric space viewpoint, the complex numbers are very like $(\mathbb{R}^2, d_2)$.

Example 5.5 Let $X = \mathbb{C}$ and for z_1, z_2 in $\mathbb{C}$ let $d(z_1, z_2) = |z_1 - z_2|$. Again (M1) and (M2) clearly hold. When we express each complex number in terms of its real and imaginary parts, the triangle inequality for $\mathbb{C}$, $|z_1 - z_3| \leqslant |z_1 - z_2| + |z_2 - z_3|$ coincides with the triangle inequality for $(\mathbb{R}^2, d_2)$. The close relationship between these metric spaces will be made precise in Example 6.40.

The next example is a much stranger one.

Example 5.6 Let X be any non-empty set and define d by

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

It is easy to check that (M1) and (M2) hold. To check (M3), note that if $x = z$ then $d(x, z) = 0$ and certainly $d(x, y) + d(y, z) \geqslant d(x, z)$ since both of $d(x, y)$, $d(y, z)$ are non-negative. If $x \neq z$ then at least one of $x \neq y$, $y \neq z$ must be true, so $d(x, y) + d(y, z)$ is 1 or 2, while $d(x, z) = 1$ so again the triangle inequality holds.

For any set X this metric is called *the discrete metric*. Such 'pathological' examples, as they are nicknamed, are not normally used in applications in analysis. They serve as a warning to check by rigorous proofs that results suggested by intuition really hold in general metric spaces. In

other words, they are potential counterexamples; they explore the boundaries of the concept of a metric space. Nevertheless, metric spaces built up from discrete metric spaces have been used in problems in combinatorics.

The next set of examples show that there are metrics for $\mathbb{R}^n$ other than the Euclidean metric of Example 5.4. We illustrate this in the plane.

Example 5.7 Let $X = \mathbb{R}^2$ and for $x = (x_1, x_2)$, $y = (y_1, y_2)$ let

$$d_1(x, y) = |x_1 - y_1| + |x_2 - y_2|,$$
$$d_2(x, y) = [(x_1 - y_1)^2 + (x_2 - y_2)^2]^{\frac{1}{2}},$$
$$d_\infty(x, y) = \max\{|x_1 - y_1|, |x_2 - y_2|\}.$$

The choice of subscripts is explained in the web site.

We already know that d_2 satisfies the metric space axioms. It is easy to see that d_1 and d_∞ satisfy (M1) and (M2). To check (M3), let $z = (z_1, z_2)$. Then

$$\begin{aligned} d_1(x, y) + d_1(y, z) &= |x_1 - y_1| + |x_2 - y_2| + |y_1 - z_1| + |y_2 - z_2| \\ &= |x_1 - y_1| + |y_1 - z_1| + |x_2 - y_2| + |y_2 - z_2| \\ &\geqslant |x_1 - z_1| + |x_2 - z_2| \\ &= d_1(x, z). \end{aligned}$$

Also, for $i = 1, 2$ we have

$$|x_i - z_i| \leqslant |x_i - y_i| + |y_i - z_i| \leqslant d_\infty(x, y) + d_\infty(y, z),$$

so
$$d_\infty(x, z) \leqslant d_\infty(x, y) + d_\infty(y, z).$$

We could also let d_0 be the discrete metric on $\mathbb{R}^2$. So there may be several distinct metric spaces with the same underlying set. We shall see later that d_1, d_2, d_∞ are all equivalent in a certain sense (and each one deserves to be called a 'product metric'), but that they are not equivalent to d_0.

It is clear how to define analogues of these three metrics on $\mathbb{R}^n$ for any $n \in \mathbb{N}$. The proofs that the axioms hold are similar to the above.

Next we are going to look ways of 'getting new metric spaces from old', which have counterparts for many other mathematical structures. Here they are called (metric) subspaces and products.

Example 5.8 *Metric subspaces.* Suppose that (X, d) is a metric space and that A is a non-empty subset of X. Let $d_A : A \times A \to \mathbb{R}$ be the restriction of d to $A \times A$ (recall that this means $d_A(x, y) = d(x, y)$ for any x, y in A). The metric space axioms hold for d_A since they hold for d.

The metric space (A, d_A) is called a *(metric) subspace* of (X, d) and d_A is called *the metric on A induced by d*. When it is agreed which metric d is intended we may just say that A is a subspace of X. In this looser terminology we call A either a subset or a subspace of X according to the emphasis desired at the time. If A is a non-empty subset of $\mathbb{R}^n$ then in referring to A as a metric space we assume the metric induced by the Euclidean metric on $\mathbb{R}^n$ unless some other is specified. In particular this applies to subsets of $\mathbb{R}$.

When we have metric spaces (X, d), (X', d'), and a map $f : X \to X'$ then for any subset $A \subseteq X$ and $a \in A$ we can talk about continuity (more precisely, (d_A, d')-continuity) of $f|A$ at a. It is important to distinguish this from continuity (more precisely (d, d')-continuity) of f at a. An extreme example of this is

Example 5.9 Consider $f : \mathbb{R} \to \mathbb{R}$ given by

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

Then $f|\mathbb{Q} : \mathbb{Q} \to \mathbb{R}$ is the constant function with value 0, and is continuous at every point of $\mathbb{Q}$, whereas f is not continuous at any point. This example might suggest that continuity of $f|A$ is not very useful. However, suppose that a real-valued function f is defined on some subset of $\mathbb{R}$ containing $[a, b]$. Let us see what continuity of $f|[a, b]$ means. At the point a it means: given any $\varepsilon > 0$ there exists $\delta > 0$ such that $|f(x) - f(a)| < \varepsilon$ for all x satisfying $|x - a| < \delta$ *and also* $x \in [a, b]$, or equivalently, for all x satisfying $a \leqslant x < a + \delta$. This means continuity from the right of f at a. Similarly at b it means continuity from the left. Finally, for any $c \in (a, b)$ it means ordinary (two-sided) continuity. Thus continuity of $f|[a, b]$ *is* of interest, as the reader who has studied the mean value theorem in differential calculus knows.

Example 5.10 *Product spaces.* This generalizes Example 5.7. Given two metric spaces (X, d_X) and (Y, d_Y) we can define several metrics on $X \times Y$. For points (x_1, y_1) and $y = (x_2, y_2)$ in $X \times Y$ let

$$\begin{aligned} d_1((x_1, y_1), (x_2, y_2)) &= d_X(x_1, x_2) + d_Y(y_1, y_2), \\ d_2((x_1, y_1), (x_2, y_2)) &= [d_X(x_1, x_2)^2 + d_Y(y_1, y_2)^2]^{\frac{1}{2}}, \\ d_\infty((x_1, y_1), (x_2, y_2)) &= \max\{d_X(x_1, x_2), d_Y(y_1, y_2)\}. \end{aligned}$$

These may be proved to be metrics on $X \times Y$ just as in the case $X = Y = \mathbb{R}$ (see Exercise 5.16). As in Example 5.7 any one of these deserves to be called a product metric. We shall see in the next chapter that they are all equivalent in a certain sense. The definition may be extended to the product of any finite number of metric spaces.

If the only examples of metric spaces were Examples 5.4, 5.6, 5.7, together with metric subspaces and products formed from them, it is doubtful whether general metric space theory would be worthwhile. The examples below indicate the wide range of metric space theory (but do not exhaust it). First we sketch examples arising in number theory and group theory, respectively.

Example 5.11 Let p be a fixed prime number, and define $d : \mathbb{Z} \times \mathbb{Z} \to \mathbb{R}$ by $d(m, m) = 0$ and for $m \neq n$, $d(m, n) = 1/r$ where p^{r-1} is the highest power of p which divides $m - n$.

(M1) and (M2) are easy to check. For (M3), suppose that $m - n = p^{r-1}k$ and $n - q = p^{s-1}k'$, where k and k' are not divisible by p. We can check that $m - q = p^{t-1}k''$ where $t \geqslant \min\{r, s\}$ (equality holds when $r \neq s$) and k'' is not divisible by p. So

$$d(m, q) = 1/t \leqslant 1/(\min\{r, s\}) = \max\{1/r, 1/s\}$$
$$= \max\{d(m, n), d(n, q)\} \leqslant d(m, n) + d(n, q).$$

Example 5.12 ▶ This example will make sense only if you know about groups and generating sets. Suppose G is a finitely generated group and $\mathcal{A}$ is a generating set for G. Let $F(\mathcal{A})$ be the free group on $\mathcal{A}$, and let $p : F(\mathcal{A}) \to G$ be the natural (onto) map. *The word metric $d_{\mathcal{A}}$ on G associated to $\mathcal{A}$* is defined as follows: for $g_1, g_2 \in G$ let $d_{\mathcal{A}}(g_1, g_2)$ be the length of the shortest word in $p^{-1}(g_1^{-1}g_2)$.

Again (M1) and (M2) are easy to check. For (M3), if w is a word of shortest length in $p^{-1}(g_1^{-1}g_2)$ and w' is a word of shortest length in $p^{-1}(g_2^{-1}g_3)$ then ww' is a word in $p^{-1}(g_1^{-1}g_3)$, and

$$d_{\mathcal{A}}(g_1, g_3) \leqslant \text{length } (ww') = \text{length}(w) + \text{length}(w')$$
$$= d_{\mathcal{A}}(g_1, g_2) + d_{\mathcal{A}}(g_2, g_3). \qquad \blacktriangleleft$$

There are two further kinds of metric spaces used in analysis: sequence spaces and function spaces. We discuss sequence spaces on the web site, and introduce function spaces here. We take some collection of functions and decide to treat it as a 'space', calling the individual functions 'points' and putting a metric on the collection.

To indicate how this might be useful, let us consider a classical problem in one of the first areas where function space language was seen to be appropriate, calculus of variations. The brachistocrone problem is roughly as follows. Suppose we have any two points x, y in a vertical plane, with x higher than, but not vertically above, y. What is the shape of the curve in this plane along which a heavy particle will take the least time to slide from x to y under the action of gravity (with no friction)? It is not important for this illustration to be precise about what kind of curves are intended, but we could take 'curve' to mean one defined by a function $h : [0, 1] \to \mathbb{R}^2$ given by $h(t) = (f(t), g(t))$ where $f, g : [0, 1] \to \mathbb{R}$ are continuously differentiable functions. For any given curve λ we can use integration to calculate the time $T(\lambda)$ the particle takes to make the journey from x to y along λ. Thus we get a real-valued function T defined on the collection of all curves from x to y, and the brachistochrone problem is to find the 'point' λ_0 (if there is one) at which T takes a minimum value. (The answer turns out to be part of a cycloid.) This is like looking for the minimum of a function $T : \mathbb{R} \to \mathbb{R}$, except that the domain $\mathbb{R}$ is replaced by the 'space' of curves. The calculus of variations develops analogues of ordinary calculus for solving such problems. Even to begin calculus, we need continuity of T, and this motivates putting a metric on the collection of curves. (The collection of curves is more or less the set of above functions such as h, satisfying $h(0) = x$, $h(1) = y$, except that distinct functions may define the same geometric curve—for example, h_1, $h_2 : [0, 1] \to \mathbb{R}$ given by $h_1(t) = (t, t)$ and $h_2(t) = (t^2, t^2)$ both describe a straight line segment joining the points $(0, 0)$ and $(1, 1)$.)

In this and other contexts where maps are defined on collections of functions, the language of function spaces is useful. The possibilities it allows for the use of geometric intuition have proved to be fruitful. We now give a few examples of function spaces with metrics on them.

Example 5.13 Let X be the set of all bounded functions $f : [a, b] \to \mathbb{R}$. Given two points f and g in X, let

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

The right-hand side exists, since f and g are bounded, so there are constants K, L such that $|f(x)| \leqslant K$, $|g(x)| \leqslant L$ for all $x \in [a, b]$ and we have

$$|f(x) - g(x)| \leqslant |f(x)| + |g(x)| \leqslant K + L \text{ for all } x \in [a, b].$$

We shall now check in detail that the metric space axioms hold.

(M1) It is clear that $d(f, g) \geqslant 0$ since it is the sup of a (bounded, non-empty) set of non-negative real numbers. Also, if f and g are the same point in X, this means they are identical as functions from $[a, b]$ to $\mathbb{R}$, so $f(x) = g(x)$ for all $x \in [a, b]$, and from its definition $d(f, g) = 0$. Finally, $d(f, g) = 0$ says that $\sup_{x\in[a,b]} |f(x) - g(x)| = 0$, so $f(x) = g(x)$ for all $x \in [a, b]$ which says that $f = g$.

(M2) For any $f, g \in X$ we have $|f(x) - g(x)| = |g(x) - f(x)|$ for all $x \in [a, b]$ so

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

(M3) Let $f, g, h \in X$. For any $c \in [a, b]$,

$$\begin{aligned} |f(c) - h(c)| &\leqslant |f(c) - g(c)| + |g(c) - h(c)| \\ &\leqslant \sup_{x\in[a,b]} |f(x) - g(x)| + \sup_{x\in[a,b]} |g(x) - h(x)| \\ &= d(f, g) + d(g, h), \end{aligned}$$

where the first inequality is just the triangle inequality in $\mathbb{R}$. The above holds for any $c \in [a, b]$, so $d(f, g) + d(g, h)$ is an upper bound for the set

$$S = \{|f(c) - h(c)| : c \in [a, b]\}.$$

Hence $d(f, g) + d(g, h) \geqslant \sup S = d(f, h)$ as required.

This metric is called the *sup metric* or the *uniform metric*. We denote the resulting metric space by $(\mathcal{B}([a, b], \mathbb{R}), d_\infty)$, but note that this notation is not universally agreed.

Any continuous function $f : [a, b] \to \mathbb{R}$ is bounded, by a theorem quoted in the introduction, so the set of all such continuous functions forms a subspace of $\mathcal{B}([a, b], \mathbb{R})$, sometimes written $\mathcal{C}[a, b]$.

Example 5.14 Let X be the set of all continuous functions $f : [a, b] \to \mathbb{R}$ but this time let

$$d(f, g) = \int_a^b |f(t) - g(t)| \, \mathrm{d}t.$$

To check the metric space axioms we need the following lemma from integration theory.

Lemma 5.15 *Suppose that $h : [a, b] \to \mathbb{R}$ is continuous, that $h(t) \geqslant 0$ for all $t \in [a, b]$ and that $\int_a^b h(t) \, \mathrm{d}t = 0$. Then $h(t) = 0$ for all $t \in [a, b]$.*

The idea of the proof is that if $h(c) > 0$ for some $c \in [a, b]$ then by continuity $h(t)$ exceeds some fixed positive number, for example $\frac{1}{2}h(c)$, throughout an interval of non-zero length around c. This makes a strictly positive contribution to the integral which cannot be cancelled out elsewhere since $h(t)$ is never negative.

(M1) It is clear that $d(f, g) \geqslant 0$ for all $f, g \in X$, and that if $f = g$ then $d(f, g) = 0$. If $d(f, g) = 0$ then by Lemma 5.15 applied with $h = |f - g|$ we get $f = g$.

(M2) Symmetry of $d(f, g)$ is clear.

(M3) For any continuous $f, g, h : [a, b] \to \mathbb{R}$ and any $t \in [a, b]$,

$$|f(t) - h(t)| \leqslant |f(t) - g(t)| + |g(t) - h(t)|.$$

Hence by integration theory,

$$\int_a^b |f(t) - h(t)| \, \mathrm{d}t \leqslant \int_a^b |f(t) - g(t)| \, \mathrm{d}t + \int_a^b |g(t) - h(t)| \, \mathrm{d}t,$$

as required. This metric is called the L^1 metric and sometimes written d_1.

Examples 5.13 and 5.14 give us a choice of two metrics, d_∞ and d_1, on the set of continuous functions $f : [a, b] \to \mathbb{R}$. The metric used in any particular situation depends on which properties of the functions are of interest at the time. When we regard g as a good approximation to f iff $g(t)$ is uniformly close to $f(t)$ for all $t \in [a, b]$, we use d_∞ (this will be studied in Chapter 16). On the other hand, we might not be as much interested in the difference in values of the functions at each point as in their average deviation from one another over the range $[a, b]$. We might then use d_1 or some other metric involving integration, such as in the next example.

Example 5.16 Let X be as in Example 5.14 and let

$$d_2(f, g) = \left\{ \int_a^b (f(t) - g(t))^2 \, \mathrm{d}t \right\}^{\frac{1}{2}}.$$

Again (M1) follows from Lemma 5.15, and (M2) clearly holds. The proof that (M3) holds is similar to the proof in Example 5.4 with Cauchy's inequality replaced by its analogue for integrals, the Cauchy–Schwarz inequality:

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

which is proved on the companion web site. The metric d_2 is called the L^2 metric.

Results about continuous functions on metric spaces

Here is a generalization of Proposition 4.31. If $f, g : X \to \mathbb{R}$ are real-valued functions on a metric space X then we can define associated functions $|f|$, $f+g$, $f.g : X \to \mathbb{R}$ where, for all $x \in X$,

$$|f|(x) = |f(x)|, \quad (f+g)(x) = f(x) + g(x), \quad (f.g)(x) = f(x)g(x).$$

Also, if g never takes the value 0 on X then we may define $1/g : X \to \mathbb{R}$ by $(1/g)(x) = 1/g(x)$ for all $x \in X$.

Proposition 5.17 *Suppose that $f, g : X \to \mathbb{R}$ are continuous real-valued functions on a metric space (X, d). Then so are (a) $|f|$, (b) $f+g$, and (c) $f.g$. (d) Also, if g is never zero on X, then $1/g$ is continuous on X.*

Proof Let d be the metric on X. Then in (a)–(d) continuity at any point $a \in X$ can be proved by an exact replica of the proof of Proposition 4.31: we simply replace the domain $\mathbb{R}$ of the functions by X and every occurrence of '$|x-a| < \delta$' by '$d(x, a) < \delta$'. □

An alternative proof will be given shortly.

The next four results will be generalized in Chapter 8.

Proposition 5.18 *Suppose that $f : X \to Y$ and $g : Y \to Z$ are maps of metric spaces with metrics d_X, d_Y, d_Z, that f is continuous at $a \in X$ and that g is continuous at $f(a)$. Then $g \circ f$ is continuous at a.*

Proof Let $\varepsilon > 0$. Since g is continuous at $f(a)$ there exists $\delta_1 > 0$ such that $d_Z(g(y), g(f(a))) < \varepsilon$ whenever $d_Y(y, f(a)) < \delta_1$, and then by continuity of f at a, there exists $\delta_2 > 0$ such that $d_Y(f(x), f(a)) < \delta_1$ whenever $d_X(x, a) < \delta_2$. Then whenever $d_X(x, a) < \delta_2$ we have $d_Y(f(x), f(a)) < \delta_1$ so $d_Z(g(f(x)), g(f(a))) < \varepsilon$. This gives continuity of $g \circ f$ at a. □

The next three results involve product metric spaces. As we have already mentioned, we shall see in the next chapter that the metrics in Example 5.10 are all equivalent in a sense which means that using any one of them would make the next three propositions true. But in the meantime we shall use the metric called d_1 in Example 5.10 whenever we consider a product of metric spaces.

Proposition 5.19 *Suppose that $f : X \to X'$, $g : Y \to Y'$ are maps of metric spaces which are continuous at $a \in X$, $b \in Y$ respectively. Then the map $f \times g : X \times Y \to X' \times Y'$ given by $(f \times g)(x, y) = (f(x), g(y))$, for all $(x, y) \in X \times Y$, is continuous at (a, b).*

Proof Let d_X, d_Y, $d_{X'}$, $d_{Y'}$ be the metrics on X, Y, X', Y'. Recall we are using the metrics d_1, d_1' on $X \times Y$, $X' \times Y'$, where $d_1((x_1, y_1), (x_2, y_2))$ is defined to be $d_X(x_1, x_2) + d_Y(y_1, y_2)$ and similarly for d_1' (see Example 5.10).

Let $\varepsilon > 0$. It follows from continuity of f at a and g at b that there exist $\delta_1 > 0$, $\delta_2 > 0$ such that $d_{X'}(f(x), f(a)) < \varepsilon/2$ whenever $d_X(x, a) < \delta_1$, and $d_{Y'}(g(y), g(b)) < \varepsilon/2$ whenever $d_Y(y, b) < \delta_2$. Put $\delta = \min\{\delta_1, \delta_2\}$. If $d_1((x, y), (a, b)) < \delta$ then $d_X(x, a) \leqslant d_1((x, y), (a, b)) < \delta \leqslant \delta_1$ and similarly $d_Y(y, b) < \delta_2$ so

$$d_1'((f(x), g(y)), (f(a), g(b))) = d_{X'}(f(x), f(a)) + d_{Y'}(g(y), g(b)) < \varepsilon.$$

This proves that $f \times g$ is continuous at (a, b). □

Proposition 5.20 *The projections $p_X : X \times Y \to X$ $p_Y : X \times Y \to Y$ of a metric product onto its factors, defined by $p_X(x, y) = x$, $p_Y(x, y) = y$, are continuous.*

Proof We again use the metric d_1 on $X \times Y$ as in the proof of Proposition 5.19. We check continuity of p_X at $(a, b) \in X \times Y$. Let $\varepsilon > 0$ and choose $\delta = \varepsilon$. Then whenever $d_1((x, y), (a, b)) < \delta$ we have

$$d_X(p_X(x, y), p_X(a, b)) = d_X(x, a) \leqslant d_1((x, y), (a, b)) < \delta = \varepsilon,$$

so p_X is continuous at (a, b), and similarly for p_Y. □

Definition 5.21 *The* diagonal map $\Delta : X \to X \times X$ *of any set X is the map defined by $\Delta(x) = (x, x)$.*

Proposition 5.22 *The diagonal map $\Delta : X \to X \times X$ of any metric space X is continuous.*

Proof As before we use the metric d_1 on $X \times X$ defined by

$$d_1((x_1, x_2), (x_1', x_2')) = d_X(x_1, x_1') + d_X(x_2, x_2').$$

Let $\varepsilon > 0$. Put $\delta = \varepsilon/2$. Then whenever $d_X(x, x') < \delta$ we have

$$d_1(\Delta(x), \Delta(x')) = d_1((x, x), (x', x')) = d_X(x, x') + d_X(x, x') < \varepsilon.$$

This establishes continuity of Δ. □

We can use these results to give a slightly different proof of Proposition 5.17. Note that as special cases of Proposition 4.31 the functions $\mathbb{R} \to \mathbb{R}$ given by $x \mapsto |x|$ and $\mathbb{R} \setminus \{0\} \to \mathbb{R}$ given by $x \mapsto 1/x$ are

both continuous. Hence if $f, g : X \to \mathbb{R}$ are continuous real-valued functions on a metric space X with $g(x)$ never 0, then by Proposition 5.18 the compositions $x \mapsto f(x) \mapsto |f(x)|$ and $x \mapsto g(x) \mapsto 1/g(x)$ are continuous.

Next, the projections $p_1, p_2 : \mathbb{R}^2 = \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ of $\mathbb{R}^2$ onto the coordinate axes are continuous as a special case of Proposition 5.20, hence by Proposition 4.31 so is their sum $(x, y) \mapsto x + y$ and their product $(x, y) \mapsto xy$. Now suppose $f, g : X \to \mathbb{R}$ are real-valued maps on a metric space X which are continuous at $a \in X$. The sum/product of f and g is the composition

$$X \xrightarrow{\Delta} X \times X \xrightarrow{f \times g} \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R},$$

where the third map is either $(x, y) \mapsto x + y$ or $(x, y) \mapsto xy$. For the composition is $x \mapsto (x, x) \mapsto (f(x), g(x)) \mapsto f(x) + g(x)$ or $f(x)g(x)$. By the above results this composition is continuous at a.

Bounded sets in metric spaces

One topic in metric spaces which intuition guides us to generalize easily from Euclidean spaces is that of bounded sets.

Definition 5.23 *A subset S of a metric space (X, d) is* bounded *if there exist $x_0 \in X$ and $K \in \mathbb{R}$ such that $d(x, x_0) \leqslant K$ for all $x \in S$.*

If S satisfies the definition for some $x_0 \in X$ and $K \in \mathbb{R}$, then it also satisfies the definition with x_0 replaced by any other point $x_1 \in X$ and K replaced by $K + d(x_0, x_1)$. For if $d(x, x_0) \leqslant K$ then

$$d(x, x_1) \leqslant d(x, x_0) + d(x_0, x_1) \leqslant K + d(x_0, x_1).$$

If S satisfies 5.23 then $d(x, y) \leqslant d(x, x_0) + d(x_0, y) \leqslant 2K$ for all $x, y \in S$. The following definition therefore makes sense.

Definition 5.24 *If S is a non-empty bounded subset of a metric space with metric d, then the* diameter of S *is* $\sup\{d(x, y) : x, y \in S\}$. *The diameter of the empty set is* 0.

Definition 5.25 *If $f : S \to X$ is a map from a set S to a metric space X, then we say f* is bounded *if the subset $f(S)$ of X is bounded.*

Here is a sample of the kind of result that our intuition about bounded sets in Euclidean spaces suggests.

Proposition 5.26 *The union of any finite number of bounded subsets of a metric space is bounded.*

Proof It is enough to prove this for two bounded sets, since the result then follows by induction. Before reading on, try to think how the proof should work by contemplating two bounded sets in the plane say. Suppose that S_1, S_2 are bounded subsets of a metric space X with metric d. Then there exist points $x_1, x_2 \in X$ and real numbers K_1, K_2 such that $d(x, x_1) \leqslant K_1$ for all $x \in S_1$ and $d(x, x_2) \leqslant K_2$ for all $x \in S_2$. Put $K = \max\{K_1, K_2 + d(x_2, x_1)\}$. Then for any $x \in S_1 \cup S_2$ we have either $x \in S_1$, so $d(x, x_1) \leqslant K_1 \leqslant K$, or else $x \in S_2$, in which case $d(x, x_1) \leqslant d(x, x_2) + d(x_2, x_1) \leqslant K_2 + d(x_2, x_1) \leqslant K$. This shows that $S_1 \cup S_2$ is bounded. □

Open balls in metric spaces

In this section we develop some terminology which is useful for discussing continuity in metric spaces, and will lead us towards a more general framework in which to discuss continuity.

Definition 5.27 *Let (X, d) be a metric space, $x_0 \in X$, and $r > 0$ a real number. The* open ball in X of radius r centred on x_0 *is the set*

$$B_r(x_0) = \{x \in X : d(x, x_0) < r\}.$$

If we are considering more than one metric on X then we write $B_r^d(x_0)$.

Both name and notation vary. Sometimes it is called an 'open spherical neighbourhood'. Notation: we are using B for 'ball'; some others use D for 'disc'.

Example 5.28 (a) In $\mathbb{R}$ (with its usual metric) $B_r(x_0)$ is the open interval $(x_0 - r, x_0 + r)$.
(b) Let $X = \mathbb{R}^2$ and $d = d_2$, the Euclidean metric. Then $B_r(x_0)$ is the open disc of radius r centred on x_0 (the set of all points strictly inside the circle of radius r centred on x_0).
(c) Let $X = \mathbb{R}^3$, $d = d_2$. Then $B_r(x_0)$ is the open ball of radius r centred on x_0 (the set of all points strictly inside the sphere of radius r centred on x_0).
(d) Let $X = \mathbb{R}^2$, $d = d_1$ (see Example 5.7). Then $B_r(x_0)$ is the inside of the square centred on x_0 with diagonals of length $2r$ parallel to the axes, as in Figure 5.2.

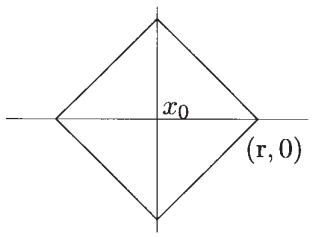

Figure 5.2. $B_r^{d_1}(x_0)$

(e) Let $X = \mathbb{R}^2$ and let d be the discrete metric. Then

$$B_r^d(x_0) = \begin{cases} \{x_0\} & \text{if } r \leqslant 1, \\ \mathbb{R}^2 & \text{if } r > 1. \end{cases}$$

(f) Let X be the set $\mathcal{B}([0, 1], \mathbb{R})$ of all bounded real-valued functions on $[0, 1]$, and let d be the sup metric d_∞. Then for $f_0 \in X$ and $r > 0$ a real number, $B_r^d(f_0)$ is the set of all functions $f \in X$ whose graphs lie inside a ribbon of vertical width $2r$ centred on the graph of f_0 (see Figure 5.3).

Examples 5.28(d) (e) warn us not to take the name 'ball' too seriously—balls are not always round. Examples 5.28(b), (d), (e) show that $B_r^d(x_0)$ depends in general on d. It also depends on the underlying set in the way shown by the next example.

Example 5.29 Let $A = [0, 1] \subseteq \mathbb{R}$ with the Euclidean metric d on $\mathbb{R}$ and the induced metric d_A on A. Then we have $B_1^d(1) = (0, 2)$ while on the other hand $B_1^{d_A}(1) = (0, 1]$.

We mention two things that can be done with open balls before going on. First, we may rephrase the definition of a bounded set: a subset S of

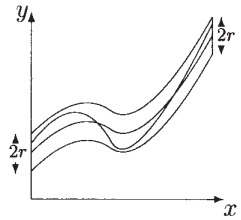

Figure 5.3. Open ball in $(\mathcal{B}([0, 1], \mathbb{R}), d_\infty)$

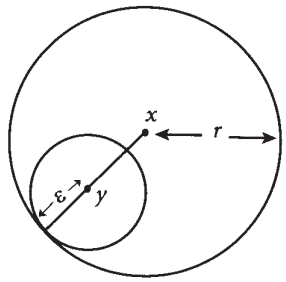

Figure 5.4. Proof of Proposition 5.3

a metric space X is bounded iff $S \subseteq B_r(x_0)$ for some $x_0 \in X$ and $r > 0$. Secondly, we can re-express the definition of continuity in terms of open balls.

Proposition 5.30 *With notation as in Definition 5.3, f is continuous at x_0 iff given $\varepsilon > 0$ there exists $\delta > 0$ such that $f(B_\delta^{d_X}(x_0)) \subseteq B_\varepsilon^{d_Y}(f(x_0))$.*

Proof This is an immediate translation of Definition 5.3. □

We end this section with an important property of open balls whose proof illustrates how geometric intuition and analytic rigour both play a role in metric space theory.

Proposition 5.31 *Given an open ball $B_r(x)$ in a metric space (X, d) and a point $y \in B_r(x)$, there exists $\varepsilon > 0$ such that $B_\varepsilon(y) \subseteq B_r(x)$.*

Proof In the plane, this asserts that we can draw a disc around y lying entirely within the larger disc in Figure 5.4. This is obvious in the picture, but the proof for general metric spaces will have to use the axioms only. However, the picture helps by suggesting what size to try taking ε, namely such that $\varepsilon + d(y, x) \leqslant r$.

Here is the formal proof. Take $\varepsilon = r - d(y, x)$ We note that then $\varepsilon > 0$ since $y \in B_r(x)$ so $d(y, x) < r$. We shall prove that $B_\varepsilon(y) \subseteq B_r(x)$. For if $z \in B_\varepsilon(y)$ then $d(z, y) < \varepsilon$, so $d(z, x) \leqslant d(z, y) + d(y, x) < \varepsilon + d(y, x) = r$, and $z \in B_r(x)$ as required. □

Open sets in metric spaces

Despite the usefulness of open balls, we want a similar but more widely applicable concept generalizing them. Specifically, we generalize the property of open balls expressed in Proposition 5.31, which has been described as 'having some elbow-room around each point'.

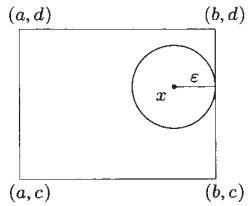

Figure 5.5. Open rectangle

Definition 5.32 *Let (X, d) be a metric space and $U \subseteq X$. We say that U is* open in X *if for every $x \in U$ there exists $\varepsilon_x > 0$ such that $B_{\varepsilon_x}(x) \subseteq U$.*

We have put a suffix x on the ε here to emphasize that in general the size of ε that does the trick will depend on the position of x in the set U. But hereafter we nearly always leave off this x—it is understood that ε depends on x, but the notation becomes clumsy if we insist on emphasising it.

Example 5.33 By Proposition 5.31, any open ball in a metric space X is open in X. In particular any 'open interval' in $\mathbb{R}$ is open in $\mathbb{R}$. On the other hand, intervals in $\mathbb{R}$ such as $[a, b]$, $[a, b)$, $(a, b]$ are not open in $\mathbb{R}$: for $a \in [a, b)$, but no matter how small a positive ε we choose, $B_\varepsilon(a)$ contains points, such as $a - \varepsilon/2$, to the left of a, which are not in $[a, b)$. Note that not every open set is an open ball: for example, in $\mathbb{R}^2$ let U be the interior of a rectangle, say

$$U = \{(x_1, x_2) \in \mathbb{R}^2 : a < x_1 < b,\ c < x_2 < d\}.$$

If $x = (x_1, x_2) \in U$ and we set $\varepsilon = \min\{x_1 - a, b - x_1, x_2 - c, d - x_2\}$, it is easily seen (compare Figure 5.5) that $B_\varepsilon(x) \subseteq U$.

As these Euclidean examples suggest, there are no 'boundary points' in a set U which is open in a metric space—from any point in U one can 'go' some positive distance without going outside U—each point in U has some elbow-room around it, within U.

Example 5.34 For any metric space X, the whole set X and the empty set $\emptyset$ are both open in X. This follows trivially from the definition of 'open'. For example $[a, b]$ is open in $[a, b]$.

Example 5.35 In a discrete metric space X, *any* subset $A \subseteq X$ is open in X. For if $x \in A$ we can choose ε_x to be 1 say, and then $B_{\varepsilon_x}(x) = \{x\} \subseteq A$.

The next examples show that when we say a set is open we have to be careful about which metric space we mean, both about the underlying set and also about the metric.

Example 5.36 A singleton set such as $\{0\}$ is open in $\mathbb{R}$ with the discrete metric, but not in $\mathbb{R}$ with the usual (Euclidean) metric. When necessary we say that a set is 'd-open'. The interval $[a, b]$ is open in $[a, b]$ with its usual metric, as in Example 5.34 above, but not in the larger space $\mathbb{R}$. The interval (a, b) is open in $\mathbb{R}$, but not in $\mathbb{R}^2$ when we identify (a, b) with $(a, b) \times \{0\}$): for $x \in (a, b) \times \{0\}$ there is no $\varepsilon > 0$ such that the disc $B_\varepsilon(x)$ in $\mathbb{R}^2$ is contained in $(a, b) \times \{0\}$—any such disc contains points which are off the x_1-axis—see Figure 5.6.

Next we derive yet another criterion for a map between metric spaces to be continuous, this time in terms of open sets. The reader may think we are not making much progress, but merely juggling with definitions. This is true, but eventually this criterion will lead us to generalizing our whole framework to topological spaces. The criterion says that everything about continuity in metric spaces is entirely encoded in the open sets of the spaces: if we know what the open sets in the spaces are, then a function from one metric space to another is continuous iff the inverse image of any open set is open.

If you feel at all shaky about inverse images of sets, before reading the next definition would be a good time to study Chapter 3.

Proposition 5.37 *Suppose that $f : X \to Y$ is a map of metric spaces. Then f is continuous iff $f^{-1}(U)$ is open in X whenever U is open in Y.*

Proof First suppose that f is continuous and that $U \subseteq Y$ is open in Y. We want to show that $f^{-1}(U)$ is open in X. So let $x_0 \in f^{-1}(U)$. Then $f(x_0) \in U$, and since U is open in Y there exists $\varepsilon > 0$ such that $B_\varepsilon(f(x_0)) \subseteq U$. Since f is continuous at x_0, there exists $\delta > 0$ such that $f(B_\delta(x_0)) \subseteq B_\varepsilon(f(x_0))$. From this we get $f(B_\delta(x_0)) \subseteq U$, so $B_\delta(x_0) \subseteq f^{-1}(U)$ and $f^{-1}(U)$ is open in X as required.

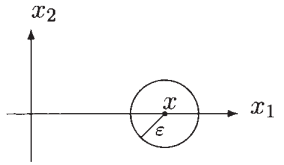

Figure 5.6. $(a, b) \times \{0\}$ not open in $\mathbb{R}^2$

Conversely suppose that $f^{-1}(U)$ is open in X whenever U is open in Y. We shall prove that f is continuous at any $x_0 \in X$. For let $\varepsilon > 0$. Then $B_\varepsilon(f(x_0))$ is open in Y by Proposition 5.31 so $f^{-1}(B_\varepsilon(f(x_0)))$ is open in X. Also, $x_0 \in f^{-1}(B_\varepsilon(f(x_0)))$ since $f(x_0) \in B_\varepsilon(f(x_0))$. So there exists $\delta > 0$ such that $B_\delta(x_0) \subseteq f^{-1}(B_\varepsilon(f(x_0)))$. Then $f(B_\delta(x_0)) \subseteq B_\varepsilon(f(x_0))$, and f is continuous at x_0 by Proposition 5.30. □

Example 5.38 The reader should be warned that when $f : X \to Y$ is a continuous map of metric spaces, it is not necessarily true that the *forwards* image of an open set is open, that is to say, U may be open in X without $f(U)$ being open in Y. For example if we let $f : \mathbb{R} \to \mathbb{R}$ be a constant map, say $f(x) = 0$ for all $x \in \mathbb{R}$, then certainly f is continuous, but for example $(0, 1)$ is open in $\mathbb{R}$ while $f((0, 1)) = \{0\}$ is not open in $\mathbb{R}$.

We end this chapter with two results which show that 'open set' is a more flexible concept than 'open ball'. They feature again later.

Proposition 5.39 *If $U_1, U_2, \ldots, U_m$ are open in a metric space X then so is $\bigcap_{i=1}^{m} U_i$.*

Proof Let $x \in \bigcap_{i=1}^{m} U_i$. Then $x \in U_i$ for each $i = 1, 2, \ldots, m$, so there exists $\varepsilon_i > 0$ such that $B_{\varepsilon_i}(x) \subseteq U_i$. Put $\varepsilon = \min\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m\}$. Then

$$B_\varepsilon(x) \subseteq B_{\varepsilon_i}(x) \subseteq U_i \text{ for each } i = 1, 2, \ldots, m, \text{ so } B_\varepsilon(x) \subseteq \bigcap_{i=1}^{m} U_i$$

and $\bigcap_{i=1}^{m} U_i$ is open as required. □

Thus the intersection of a finite number of open sets is open. Without finiteness, the result is false in general.

Example 5.40 In $\mathbb{R}$, the interval $(-1/n, 1/n)$ is open for each $n \in \mathbb{N}$. But $\bigcap_{n=1}^{\infty} (-1/n, 1/n) = \{0\}$. To see this, note that $0 \in (-1/n, 1/n)$ for every $n \in \mathbb{N}$, so $0 \in \bigcap_{n=1}^{\infty} (-1/n, 1/n)$. On the other hand if $x \neq 0$ then x is not in this intersection, since for sufficiently large n, $x \notin (-1/n, 1/n)$. Since $\{0\}$ is not open in $\mathbb{R}$, we have the required example.

Proposition 5.41 *The union of any collection of sets open in a metric space X is open in X.*

Proof Let I be an indexing set, and for each $i \in I$ let U_i be an open subset of the metric space X. We shall show that $\bigcup_{i \in I} U_i$ is open in X. Let $x \in \bigcup_{i \in I} U_i$. We have $x \in U_{i_0}$ for some $i_0 \in I$, so there exists $\varepsilon > 0$ such that $B_\varepsilon(x) \subseteq U_{i_0}$. Then $B_\varepsilon(x) \subseteq \bigcup_{i \in I} U_i$, and the latter is open in X. □

We note that in general neither an intersection nor a union of open balls is again an open ball: this illustrates the greater flexibility of open sets.

Exercise 5.1 Given points x, y, z in a metric space (X, d) prove that

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

Exercise 5.2 Given points x, y, z, t in a metric space (X, d) prove that

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

Exercise 5.3 Given points $x_1, x_2, \ldots, x_n$ in a metric space (X, d) prove that

$$d(x_1, x_n) \leqslant d(x_1, x_2) + d(x_2, x_3) + \ldots + d(x_{n-1}, x_n).$$

Exercise 5.4 Show that each of the following formulas defines a metric for $\mathbb{R}$:

$(a)\, d(x, y) = |x^3 - y^3|$, $(b)\, d(x, y) = |e^x - e^y|$, $(c)\, d(x, y) = |\tan^{-1}(x) - \tan^{-1}(y)|$.

Which property of the maps $x \mapsto x^3$, $x \mapsto e^x$, $x \mapsto \tan^{-1}(x)$ makes this work?

Exercise 5.5 Suppose that x, y are distinct points in a metric space (X, d) and let $\varepsilon = d(x, y)/2$. Prove that $B_\varepsilon(x)$ and $B_\varepsilon(y)$ are disjoint.

Exercise 5.6 Suppose that x, y are points in a metric space and that $\varepsilon > 0$. Show that if $y \in B_{\varepsilon/2}(x)$ then $B_{\varepsilon/2}(y) \subseteq B_\varepsilon(x)$.

Exercise 5.7 Show that if S is a bounded set in $\mathbb{R}^n$ then S is contained in $[a, b] \times [a, b] \times \ldots \times [a, b]$ for some $a, b \in \mathbb{R}$.

Exercise 5.8 Suppose that (X, d) is a metric space, $A \subseteq X$. Show that A is bounded iff there is some constant Δ such that $d(a, a') \leqslant \Delta$ for all $a, a' \in A$.

Exercise 5.9 Suppose that $A \subseteq B$ where B is a bounded subset of a metric space. Prove that A is bounded and diam $A \leqslant$ diam B.

Exercise 5.10 Prove that if A, B are bounded subsets of a metric space and $A \cap B \neq \emptyset$ then diam $(A \cup B) \leqslant$ diam A + diam B.

Exercise 5.11 Sketch the open ball $B_1^{d_\infty}((0, 0))$ in $\mathbb{R}^2$.

Exercise 5.12 Suppose that d is a metric for a non-empty set X, and for any $x, y \in X$ define

$d^{(1)}(x, y) = kd(x, y)$, where k is a positive constant, $d^{(2)}(x, y) = \min\{1, d(x, y)\}$

$d^{(3)}(x, y) = d(x, y)/(1+d(x, y))$, $d^{(4)}(x, y) = d(x, y)^2$,

Prove that $d^{(1)}$, $d^{(2)}$, $d^{(3)}$ are metrics for X but $d^{(4)}$ may not be a metric for X.

Exercise 5.13 Prove that a subset of a metric space is open iff it is a union of open balls.

Exercise 5.14 Show that for any $x, y \in \mathbb{R}^n$,

$$d_\infty(x, y) \leqslant d_2(x, y) \leqslant d_1(x, y) \leqslant nd_\infty(x, y).$$

Exercise 5.15 Suppose that X is a non-empty set and that d, d' are metrics on X such that $d(x_1, x_2) \leqslant kd'(x_1, x_2)$ for all $x_1, x_2 \in X$ and some positive constant k.

(a) Show that $B_{\varepsilon/k}^{d'}(x) \subseteq B_\varepsilon^d(x)$ for any $x \in X$ and any $\varepsilon > 0$.

(b) Deduce that any subset of X which is d-open is also d'-open.

(c) Show that the open sets in $\mathbb{R}^n$ are the same for the metrics d_1, d_2, d_∞.

Exercise 5.16 Let (X, d_X) and (Y, d_Y) be metric spaces. As in Example 5.10, for $(x_1, y_1), (x_2, y_2) \in X \times Y$ let

$$d_1((x_1, y_1), (x_2, y_2)) = d_X(x_1, x_2) + d_Y(y_1, y_2),$$

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

$$d_\infty((x_1, y_1), (x_2, y_2)) = \max\{d_X(x_1, x_2), d_Y(y_1, y_2)\}$$

(a) Prove that each of d_1, d_2, d_∞ is a metric for $X \times Y$.

(b) Prove that for any $p, q \in X \times Y$,

$$d_\infty(p, q) \leqslant d_2(p, q) \leqslant d_1(p, q) \leqslant 2d_\infty(p, q).$$

(c) Show that the open sets in $X \times Y$ are the same for d_1, d_2, d_∞.

(d) Let U, V be open subsets of X, Y respectively. Show that $U \times V$ is d_i-open in $X \times Y$ for $i = 1, 2, \infty$.

Exercise 5.17 Let (X, d) be a metric space and consider $X \times X$ as a metric space with the metric d_1 of Exercise 5.16. Show that $d : X \times X \to \mathbb{R}$ is continuous. (Hint: you could use Exercise 5.2.)

Exercise 5.18 Suppose that in a metric space X we have $B_r(x) = B_s(y)$ for some $x, y \in X$ and some positive real numbers r, s. Is $x = y$? Is $r = s$?