This is a rigorous introduction to real analysis for undergraduate students, starting from the axioms for a complete ordered field and a little set theory. The book avoids any preconceptions about the real numbers and takes them to be nothing but the elements of a complete ordered field. All of the standard topics are included, as well as a proper treatment of the trigonometric functions, which many authors take for granted. The final chapters of the book provide a gentle, example-based introduction to metric spaces with an application to differential equations on the real line.

The author's exposition is concise and to the point, helping students focus on the essentials. Over 200 exercises of varying difficulty are included, many of them adding to the theory in the text. The book is ideal for second-year undergraduates and for more advanced students who need a foundation in real analysis.

CAMBRIDGE
UNIVERSITY PRESS
www.cambridge.org

ISBN 978-1-107-02678-0

9 781107 026780

Lárusson

Lectures on Real Analysis

CAMBRIDGE

Australian Mathematical Society Lecture Series 21

Lectures on Real Analysis

Finnur Lárusson

CAMBRIDGE

Lectures on Real Analysis

This is a rigorous introduction to real analysis for undergraduate students, starting from the axioms for a complete ordered field and a little set theory. The book avoids any preconceptions about the real numbers and takes them to be nothing but the elements of a complete ordered field. All of the standard topics are included, as well as a proper treatment of the trigonometric functions, which many authors take for granted. The final chapters of the book provide a gentle, example-based introduction to metric spaces with an application to differential equations on the real line.

The author's exposition is concise and to the point, helping students focus on the essentials. Over 200 exercises of varying difficulty are included, many of them adding to the theory in the text. The book is ideal for second-year undergraduates and for more advanced students who need a foundation in real analysis.

AUSTRALIAN MATHEMATICAL SOCIETY LECTURE SERIES

1 Introduction to Linear and Convex Programming, N. CAMERON
2 Manifolds and Mechanics, A. JONES, A. GRAY & R. HUTTON
3 Introduction to the Analysis of Metric Spaces, J. R. GILES
4 An Introduction to Mathematical Physiology and Biology, J. MAZUMDAR
5 2-Knots and their Groups, J. HILLMAN
6 The Mathematics of Projectiles in Sport, N. DE MESTRE
7 The Petersen Graph, D. A. HOLTON & J. SHEEHAN
8 Low Rank Representations and Graphs for Sporadic Groups,
C. E. PRAEGER & L. H. SOICHER
9 Algebraic Groups and Lie Groups, G. I. LEHRER (ed.)
10 Modelling with Differential and Difference Equations,
G. FULFORD, P. FORRESTER & A. JONES
11 Geometric Analysis and Lie Theory in Mathematics and Physics,
A. L. CAREY & M. K. MURRAY (eds.)
12 Foundations of Convex Geometry, W. A. COPPEL
13 Introduction to the Analysis of Normed Linear Spaces, J. R. GILES
14 Integral: An Easy Approach after Kurzweil and Henstock, L. P. YEE & R. VYBORNY
15 Geometric Approaches to Differential Equations, P. J. VASSILIOU & I. G. LISLE (eds.)
16 Industrial Mathematics, G. R. FULFORD & P. BROADBRIDGE
17 A Course in Modern Analysis and its Applications, G. COHEN
18 Chaos: A Mathematical Introduction, J. BANKS, V. DRAGAN & A. JONES
19 Quantum Groups, R. STREET
20 Unitary Reflection Groups, G. I. LEHRER & D. E. TAYLOR

Australian Mathematical Society Lecture Series: 21

Lectures on Real Analysis

FINNUR LÁRUSSON
University of Adelaide

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9781107026780

First published 2012

Printed in the United Kingdom at the University Press, Cambridge

A catalog record for this publication is available from the British Library

Library of Congress Cataloging in Publication data
Lárusson, Finnur, 1966–
Lectures on real analysis / Finnur Lárusson.
pages cm. – (Australian Mathematical Society lecture series ; 21)
ISBN 978-1-107-02678-0 (hardback)
1. Mathematical analysis. I. Title.
QA300.5.L37 2012
515 – dc23 2012005596

ISBN 978-1-107-02678-0 Hardback
ISBN 978-1-107-60852-8 Paperback

Contents

Preface vii

To the student ix

Chapter 1. Numbers, sets, and functions 1
1.1. The natural numbers, integers, and rational numbers 1
1.2. Sets 6
1.3. Functions 9
More exercises 11

Chapter 2. The real numbers 15
2.1. The complete ordered field of real numbers 15
2.2. Consequences of completeness 17
2.3. Countable and uncountable sets 19
More exercises 21

Chapter 3. Sequences 23
3.1. Convergent sequences 23
3.2. New limits from old 25
3.3. Monotone sequences 27
3.4. Series 28
3.5. Subsequences and Cauchy sequences 32
More exercises 35

Chapter 4. Open, closed, and compact sets 39
4.1. Open and closed sets 39

4.2. Compact sets 41
More exercises 42

Chapter 5. Continuity 45
5.1. Limits of functions 45
5.2. Continuous functions 47
5.3. Continuous functions on compact sets and intervals 49
5.4. Monotone functions 51
More exercises 53

Chapter 6. Differentiation 55
6.1. Differentiable functions 55
6.2. The mean value theorem 59
More exercises 60

Chapter 7. Integration 63
7.1. The Riemann integral 63
7.2. The fundamental theorem of calculus 67
7.3. The natural logarithm and the exponential function 69
More exercises 71

Chapter 8. Sequences and series of functions 73
8.1. Pointwise and uniform convergence 73
8.2. Power series 76
8.3. Taylor series 80
8.4. The trigonometric functions 83
More exercises 87

Chapter 9. Metric spaces 91
9.1. Examples of metric spaces 91
9.2. Convergence and completeness in metric spaces 95
More exercises 99

Chapter 10. The contraction principle 103
10.1. The contraction principle 103
10.2. Picard's theorem 107
More exercises 111

Index 113

Preface

This book is a rigorous introduction to real analysis, suitable for a one-semester course at the second-year undergraduate level, based on my experience of teaching this material many times in Australia and Canada. My aim is to give a treatment that is brisk and concise, but also reasonably complete and as rigorous as is practicable, starting from the axioms for a complete ordered field and a little set theory.

Along with epsilons and deltas, I emphasise the alternative language of neighbourhoods, which is geometric and intuitive and provides an introduction to topological ideas. I have included a proper treatment of the trigonometric functions. They are sophisticated objects, not to be taken for granted. This topic is an instructive application of the theory of power series and other earlier parts of the book. Also, it involves the concept of a group, which most students won't have seen in the context of analysis before.

There may be some novelty in the gentle, example-based introduction to metric spaces at the end of the book, emphasising how straightforward the generalisation of many fundamental notions from the real line to metric spaces really is. The goal is to develop just enough metric space theory to be able to prove Picard's theorem, showing how a detour through some abstract territory can contribute back to analysis on the real line.

Needless to say, I claim no originality whatsoever for the material in this book. My contribution, such as it is, lies in the selection and presentation of the material. I thank the American Mathematical Society for allowing the book to be formatted with one of their class files.

Finnur Lárusson

To the student

The purpose of this course is twofold. First, to give a careful treatment of calculus from first principles. In first-year calculus we learn methods for solving specific problems. We focus on how to use these methods more than why they work. To pave the way for further studies in pure and applied mathematics we need to deepen our understanding of *why*, as opposed to *how*, calculus works. This won't be a simple rehashing of first-year calculus at all. Calculus done this way is called *real analysis*.

In particular, we will consider what it is about the real numbers that makes calculus work. Why can't we make do with the rationals? We will identify the key property of the real numbers, called *completeness*, that distinguishes them from the rationals and permeates all of mathematical analysis. Completeness will be our main theme through the whole course.

The second goal of the course is to practise reading and writing mathematical proofs. The course is proof-oriented throughout, not to encourage pedantry, but because proof is the only way that mathematical truth can be known with certainty. Mathematical knowledge is accumulated through long chains of reasoning. We can't rely on this knowledge unless we're sure that every link in the chain is sound. In many future endeavours, you will find that being able to construct and communicate solid arguments is a very useful skill.

With the emphasis on rigorous arguments comes the need to make our fundamental assumptions, from which our reasoning begins, clear and explicit. We shall list ten axioms that describe the real numbers and that can in fact be shown to *characterise* the real numbers. Our development of real analysis will be based on these axioms, along with a bit of set theory.

Towards the end of the course we extend some of the concepts we will have developed in the context of the real numbers to the much more general setting of metric spaces. To demonstrate the power of abstraction, the course ends with the proof, using metric space theory, of an existence and uniqueness theorem for solutions of differential equations.

Numbers, sets, and functions

1.1. The natural numbers, integers, and rational numbers

We assume that you are familiar with the set of natural numbers

$$\mathbb{N} = \{1, 2, 3, \dots\},$$

the set of integers

$$\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\},$$

and the set of rational numbers

$$\mathbb{Q} = \{p/q : p, q \in \mathbb{Z}, q \neq 0\}.$$

We also assume that you are familiar with the important method of proof known as the *principle of induction*. It says that if we have a property $P(n)$ that each natural number n may or may not have, such that:

(a) $P(1)$ is true, and

(b) if $k \in \mathbb{N}$ and $P(k)$ is true, it follows that $P(k+1)$ is true,

then $P(n)$ is true for all $n \in \mathbb{N}$. There is another way to state the principle of induction that shows it to be a fundamental property of the natural numbers.

1.1. Theorem. The following are equivalent.

(1) The principle of induction.

(2) Every nonempty subset of $\mathbb{N}$ has a smallest element.

Property (2) is called the *well-ordering property* of $\mathbb{N}$. We say that $\mathbb{N}$ is *well ordered*.

Proof. To show that the two statements are equivalent, we must prove that each implies the other.

(1) $\Rightarrow$ (2): Let S be a subset of $\mathbb{N}$ with no smallest element. Let $P(n)$ be the property that $k \notin S$ for all $k \leq n$. Since S has no smallest element, $1 \notin S$, so $P(1)$ is true. Also, if $P(n)$ is true, $P(n+1)$ must be true as well, for otherwise $n+1$ would be the smallest element of S. Thus $P(n)$ satisfies (a) and (b), so by assumption, $P(n)$ holds for all $n \in \mathbb{N}$ and S is empty.

(2) $\Rightarrow$ (1): Let $P(n)$ be a property of natural numbers satisfying (a) and (b). Define S to be the set of those $n \in \mathbb{N}$ for which $P(n)$ is false. Then (a) says that $1 \notin S$, and (b) (or rather its contrapositive) says that if $k \in S$, $k > 1$, then $k - 1 \in S$. Therefore S has no smallest element, so by assumption S must be empty, which means that $P(n)$ is true for all $n \in \mathbb{N}$. □

1.2. Remark. The *contrapositive* of an implication $P \Rightarrow Q$ is the implication not-$Q \Rightarrow$ not-P. These two implications are logically equivalent. Thus, if we want to prove that P implies Q, then we can instead prove that not-Q implies not-P. This is sometimes convenient. Do not confuse the contrapositive with the *converse* of $P \Rightarrow Q$, which is the implication $Q \Rightarrow P$. An implication and its converse are in general *not* equivalent.

We can think of $\mathbb{Z}$ as an extension of $\mathbb{N}$ that allows us to do subtraction without any restrictions, and of $\mathbb{Q}$ as an extension of $\mathbb{Z}$ that allows us to do division with the sole restriction that division by zero cannot be reasonably defined. The set $\mathbb{Q}$ with addition and multiplication and all the familiar rules satisfied by these operations is a mathematical structure called a *field*.

1.3. Definition. A *field* is a set F with two operations, *addition*, denoted $+$, and *multiplication*, denoted $\cdot$, such that the following axioms are satisfied.

- A1 *Associativity:* $a + (b + c) = (a + b) + c$, $a \cdot (b \cdot c) = (a \cdot b) \cdot c$ for all $a, b, c \in F$.
- A2 *Commutativity:* $a + b = b + a$, $a \cdot b = b \cdot a$ for all $a, b \in F$.
- A3 *Distributivity:* $a \cdot (b + c) = a \cdot b + a \cdot c$ for all $a, b, c \in F$.
- A4 *Additive identity.* There is an element called 0 in F such that $a + 0 = a$ for all $a \in F$.
  Multiplicative identity. There is an element called 1 in F such that $1 \neq 0$ and $a \cdot 1 = a$ for all $a \in F$.
- A5 *Additive inverses.* For every $a \in F$, there is an element called $-a$ in F such that $a + (-a) = 0$.
  Multiplicative inverses. For every $a \in F$, $a \neq 0$, there is an element called a^{-1} in F such that $a \cdot a^{-1} = 1$.

We usually write $a \cdot b$ as ab, $a + (-b)$ as $a - b$, a^{-1} as $1/a$, and ab^{-1} as a/b.

From the field axioms we can derive many familiar properties of fields. It is a good exercise to work out careful proofs of some of these properties based only on the axioms. Here are a few examples. If you prefer, you can simply take $F = \mathbb{Q}$.

1.4. Example. From A2 and A4 we see that $0 + a = a$ and $1 \cdot a = a$ for all $a \in F$.

1.5. Example. The additive identity 0 is unique. Namely, assume $0'$ is another additive identity. By A4, $a + 0 = a$ for all $a \in F$. In particular, taking $a = 0'$, we see that $0' + 0 = 0'$, so by A2, $0 + 0' = 0'$. On the other hand, by assumption, $a + 0' = a$ for all $a \in F$, so taking $a = 0$, we see that $0 + 0' = 0$. We conclude that $0' = 0 + 0' = 0$. Similarly, the multiplicative identity is unique.

Exercise 1.1. Using only the axioms A1–A5, show that the additive inverse of $x \in F$ is unique, that is, if $x + y = 0$ and $x + z = 0$, then $y = z$ (so talking about *the* additive inverse of x is justified). Show also that the multiplicative inverse of $x \in F$, $x \neq 0$, is unique.

1.6. Example. From A2 and A5 we see that for $x \in F$, $(-x) + x = 0$. By Exercise 1.1, we conclude that the additive inverse of $-x$ must be x, that is, $-(-x) = x$. Similarly, for $x \neq 0$, $(x^{-1})^{-1} = x$.

1.7. Example. For every $x \in F$,

$$0 \cdot x \overset{\text{A4}}{=} (0 + 0) \cdot x \overset{\text{A2, A3}}{=} 0 \cdot x + 0 \cdot x.$$

Adding the additive inverse $-(0 \cdot x)$ of $0 \cdot x$ to both sides, we get $0 = 0 \cdot x$. By A2, $x \cdot 0 = 0$ as well.

Exercise 1.2. In A5, $-x$ was introduced as a symbol for the additive inverse of $x \in F$. Using Example 1.7, show that $-x$ is in fact the product of x and the additive inverse -1 of the multiplicative identity 1. In particular,

$$(-1)(-1) = -(-1) = 1.$$

If $x \in F$ and $n \in \mathbb{N}$, $n \geq 2$, we write x^n for the product of n factors of x. By A1, it does not matter how we bracket the product. For example, $x^3 = (x \cdot x) \cdot x = x \cdot (x \cdot x)$. We set $x^0 = 1$ and $x^1 = x$. If $x \neq 0$, we write x^{-n} for $(x^{-1})^n$, which equals $(x^n)^{-1}$. Then $x^{m+n} = x^m x^n$ and $(x^m)^n = x^{mn}$ for all $m, n \in \mathbb{Z}$.

There is more to the rationals than addition and multiplication. The rationals are also *ordered* in a way that interacts well with addition and multiplication. This structure is called an *ordered field*.

1.8. Definition. An *ordered field* is a field F with a relation $<$ (read 'less than') such that the following axioms are satisfied.

A6 For every $a, b \in F$, precisely one of the following holds: $a < b$, $b < a$, or $a = b$.

A7 If $a < b$ and $b < c$, then $a < c$ (the order relation is *transitive*).

A8 If $a < b$, then $a + c < b + c$ for all $c \in F$.

A9 If $a < b$ and $0 < c$, then $ac < bc$.

We take $a \le b$ to mean that $a < b$ or $a = b$; $a > b$ to mean that $b < a$; and $a \ge b$ to mean that $b \le a$. We say that a is *positive* if $a > 0$, and *negative* if $a < 0$.

Again, the axioms imply many further properties.

1.9. Example. We claim that 1 is positive. Note that if $1 < 0$, then adding -1 to both sides gives $0 < -1$ by A8, so multiplying both sides by -1 gives $0 = 0(-1) < (-1)(-1) = 1$ by A9, Example 1.7, and Exercise 1.2, but having both $1 < 0$ and $0 < 1$ contradicts A6.

Having derived a contradiction from the assumption that $1 < 0$, we must reject the assumption as false. Since $0 \ne 1$ by A4, the one remaining possibility by A6 is $0 < 1$.

Exercise 1.3. (a) Show that if $x > 0$, then $-x < 0$. Likewise, if $x < 0$, then $-x > 0$. In particular, by Example 1.9, $-1 < 0$.

(b) Show that if $x > 0$, then $x^{-1} > 0$. Show that if $x > 1$, then $x^{-1} < 1$.

1.10. Definition. An *interval* in an ordered field F is a subset of F of one of the following types, where $a, b \in F$.

$$\begin{aligned}
(a, b) &= \{x : a < x < b\} \\
[a, b] &= \{x : a \le x \le b\} \\
(a, b] &= \{x : a < x \le b\} \\
[a, b) &= \{x : a \le x < b\} \\
(a, \infty) &= \{x : x > a\} \\
(-\infty, a) &= \{x : x < a\} \\
[a, \infty) &= \{x : x \ge a\} \\
(-\infty, a] &= \{x : x \le a\} \\
(-\infty, \infty) &= F
\end{aligned}$$

The intervals (a, b), (a, ∞), $(-\infty, a)$, and F itself are said to be *open*. The intervals $[a, b]$, $[a, \infty)$, $(-\infty, a]$, and F itself are said to be *closed*. Taking $a > b$, we see that the empty set is an interval which is both open and closed. One-point sets $[a, a]$ and the empty set are called *degenerate* intervals. Thus an interval is *nondegenerate* if it contains at least two points.

Exercise 1.4. Show that a nondegenerate interval contains infinitely many points.

1.11. Remark. By A7, if I is an interval, $x < y < z$, and $x, z \in I$, then $y \in I$. In other words, along with any two of its points, an interval contains all the points in between. Conversely, when F is the field of real numbers, a set satisfying this property is an interval (Exercise 2.12).

1.12. Definition. If a and b are elements of an ordered field and $a \leq b$, then we write $\min\{a, b\} = a$ for the *minimum* of a and b, and $\max\{a, b\} = b$ for the *maximum*.

1.13. Definition. The *absolute value* of an element a in an ordered field is the nonnegative element

$$|a| = \max\{a, -a\} = \begin{cases} a & \text{if } a \geq 0, \\ -a & \text{if } a < 0. \end{cases}$$

1.14. Theorem (triangle inequality)**.** For all elements a and b in an ordered field,

$$|a + b| \leq |a| + |b|.$$

For all elements x, y, z in an ordered field,

$$|x - z| \leq |x - y| + |y - z|.$$

Proof. Three cases need to be considered: $a, b \geq 0$; $a \geq 0$ and $b < 0$ (the case when $a < 0$ and $b \geq 0$ is analogous and does not need to be written out in detail); and $a, b < 0$. Let us treat the second case and leave the others as an exercise.

Since $a \geq 0$, we have $-a \leq 0 \leq a$, so, adding $-b$, we get $-(a + b) \leq a - b = |a| + |b|$. Since $b < 0$, we have $b < 0 < -b$, so, adding a, we get $a + b < a - b = |a| + |b|$. These two inequalities together give

$$|a + b| = \max\{a + b, -(a + b)\} \leq |a| + |b|.$$

To get the second inequality, take $a = x - y$ and $b = y - z$. □

Although the rational numbers have a rich structure, they suffer from limitations that call for a larger number system. The following result is attributed to Pythagoras and his associates some 2500 years ago.

1.15. Theorem. There is no rational number with square 2.

Proof. Suppose there are $p, q \in \mathbb{N}$ with $(p/q)^2 = 2$. Choose q to be as small as possible. Now $q < p < 2q$, so $0 < p - q < q$ and $2q - p > 0$. It is easily computed that $\left(\dfrac{2q - p}{p - q}\right)^2 = 2$, contradicting the minimality of q. □

1.16. Remark. Theorem 1.15 has many different proofs. Here is another one. Suppose there was $r \in \mathbb{Q}$ with $r^2 = 2$. We can write $r = p/q$, where p and q are integers with no common factors. We will derive a contradiction from this assumption.

Now $2 = r^2 = p^2/q^2$, so $p^2 = 2q^2$ and p^2 is even. Hence p is even, say $p = 2k$, where k is an integer. Then $2q^2 = p^2 = (2k)^2 = 4k^2$, so $q^2 = 2k^2$ and q^2 is even. Hence q is even, so p and q are both divisible by 2, contrary to our assumption.

Exercise 1.5. Show that there is no rational number with square 3 by modifying the proof of Theorem 1.15 given in Remark 1.16. Where does the proof fail if you try to carry it out for 4? For which $n \in \mathbb{N}$ can you show by the same method that there is no rational number with square n?

This deficiency of $\mathbb{Q}$ leads us to a larger and more sophisticated number system. The real number system has a crucial property called *completeness* which implies, among many other consequences, that every positive real number has a real square root.

A small amount of set theory is essential for real analysis, so before turning to the real numbers we will review some basic concepts to do with sets and functions.

1.2. Sets

The notion of a *set* is a (many would say *the*) fundamental concept of modern mathematics. It cannot be defined in terms of anything more fundamental. Rather, the notion of a set is circumscribed by axioms (usually the so-called *Zermelo-Fraenkel axioms* along with the *axiom of choice*) from which virtually all of mathematics can be derived, at least in principle.

Our approach will be informal. We think of a set as any collection of objects. The objects are called the *elements* of the set. If x is an element of a set A, we write $x \in A$. A set is determined by its elements, that is, two sets are the same if and only if they have the same elements. Thus the most common way to show that sets A and B are equal is to prove, first, that if $x \in A$, then $x \in B$, and second, that if $x \in B$, then $x \in A$.

1.17. Definition. Let A and B be sets. We say that A is a *subset* of B and write $A \subset B$ (some write $A \subseteq B$) if every element of A is also an element of B. We say that A is a *proper subset* of B if $A \subset B$ and $A \neq B$. The *union* of A and B is the set

$$A \cup B = \{x : x \in A \text{ or } x \in B\}.$$

The *intersection* of A and B is the set

$$A \cap B = \{x : x \in A \text{ and } x \in B\}.$$

We say that A and B are *disjoint* if they have no elements in common. The *complement* of A in B is the set

$$B \setminus A = \{x \in B : x \notin A\}.$$

Sometimes $B \setminus A$ is written as $B - A$, or as A^c if B is understood.

1.18. Remark. In mathematics, the conjunction *or* (as in the definition of the union $A \cup B$) is always understood in the inclusive sense: 'p or q' always means 'p or q *or both*'. If we want the exclusive *or*, then we must say so explicitly by adding the phrase 'but not both'.

1.19. Remark. The operations on sets in Definition 1.17 satisfy various identities reminiscent of the laws of arithmetic. There are the associative laws

$$A \cup (B \cup C) = (A \cup B) \cup C, \qquad A \cap (B \cap C) = (A \cap B) \cap C,$$

the commutative laws

$$A \cup B = B \cup A, \qquad A \cap B = B \cap A,$$

the distributive laws

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C), \qquad A \cup (B \cap C) = (A \cup B) \cap (A \cup C),$$

and *De Morgan's laws*

$$A \setminus (B \cup C) = (A \setminus B) \cap (A \setminus C), \qquad A \setminus (B \cap C) = (A \setminus B) \cup (A \setminus C).$$

Let us prove the second De Morgan's law. There are two implications to be proved: first, the implication that if $x \in A \setminus (B \cap C)$, then $x \in (A \setminus B) \cup (A \setminus C)$, and second, the converse implication. So suppose that $x \in A \setminus (B \cap C)$. This means that $x \in A$ but $x \notin B \cap C$. Now $x \notin B \cap C$ means that $x \notin B$ *or* $x \notin C$, so we conclude that either $x \in A$ and $x \notin B$, or $x \in A$ and $x \notin C$ (*either ... or* is still the inclusive *or*). Hence $x \in A \setminus B$ or $x \in A \setminus C$, that is, $x \in (A \setminus B) \cup (A \setminus C)$. We leave the converse implication to you.

Note that this proof required three things:

- knowing how to prove that two sets are equal,
- unravelling the definitions of the sets $A \setminus (B \cap C)$ and $(A \setminus B) \cup (A \setminus C)$,
- being able to *negate* the statement $x \in B \cap C$, that is, realising that $x \notin B \cap C$ means that $x \notin B$ or $x \notin C$.

1.20. Definition. The *empty set* is *the* set with no elements, denoted $\varnothing$.

1.21. Remark. To say that A is a subset of B is to say that if $x \in A$, then $x \in B$. Hence, to say that A is *not* a subset of B is to say that there is $x \in A$ with $x \notin B$. It follows that the empty set $\varnothing$ is a subset of *every* set B. Otherwise, there would be an element $x \in \varnothing$ with $x \notin B$, but $\varnothing$ has no elements at all.

Exercise 1.6. Prove that if $A \subset B$, then $A \setminus B = \varnothing$.

We can take unions and intersections not just of two sets, but of arbitrary collections of sets.

1.22. Definition. Let $(A_i)_{i\in I}$ be a *family* of sets, that is, we have a set I (called an *index set*), and associated to every $i \in I$, we have a set called A_i. The *union* of the family is the set

$$\bigcup_{i\in I} A_i = \{x : x \in A_i \text{ for some } i \in I\}.$$

The *intersection* of the family is the set

$$\bigcap_{i\in I} A_i = \{x : x \in A_i \text{ for all } i \in I\}.$$

1.23. Example. Define a family $(A_n)_{n\in\mathbb{N}}$ of sets by setting $A_1 = \mathbb{N}$, $A_2 = \{2, 3, 4, \dots\}$, $A_3 = \{3, 4, 5, \dots\}$, and so on, that is, $A_n = \{n, n+1, n+2, \dots\}$ for each $n \in \mathbb{N}$. Then $A_1 \supset A_2 \supset A_3 \supset \cdots$, so we have

$$\bigcup_{n\in\mathbb{N}} A_n = A_1 \cup A_2 \cup A_3 \cup \cdots = A_1 = \mathbb{N}.$$

Also,

$$\bigcap_{n\in\mathbb{N}} A_n = A_1 \cap A_2 \cap A_3 \cap \cdots = \varnothing,$$

because there is no natural number that belongs to A_n for all $n \in \mathbb{N}$. Indeed, if $k \in A_1 = \mathbb{N}$, then $k \notin A_{k+1}$.

1.24. Definition. The *product* of sets A and B, denoted $A \times B$, is the set of all *ordered pairs* (a, b) with $a \in A$ and $b \in B$.

What is an ordered pair, you may ask. All you need to know is that $(a_1, b_1) = (a_2, b_2)$ if and only if $a_1 = a_2$ and $b_1 = b_2$. But you may be interested to also know that we do not need to take an ordered pair as a new fundamental notion. If we define (a, b) to be the set $\{\{a\}, \{a, b\}\}$, then we can prove that $(a_1, b_1) = (a_2, b_2)$ if and only if $a_1 = a_2$ and $b_1 = b_2$.

It is unfortunate that the same notation is used for an ordered pair and an open interval, but the intended meaning should always be clear from the context.

1.3. Functions

1.25. Definition. A *function* (or a *map* or a *mapping*—these are synonyms) f consists of three things:

- a set A called the *source* or *domain* of f,
- a set B called the *target* or *codomain* of f,
- a *rule* that assigns to each element x of A a unique element of B. This element is called the *image* of x by f or the *value* of f at x, and denoted $f(x)$.

We write $f : A \to B$ to indicate that f is a function with source A and target B, that is, a function *from A to B*.

1.26. Remark. Note that the source and the target of the function must be specified for the function to be well defined. Also, the rule does not have to be a formula. Any unambiguous description will do.

1.27. Definition. The *identity function* of a set A is the function $\mathrm{id}_A : A \to A$ with $\mathrm{id}_A(x) = x$ for all $x \in A$.

1.28. Definition. Let $f : A \to B'$ and $g : B \to C$ be functions such that $B' \subset B$. The *composition* of f and g is the function $g \circ f : A \to C$ with $(g \circ f)(x) = g(f(x))$ for all $x \in A$ ('first apply f, then g').

1.29. Definition. Let $f : A \to B$ be a function. The *image* by f of a subset $C \subset A$ is the subset

$$f(C) = \{f(x) : x \in C\}$$

of B. The *image* or *range* of f is the set $f(A)$. The *preimage* or *inverse image* by f of a subset $D \subset B$ is the subset

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

of A, that is, the set of elements of A that f maps into D. If D consists of only one element, say $D = \{y\}$ for some $y \in B$, then, for simplicity, we write $f^{-1}(y)$ for $f^{-1}(\{y\})$, and call $f^{-1}(y)$ the *fibre* of f over y.

1.30. Example. Assuming for the purposes of this example that we know about the real numbers, consider the function $f : \mathbb{R} \to \mathbb{R}$ defined by the formula $f(x) = x^2$. Instead of $f(x) = x^2$, we can write $f : x \mapsto x^2$ (the arrow $\mapsto$ is read 'maps to'). The range of f consists of all the nonnegative real numbers, that is, $f(\mathbb{R}) = [0, \infty)$. We have

$$f^{-1}(0) = \{0\}, \qquad f^{-1}(1) = \{1, -1\}, \qquad f^{-1}(\{1, 4\}) = \{1, -1, 2, -2\}.$$

The function $g : \mathbb{R} \to [0, \infty)$, $x \mapsto x^2$, is not the same function as f because its target is different. And the function $h : [0, \infty) \to [0, \infty)$, $x \mapsto x^2$, is different still, because its source is different. All three functions are defined by the same formula and have the same range $[0, \infty)$.

Images and preimages interact with unions, intersections, and complements to a certain extent. Note that preimages are better behaved than images.

1.31. Theorem. Let $f : A \to B$ be a function. For subsets $K, L \subset A$ and $M, N \subset B$, the following hold.

(1) $f(K \cup L) = f(K) \cup f(L)$.
(2) $f^{-1}(M \cup N) = f^{-1}(M) \cup f^{-1}(N)$.
(3) $f^{-1}(M \cap N) = f^{-1}(M) \cap f^{-1}(N)$.
(4) $f^{-1}(M \setminus N) = f^{-1}(M) \setminus f^{-1}(N)$.

Proof. We shall prove (4) and leave the other parts as an exercise. Normally we prove the equality of two sets as two separate implications, but here things are simple enough that we can prove both implications at the same time. Namely, we have $x \in f^{-1}(M \setminus N)$ if and only if $f(x) \in M \setminus N$ if and only if $f(x) \in M$ and $f(x) \notin N$ if and only if $x \in f^{-1}(M)$ and $x \notin f^{-1}(N)$ if and only if $x \in f^{-1}(M) \setminus f^{-1}(N)$. □

Exercise 1.7. Finish the proof of Theorem 1.31.

1.32. Remark. It is not true in general that $f(K \cap L) = f(K) \cap f(L)$ or $f(K \setminus L) = f(K) \setminus f(L)$. For example, take f as in Example 1.30, $K = \{1\}$, and $L = \{-1\}$. Then $f(K \cap L) = f(\varnothing) = \varnothing$, but $f(K) \cap f(L) = \{1\} \cap \{1\} = \{1\}$. Also, $f(K \setminus L) = f(\{1\}) = \{1\}$, but $f(K) \setminus f(L) = \{1\} \setminus \{1\} = \varnothing$.

1.33. Definition. A function $f : A \to B$ is called:

- *injective* (or *one-to-one*) if it takes distinct elements to distinct elements, that is, if $x, y \in A$ and $f(x) = f(y)$, then $x = y$;
- *surjective* (or *onto*) if $f(A) = B$, that is, every element of B is the image by f of some element of A;
- *bijective* if f is both injective and surjective.

An injective function is also called an *injection*, a surjective function is called a *surjection*, and a bijective function is called a *bijection*.

1.34. Remark. Note that a function $f : A \to B$ is:

- injective if and only if the fibre $f^{-1}(y)$ contains *at most* one element for every $y \in B$,
- surjective if and only if the fibre $f^{-1}(y)$ contains *at least* one element for every $y \in B$,
- bijective if and only if the fibre $f^{-1}(y)$ contains *precisely* one element for every $y \in B$.

Exercise 1.8. Of the functions f, g, and h in Example 1.30, show that only h is injective and only g and h are surjective.

Exercise 1.9. Returning to Theorem 1.31 and Remark 1.32, show that a function $f : A \to B$ is injective if and only if $f(K \cap L) = f(K) \cap f(L)$ for all subsets $K, L \subset A$. Does a similar result hold for complements?

1.35. Definition. Let $f : A \to B$ be a bijection. The *inverse function* of f is the function $f^{-1} : B \to A$ defined by letting $f^{-1}(y)$ for $y \in B$ be the unique element x of A for which $f(x) = y$. Thus, for $x \in A$ and $y \in B$,

$$f^{-1}(y) = x \text{ if and only if } f(x) = y.$$

In other words,

$$f^{-1} \circ f = \mathrm{id}_A, \qquad f \circ f^{-1} = \mathrm{id}_B.$$

1.36. Remark. Sometimes we speak of the inverse of a function $f : A \to B$ that is merely injective. By this we mean the inverse of the bijection obtained from f by replacing its target B by its image $f(A)$. In other words, the inverse of the injection $f : A \to B$ is the function $f^{-1} : f(A) \to A$ defined by letting $f^{-1}(y)$ for $y \in f(A)$ be the unique $x \in A$ for which $f(x) = y$. Thus, for $x \in A$ and $y \in f(A)$, $f^{-1}(y) = x$ if and only if $f(x) = y$.

1.37. Example. The function h in Example 1.30 is bijective, so it has an inverse function $h^{-1} : [0, \infty) \to [0, \infty)$. For $x \in [0, \infty)$, $h^{-1}(x)$ is the unique nonnegative square root of x.

1.38. Definition. The *graph* of a function $f : A \to B$ is the subset $\{(a, f(a)) : a \in A\}$ of $A \times B$.

1.39. Remark. The graph G of a function $f : A \to B$ has the property that for every $a \in A$, there is a unique $b \in B$ such that $(a, b) \in G$ (namely $b = f(a)$). We can rigorously define a *rule* as in Definition 1.25 to be a subset of $A \times B$ with this property.

More exercises

1.10. Prove that the product of a nonzero rational number and an irrational number is irrational.

1.11. Let $x_1 = 1$, and for each $n \in \mathbb{N}$, let $x_{n+1} = \frac{2}{3}x_n + 1$. Prove by induction that $x_n < 3$ for all $n \in \mathbb{N}$.

1.12. Prove by induction that for every $n \in \mathbb{N}$,

$$\sum_{k=1}^{n} k^2 = \tfrac{1}{6}n(n+1)(2n+1).$$

1.13. Using only the field axioms A1–A5, give a careful proof of the following two *cancellation laws*.

(a) If $a + c = b + c$, then $a = b$.

(b) If $ac = bc$ and $c \neq 0$, then $a = b$.

1.14. (a) In how many ways can a sum of four terms $a+b+c+d$ be bracketed? Use the associative law A1 to show that the different bracketings all give the same sum.

(b) More generally, the associative law implies that the different ways to bracket a sum of three or more terms give the same result, so it is unambiguous to write $a_1 + a_2 + \cdots + a_n$ without brackets, and similarly for products. The commutative law A2 implies that changing the order of the terms of a sum or the factors in a product does not affect the result. Prove these statements.

1.15. Use the triangle inequality $|x + y| \leq |x| + |y|$ to prove that

$$||x| - |y|| \leq |x - y|$$

for all elements x and y of an ordered field, say $x, y \in \mathbb{Q}$.

1.16. (a) Prove that every nonempty finite subset A of an ordered field has a smallest element. *Hint.* Let $a_1 \in A$. If a_1 is not the smallest element of A, then there is $a_2 \in A$ with $a_2 < a_1$. If a_2 is not the smallest element of A, then ....

(b) Deduce that every nonempty finite subset of an ordered field has a largest element.

1.17. For this exercise you need to know what the complex numbers are. The complex numbers form a field $\mathbb{C}$ satisfying the field axioms A1–A5. Show that there is no way to turn $\mathbb{C}$ into an ordered field, that is, there is no order relation on $\mathbb{C}$ that makes axioms A6–A9 true. *Hint.* Assume that A6–A9 hold and try to derive a contradiction.

1.18. Prove the following statement, or disprove it by a counterexample. If A, B, and C are sets, then $A \cap (B \cup C) = (A \cap B) \cup C$.

1.19. (a) Suppose we have a set A_n for each $n \in \mathbb{N}$. Fill in the blanks with two words so as to get a true statement. Justify your answer.

$$x \in \bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_n \text{ if and only if } x \in A_n \text{ for } \_\_\_\_ \ \_\_\_\_ \ n \in \mathbb{N}$$

(b) Fill in the blanks with four words so as to get a true statement.

$$x \in \bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} A_n \text{ if and only if } x \in A_n \text{ for } \_\_\_\_ \ \_\_\_\_ \ \_\_\_\_ \ \_\_\_\_ \ n \in \mathbb{N}$$

1.20. Consider the function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^2 - 2$. What is the image $f([-2, 3])$? Justify your answer and express it in interval notation.

1.21. Let A be a set and $f, g, h : A \to A$ be functions such that $f \circ g = h \circ f = \mathrm{id}_A$. Show that $g = h$.

1.22. Let $t : \mathbb{N} \to \mathbb{N}$, $n \mapsto n + 1$. How many surjections $g : \mathbb{N} \to \mathbb{N}$ with $g \circ t = t \circ g$ are there?

1.23. Prove the following *cancellation laws* for functions.

(a) Let $f : A \to B$ and $g, g' : B \to C$ be functions. Show that if $g \circ f = g' \circ f$ and f is surjective, then $g = g'$.

(b) Let $g, g' : B \to C$ and $h : C \to D$ be functions. Show that if $h \circ g = h \circ g'$ and h is injective, then $g = g'$.

(c) What if f is not surjective? What if h is not injective?

1.24. Let X and Y be sets and $f : X \to Y$ be a function.

(a) Prove that if $A \subset X$, then $A \subset f^{-1}(f(A))$. Give an example for which $f^{-1}(f(A)) \neq A$.

(b) Prove that if $B \subset Y$, then $f(f^{-1}(B)) \subset B$. Give an example for which $f(f^{-1}(B)) \neq B$.

Chapter 2

The real numbers

2.1. The complete ordered field of real numbers

The real numbers form an ordered field $\mathbb{R}$ containing the rationals with an additional property called *completeness* that the rationals do not satisfy. We need some preliminary definitions to be able to say what completeness means.

2.1. Definition. An *upper bound* for a subset $A \subset \mathbb{R}$ is an element $b \in \mathbb{R}$ such that $a \leq b$ for all $a \in A$. If A has an upper bound, then A is said to be *bounded above*.

A *lower bound* for a subset $A \subset \mathbb{R}$ is an element $b \in \mathbb{R}$ such that $b \leq a$ for all $a \in A$. If A has a lower bound, then A is said to be *bounded below*.

If A is bounded above and bounded below, then A is said to be *bounded*.

2.2. Example. Consider the interval $[0,1] = \{x \in \mathbb{R} : 0 \leq x \leq 1\}$. It is bounded above, for example by the upper bound 1. The upper bounds for $[0,1]$ are precisely the numbers b with $b \geq 1$. Thus 1 is the smallest upper bound for $[0,1]$, and it is of course also the largest element of $[0,1]$.

Now consider the interval $(0,1) = \{x \in \mathbb{R} : 0 < x < 1\}$, also bounded above, for example by 1. It has the *same* upper bounds as $[0,1]$. Namely, if $b \geq 1$ and $x \in (0,1)$, then $x < 1 \leq b$, so b is an upper bound for $(0,1)$. Conversely, if b is an upper bound for $(0,1)$, so in particular $b \geq \frac{1}{2} \in (0,1)$, then we must have $b \geq 1$, for if $b < 1$, then $x = \frac{1}{2}(b+1) \in (0,1)$ and $b < x$, so b would not be an upper bound for $(0,1)$. Thus 1 is also the smallest upper bound for $(0,1)$. Note that $(0,1)$ has no largest element.

Similarly, both $[0,1]$ and $(0,1)$ are bounded below and the largest lower bound of both is 0.

2.3. Definition. If $A \subset \mathbb{R}$ is bounded above and A has an upper bound s that is smaller than every other upper bound for A, then s is called the *supremum* (plural: *suprema*) or the *least upper bound* of A, denoted $\sup A$.

If $A \subset \mathbb{R}$ is bounded below and A has a lower bound t that is larger than every other lower bound for A, then t is called the *infimum* (plural: *infima*) or the *greatest lower bound* of A, denoted $\inf A$.

2.4. Example. By Example 2.2, $\sup[0,1] = \sup(0,1) = 1$ and $\inf[0,1] = \inf(0,1) = 0$.

2.5. Remark. Note that if $A \subset \mathbb{R}$ has a largest element (*maximum*), then the maximum is the supremum of A. By Example 2.2, $(0,1)$ has a supremum without having a maximum. When there is no maximum, we can think of the supremum as the next-best thing.

Similarly, if $A \subset \mathbb{R}$ has a smallest element (*minimum*), then the minimum is the infimum of A, but A can have an infimum without having a minimum.

The following lemma provides a handy criterion for an upper bound to be the supremum.

2.6. Lemma. Let s be an upper bound for a subset A of $\mathbb{R}$. Then $s = \sup A$ if and only if for every $\epsilon > 0$, there is $a \in A$ with $s - \epsilon < a$.

Proof. The lemma says precisely that $s = \sup A$ if and only if no smaller number is an upper bound for A. □

2.7. Example. Consider the bounded set $A = \{\frac{1}{n} : n \in \mathbb{N}\} = \{1, \frac{1}{2}, \frac{1}{3}, \dots\}$. Clearly, 1 is the largest element of A, so $\sup A = 1$, but A has no smallest element. It looks like the infimum of A should be 0. By the analogue of Lemma 2.6 for infima (which you can formulate for yourself), we have $\inf A = 0$ if and only if for every $\epsilon > 0$, there is $n \in \mathbb{N}$ with $\frac{1}{n} < \epsilon$. This is the so-called *Archimedean property* of $\mathbb{R}$, which we shall prove as a consequence of the completeness of $\mathbb{R}$ (Theorem 2.8).

We now come to the axiom that, in addition to the axioms A1–A9 for an ordered field, is satisfied by the real numbers.

Axiom of completeness. Every nonempty set of real numbers that is bounded above has a least upper bound.

We will soon see that the rational numbers do not satisfy the axiom of completeness (Remark 2.10).

The following exercise shows that the existence of suprema for nonempty sets that are bounded above implies the existence of infima for nonempty

sets that are bounded below. Thus the latter does not have to be taken as a separate axiom.

Exercise 2.1. (a) For $A \subset \mathbb{R}$, let $-A = \{-x : x \in A\}$. Show that if A is bounded below, then $-A$ is bounded above.

(b) Use (a) and the axiom of completeness to show that if $A \subset \mathbb{R}$ is nonempty and bounded below, then A has an infimum and $\inf A = -\sup(-A)$.

Here is where our course really begins. We take as our starting point a complete ordered field denoted $\mathbb{R}$. We call its elements real numbers. We now proceed to a careful development of the various topics of calculus, assuming only the ten axioms that describe the structure of $\mathbb{R}$. We will try not to rely on any preconceptions about the real numbers. For us, a real number is nothing but an element of a complete ordered field.

We do not need the following facts, and to explain them is beyond the scope of the course, but it is certainly of interest to know that:

- $\mathbb{R}$ 'exists' in the sense that it can be constructed from the rationals. We do not need to assume any new fundamental notions to produce from the rationals a complete ordered field containing them. There are several ways to do this. The two most popular methods use so-called *Dedekind cuts* and *Cauchy sequences*.
- $\mathbb{R}$ is unique, in the sense that if F is another complete ordered field, then there is a bijection $\phi : \mathbb{R} \to F$ that is an *isomorphism of ordered fields* in the sense that:
  - ϕ preserves addition: $\phi(x + y) = \phi(x) + \phi(y)$ for all $x, y \in \mathbb{R}$,
  - ϕ preserves multiplication: $\phi(xy) = \phi(x)\phi(y)$ for all $x, y \in \mathbb{R}$,
  - ϕ preserves the identity elements: $\phi(0) = 0$, $\phi(1) = 1$,
  - and ϕ preserves order: if $x < y$ in $\mathbb{R}$, then $\phi(x) < \phi(y)$ in F.

  In other words, any two complete ordered fields are 'the same' as ordered fields.

2.2. Consequences of completeness

The axiom of completeness has massive consequences, some of which we explore in the remainder of this chapter. First come three versions of the Archimedean property mentioned already in Example 2.7.

2.8. Theorem (Archimedean property)**.** (1) $\mathbb{N}$ is not bounded above.

(2) For every $y \in \mathbb{R}$, $y > 0$, there is $n \in \mathbb{N}$ such that $\frac{1}{n} < y$.

(3) If $x, y \in \mathbb{R}$, $y > 0$, then there is $n \in \mathbb{N}$ such that $ny > x$.

Proof. We prove (1) and leave (2) and (3) as exercises. Suppose $\mathbb{N}$ was bounded above. Then $\mathbb{N}$ would have a supremum s by the axiom of completeness. If $n \in \mathbb{N}$, then also $n+1 \in \mathbb{N}$, so $n+1 \le s$, so $n \le s-1$. Thus $s-1$ is an upper bound for $\mathbb{N}$, contradicting s being the smallest one. □

Exercise 2.2. Finish the proof of Theorem 2.8.

2.9. Theorem. There is $s \in \mathbb{R}$ with $s^2 = 2$.

Proof. We shall obtain s as the supremum of a suitable set, namely the set $A = \{x \in \mathbb{R} : x^2 < 2\}$. Since $1 \in A$, $A \neq \varnothing$. Also, A is bounded above, for example by 2, because if $x > 2$, then $x^2 > 4 > 2$, so $x \notin A$. Hence A has a supremum s by the axiom of completeness. We need to show that $s^2 = 2$.

Suppose $s^2 > 2$. Choose $\epsilon \in (0, s)$ with $\epsilon < \dfrac{s^2-2}{2s}$. Then

$$(s-\epsilon)^2 = s^2 - 2s\epsilon + \epsilon^2 > s^2 - 2s\epsilon > 2.$$

We claim that $s - \epsilon$ is an upper bound for A. If not, there is $x \in A$ with $0 < s - \epsilon < x$, but then $(s-\epsilon)^2 < x^2 < 2$. Thus $s^2 > 2$ contradicts s being the least upper bound of A.

Finally, suppose $s^2 < 2$. Choose $\epsilon \in (0, 1)$ with $\epsilon < \dfrac{2-s^2}{2s+1}$. Then

$$(s+\epsilon)^2 = s^2 + 2s\epsilon + \epsilon^2 < s^2 + (2s+1)\epsilon < 2,$$

so $s + \epsilon \in A$, contradicting s being an upper bound for A. □

2.10. Remark. The only property of $\mathbb{R}$, in addition to the axioms for an ordered field, that was used to prove Theorem 2.9 was completeness. By Theorem 1.15, Theorem 2.9 fails for $\mathbb{Q}$. We conclude that $\mathbb{Q}$ does not satisfy the axiom of completeness. It also follows that the number s in Theorem 2.9 is irrational. In particular, *irrational numbers exist.*

2.11. Remark. The proof of Theorem 2.9 can be generalised to show that if $x \in \mathbb{R}$, $x > 0$, and $n \in \mathbb{N}$, then x has a positive n^{th} root, denoted $\sqrt[n]{x}$ or $x^{1/n}$, which is unique because the function $(0,\infty) \to (0,\infty)$, $x \mapsto x^n$, is strictly increasing and hence injective. Later, we will be able to prove the existence of n^{th} roots much more easily using the intermediate value theorem (Exercise 5.17).

2.12. Definition. A subset D of $\mathbb{R}$ is *dense* in $\mathbb{R}$ if every nonempty open interval intersects D. In other words, for every $a, b \in \mathbb{R}$ with $a < b$, there is $x \in D$ with $a < x < b$.

2.13. Theorem. $\mathbb{Q}$ is dense in $\mathbb{R}$.

Proof. Suppose for convenience that $0 \le a < b$ (can you reduce the general case to this special case?). By the Archimedean property there is $n \in \mathbb{N}$ with $b - a > \frac{1}{n}$. Consider the set $\{k \in \mathbb{N} : k > na\}$. Again by the Archimedean property, this set is nonempty, so it has a smallest element m by the well-ordering property of $\mathbb{N}$ (see Exercise 2.17). Then $m - 1 \le na < m$, so $m \le na + 1 < nb$ by the choice of n. Hence $a < \frac{m}{n} < b$. □

Exercise 2.3. We have just learned that every nonempty open interval $I \subset \mathbb{R}$ contains a rational number. Prove that in fact I contains infinitely many rational numbers.

Exercise 2.4. Prove that the set $\mathbb{R} \setminus \mathbb{Q}$ of irrationals is dense in $\mathbb{R}$. *Hint.* Since $\mathbb{Q}$ is dense, if $a < b$ in $\mathbb{R}$, then the interval $(a - \sqrt{2}, b - \sqrt{2})$ contains a rational.

Finally, this consequence of completeness looks technical, but turns out to be useful.

2.14. Theorem (nested interval property)**.** Let $I_1 \supset I_2 \supset I_3 \supset \cdots$ be a decreasing sequence of closed, bounded, and nonempty intervals. Then

$$\bigcap_{n=1}^{\infty} I_n \neq \varnothing.$$

Proof. Say $I_n = [a_n, b_n]$ with $a_n \le b_n$. Then $a_1 \le a_2 \le \cdots \le b_2 \le b_1$. Let $A = \{a_1, a_2, \dots\}$. Then A is nonempty and bounded above, for example by each b_n. By the axiom of completeness, A has a supremum c. Since c is an upper bound for A, $a_n \le c$ for all $n \in \mathbb{N}$. Also, for each $n \in \mathbb{N}$, since b_n is an upper bound for A, $c \le b_n$. This shows that $c \in I_n$ for all $n \in \mathbb{N}$, so the intersection is not empty. □

2.15. Remark. The nested interval property fails for open intervals. For example, $\bigcap_{n=1}^{\infty} (0, \frac{1}{n}) = \varnothing$ (this is nothing but the Archimedean property). It also fails for unbounded intervals. For example, $\bigcap_{n=1}^{\infty} [n, \infty) = \varnothing$ (this is also nothing but the Archimedean property).

2.3. Countable and uncountable sets

We can establish that two finite sets have the same size, without knowing anything about numbers or counting, by setting up a bijective correspondence between their elements. This simple idea can be applied to infinite sets too.

2.16. Definition. Sets A and B are *equinumerous*, or have the same *cardinality*, denoted $A \sim B$, if there is a bijection $A \to B$.

Exercise 2.5. Let A, B, and C be sets.

(a) Show that $A \sim A$. (We say that the relation $\sim$ is *reflexive*.)

(b) Show that if $A \sim B$, then $B \sim A$ ($\sim$ is *symmetric*).

(c) Show that if $A \sim B$ and $B \sim C$, then $A \sim C$ ($\sim$ is *transitive*).

2.17. Example. (a) The set $S = \{1, 4, 9, \dots\}$ of squares is equinumerous to $\mathbb{N}$. An example of a bijection $\mathbb{N} \to S$ is the function $n \mapsto n^2$. So an infinite set can be equinumerous to a proper subset of itself. This observation was made by Galileo Galilei in the early seventeenth century. Richard Dedekind turned it into an elegant definition of an infinite set: a set is infinite if and only if it is equinumerous to some proper subset of itself.

(b) Similarly, $\mathbb{Z}$ is equinumerous to $\mathbb{N}$. We can set up a bijection $\mathbb{N} \to \mathbb{Z}$, for example by mapping $1, 2, 3, 4, 5, \dots$ to $0, 1, -1, 2, -2, \dots$. (Recall that a function does not have to be defined by a formula: we only need to describe it unambiguously.)

2.18. Definition. A set A is *countably infinite* if $A \sim \mathbb{N}$. A set is *countable* if it is finite or countably infinite. A set that is not countable is called *uncountable*.

By Example 2.17, the squares and the integers are countable.

2.19. Theorem. (1) The set $\mathbb{Q}$ of all rational numbers is countable.

(2) The set $\mathbb{R}$ of all real numbers is uncountable.

Georg Cantor, the founder of set theory, discovered the uncountability of the reals in December 1873. He concluded that, since $\mathbb{Q}$ and $\mathbb{R}$ are not equinumerous, there must be irrational numbers. This is a pure existence proof. It does not exhibit a single irrational or tell us how to find one. Objections were raised to such arguments at the time, but their validity is now firmly accepted.

Proof. (1) First list the integers: $0, 1, -1, 2, -2, \dots$. Divide the integers by 2, discarding those quotients that are integers: $\frac{1}{2}, -\frac{1}{2}, \frac{3}{2}, -\frac{3}{2}, \frac{5}{2}, \dots$. Divide the integers by 3, discarding those quotients that have already appeared: $\frac{1}{3}, -\frac{1}{3}, \frac{2}{3}, -\frac{2}{3}, \frac{4}{3}, \dots$. Continuing, we obtain a list of lists such that each rational number appears precisely once on precisely one of the lists. Denote the n^{th} number on the m^{th} list by a_{mn}. Now we simply arrange this two-dimensional array into a sequence in which every rational number appears precisely once, for example like this: $a_{11}, a_{12}, a_{21}, a_{13}, a_{22}, a_{31}, a_{14}, \dots$.

(2) Let A be a countably infinite subset of $\mathbb{R}$. We will show that $A \neq \mathbb{R}$. Let $f : \mathbb{N} \to A$ be a bijection. Find an interval $I_1 = [a_1, b_1]$ with $a_1 < b_1$ such that $f(1) \notin I_1$ (we could for instance take $a_1 = f(1) + 1$ and $b_1 = f(1) + 2$).

Next find an interval $I_2 = [a_2, b_2]$ with $a_2 < b_2$ such that $I_2 \subset I_1$ and $f(2) \notin I_2$.

Continuing, we obtain a decreasing sequence $I_1 \supset I_2 \supset I_3 \supset \cdots$ of closed, bounded, and nonempty intervals such that $f(n) \notin I_n$ for each $n \in \mathbb{N}$. By the nested interval property (Theorem 2.14), the intersection $\bigcap_{n=1}^{\infty} I_n$ is not empty, say $c \in \bigcap_{n=1}^{\infty} I_n$. Then, for each $n \in \mathbb{N}$, $c \in I_n$, so $c \neq f(n)$. Thus $c \in \mathbb{R} \setminus A$. This shows that $A \neq \mathbb{R}$, so $\mathbb{R}$ is not countable. □

You may have seen a proof of the uncountability of $\mathbb{R}$ using decimal expansions. The proof we have just given is much more elementary. It shows how uncountability follows quite directly from completeness, via the nested interval property.

More exercises

2.6. Find the suprema and infima of the following sets. You do not have to give formal proofs of your answers, but do give a brief explanation, perhaps including a picture.

(a) $\{n/(n+1) : n \in \mathbb{N}\}$

(b) $\{x/(x+1) : x \in \mathbb{R}, x > 0\}$

(c) $\{1/(3n^2+5) : n \in \mathbb{N}\}$

(d) $\{n/m : m, n \in \mathbb{N}, m + n \leq 10\}$

(e) $\{2x - x^2 : 0 < x < 2\}$

2.7. Let $A \subset \mathbb{R}$ be nonempty and bounded above. Let $c \in \mathbb{R}$. Show that $B = \{x + c : x \in A\}$ is bounded above and that $\sup B = \sup A + c$.

2.8. Suppose $A, B \subset \mathbb{R}$ are nonempty and bounded above, with $B \subset A$. Show that $\sup B \leq \sup A$.

2.9. (a) Let $A, B \subset \mathbb{R}$. If $\sup A < \sup B$, prove that B contains an upper bound for A.

(b) If $\sup A \leq \sup B$, is it true that B contains an upper bound for A?

2.10. Let A and B be nonempty subsets of $\mathbb{R}$, bounded above, such that for every $a \in A$ there is $b \in B$ with $a \leq b$. Show that $\sup A \leq \sup B$.

2.11. Let $A \subset \mathbb{R}$ be nonempty and bounded below. Let B be the set of lower bounds of A. Show that B is bounded above. What is $\sup B$?

2.12. (a) Let $I \subset \mathbb{R}$ have the property that if $x, z \in I$, $y \in \mathbb{R}$, and $x < y < z$, then $y \in I$. Prove that I is an interval in $\mathbb{R}$ (see Definition 1.10 and Remark 1.11).

(b) Show that (a) fails if $\mathbb{R}$ is replaced by $\mathbb{Q}$.

2.13. Prove that the bounded interval $(0, 1)$ and the unbounded interval $(1, \infty)$ are equinumerous.

2.14. Show that $\mathbb{R}$ and $\mathbb{R} \setminus \{0\}$ are equinumerous.

2.15. (a) Prove that the set of all functions $\mathbb{N} \to \{0, 1\}$ is uncountable. *Hint.* Let $f_1, f_2, f_3, \ldots$ be functions $\mathbb{N} \to \{0, 1\}$. Define $g : \mathbb{N} \to \{0, 1\}$ by $g(n) = 1 - f_n(n)$.

(b) Is the set of all functions $\{0, 1\} \to \mathbb{N}$ countable or uncountable?

2.16. A real number x is said to be *algebraic* if there are integers $a_0, \ldots, a_n$, not all zero, such that $a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0 = 0$. We say that x is *transcendental* if x is not algebraic.

(a) Show that every rational number, $\sqrt{2}$, and $\sqrt{2} + \sqrt{3}$ are algebraic.

(b) Show that the set of all algebraic numbers is countable. You may use the fact that a polynomial of degree n has at most n roots.

It follows that transcendental numbers exist. It is quite difficult, and beyond the scope of this course, to prove that any particular number is transcendental.

2.17. This exercise shows how we can introduce the natural numbers and prove the well-ordering property (stated in Theorem 1.1) from the axioms for an ordered field. (We have already used the well-ordering property in the proof of Theorem 2.13.) The natural numbers are the multiplicative identity 1 provided by axiom A4; the number 2, defined as 1+1; the number 3, defined as $2 + 1$; and so on. Since $0 < 1$ (Example 1.9), we have $1 < 2 < 3 < \cdots$ by axiom A8.

If you wonder what the phrase 'and so on' really means, you will appreciate the following rigorous definition. We say that a subset $A \subset \mathbb{R}$ is *inductive* if $1 \in A$ and, for every $x \in A$, $x + 1 \in A$. Examples of inductive sets are $\mathbb{R}$ itself, $[1, \infty)$, and $[2, \infty) \cup \{1\}$. Note that the intersection of any collection of inductive sets is an inductive set. We define the set $\mathbb{N}$ of natural numbers to be the smallest inductive set, that is, the intersection of all inductive subsets of $\mathbb{R}$. Since $[1, \infty)$ is inductive, $\mathbb{N} \subset [1, \infty)$, so 1 is the smallest natural number. Since $[2, \infty) \cup \{1\}$ is inductive, $\mathbb{N} \subset [2, \infty) \cup \{1\}$, so there is no natural number strictly between 1 and 2.

(a) Prove that for every $n \in \mathbb{N}$, $(n - 1, n + 1) \cap \mathbb{N} = \{n\}$. *Hint.* Show that the set of such n is inductive, so it must be all of $\mathbb{N}$.

(b) Prove that for every $m \in \mathbb{N}$, the set $\{n \in \mathbb{N} : n \leq m\}$ is finite.

(c) Prove that every nonempty subset S of $\mathbb{N}$ has a smallest element. *Hint.* Take $m \in S$. By (b), the set $\{n \in S : n \leq m\}$ is finite. Use Exercise 1.16.

Sequences

3.1. Convergent sequences

3.1. Definition. A *sequence* in a set A is a function $a : \mathbb{N} \to A$. We usually write a_n for $a(n)$, and write $(a_n)_{n\in\mathbb{N}}$ or simply (a_n) for a. We call a_n the n^{th} *term* of the sequence a.

Until we reach Chapter 8, we will only consider sequences of real numbers, so by a sequence we shall always mean a sequence in $\mathbb{R}$.

3.2. Definition. A sequence (a_n) in $\mathbb{R}$ *converges* to $b \in \mathbb{R}$ if for every $\epsilon > 0$, there is $N \in \mathbb{N}$ such that $|a_n - b| < \epsilon$ for all $n \geq N$. We call b the *limit* of (a_n) and write $b = \lim_{n\to\infty} a_n$ or $a_n \to b$.

A sequence that does not converge is called *divergent*.

Exercise 3.1. Show that $a_n \to b$ if and only if $|a_n - b| \to 0$.

3.3. Proposition. The limit of a convergent sequence is unique, that is, if (a_n) is a sequence such that $a_n \to b$ and $a_n \to c$, then $b = c$.

Proof. Let $\epsilon > 0$. There is $N_1 \in \mathbb{N}$ such that $|a_n - b| < \epsilon/2$ for all $n \geq N_1$. Also, there is $N_2 \in \mathbb{N}$ such that $|a_n - c| < \epsilon/2$ for all $n \geq N_2$. Hence, for $n \geq \max\{N_1, N_2\}$,

$$|b - c| \leq |a_n - b| + |a_n - c| < \epsilon.$$

We have shown that $|b-c| < \epsilon$ for every $\epsilon > 0$. We conclude that $|b-c| = 0$, that is, $b = c$. □

An equivalent but more geometric definition of a convergent sequence is obtained via the concept of a neighbourhood.

3.4. Definition. For $\epsilon > 0$, the *ϵ-neighbourhood* of $b \in \mathbb{R}$ is the open interval $(b-\epsilon, b+\epsilon)$. A *neighbourhood* of b is any subset of $\mathbb{R}$ that contains the ϵ-neighbourhood of b for some $\epsilon > 0$.

3.5. Remark. Note that $|a_n - b| < \epsilon$ means that $a_n \in (b-\epsilon, b+\epsilon)$. Thus Definition 3.2 can be reformulated as follows. A sequence (a_n) in $\mathbb{R}$ converges to $b \in \mathbb{R}$ if for every neighbourhood V of b, there is $N \in \mathbb{N}$ such that $a_n \in V$ for all $n \geq N$. In other words, each neighbourhood of b contains a_n for all but finitely many $n \in \mathbb{N}$.

Let us prove the uniqueness of the limit of a convergent sequence using neighbourhoods. We need the fundamental fact that distinct points in $\mathbb{R}$ have disjoint neighbourhoods. Namely, take $b, c \in \mathbb{R}$, $b \neq c$. Let $\epsilon = \frac{1}{2}|b-c| > 0$. Then $(b-\epsilon, b+\epsilon)$ and $(c-\epsilon, c+\epsilon)$ are disjoint.

Let (a_n) be a sequence such that $a_n \to b$ and $a_n \to c$. Let U be a neighbourhood of b and V be a neighbourhood of c. Since $a_n \to b$, U contains a_n for all but finitely many $n \in \mathbb{N}$. Similarly, V contains a_n for all but finitely many n. Hence $U \cap V$ contains a_n for all but finitely many n. In particular, $U \cap V$ is not empty.

We have shown that every neighbourhood of b intersects every neighbourhood of c. This implies that $b = c$.

3.6. Example. (a) Let us prove that $\lim\limits_{n\to\infty} \frac{1}{n} = 0$. Take $\epsilon > 0$. We need to show that there is $N \in \mathbb{N}$ such that $\frac{1}{n} = |\frac{1}{n} - 0| < \epsilon$ for all $n \geq N$ or, equivalently, $\frac{1}{N} < \epsilon$. This is nothing but the Archimedean property (Theorem 2.8).

(b) Prove that $\lim\limits_{n\to\infty} \dfrac{6n+5}{3n+2} = 2$. Let $\epsilon > 0$. Note that for $n \in \mathbb{N}$,

$$\left|\frac{6n+5}{3n+2} - 2\right| = \frac{1}{3n+2} < \epsilon$$

if and only if $3n+2 > \frac{1}{\epsilon}$, that is, $n > \frac{1}{3}(\frac{1}{\epsilon} - 2)$. By the Archimedean property, there is $N \in \mathbb{N}$ with $N > \frac{1}{3}(\frac{1}{\epsilon} - 2)$. Then, if $n \geq N$, we have $n > \frac{1}{3}(\frac{1}{\epsilon} - 2)$, so by the calculation above, $\left|\dfrac{6n+5}{3n+2} - 2\right| < \epsilon$. This shows that $\lim\limits_{n\to\infty} \dfrac{6n+5}{3n+2} = 2$.

(c) The sequence $0, 1, 0, 1, 0, 1, \ldots$ diverges. Namely, it does not have limit 0 because infinitely many of its terms lie outside the neighbourhood $(-1, 1)$ of 0. It does not have limit 1 because infinitely many of its terms lie outside the neighbourhood $(0, 2)$ of 1. And, for $b \neq 0, 1$, the sequence does not converge to b because all of of its terms lie outside the neighbourhood $(b-\epsilon, b+\epsilon)$ of b, where $\epsilon = \min\{|b|, |b-1|\} > 0$.

Exercise 3.2. (a) Deduce from the Archimedean property (Theorem 2.8) that the set $\{2^n : n \in \mathbb{N}\}$ is unbounded above, and conclude that $2^{-n} = 1/2^n \to 0$ as $n \to \infty$. *Hint.* Prove that $n < 2^n$ for all $n \in \mathbb{N}$.

(b) Prove along the lines of the proof of the Archimedean property that if $a \in \mathbb{R}$, $a > 1$, then $a^{-n} = 1/a^n \to 0$ as $n \to \infty$.

3.7. Definition. A sequence (a_n) is *bounded* if the set $\{a_n : n \in \mathbb{N}\}$ of its terms is bounded (Definition 2.1). Equivalently, there is $M > 0$ such that $|a_n| \leq M$ for all $n \in \mathbb{N}$. We say that (a_n) is *bounded above* if $\{a_n : n \in \mathbb{N}\}$ is bounded above, and that (a_n) is *bounded below* if $\{a_n : n \in \mathbb{N}\}$ is bounded below.

3.8. Theorem. A convergent sequence is bounded. In other words, an unbounded sequence diverges.

Proof. Say $a_n \to b$. Find $N \in \mathbb{N}$ such that $|a_n - b| < 1$, so $|a_n| < 1 + |b|$, for all $n \geq N$. Then

$$|a_n| \leq \max\{|a_1|, \ldots, |a_{N-1}|, 1 + |b|\}$$

for all $n \in \mathbb{N}$. □

3.9. Remark. The converse of Theorem 3.8 fails: a bounded sequence need not converge, and a divergent sequence need not be unbounded (Example 3.6 (c)).

3.2. New limits from old

3.10. Theorem (squeeze theorem)**.** If $a_n \to s$, $c_n \to s$, and $a_n \leq b_n \leq c_n$ for all but finitely many $n \in \mathbb{N}$, then $b_n \to s$.

Proof. Let $\epsilon > 0$ and $I = (s - \epsilon, s + \epsilon)$. By assumption, for n sufficiently large, $a_n \in I$. Also, for n sufficiently large, $c_n \in I$. Hence, since I is an interval and b_n lies between a_n and c_n for n sufficiently large, we have $b_n \in I$ for n sufficiently large. This shows that $b_n \to s$. □

3.11. Theorem (algebraic limit theorem)**.** If $a_n \to a$ and $b_n \to b$, then:

(1) $ca_n \to ca$ for all $c \in \mathbb{R}$.

(2) $a_n + b_n \to a + b$.

(3) $a_n b_n \to ab$.

(4) $a_n/b_n \to a/b$ if $b_n \neq 0$ for all $n \in \mathbb{N}$ and $b \neq 0$.

(5) $|a_n| \to |a|$.

Proof. We prove (2), (4), and (5), and leave (1) and (3) as exercises.

(2) Let $\epsilon > 0$. Find N_1 such that $|a_n - a| < \epsilon/2$ for $n \geq N_1$. Find N_2 such that $|b_n - b| < \epsilon/2$ for $n \geq N_2$. Then, for $n \geq \max\{N_1, N_2\}$,

$$|(a_n + b_n) - (a + b)| \leq |a_n - a| + |b_n - b| < \epsilon/2 + \epsilon/2 = \epsilon.$$

(4) We have

$$\left|\frac{a_n}{b_n} - \frac{a}{b}\right| = \frac{|a_n b - ab_n|}{|b_n||b|} = \frac{|a_n b - ab + ab - ab_n|}{|b_n||b|} \leq \frac{|b||a_n - a| + |a||b_n - b|}{|b_n||b|}.$$

Find N such that $|b_n - b| < \frac{1}{2}|b|$ for $n \geq N$. Then, for $n \geq N$, $|b_n| > \frac{1}{2}|b|$, so

$$\left|\frac{a_n}{b_n} - \frac{a}{b}\right| \leq \frac{2}{|b|^2}\big(|b||a_n - a| + |a||b_n - b|\big).$$

By assumption, using (1) and (2), the right-hand side goes to 0 as $n \to \infty$, so by the squeeze theorem, the left-hand side does as well.

(5) By Exercise 1.15, $||a_n| - |a|| \leq |a_n - a| \to 0$, so $||a_n| - |a|| \to 0$ by the squeeze theorem. □

Exercise 3.3. Finish the proof of Theorem 3.11.

3.12. Proposition. If $a_n \to a$, where $a_n \geq 0$ for all $n \in \mathbb{N}$, then $\sqrt{a_n} \to \sqrt{a}$.

Proof. We consider the case of $a > 0$. Then

$$|\sqrt{a_n} - \sqrt{a}| = \frac{|a_n - a|}{\sqrt{a_n} + \sqrt{a}} \leq \frac{1}{\sqrt{a}}|a_n - a|,$$

and $\dfrac{1}{\sqrt{a}}|a_n - a| \to 0$ by Theorem 3.11 (1), so $|\sqrt{a_n} - \sqrt{a}| \to 0$ by the squeeze theorem. □

Exercise 3.4. Prove Proposition 3.12 for $a = 0$.

3.13. Theorem (order limit theorem)**.** If $a_n \to a$, $b_n \to b$, and $a_n \leq b_n$ for infinitely many $n \in \mathbb{N}$, then $a \leq b$.

Proof. We prove the contrapositive. Suppose $b < a$. Let $\epsilon = \frac{1}{2}(a - b) > 0$. Find N_1 such that $|a_n - a| < \epsilon$ for $n \geq N_1$. Find N_2 such that $|b_n - b| < \epsilon$ for $n \geq N_2$. Then, for $n \geq \max\{N_1, N_2\}$,

$$b_n < b + \epsilon = \tfrac{1}{2}(a + b) = a - \epsilon < a_n,$$

so it is not the case that $a_n \leq b_n$ for infinitely many $n \in \mathbb{N}$. □

3.14. Remark. The following is a special case of Theorem 3.13. If $b_n \to b$ and $b_n \geq 0$ for infinitely many $n \in \mathbb{N}$, then $b \geq 0$. Even if $b_n > 0$ for all $n \in \mathbb{N}$, we can still only conclude that $b \geq 0$: consider for example $b_n = \frac{1}{n} \to 0$.

3.3. Monotone sequences

3.15. Definition. A sequence (a_n) is:

- *increasing* if $a_n \leq a_{n+1}$ for all $n \in \mathbb{N}$,
- *strictly increasing* if $a_n < a_{n+1}$ for all $n \in \mathbb{N}$,
- *decreasing* if $a_n \geq a_{n+1}$ for all $n \in \mathbb{N}$,
- *strictly decreasing* if $a_n > a_{n+1}$ for all $n \in \mathbb{N}$,
- *monotone* if it is increasing or decreasing,
- *strictly monotone* if it is strictly increasing or strictly decreasing.

Monotone sequences have the advantage that for them, boundedness is not only a necessary but also a sufficient condition for convergence. Notice how completeness is used in the proof of the following theorem.

3.16. Theorem (monotone convergence theorem)**.** A bounded monotone sequence converges.

Proof. Let (a_n) be bounded and monotone, say increasing (the decreasing case is analogous). Then the set $A = \{a_n : n \in \mathbb{N}\}$ is nonempty and bounded. Let $s = \sup A$. We claim that $s = \lim a_n$. Let $\epsilon > 0$. Then $s - \epsilon$ is not an upper bound for A, so there is $N \in \mathbb{N}$ with $s - \epsilon < a_N$. But then, if $n \geq N$, $s - \epsilon < a_N \leq a_n \leq s$, so $|a_n - s| < \epsilon$. □

3.17. Example. The sequence $1, 2, 3, \ldots$ is monotone, not bounded, and not convergent. The sequence $0, 1, 0, 1, 0, 1, \ldots$ is bounded, not monotone, and not convergent.

Sequences of the following important kind tend to be monotone.

3.18. Definition. A *recursively* or *inductively defined sequence* is a sequence (x_n) defined by specifying x_1 and giving a *recursion formula*

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

where f is a function, that allows us to compute x_2 from x_1, x_3 from x_2, and so on.

Often we can use the monotone convergence theorem to show that such a sequence converges, even though we do not have an explicit formula for x_n in terms of n. We illustrate this by an example.

3.19. Example. Let $x_1 = 1$ and $x_{n+1} = 3 - 1/x_n$. Then $x_2 = 2$ and $x_3 = 2\frac{1}{2}$, so it looks like (x_n) may be increasing. Let us verify this guess. We want to prove by induction that $0 < x_n < x_{n+1}$ for all $n \in \mathbb{N}$. This is

clear for $n = 1$. Suppose $0 < x_n < x_{n+1}$. Then $1/x_n > 1/x_{n+1}$, so

$$x_{n+1} = 3 - \frac{1}{x_n} < 3 - \frac{1}{x_{n+1}} = x_{n+2}.$$

To be able to apply the monotone convergence theorem, we also need to show that (x_n) is bounded above (boundedness below is obvious because (x_n) is increasing). We need to guess a suitable upper bound M and prove by induction that it works. Let us try $M = 3$. We need to prove that $x_n \leq 3$ for all n. This is clear for $n = 1$. Suppose $x_n \leq 3$. Then $1/x_n \geq 1/3$ (recall that $x_n > 0$), so $x_{n+1} = 3 - 1/x_n \leq 3 - 1/3 < 3$.

The monotone convergence theorem now implies that (x_n) converges, but it does not tell us what the limit is. Here the recursion formula can help. Let us call the limit a. Then

$$a = \lim x_n = \lim x_{n+1} = \lim \left(3 - \frac{1}{x_n}\right) = 3 - \frac{1}{\lim x_n} = 3 - \frac{1}{a}$$

(for the second equality, see Exercise 3.12), so $a^2 - 3a + 1 = 0$ and $a = \frac{1}{2}(3 \pm \sqrt{5})$. Finally, since $a > x_1 = 1$, we have $x_n \to \frac{1}{2}(3 + \sqrt{5})$.

Exercise 3.5. Show that the sequence $\sqrt{2}$, $\sqrt{2\sqrt{2}}$, $\sqrt{2\sqrt{2\sqrt{2}}}, \ldots$ converges and find its limit.

3.4. Series

3.20. Definition. Let $(a_n)_{n\in\mathbb{N}}$ be a sequence of real numbers. The *series* associated to (a_n) is the sequence (s_n) of *partial sums* $s_n = a_1 + \cdots + a_n$, $n \in \mathbb{N}$. We write $\sum_{n=1}^{\infty} a_n$ or simply $\sum a_n$ for (s_n). We say that the series $\sum a_n$ *converges* with *sum* s if $\lim_{n\to\infty} s_n = s$. We then also use $\sum_{n=1}^{\infty} a_n$ or $\sum a_n$ as a symbol for s. A series that does not converge is said to *diverge*.

The following is an immediate consequence of the algebraic limit theorem for sequences (Theorem 3.11).

3.21. Proposition. If $\sum a_n$ converges with sum s, and $\sum b_n$ converges with sum t, then:

(1) $\sum(a_n + b_n)$ converges with sum $s + t$.
(2) $\sum(ca_n)$ converges with sum cs for every $c \in \mathbb{R}$.

3.22. Remark. If $a_n \geq 0$ for all $n \in \mathbb{N}$, then the sequence (s_n) of partial sums is increasing, so by the monotone convergence theorem (Theorem 3.16), $\sum a_n$ converges if and only if (s_n) is bounded above.

3.23. Example. (a) Consider the *harmonic series* $\sum 1/n$. Since the subsequence of partial sums

$$1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{2^k}=1+\frac{1}{2}+\left(\frac{1}{3}+\frac{1}{4}\right)+\cdots+\left(\frac{1}{2^{k-1}+1}+\cdots+\frac{1}{2^k}\right)$$
$$\geq 1+\frac{1}{2}+2\cdot\frac{1}{4}+\cdots+2^{k-1}\frac{1}{2^k}=1+\frac{k}{2}$$

is unbounded, the harmonic series diverges.

(b) For $n\geq 2$, the n^{th} partial sum of the series $\sum 1/n^2$ is

$$1+\frac{1}{2^2}+\cdots+\frac{1}{n^2}<1+\frac{1}{2\cdot 1}+\frac{1}{3\cdot 2}+\cdots+\frac{1}{n(n-1)}$$
$$=1+\left(1-\frac{1}{2}\right)+\left(\frac{1}{2}-\frac{1}{3}\right)+\cdots+\left(\frac{1}{n-1}-\frac{1}{n}\right)$$
$$=2-\frac{1}{n}<2.$$

Hence $\sum 1/n^2$ converges with sum at most 2.

3.24. Proposition. If $\sum a_n$ converges, then $a_n\to 0$. The converse fails in general: even if $a_n\to 0$, $\sum a_n$ may diverge.

Proof. If $\sum a_n$ converges with sum s, then $a_n=s_n-s_{n-1}\to s-s=0$. The harmonic series $\sum 1/n$ diverges although $1/n\to 0$ (Example 3.23). □

3.25. Example. The most important series of all is the *geometric series* $\sum_{n=0}^{\infty} r^n$, where r is a fixed real number (note that we start the summation with $n=0$ here). If $|r|\geq 1$, then $r^n\not\to 0$, so $\sum r^n$ diverges by Proposition 3.24. If $|r|<1$, then

$$1+r+\cdots+r^n=\frac{1-r^{n+1}}{1-r}\to\frac{1}{1-r}$$

as $n\to\infty$, so $\sum r^n$ converges with sum $\dfrac{1}{1-r}$ (for the fact that $r^{n+1}\to 0$, see Exercise 3.2).

The following simple result is one of the most useful in the theory of series.

3.26. Proposition (comparison test)**.** Let (a_n) and (b_n) be sequences with $0\leq a_n\leq b_n$ for all $n\in\mathbb{N}$. If $\sum b_n$ converges, then $\sum a_n$ converges. Equivalently, if $\sum a_n$ diverges, then $\sum b_n$ diverges.

Proof. Let $s_n=a_1+\cdots+a_n$ and $t_n=b_1+\cdots+b_n$. Then $0\leq s_n\leq t_n$ for all $n\in\mathbb{N}$, so if (t_n) is bounded above, so is (s_n). Now apply Remark 3.22. □

Exercise 3.6. Show that the comparison test is still valid with the weaker hypothesis that there is $m \in \mathbb{N}$ such that $0 \le a_n \le b_n$ for all $n \ge m$.

Exercise 3.7 (limit comparison test)**.** Let (a_n) and (b_n) be sequences of positive terms such that (a_n/b_n) converges. Prove that if $\sum b_n$ converges, then $\sum a_n$ converges. *Hint.* If $a_n/b_n \to c$, then $a_n \le (c+1)b_n$ for n large enough.

3.27. Definition. A series $\sum a_n$ is *absolutely convergent* if $\sum |a_n|$ is convergent. A series that is convergent but not absolutely convergent is called *conditionally convergent.*

This terminology is justified by the following result.

3.28. Proposition. An absolutely convergent series converges.

Proof. Say $\sum a_n$ is absolutely convergent. Now, for every $n \in \mathbb{N}$, $0 \le a_n + |a_n| \le 2|a_n|$, so $\sum(a_n + |a_n|)$ converges by the comparison test (Proposition 3.26). Hence $\sum a_n = \sum((a_n + |a_n|) - |a_n|)$ converges. □

Next we present three commonly used convergence tests. One more test, the integral test, appears later, in Exercise 7.19.

3.29. Theorem (ratio test)**.** Let (a_n) be a sequence of positive terms. Suppose $\dfrac{a_{n+1}}{a_n} \to c$ as $n \to \infty$.

(1) If $c < 1$, then $\sum a_n$ converges.

(2) If $c > 1$, then $\sum a_n$ diverges.

(3) If $c = 1$, then $\sum a_n$ may or may not converge.

Proof. (1) Choose r with $c < r < 1$. There is $N \in \mathbb{N}$ such that $a_{n+1}/a_n < r$ for all $n \ge N$. Then, for $n > N$,

$$a_n < ra_{n-1} < \cdots < r^{n-N}a_N = (a_N/r^N)r^n.$$

By comparison with the geometric series $\sum r^n$, which converges, we see that $\sum a_n$ converges.

(2) There is $N \in \mathbb{N}$ such that $a_{n+1}/a_n > 1$, that is, $a_n < a_{n+1}$, for all $n \ge N$. Thus $a_n \not\to 0$, so $\sum a_n$ diverges.

(3) The series in Example 3.23 both have $c = 1$, but one converges and the other diverges. □

3.30. Corollary. Let (a_n) be a sequence with $a_n \ne 0$ for all $n \in \mathbb{N}$ such that $\dfrac{|a_{n+1}|}{|a_n|} \to c$ as $n \to \infty$. If $c < 1$, then $\sum a_n$ converges absolutely.

The proof of the next test is very similar to the proof of the ratio test.

3.31. Theorem (root test)**.** Let (a_n) be a sequence of nonnegative terms. Suppose $a_n^{1/n} \to c$ as $n \to \infty$.

(1) If $c < 1$, then $\sum a_n$ converges.

(2) If $c > 1$, then $\sum a_n$ diverges.

(3) If $c = 1$, then $\sum a_n$ may or may not converge.

Proof. (1) Choose r with $c < r < 1$. There is $N \in \mathbb{N}$ such that $a_n^{1/n} < r$ and thus $a_n < r^n$ for all $n \geq N$. By comparison with the geometric series $\sum r^n$, which converges, we see that $\sum a_n$ converges.

(2) There is $N \in \mathbb{N}$ such that $a_n^{1/n} > 1$ and thus $a_n > 1$ for all $n \geq N$. Hence $a_n \not\to 0$, so $\sum a_n$ diverges.

(3) By Exercise 3.17, the series in Example 3.23 both have $c = 1$. □

3.32. Theorem (alternating series test)**.** Let (a_n) be a decreasing sequence of positive terms with $a_n \to 0$. Then $\sum(-1)^{n+1}a_n$ converges.

Proof. Let $s_n = a_1 - a_2 + a_3 - \cdots + (-1)^{n+1}a_n$. Note that

$$s_2 \leq s_4 \leq s_6 \leq \cdots \leq s_5 \leq s_3 \leq s_1.$$

By the monotone convergence theorem (Theorem 3.16), (s_{2n-1}) and (s_{2n}) converge to, say, s and t, respectively. Since $s_{2n} = s_{2n-1} - a_{2n}$ and $a_{2n} \to 0$, we have $s = t$. By the following exercise, $s_n \to s$, so $\sum(-1)^{n+1}a_n$ converges with sum s. □

Exercise 3.8. Let (s_n) be a sequence such that the sequences (s_{2n-1}) and (s_{2n}) converge to the same limit b. Show that $s_n \to b$.

Exercise 3.9. Show that the sum s of the alternating series above satisfies $|s - s_n| \leq a_{n+1}$ for every $n \in \mathbb{N}$. Thus, if we approximate the true sum s by the partial sum s_n, the error is at most a_{n+1}.

3.33. Example. By the alternating series test, the *alternating harmonic series* $1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots$ converges. Since the harmonic series diverges (Example 3.23), the alternating harmonic series converges conditionally. Call the sum s. From the error estimate in Exercise 3.9 we see, for example, that $|s - (1 - \frac{1}{2} + \frac{1}{3})| \leq \frac{1}{4}$, that is, $\frac{7}{12} \leq s \leq \frac{13}{12}$.

We conclude this section by touching on the topic of rearrangements. The terms of a finite sum can be added up in any order: the result is always the same. This is not so straightforward for infinite sums.

3.34. Definition. A *rearrangement* of a series $\sum a_n$ is a series of the form $\sum a_{\sigma(n)}$, where $\sigma : \mathbb{N} \to \mathbb{N}$ is a bijection.

3.35. Theorem. If $\sum a_n$ is absolutely convergent, then every rearrangement $\sum a_{\sigma(n)}$ is also absolutely convergent and has the same sum.

Proof. Let $s_n = a_1 + \cdots + a_n$ and $t_n = a_{\sigma(1)} + \cdots + a_{\sigma(n)}$. Let $\epsilon > 0$. Since $\sum a_n$ is absolutely convergent, there is $p \in \mathbb{N}$ with $\sum\limits_{n>p} |a_n| < \epsilon$. There are exactly p values of n with $\sigma(n) \leq p$. Let q be the largest of them, so that $n \leq q$ if $\sigma(n) \leq p$. Then $\sigma(n) > p$ if $n > q$, so $\sum\limits_{n>q} |a_{\sigma(n)}| \leq \sum\limits_{n>p} |a_n| < \epsilon$. This shows that $\sum a_{\sigma(n)}$ converges absolutely.

Finally, if $n \geq p$ and $n \geq q$, then, in the difference $s_n - t_n$, the terms $a_1, \ldots, a_p$ cancel, so $|s_n - t_n| \leq 2 \sum\limits_{n>p} |a_n| < 2\epsilon$. Hence (s_n) and (t_n) converge to the same limit. □

3.36. Example. The case of conditionally convergent series is very different. Consider the alternating harmonic series $1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots$ (Example 3.33). Note that the series $\sum \frac{1}{2n-1}$ of positive terms and the series $-\sum \frac{1}{2n}$ of negative terms both diverge. We will rearrange the alternating harmonic series in such a way that the positive terms appear in their original order, and the negative terms as well, but the negative terms are spread more thinly among the positive ones. Start by taking positive terms that add up to at least 2. Then take one negative term. Again take positive terms adding up to at least 2, followed by one negative term, and so on. The resulting series diverges.

In fact, every conditionally convergent series can be rearranged so as to converge to any sum whatsoever or to diverge (Exercise 3.24).

3.5. Subsequences and Cauchy sequences

3.37. Definition. Let $(a_n)_{n\in\mathbb{N}}$ be a sequence. Let $(n_k)_{k\in\mathbb{N}}$ be a strictly increasing sequence of natural numbers. Then the sequence $(a_{n_k})_{k\in\mathbb{N}}$, that is, $a_{n_1}, a_{n_2}, a_{n_3}, \ldots$, is called a *subsequence* of (a_n).

3.38. Remark. In the language of functions, a sequence a in $\mathbb{R}$ is a function $a : \mathbb{N} \to \mathbb{R}$. A subsequence of a is then a sequence of the form $a \circ f$, where $f : \mathbb{N} \to \mathbb{N}$ is a strictly increasing function.

3.39. Proposition. A subsequence of a convergent sequence converges to the same limit.

Proof. Say (a_{n_k}) is a subsequence of (a_n), and $a_n \to b$. Let U be a neighbourhood of b. Then $a_n \notin U$ for at most finitely many n, so $a_{n_k} \notin U$ for at most finitely many k. This shows that $a_{n_k} \to b$. □

The next result is yet another manifestation of the completeness of the real numbers. Be sure to note the two places in the proof where completeness is invoked.

3.40. Theorem (Bolzano-Weierstrass theorem)**.** A bounded sequence has a convergent subsequence.

Proof. This is the most complicated proof we have had so far. For clarity, we shall break it up into three steps. Let (a_n) be a bounded sequence. Take $M > 0$ with $|a_n| \leq M$ for all n.

Step 1. One half of $[-M, M]$, either $[-M, 0]$ or $[0, M]$, call it I_1, contains a_n for infinitely many n (if both halves do, then it does not matter which one we pick). One half of I_1, call it I_2, contains a_n for infinitely many n. Continuing, we obtain a sequence $I_1 \supset I_2 \supset I_3 \supset \cdots$ of closed bounded intervals such that each I_k contains a_n for infinitely many n.

Step 2. Choose $n_1 \in \mathbb{N}$ with $a_{n_1} \in I_1$. Choose $n_2 > n_1$ with $a_{n_2} \in I_2$. This is possible because $a_n \in I_2$ for infinitely many n: these n cannot all be less than or equal to n_1. Continue and obtain a subsequence (a_{n_k}) of (a_n) with $a_{n_k} \in I_k$ for all k. We want to prove that this subsequence converges.

Step 3. There is $b \in \bigcap_{k=1}^{\infty} I_k$ by the nested interval property (Theorem 2.14). We claim that $a_{n_k} \to b$. Let $\epsilon > 0$. By the Archimedean property (Exercise 3.2), there is $N \in \mathbb{N}$ such that the length of I_N, which is $M/2^N$, is smaller than ϵ. Thus, for $k \geq N$, $|a_{n_k} - b| < \epsilon$ since a_{n_k} and b both lie in $I_k \subset I_N$. □

We will develop the theory of the trigonometric functions in Section 8.4. Until then, when required in examples and exercises, let us take for granted what we know about them from first year and high school.

3.41. Example. The sequence $(\sin n)_{n\in\mathbb{N}}$ is bounded, so it has a convergent subsequence. However, an explicit example of such a subsequence is not easy to find. Would you like to try?

3.42. Definition. A sequence (a_n) is a *Cauchy sequence* if for every $\epsilon > 0$, there is $N \in \mathbb{N}$ such that if $m, n \geq N$, then $|a_m - a_n| < \epsilon$.

3.43. Theorem (Cauchy criterion)**.** A sequence is Cauchy if and only if it converges.

Note that the Cauchy criterion characterises convergence without any mention of a limit.

Sketch of proof. $\Leftarrow$ is the easy direction. If $a_n \to a$, just use the triangle inequality $|a_m - a_n| \leq |a_m - a| + |a_n - a|$.

$\Rightarrow$ First show that a Cauchy sequence is bounded. Second, invoke the Bolzano-Weierstrass theorem to obtain a convergent subsequence. Third, show that a Cauchy sequence with a convergent subsequence is itself convergent (to the same limit). $\square$

3.44. Remark. For a sequence to be convergent, it is not enough to assume that successive terms get arbitrarily close. For example, take $a_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n}$. Then $|a_{n+1} - a_n| = \frac{1}{n+1} \to 0$ as $n \to \infty$, but (a_n) is unbounded and thus divergent.

The final theorem of the chapter is the culmination of the work we have done so far. It gives five different characterisations, all important, of the fundamental property of completeness.

3.45. Theorem. For an ordered field, the following are equivalent.

(1) The axiom of completeness.

(2) The nested interval property and the Archimedean property.

(3) The monotone convergence theorem.

(4) The Bolzano-Weierstrass theorem.

(5) The Cauchy criterion and the Archimedean property.

Proof. We know that (1) $\Rightarrow$ (2) $\Rightarrow$ (4), and (1) $\Rightarrow$ (3). We have sketched a proof that (4) implies the Cauchy criterion.

The Archimedean property follows from (4). Namely, if $\mathbb{N}$ was bounded above, then the Bolzano-Weierstrass theorem would provide a limit for a subsequence of $1, 2, 3, \ldots$. This limit would then be a supremum for $\mathbb{N}$, leading to a contradiction as in our proof of Theorem 2.8. Thus (4) $\Rightarrow$ (5).

Next, (3) $\Rightarrow$ (4). Namely, assume (3) and let (a_n) be a bounded sequence. By Exercise 3.10, (a_n) has a monotone subsequence, which is also bounded, and hence convergent by (3).

To complete the circuit of implications, we need to show that (5) $\Rightarrow$ (1). Assume (5) and let A be a subset of the ordered field, nonempty and bounded above. Let $a \in A$ and let $b > a$ be an upper bound for A. Let $I = [a, b]$. For each $n \in \mathbb{N}$, divide I into 2^n intervals $I_k = [a + (k-1)\frac{b-a}{2^n}, a + k\frac{b-a}{2^n}]$, $k = 1, \ldots, 2^n$, of equal length $\frac{b-a}{2^n}$. Choose $a_n \in A \cap I_k$ for the largest k for which $A \cap I_k \neq \varnothing$, and let $b_n \geq a_n$ be the right end point $a + k\frac{b-a}{2^n}$ of I_k. Then b_n is an upper bound for A, and $b_n - a_n \leq \frac{b-a}{2^n}$. The sequence (b_n) is decreasing. For $n \leq m$, $a_n \leq b_m \leq b_n$, so $b_n - b_m \leq \frac{b-a}{2^n}$.

By the Archimedean property (Exercise 3.2), $\frac{b-a}{2^n} \to 0$ as $n \to \infty$. Hence (b_n) is a Cauchy sequence and therefore convergent with limit c by the

Cauchy criterion. Since each b_n is an upper bound for A, so is c. Furthermore, for each $n \in \mathbb{N}$, $a_n \leq c \leq b_n$, so $c - a_n \leq \frac{b-a}{2^n}$, that is, applying the Archimedean property again, there are elements of A arbitrarily close to c. Hence c is the least upper bound of A. □

Exercise 3.10. Show that every sequence has a monotone subsequence. *Hint.* This is surprisingly tricky to prove. Proceed as follows.

Step 1. Show that a sequence with no largest term has an increasing subsequence. Therefore, if (a_n) is a sequence, and for some $N \in \mathbb{N}$, the 'tail' $a_N, a_{N+1}, a_{N+2}, \ldots$ has no largest term, then that tail, and hence (a_n) itself, has an increasing subsequence.

Step 2. Show that if every tail of (a_n) has a largest term, then (a_n) has a decreasing subsequence.

More exercises

3.11. Using only the definition of the limit, show that $\lim_{n\to\infty} \dfrac{3n-1}{n+2} = 3$.

3.12. Let (a_n) be a sequence. Define a sequence (b_n) by the formula $b_n = a_{n+1}$ for each $n \in \mathbb{N}$. In other words, (b_n) is (a_n) with the first term removed. Prove that $a_n \to c$ if and only if $b_n \to c$.

3.13. Prove the following statement or disprove it by a counterexample. If (a_n) and (b_n) are sequences such that (a_n) and $(a_n b_n)$ converge, then (b_n) converges.

3.14. Prove the following statement or disprove it by a counterexample. If (a_n) and (b_n) are sequences of positive numbers such that (a_n) and (a_n/b_n) converge, then (b_n) converges.

3.15. Let (a_n) be a bounded (not necessarily convergent) sequence and (b_n) be a sequence such that $b_n \to 0$. Show that $a_n b_n \to 0$.

3.16. (a) Let $a \geq 0$ and $n \in \mathbb{N}$. Show that $(1+a)^n \geq na$.

(b) Let $c > 0$. Show that $c^{1/n} \to 1$ as $n \to \infty$. *Hint.* For $c > 1$, let $a = c^{1/n} - 1$.

3.17. (a) Let $a \geq 0$ and $n \in \mathbb{N}$, $n \geq 2$. Show that $(1+a)^n \geq \frac{1}{2}n(n-1)a^2$.

(b) Show that $n^{1/n} \to 1$ as $n \to \infty$.

3.18. Consider the recursively defined sequence (a_n) with $a_1 = 3$ and $a_{n+1} = a_n/2 + 3/a_n$. Show that (a_n) converges and find its limit. *Hint.* Induction may not be the best way to show that (a_n) is bounded and monotone.

3.19. Determine whether the following series converge:

$$\sum \frac{1}{\sqrt{n}}, \quad \sum \frac{2}{3n^2+n+5}, \quad \sum \frac{(-1)^n n}{n+1}, \quad \sum \frac{n^2}{2^n}.$$

3.20. Let (a_n) be a sequence such that the series $\sum |a_{n+1} - a_n|$ converges. Show that (a_n) converges.

3.21. Here are two applications of the useful inequality $xy \le \frac{1}{2}(x^2+y^2)$.

(a) Show that if $\sum a_n^2$ and $\sum b_n^2$ converge, then $\sum a_n b_n$ converges absolutely.

(b) Let $\sum a_n$ be a convergent series of nonnegative terms. Show that $\sum \sqrt{a_n}/n$ converges.

3.22. (a) Find a convergent series $\sum a_n$ such that $\sum a_n^2$ diverges.

(b) Show that if $\sum a_n$ converges absolutely, then $\sum a_n^2$ converges.

3.23. Let $\sum a_n$ be a series. Set $a_n^+ = \max\{0, a_n\}$ and $a_n^- = \min\{0, a_n\}$. Consider the series $\sum a_n^+$ and $\sum a_n^-$. In $\sum a_n^+$ the negative terms in $\sum a_n$ have been changed to 0. In $\sum a_n^-$ the positive terms in $\sum a_n$ have been changed to 0.

(a) Prove that $\sum a_n$ is absolutely convergent if and only if $\sum a_n^+$ and $\sum a_n^-$ both converge. Then $\sum a_n = \sum a_n^+ + \sum a_n^-$.

(b) Prove that if $\sum a_n$ is conditionally convergent, then $\sum a_n^+$ and $\sum a_n^-$ both diverge.

3.24. Let $\sum a_n$ be a conditionally convergent series. Prove that for every $s \in \mathbb{R}$, there is a rearrangement of $\sum a_n$ converging to s. Prove also that there is a divergent rearrangement of $\sum a_n$. *Hint.* Use Exercise 3.23.

3.25. Give an example of each of the following or show that it does not exist.

(a) A sequence not containing 0 or 1 as a term but containing subsequences converging to each of these values.

(b) A monotone sequence that diverges but has a convergent subsequence.

(c) An unbounded sequence with a convergent subsequence.

(d) A sequence with a bounded subsequence but without a convergent subsequence.

3.26. Show that there is a strictly increasing sequence $(n_k)_{k\in\mathbb{N}}$ of natural numbers such that the sequences $(\cos n_k)_{k\in\mathbb{N}}$ and $(\sin n_k)_{k\in\mathbb{N}}$ both converge.

3.27. Let (a_n) be a sequence and c be a number such that every subsequence of (a_n) has a subsequence converging to c. Show that (a_n) itself converges to c.

3.28. Show that a bounded sequence is divergent if and only if it has two subsequences with different limits.

3.29. Let (a_n) be a sequence with $a_n \to 0$. Show that there is a subsequence (a_{n_k}) such that the series $\sum_{k=1}^{\infty} a_{n_k}$ is absolutely convergent.

3.30. Show that there is a sequence (a_n) such that for every real number x, there is a subsequence of (a_n) converging to x. *Hint.* Start with a bijection $a : \mathbb{N} \to \mathbb{Q}$.

3.31. Prove directly from the definition of a Cauchy sequence that a Cauchy sequence is bounded. Do not use the result that a Cauchy sequence converges: the first step in the proof of that result is to show that a Cauchy sequence is bounded.

3.32. Show directly from the definition of a Cauchy sequence that a sum of Cauchy sequences is Cauchy. Do not use the Cauchy criterion.

3.33. (a) Fix a natural number $b \geq 2$. Let (a_n) be a sequence of integers with $0 \leq a_n < b$. Show that the series $\sum_{n=1}^{\infty} \frac{a_n}{b^n}$ converges with sum in $[0, 1]$.

(b) Conversely, let $x \in [0, 1]$. Prove that there is a sequence (a_n) of integers with $0 \leq a_n < b$ such that $\sum_{n=1}^{\infty} \frac{a_n}{b^n} = x$. The expression $0.a_1a_2a_3\ldots$ is called the *expansion of x to base b* (although for some x it is not quite unique: see (c)). If $b = 2$, it is also called the *binary expansion* of x, and if $b = 10$, it is also called the *decimal expansion* of x.

Hint. Let a_1 be the largest number in $\{0, \ldots, b-1\}$ with $a_1/b \leq x$. Having chosen $a_1, \ldots, a_n$, let a_{n+1} be the largest number in $\{0, \ldots, b-1\}$ with $\sum_{k=1}^{n+1} \frac{a_k}{b^k} \leq x$.

(c) Suppose (a_n) and (c_n) are distinct sequences in $\{0, \ldots, b-1\}$ with $\sum_{n=1}^{\infty} \frac{a_n}{b^n} = \sum_{n=1}^{\infty} \frac{c_n}{b^n}$. Prove that, after possibly interchanging (a_n) and (c_n), there is $m \in \mathbb{N}$ such that $a_n = c_n$ for $n < m$, $a_m = c_m + 1$, and $a_n = 0$ and $c_n = b - 1$ for $n > m$.

3.34. (a) Let (x_n) be a bounded sequence. For each $k \in \mathbb{N}$, let

$$y_k = \sup_{n \geq k} x_n = \sup\{x_k, x_{k+1}, x_{k+2}, \ldots\}.$$

Show that the sequence (y_k) is decreasing and bounded below. Conclude that (y_k) converges. The limit of (y_k) is called the *limit superior* of the original sequence (x_n). In other words,

$$\limsup_{n\to\infty} x_n = \lim_{k\to\infty} \sup_{n \geq k} x_n.$$

(b) Show that if (x_n) is convergent, then $\limsup x_n = \lim x_n$.

(c) Show that (x_n) has a subsequence converging to $\limsup x_n$.

(d) Show that no convergent subsequence of (x_n) has a limit larger than $\limsup x_n$.

Thus $\limsup x_n$ is the largest limit that a convergent subsequence of (x_n) can have.

(e) Formulate a definition of the *limit inferior* of (x_n). State and prove the analogues for $\liminf$ of the above properties of $\limsup$.

Chapter 4

Open, closed, and compact sets

The three concepts introduced in this chapter belong to the subject of topology. Of the major branches of mathematics, topology is the youngest. It emerged as a subject in its own right in the early twentieth century. Topology is heavily used in modern analysis.

4.1. Open and closed sets

4.1. Definition. A subset U of $\mathbb{R}$ is *open* if it is a neighbourhood of each of its points. That is, for every $a \in U$, there is $\epsilon > 0$ such that $(a-\epsilon, a+\epsilon) \subset U$.

4.2. Proposition. (1) $\mathbb{R}$ and $\varnothing$ are open. An open interval is open.

(2) The union of an arbitrary collection of open sets is open.

(3) The intersection of finitely many open sets is open.

Proof. (1) It should be clear that every nonempty open interval, including $\mathbb{R}$ itself, is open. It may not be quite so obvious why $\varnothing$ is open. What would it mean for $\varnothing$ not to be open? It would mean not being a neighbourhood of one of its elements. But $\varnothing$ has no elements at all, in particular, no elements that could refute $\varnothing$ being open. Hence $\varnothing$ is open.

(2) Let $(U_i)_{i\in I}$ be a family of open subsets of $\mathbb{R}$. We want to show that $\bigcup_{i\in I} U_i$ is open. Let $a \in \bigcup_{i\in I} U_i$. This means that there is $j \in I$ such that $a \in U_j$. Since U_j is open, there is $\epsilon > 0$ such that $(a-\epsilon, a+\epsilon) \subset U_j$. Since $U_j \subset \bigcup_{i\in I} U_i$, we conclude that $(a-\epsilon, a+\epsilon) \subset \bigcup_{i\in I} U_i$. Hence $\bigcup_{i\in I} U_i$ is open.

(3) Let $U_1, \dots, U_n$ be open subsets of $\mathbb{R}$. Let $a \in U_1 \cap \dots \cap U_n$. For each $i = 1, \dots, n$, since U_i is open, there is $\epsilon_i > 0$ such that $(a - \epsilon_i, a + \epsilon_i) \subset U_i$. Let $\epsilon = \min\{\epsilon_1, \dots, \epsilon_n\} > 0$. Then $(a - \epsilon, a + \epsilon) \subset (a - \epsilon_i, a + \epsilon_i) \subset U_i$ for every $i = 1, \dots, n$, so $(a - \epsilon, a + \epsilon) \subset U_1 \cap \dots \cap U_n$. This shows that $U_1 \cap \dots \cap U_n$ is open. □

4.3. Definition. A subset A of $\mathbb{R}$ is *closed* if its complement $\mathbb{R} \setminus A$ is open.

The following proposition is dual to Proposition 4.2.

4.4. Proposition. (1) $\mathbb{R}$ and $\varnothing$ are closed. A closed interval is closed.

(2) The intersection of an arbitrary collection of closed sets is closed.

(3) The union of finitely many closed sets is closed.

Proof. We prove (2) and leave (1) and (3) as exercises. The proof is an application of one of De Morgan's laws (Remark 1.19) to Proposition 4.2. Namely, if A_i is a closed subset of $\mathbb{R}$ for each $i \in I$, then $\mathbb{R} \setminus A_i$ is open, so by Proposition 4.2 and De Morgan,

$$\bigcup_{i \in i} \mathbb{R} \setminus A_i = \mathbb{R} \setminus \bigcap_{i \in I} A_i$$

is open. This shows that $\bigcap_{i \in I} A_i$ is closed. □

Exercise 4.1. Finish the proof of Proposition 4.4.

4.5. Example. (a) Note that the sets $\mathbb{R}$ and $\varnothing$ are both open and closed.

(b) An example of a set that is neither open nor closed is the interval $I = [0, 1)$. It is not open because $0 \in I$ but I is not a neighbourhood of 0. It is not closed, that is, $\mathbb{R} \setminus I$ is not open, because $1 \in \mathbb{R} \setminus I$, but $\mathbb{R} \setminus I$ is not a neighbourhood of 1.

(c) An arbitrary intersection of open sets need not be open. For example,

$$\bigcap_{n=1}^{\infty} (-\tfrac{1}{n}, \tfrac{1}{n}) = \{0\}.$$

Exercise 4.2. Show by an example that an arbitrary union of closed sets need not be closed.

Exercise 4.3. Show that the only subsets of $\mathbb{R}$ that are both open and closed are $\mathbb{R}$ itself and $\varnothing$.

Closed sets can be characterised in terms of convergent sequences.

4.6. Theorem. A subset A of $\mathbb{R}$ is closed if and only if whenever $a_n \in A$ for all $n \in \mathbb{N}$, and $a_n \to c$, we also have $c \in A$.

Proof. $\Rightarrow$ Say $a_n \in A$, $n \in \mathbb{N}$, and $a_n \to c$. If $c \notin A$, then, since $\mathbb{R} \setminus A$ is open, $\mathbb{R} \setminus A$ is a neighbourhood of c, so $a_n \in \mathbb{R} \setminus A$ for all but finitely many n, which is absurd.

$\Leftarrow$ We prove the contrapositive. Suppose A is not closed, that is, $\mathbb{R} \setminus A$ is not open. This means that there is $c \in \mathbb{R} \setminus A$ such that $\mathbb{R} \setminus A$ is not a neighbourhood of c. Thus, for each $n \in \mathbb{N}$, $\mathbb{R} \setminus A$ does not contain the $\frac{1}{n}$-neighbourhood of c, so there is $a_n \in A$ in this neighbourhood, that is, $|a_n - c| < \frac{1}{n}$. Then $a_n \to c$. □

4.7. Remark. It may be shown that every open subset of $\mathbb{R}$ is the union of countably many mutually disjoint open intervals. There is no equally simple description of what a closed subset looks like. The structure of a closed set can be very complicated (for an example, see Exercise 4.12).

4.2. Compact sets

4.8. Definition. A subset K of $\mathbb{R}$ is *compact* if every sequence in K has a subsequence that converges to a limit that is also in K.

It is not meant to be obvious that this is a useful definition. The importance of the notion of compactness will emerge when we get to Theorems 5.17 and 5.20.

The following result gives a supply of compact sets.

4.9. Proposition. A closed and bounded interval is compact.

Proof. The empty set is clearly compact, because it contains no sequences at all. Let (x_n) be a sequence in a nonempty, closed, and bounded interval $[a, b]$. By the Bolzano-Weierstrass theorem (Theorem 3.40), (x_n) has a convergent subsequence (x_{n_k}). Call its limit c. Since $a \leq x_{n_k} \leq b$ for all k, we have $a \leq c \leq b$, that is, $c \in [a, b]$, by the order limit theorem (Theorem 3.13). □

The next theorem strengthens Proposition 4.9 and gives a useful characterisation of compactness that works for arbitrary sets.

4.10. Theorem (Heine-Borel theorem)**.** A subset of $\mathbb{R}$ is compact if and only if it is closed and bounded.

Proof. $\Leftarrow$ We argue as in the proof of Proposition 4.9, except we use Theorem 4.6 and the assumption that our set is closed to conclude that it contains the limit of the subsequence.

$\Rightarrow$ We prove the contrapositive, namely, that if $A \subset \mathbb{R}$ is either not closed or not bounded, then A is not compact. Suppose A is not closed. Then, by Theorem 4.6, there is a sequence (a_n) in A with $a_n \to c \notin A$. Then (a_n)

cannot have a subsequence with a limit in A, because every subsequence of (a_n) converges to c.

Finally, suppose A is not bounded, say not bounded above. This means that A contains an increasing sequence that is not bounded above. Such a sequence has no bounded subsequence, let alone a convergent one with limit in A. □

4.11. Corollary. The union of finitely many compact sets is compact.

Proof. This follows from Theorem 4.10 because a finite union of closed sets (meaning the union of a finite number of closed sets) is closed, and a finite union of bounded sets is bounded. □

4.12. Example. (a) Every finite set is compact: it is a finite union of one-point sets, and a one-point set is clearly both closed and bounded.

(b) The set $A = \{1, \frac{1}{2}, \frac{1}{3}, \dots\}$ is bounded but not closed and thus not compact. Indeed, the sequence $1, \frac{1}{2}, \frac{1}{3}, \dots$ in A converges to $0 \notin A$, so it has no subsequence that converges to a limit in A.

(c) The set $B = [0, \infty)$ is closed but not bounded and thus not compact. Indeed, the sequence $1, 2, 3, \dots$ in B has no convergent subsequence at all, let alone one with a limit in B.

Finally, we prove a generalisation of the nested interval property (Theorem 2.14).

4.13. Theorem. Let $K_1 \supset K_2 \supset K_3 \supset \cdots$ be nonempty compact subsets of $\mathbb{R}$. Then the intersection $\bigcap_{n=1}^{\infty} K_n$ is not empty.

Proof. For each $n \in \mathbb{N}$, choose $a_n \in K_n$. Then (a_n) is a sequence in K_1. Since K_1 is compact, (a_n) has a convergent subsequence (a_{n_k}) with limit $c \in K_1$. For each $m \geq 2$ and $k \geq m$, we have $n_k \geq k \geq m$, so $a_{n_k} \in K_{n_k} \subset K_m$. Thus the sequence $a_{n_m}, a_{n_{m+1}}, \dots$ lies in K_m, and it converges to c. Since K_m is closed, $c \in K_m$. This shows that $c \in \bigcap_{n=1}^{\infty} K_n$. □

More exercises

4.4. Let $A \subset \mathbb{R}$. The *closure* of A, denoted $\bar{A}$, is the intersection of all closed sets containing A.

(a) Show that $\bar{A}$ is closed.

(b) Show that $\bar{A}$ is the smallest closed set containing A, that is, if E is closed and $A \subset E$, then $\bar{A} \subset E$.

(c) Show that $x \in \bar{A}$ if and only if there is a sequence in A converging to x.

(d) Show that A is dense in $\mathbb{R}$ if and only if $\bar{A} = \mathbb{R}$.

(e) Is $\overline{A \cup B} = \bar{A} \cup \bar{B}$ for all $A, B \subset \mathbb{R}$? How about $\overline{A \cap B} = \bar{A} \cap \bar{B}$?

4.5. (a) Let $A \subset \mathbb{R}$. The *interior* of A, denoted A°, is the union of all open sets contained in A. Show that A° is open. Show that A° is the largest open set contained in A, that is, if U is open and $U \subset A$, then $U \subset A^\circ$.

(b) Show that $\overline{\mathbb{R} \setminus A} = \mathbb{R} \setminus A^\circ$ and $(\mathbb{R} \setminus A)^\circ = \mathbb{R} \setminus \bar{A}$.

(c) Is $(A \cup B)^\circ = A^\circ \cup B^\circ$ for all $A, B \subset \mathbb{R}$? How about $(A \cap B)^\circ = A^\circ \cap B^\circ$?

4.6. Let $A \subset \mathbb{R}$. The *boundary* of A, denoted ∂A, is $\bar{A} \setminus A^\circ$.

(a) Show that ∂A is closed.

(b) Use Exercise 4.5 to show that $\partial A = \bar{A} \cap \overline{\mathbb{R} \setminus A}$.

(c) Show that $x \in \partial A$ if and only if every neighbourhood of x intersects both A and $\mathbb{R} \setminus A$.

4.7. Find the closure, interior, and boundary of each of the following sets.

(a) $[0, 1]$.

(b) $(0, 1)$.

(c) $\mathbb{Z}$.

(d) $\mathbb{Q}$.

4.8. Prove directly from the definition of compactness (that is, not using the Heine-Borel theorem) that a finite subset of $\mathbb{R}$ is compact.

4.9. Prove that the set $\{0, 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots\}$ is compact. *Hint.* To show that the set is closed, express its complement as a union of open intervals.

4.10. Show that a nonempty compact set has both a largest element and a smallest element.

4.11. Prove that if K is compact and F is closed, then $K \cap F$ is compact.

4.12. Remove the open middle third $(\frac{1}{3}, \frac{2}{3})$ from $[0, 1]$, so $C_1 = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$ remains. Remove the open middle thirds $(\frac{1}{9}, \frac{2}{9})$ and $(\frac{7}{9}, \frac{8}{9})$ from the two intervals in C_1, so $C_2 = [0, \frac{1}{9}] \cup [\frac{2}{9}, \frac{1}{3}] \cup [\frac{2}{3}, \frac{7}{9}] \cup [\frac{8}{9}, 1]$ remains. Continuing in this way, we obtain closed sets $C_1 \supset C_2 \supset C_3 \supset \cdots$, such that C_n is the union of 2^n closed intervals of length 3^{-n}. The intersection $C = \bigcap_{n=1}^{\infty} C_n$ is called the *Cantor set*.

(a) Show that C is compact and not empty.

(b) The complement $[0,1]\setminus C$ is the union of the open middle thirds that were removed from $[0,1]$ in the construction of C. Show that the sum of the lengths of these intervals is 1.

(c) Show that $x \in [0,1]$ belongs to C if and only if x has an expansion to base 3 without the digit 1 (Exercise 3.33), that is, there is a sequence (a_n) in $\{0,2\}$ with $x = \sum_{n=1}^{\infty} \frac{a_n}{3^n}$.

(d) Show that C contains no nondegenerate intervals. Conclude that the interior of C (Exercise 4.5) is empty.

(e) Show that for all $x \in C$ and $\epsilon > 0$, there is $y \in C$ with $0 < |x-y| < \epsilon$. In the language of Definition 5.8, this says that C has no isolated points.

(f) Show that C is uncountable. *Hint.* Use Exercise 2.15.

(g) Let $C + C = \{x + y : x, y \in C\}$. Show that $C + C = [0,2]$. *Hint.* Let D be the set of all numbers in $[0,1]$ that have an expansion to base 3 without the digit 2. Start by showing that $[0,1] \subset D + D$.

Continuity

5.1. Limits of functions

5.1. Definition. Let $A \subset \mathbb{R}$, let $f : A \to \mathbb{R}$ be a function, and let c be a *limit point* of A, that is, there is a sequence (x_n) in A such that $x_n \neq c$ for all $n \in \mathbb{N}$ and $x_n \to c$. We say that the *limit* of f at c is $L \in \mathbb{R}$ if for every $\epsilon > 0$, there is $\delta > 0$ such that if $x \in A$ and $0 < |x - c| < \delta$, then $|f(x) - L| < \epsilon$. Then we write $\lim\limits_{x \to c} f(x) = L$ or $f(x) \to L$ as $x \to c$.

5.2. Remark. The limit is unique if it exists. In terms of neighbourhoods, we have $\lim\limits_{x \to c} f(x) = L$ if and only if for every neighbourhood V of L, there is a neighbourhood U of c such that $f(U \cap A \setminus \{c\}) \subset V$.

5.3. Example. (1) The limit points of the open interval (a, b), $a < b$, are precisely the points of the closed interval $[a, b]$. If I is a nondegenerate interval and $c \in I$, then c is a limit point of I and (equivalently) of $I \setminus \{c\}$.

(2) Let us show that $\lim\limits_{x \to 1}(2x + 3) = 5$. First note that $|(2x + 3) - 5| = 2|x - 1|$. Hence, if $\epsilon > 0$ and $|x - 1| < \epsilon/2$, then $|(2x + 3) - 5| < \epsilon$. Thus $\delta = \epsilon/2$ satisfies the definition of the limit.

(3) Showing that $\lim\limits_{x \to 1} \dfrac{1}{x} = 1$ is a bit more complicated. Given ϵ, a corresponding δ is determined in two steps. Note that $\left|\dfrac{1}{x} - 1\right| = \dfrac{1}{|x|}|x - 1|$. We need a bound on the factor $\dfrac{1}{|x|}$ on a neighbourhood of 1, say for $|x-1| < \frac{1}{2}$. If $|x - 1| < \frac{1}{2}$, so $\frac{1}{2} < x < \frac{3}{2}$, then $\dfrac{1}{|x|} < 2$, so $\left|\dfrac{1}{x} - 1\right| \leq 2|x - 1|$. Therefore, if $\epsilon > 0$ is given and we set $\delta = \min\{\frac{\epsilon}{2}, \frac{1}{2}\}$, $0 < |x-1| < \delta$ implies

$\left|\frac{1}{x} - 1\right| \leq 2|x-1| < 2\delta \leq \epsilon$. Here, the first inequality follows from $\delta \leq \frac{1}{2}$, and the third from $\delta \leq \frac{\epsilon}{2}$.

Exercise 5.1. Let $f : A \to \mathbb{R}$ be a function and c be a limit point of A such that $\lim_{x\to c} f(x)$ exists and is not zero. Show that there is a neighbourhood U of c such that $f(x) \neq 0$ for all $x \in U \cap A \setminus \{c\}$.

It is useful to be able to characterise the limit of a function in terms of sequences.

5.4. Theorem. For a function $f : A \to \mathbb{R}$ and a limit point c of A, the following are equivalent.

(i) $\lim_{x\to c} f(x) = L$.

(ii) For every sequence (x_n) in $A\setminus\{c\}$ with $x_n \to c$, we have $f(x_n) \to L$.

Proof. (i) $\Rightarrow$ (ii): Let (x_n) be a sequence in $A \setminus \{c\}$ with $x_n \to c$. We need to show that $f(x_n) \to L$. Let $\epsilon > 0$. By (i), there is $\delta > 0$ such that if $x \in A$ and $0 < |x - c| < \delta$, then $|f(x) - L| < \epsilon$. Since $x_n \to c$, there is $N \in \mathbb{N}$ such that for all $n \geq N$ we have $|x_n - c| < \delta$ and therefore $|f(x_n) - L| < \epsilon$. This proves (ii).

(ii) $\Rightarrow$ (i): We prove the contrapositive. Suppose (i) fails. This means that there is $\epsilon > 0$ such that for every $\delta > 0$, there is $x \in A$ with $0 < |x - c| < \delta$ but $|f(x) - L| \geq \epsilon$. Taking $\delta = \frac{1}{n}$ for each $n \in \mathbb{N}$, we obtain a sequence (x_n) in $A \setminus \{c\}$ with $|x_n - c| < \frac{1}{n}$ for every $n \in \mathbb{N}$, so $x_n \to c$, but $|f(x_n) - L| \geq \epsilon$ for every $n \in \mathbb{N}$, so $f(x_n) \not\to L$. Hence (ii) fails. □

5.5. Example. Define $g : \mathbb{R} \to \mathbb{R}$ by the formula

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

We claim that $\lim_{x\to c} g(x)$ does not exist for any $c \in \mathbb{R}$. Recall that both the rationals and the irrationals are dense in $\mathbb{R}$ (Theorem 2.13 and Exercise 2.4). Hence there is a sequence (r_n) of rationals with $c \neq r_n \to c$, and a sequence (z_n) of irrationals with $c \neq z_n \to c$. Then $g(r_n) = 1 \to 1$ and $g(z_n) = 0 \to 0$, so condition (ii) in Theorem 5.4 fails for every $L \in \mathbb{R}$ and $\lim_{x\to c} g(x)$ does not exist.

By Theorem 5.4, the following two results are immediate consequences of the corresponding results for sequences (Theorems 3.10 and 3.11).

5.6. Theorem (squeeze theorem)**.** Let $f, g, h : A \to \mathbb{R}$ be functions such that $f \leq g \leq h$ on A, and let c be a limit point of A. If $f(x) \to s$ and $h(x) \to s$ as $x \to c$, then $g(x) \to s$ as $x \to c$.

5.7. Theorem (algebraic limit theorem)**.** Let $f, g : A \to \mathbb{R}$ be functions and c be a limit point of A. If $f(x) \to s$ and $g(x) \to t$ as $x \to c$, then, as $x \to c$:

(1) $kf(x) \to ks$ for all $k \in \mathbb{R}$.
(2) $f(x) + g(x) \to s + t$.
(3) $f(x)g(x) \to st$.
(4) $f(x)/g(x) \to s/t$ if $t \neq 0$.
(5) $|f(x)| \to |s|$.

Note that if $t \neq 0$, by Exercise 5.1, there is a neighbourhood U of c such that $g(x) \neq 0$ and $f(x)/g(x)$ is defined for all $x \in U \cap A \setminus \{c\}$.

Proof. We prove (3), just to show how straightforward is the reduction to the algebraic limit theorem for sequences. Let (x_n) be a sequence in $A \setminus \{c\}$ with $x_n \to c$. By assumption, $f(x_n) \to s$ and $g(x_n) \to t$. Hence, by the algebraic limit theorem for sequences, $f(x_n)g(x_n) \to st$. By Theorem 5.4, this shows that $f(x)g(x) \to st$ as $x \to c$. □

Exercise 5.2. Adapt the order limit theorem for limits of sequences (Theorem 3.13) to limits of functions.

Finally, we extend in a trivial way the definition of the limit of a function to a point of the domain that is not a limit point.

5.8. Definition. Let $f : A \to \mathbb{R}$ be a function and let $c \in A$ be an *isolated point* of A, that is, there is $\epsilon > 0$ such that $A \cap (c - \epsilon, c + \epsilon) = \{c\}$. Then we define $\lim\limits_{x \to c} f(x)$ to equal $f(c)$.

5.2. Continuous functions

5.9. Definition. A function $f : A \to \mathbb{R}$ is *continuous* at $c \in A$ if the following equivalent conditions hold.

(i) $\lim\limits_{x \to c} f(x) = f(c)$.
(ii) For every $\epsilon > 0$, there is $\delta > 0$ such that if $x \in A$ and $|x - c| < \delta$, then $|f(x) - f(c)| < \epsilon$.
(iii) For every neighbourhood V of $f(c)$, there is a neighbourhood U of c such that $f(U \cap A) \subset V$.
(iv) If (x_n) is a sequence in A and $x_n \to c$, then $f(x_n) \to f(c)$.

We say that f is *continuous* if f is continuous at each point of A.

Exercise 5.3. Let $A \subset \mathbb{R}$ and $b \in \mathbb{R}$. Prove that the identity function $A \to \mathbb{R}$, $x \mapsto x$, and the constant function $A \to \mathbb{R}$, $x \mapsto b$, are continuous.

5.10. Example. The function $g : \mathbb{R} \to \mathbb{R}$ in Example 5.5 is discontinuous at every point of $\mathbb{R}$.

The algebraic limit theorem (Theorem 5.7) immediately yields the following result.

5.11. Theorem. If $f, g : A \to \mathbb{R}$ are continuous at $c \in A$, then:

(1) kf is continuous at c for all $k \in \mathbb{R}$.

(2) $f + g$ is continuous at c.

(3) fg is continuous at c.

(4) f/g is continuous at c, provided $g(c) \neq 0$.

(5) $|f|$ is continuous at c.

5.12. Definition. A *polynomial function* is a function $P : \mathbb{R} \to \mathbb{R}$ of the form $x \mapsto a_n x^n + \cdots + a_1 x + a_0$, where the coefficients $a_0, \ldots, a_n$ are real numbers. If $a_n \neq 0$, we say that P has *degree* n.

A *rational function* is a function $\mathbb{R} \setminus Z \to \mathbb{R}$ of the form $x \mapsto P(x)/Q(x)$, where P and Q are polynomial functions and Q is not identically zero, so $Z = \{x \in \mathbb{R} : Q(x) = 0\}$ is a finite set (Exercise 6.13).

5.13. Corollary. Rational functions are continuous. In particular, polynomial functions are continuous.

Proof. A rational function is constructed from the identity function and constant functions using the operations of addition, multiplication, and division. Thus the result follows from Exercise 5.3 and Theorem 5.11. □

Exercise 5.4. Use Proposition 3.12 and Exercise 3.4 to show that the square root function $[0, \infty) \to [0, \infty)$, $x \mapsto \sqrt{x}$, is continuous.

In fact, for every $n \in \mathbb{N}$, the n^{th} root function $[0, \infty) \to [0, \infty)$, $x \mapsto \sqrt[n]{x}$, is continuous. Prove this using Theorem 5.26.

The composition of two continuous functions, when defined, is also continuous.

5.14. Theorem. Let $f : A \to \mathbb{R}$ and $g : B \to \mathbb{R}$ be functions such that $f(A) \subset B$, so the composition $g \circ f : A \to \mathbb{R}$ is defined. If f is continuous at $c \in A$, and g is continuous at $f(c)$, then $g \circ f$ is continuous at c.

Proof. Let us prove this using version (iv) of Definition 5.9 (exercise: give alternative proofs based on (ii) and (iii)). The proof is very quick. Let $x_n \to c$ in A. Since f is continuous at c, $f(x_n) \to f(c)$. Since g is continuous at $f(c)$, $(g \circ f)(x_n) = g(f(x_n)) \to g(f(c)) = (g \circ f)(c)$. □

5.15. Example. The function $g : \mathbb{R} \to \mathbb{R}$, $g(x) = \begin{cases} x \sin \frac{1}{x} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0, \end{cases}$ is continuous at 0. Namely, for every $x \in \mathbb{R}$, we have $0 \leq |g(x)| \leq |x|$, so the squeeze theorem (Theorem 5.6) implies that $g(x) \to 0 = g(0)$ as $x \to 0$.

5.3. Continuous functions on compact sets and intervals

There is a close relationship between continuity and compactness.

5.16. Theorem. Let $f : A \to \mathbb{R}$ be continuous. If $K \subset A$ is compact, then the image $f(K)$ is compact.

Proof. Let (y_n) be a sequence in $f(K)$. Say $y_n = f(x_n)$ with $x_n \in K$. Since K is compact, there is a convergent subsequence (x_{n_k}) with limit x in K. Then, since f is continuous, $y_{n_k} = f(x_{n_k}) \to f(x) \in f(K)$. □

The first main theorem of this section is the following.

5.17. Theorem (extreme value theorem)**.** A continuous function $f : K \to \mathbb{R}$ on a nonempty compact subset K of $\mathbb{R}$ has a maximum and a minimum value.

Proof. By Theorem 5.16, $f(K)$ is compact, so $f(K)$ has a largest and a smallest element by Exercise 4.10. □

The extreme value theorem shows that the problem of maximising or minimising a continuous function on a compact set always has a solution. Finding the maximum and minimum values and identifying some or all of the points at which they are taken is another matter, which normally requires differentiation and will be studied in Chapter 6.

5.18. Definition. A function $f : A \to \mathbb{R}$ is *uniformly continuous* on A if for every $\epsilon > 0$, there is $\delta > 0$ such that if $x, y \in A$ and $|x - y| < \delta$, then $|f(x) - f(y)| < \epsilon$.

5.19. Example. (a) The function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = 3x + 2$, is uniformly continuous. Namely, since $|f(x) - f(y)| = 3|x - y|$, given $\epsilon > 0$, we can take $\delta = \epsilon/3$.

(b) The continuous function $g : \mathbb{R} \to \mathbb{R}$, $g(x) = x^2$, is not uniformly continuous. To see this, we first need to negate the definition of uniform continuity. A function $f : A \to \mathbb{R}$ fails to be uniformly continuous on A if there is $\epsilon > 0$ such that for all $\delta > 0$ there are $x, y \in A$ with $|x - y| < \delta$ and $|f(x) - f(y)| \geq \epsilon$. In other words, there is $\epsilon > 0$ and and sequences (x_n), (y_n) in A such that $|x_n - y_n| \to 0$ and $|f(x_n) - f(y_n)| \geq \epsilon$ for all $n \in \mathbb{N}$.

For the function g, perhaps after some experimentation, we come up with $x_n = n$, $y_n = n + \frac{1}{n}$. Then $|x_n - y_n| = \frac{1}{n} \to 0$ and $|g(x_n) - g(y_n)| = y_n^2 - x_n^2 = 2 + \frac{1}{n^2} \geq 2$ for all n.

Uniform continuity is stronger than continuity in that, given $\epsilon > 0$, it requires the existence of a corresponding $\delta > 0$ that works at every point of the domain. However, if the domain is compact, it turns out that the two properties are equivalent.

5.20. Theorem. If $f : K \to \mathbb{R}$ is a continuous function on a compact set K, then f is uniformly continuous on K.

This result will be used later to show that a continuous function on a closed and bounded interval is integrable (Theorem 7.7).

Proof. We argue by contradiction. Suppose f is not uniformly continuous. As we saw in Example 5.19, this means that there is $\epsilon > 0$ and sequences (x_n), (y_n) in K with $|x_n - y_n| \to 0$ and $|f(x_n) - f(y_n)| \geq \epsilon$ for all $n \in \mathbb{N}$. Since K is compact, there is a convergent subsequence $x_{n_k} \to a \in K$. Then $y_{n_k} = x_{n_k} - (x_{n_k} - y_{n_k}) \to a - 0 = a$, so $f(x_{n_k}) \to f(a)$ and $f(y_{n_k}) \to f(a)$, but $|f(x_{n_k}) - f(y_{n_k})| \geq \epsilon$ for all $k \in \mathbb{N}$, which is absurd. □

We now come to the second main theorem of this section.

5.21. Theorem (intermediate value theorem)**.** If $f : [a, b] \to \mathbb{R}$ is continuous and s is a real number between $f(a)$ and $f(b)$, then there is $c \in [a, b]$ with $f(c) = s$.

Proof. Say $f(a) < s < f(b)$. Let $A = \{x \in [a, b] : f(x) \leq s\}$. Then A is nonempty (since $a \in A$) and bounded above (by b), so $c = \sup A$ exists and $c \in [a, b]$. We claim that $f(c) = s$. First, there are $x_n \in A$ with $x_n \to c$. Since $f(x_n) \leq s$ for all n, and $f(x_n) \to f(c)$, we have $f(c) \leq s$. Second, there is a sequence (y_n) in $[a, b]$ with $y_n \to c$ and $y_n \notin A$, that is, $f(y_n) > s$: otherwise, A would be a neighbourhood of c, so there would be numbers in A larger than c. Hence $f(c) = \lim f(y_n) \geq s$. □

5.22. Example. (a) The intermediate value theorem can be used to approximately locate roots of polynomials. Consider for example the polynomial $P(x) = x^7 - 3x^2 + 1$. Since $P(0) = 1$ and $P(1) = -1$, the intermediate value theorem applied to P as a continuous function $[0, 1] \to \mathbb{R}$ shows that P has a root in $(0, 1)$. Since $P(\frac{1}{2}) > 0$, there is even a root in $(\frac{1}{2}, 1)$. Continuing in this manner, we can locate the roots of P as accurately as we wish.

(b) Another application of the intermediate value theorem is the following fixed point theorem. Let $f : [0, 1] \to [0, 1]$ be continuous. Then f has a fixed point, that is, there is $p \in [0, 1]$ such that $f(p) = p$.

Namely, consider the continuous function $g : [0,1] \to \mathbb{R}$, $g(x) = f(x) - x$. Then $g(0) = f(0) \geq 0$ and $g(1) = f(1) - 1 \leq 0$, so g has a zero by the intermediate value theorem. A zero of g is nothing but a fixed point of f.

The intermediate value theorem can be rephrased as follows.

5.23. Corollary. If $f : A \to \mathbb{R}$ is continuous and $I \subset A$ is an interval, then $f(I)$ is an interval.

Proof. Suppose $r < s < t$ with $r, t \in f(I)$. We need to show that $s \in f(I)$ (recall Remark 1.11). There are $a, b \in I$ with $f(a) = r$, $f(b) = t$. Say $a < b$. Theorem 5.21 applied to f restricted to $[a,b]$ shows that there is $c \in [a,b] \subset I$ with $f(c) = s$, so $s \in f(I)$. □

The extreme value theorem says that a continuous function maps a compact set onto a compact set. The intermediate value theorem says that a continuous function maps an interval onto an interval. Together they imply the final result of the section.

5.24. Corollary. A continuous function maps a compact interval onto a compact interval.

Exercise 5.5. Does a continuous function always map an open interval onto an open interval? What about closed intervals? What about bounded intervals?

5.4. Monotone functions

5.25. Definition. A function $f : A \to \mathbb{R}$ is:

- *increasing* if $f(x) \leq f(y)$ whenever $x < y$ in A,
- *strictly increasing* if $f(x) < f(y)$ whenever $x < y$ in A,
- *decreasing* if $f(x) \geq f(y)$ whenever $x < y$ in A,
- *strictly decreasing* if $f(x) > f(y)$ whenever $x < y$ in A,
- *monotone* if it is increasing or decreasing,
- *strictly monotone* if it is strictly increasing or strictly decreasing.

The next result is an application of the intermediate value theorem. The first part illustrates the fact that proving the obvious can be hard. The second part will be used later to prove the inverse function theorem (Theorem 6.7).

5.26. Theorem. Let I be an interval and $f : I \to \mathbb{R}$ be a continuous injection. Then:

(1) f is strictly monotone.

(2) The inverse function $f^{-1} : f(I) \to I$ is continuous.

Exercise 5.6. Show that both parts of the theorem can fail if the domain of the function is not an interval.

Exercise 5.7. Show that the inverse of a strictly increasing function is strictly increasing, and the inverse of a strictly decreasing function is strictly decreasing.

Proof of Theorem 5.26. (1) If $a < b < c$ are points in I and $f(a) < f(b) > f(c)$, take a number t such that $f(a) \le t < f(b) > t \ge f(c)$, for example $t = \max\{f(a), f(c)\}$, and apply the intermediate value theorem to f on $[a, b]$ and on $[b, c]$ to conclude that t is a value of f on both of these intervals, contradicting injectivity of f. Similarly, we cannot have $f(a) > f(b) < f(c)$.

This shows that if $x \in I$ and $u < x < v$, then either $f(u) < f(x) < f(v)$ or $f(u) > f(x) > f(v)$. If there is $u_1 < x$ with $f(u_1) < f(x)$, and $u_2 < x$ with $f(u_2) > f(x)$, then either $u_1 < u_2 < x$ and $f(u_1) < f(u_2) > f(x)$, or $u_2 < u_1 < x$ and $f(u_2) > f(u_1) < f(x)$, which we have just shown to be impossible.

We conclude that if $x \in I$, then either (A) $f(u) < f(x) < f(v)$ for all $u < x < v$ in I, or (B) $f(u) > f(x) > f(v)$ for all $u < x < v$ in I. We need to prove that the same of these two alternatives holds for all x. Suppose this was not the case. Then there would be x_1 for which (A) holds and x_2 for which (B) holds. Say $x_1 < x_2$; the case $x_2 < x_1$ is analogous. Then $f(x_1) < f(x_2)$ since x_1 satisfies (A), and $f(x_1) > f(x_2)$ since x_2 satisfies (B), which is absurd.

In conclusion, we have shown that either (A) holds for all $x \in I$, in which case f is strictly increasing, or (B) holds for all $x \in I$, in which case f is strictly decreasing.

(2) Say f is strictly increasing. Let $c \in I$ and $\epsilon > 0$. To show that f^{-1} is continuous at $f(c)$, we need $\delta > 0$ such that if $y \in f(I)$ and $|y - f(c)| < \delta$, then $|f^{-1}(y) - c| < \epsilon$, that is (writing $y = f(x)$), if $x \in I$ and $|f(x) - f(c)| < \delta$, then $|x - c| < \epsilon$.

Suppose c is not the right end point of I (if I has one), that is, $f(c)$ is not the right end point of the image $f(I)$, which is an interval by the intermediate value theorem. After replacing ϵ by a smaller positive number if necessary, we have $[c, c+\epsilon] \subset I$. Let $\delta = f(c+\epsilon) - f(c) > 0$. Let $x \in I$ with $0 < f(x) - f(c) < \delta$. Then $f(x) \in (f(c), f(c + \epsilon))$. The intermediate value theorem applied to f restricted to $[c, c + \epsilon]$ shows that there is $t \in (c, c + \epsilon)$ with $f(t) = f(x)$. Since f is injective, $x = t$, so $x \in (c, c + \epsilon)$. If c is the left end point of I, this shows that f^{-1} is continuous at $f(c)$.

If c is not the left end point of I, we similarly get $\delta > 0$ such that if $x \in I$ has $0 < f(c) - f(x) < \delta$, then $x \in (c - \epsilon, c)$. This alone proves continuity of f^{-1} at $f(c)$ if c is the right end point of I. If c is not an end point of I, then the two arguments together show that f^{-1} is continuous at $f(c)$. □

More exercises

5.8. Using the ϵ-δ-definition of the limit of a function, show that:

(a) $\lim_{x \to 1}(2x - 1) = 1$.

(b) $\lim_{x \to 1}(x^2 - 1) = 0$.

5.9. Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function. Show that the set $f^{-1}(0) = \{x \in \mathbb{R} : f(x) = 0\}$ is closed.

5.10. Let D be a dense subset of $\mathbb{R}$. If $f, g : \mathbb{R} \to \mathbb{R}$ are continuous functions and $f = g$ on D, prove that $f = g$ on $\mathbb{R}$.

5.11. (a) Fix $a \in \mathbb{R}$ and define $f : \mathbb{R} \to \mathbb{R}$, $f(x) = |x - a|$. Prove that f is continuous at every $c \in \mathbb{R}$.

(b) Let K be a nonempty compact subset of $\mathbb{R}$ and let $a \in \mathbb{R}$. Prove that K has a closest point to a, that is, prove that there is $p \in K$ with $|p - a| \leq |q - a|$ for all $q \in K$.

5.12. A function $f : A \to \mathbb{R}$ is said to be *bounded*, *bounded above*, or *bounded below* if its image $f(A)$ is bounded, bounded above, or bounded below, respectively, as a subset of $\mathbb{R}$.

(a) Show that if f is continuous at $c \in A \subset \mathbb{R}$, then there is a neighbourhood U of c such that f is bounded on $U \cap A$.

(b) Show that if A is compact and f is continuous, then f is bounded.

5.13. Show that the function $g : (0, 1) \to \mathbb{R}$, $x \mapsto \sin \frac{1}{x}$, is not uniformly continuous.

5.14. Let $f, g : \mathbb{R} \to \mathbb{R}$ be uniformly continuous functions. Prove that the composition $g \circ f : \mathbb{R} \to \mathbb{R}$ is uniformly continuous.

5.15. Let $f : A \to \mathbb{R}$, $A \subset \mathbb{R}$, be a uniformly continuous function. Show that if (x_n) is a Cauchy sequence in A, then the image sequence $(f(x_n))$ is also Cauchy. What if f is merely continuous?

5.16. Let $g : [0, 1] \to [0, 1]$ be continuous. Show that there is $a \in [0, 1]$ such that $g(a) + 2a^5 = 3a^7$.

5.17. (a) Let $n \in \mathbb{N}$. Use the intermediate value theorem to show that the function $(0, \infty) \to (0, \infty)$, $x \mapsto x^n$, is surjective. Conclude that every positive real number has a unique positive n^{th} root (see Remark 2.11).

(b) Let $n \in \mathbb{N}$ be odd. Show that the function $\mathbb{R} \to \mathbb{R}$, $x \mapsto x^n$, is bijective. Thus every real number has a unique n^{th} root.

5.18. Let $g : \mathbb{R} \to \mathbb{R}$ be a continuous function such that $g(0) > g(1) < g(2)$. Show that g is not injective.

5.19. What can you say about a continuous function $\mathbb{R} \to \mathbb{R}$ that takes only rational values?

5.20. Show that the function $f : [0,1] \to \mathbb{R}$, $f(x) = \begin{cases} 0 & \text{if } x = 0, \\ \sin \frac{1}{x} & \text{if } x > 0, \end{cases}$ satisfies the intermediate value theorem even though it is not continuous.

5.21. Let $f : [0,1] \to [0,1]$ be continuous. We have seen how to use the intermediate value theorem to prove that f has a fixed point (Example 5.22 (b)). Here is a method for finding a fixed point (or approximating one as closely as we wish) that sometimes works. Let c be any point in $[0,1]$. Recursively define a sequence (x_n) in $[0,1]$ by the equations

$$x_1 = c, \qquad x_{n+1} = f(x_n).$$

Show that *if* (x_n) converges to a limit p, then $f(p) = p$.

5.22. In this exercise we introduce three variants of Definition 5.1.

(a) Let $f : (a, \infty) \to \mathbb{R}$ be a function. Say that $f(x) \to L \in \mathbb{R}$ as $x \to \infty$ if for every $\epsilon > 0$, there is $s > a$ such that if $x > s$, then $|f(x) - L| < \epsilon$. Prove that $1/x \to 0$ as $x \to \infty$.

(b) Let $f : (a, \infty) \to \mathbb{R}$ be a function. Say that $f(x) \to \infty$ as $x \to \infty$ if for every $t \in \mathbb{R}$, there is $s > a$ such that if $x > s$, then $f(x) > t$. Prove that $\sqrt{x} \to \infty$ as $x \to \infty$.

(c) Let $f : (a,b) \setminus \{c\} \to \mathbb{R}$ be a function, where $a < c < b$. Say that $f(x) \to \infty$ as $x \to c$ if for every $t \in \mathbb{R}$, there is $\delta > 0$ such that if $x \in (a,b)$ and $0 < |x - c| < \delta$, then $f(x) > t$. Prove that $1/x^2 \to \infty$ as $x \to 0$.

5.23. Let $g : A \to \mathbb{R}$ be a function on $A \subset \mathbb{R}$ (not necessarily continuous). We say that g is *locally bounded* if every $a \in A$ has a neighbourhood U such that g is bounded on $U \cap A$. Clearly, if g is bounded, then g is locally bounded. Prove that if $K \subset \mathbb{R}$ is compact, then every locally bounded function $K \to \mathbb{R}$ is bounded.

5.24. Let I be an interval. Prove that a monotone function $f : I \to \mathbb{R}$ has only countably many discontinuities, that is, the set of all $c \in I$ such that f is not continuous at c is countable. *Hint.* First show that a discontinuity of f is a 'jump'.

5.25. Let $C \subset [0,1]$ be the Cantor set (Exercise 4.12) and consider the function $h : [0,1] \to \mathbb{R}$, $h(x) = \begin{cases} 1 & \text{if } x \in C, \\ 0 & \text{if } x \notin C. \end{cases}$ Show that h is continuous at $x \in [0,1]$ if and only if $x \notin C$.

Differentiation

6.1. Differentiable functions

6.1. Definition. Let $f : I \to \mathbb{R}$ be a function defined on a nondegenerate interval I. (More generally, we could take I to be any nonempty subset of $\mathbb{R}$ without isolated points.) We say that f is *differentiable* at $a \in I$ if the limit

$$f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a}$$

exists. We call $f'(a)$ the *derivative* of f at a. We say that f is *differentiable* if f is differentiable at every point of I. Then the *derivative* of f is the function $f' : I \to \mathbb{R}$ that maps each $a \in I$ to $f'(a)$. If f' is continuous, then we say that f is *continuously differentiable*.

6.2. Example. (a) A constant function is differentiable at every point, with derivative zero.

(b) Let $n \in \mathbb{N}$. For the monomial function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^n$, we have

$$\frac{f(x) - f(a)}{x - a} = x^{n-1} + ax^{n-2} + \cdots + a^{n-2}x + a^{n-1} \to na^{n-1}$$

as $x \to a$, so f is differentiable with $f'(x) = nx^{n-1}$ for all $x \in \mathbb{R}$.

(c) The absolute value function $g : \mathbb{R} \to \mathbb{R}$, $g(x) = |x|$, is not differentiable at 0. Namely, if $x_n \to 0$ and $x_n > 0$ for all $n \in \mathbb{N}$, then $g(x_n)/x_n = 1 \to 1$, whereas if $x_n \to 0$ and $x_n < 0$ for all $n \in \mathbb{N}$, then $g(x_n)/x_n = -1 \to -1$. Thus $g(x)/x$ does not have a limit as $x \to 0$.

(d) This example shows that the derivative of a differentiable function need not be continuous. The function $h : \mathbb{R} \to \mathbb{R}$,

$$h(x) = \begin{cases} x^2 \cos \frac{1}{x} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0, \end{cases}$$

is differentiable on $\mathbb{R}$. Namely, for $x \neq 0$, $h(x)/x = x \cos \frac{1}{x} \to 0$ as $x \to 0$ by the squeeze theorem, so

$$h'(x) = \begin{cases} 2x \cos \frac{1}{x} + \sin \frac{1}{x} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0, \end{cases}$$

Note that h' is not continuous at 0: $\lim_{x \to 0} h'(x)$ does not exist.

6.3. Proposition. If $f : I \to \mathbb{R}$ is differentiable at $a \in I$, then f is continuous at a.

Proof. We have

$$f(x) - f(a) = \frac{f(x) - f(a)}{x - a}(x - a) \to f'(a) \cdot 0 = 0$$

as $x \to a$. □

The next three theorems are the primary tools that allow us to calculate new derivatives from old.

6.4. Theorem. Let $f, g : I \to \mathbb{R}$ be differentiable at $a \in I$. Then:

(1) $f + g$ is differentiable at a, and $(f + g)'(a) = f'(a) + g'(a)$.

(2) kf is differentiable at a for every $k \in \mathbb{R}$, and $(kf)'(a) = kf'(a)$.

(3) *Product rule:* fg is differentiable at a, and

$$(fg)'(a) = f'(a)g(a) + f(a)g'(a).$$

(4) *Quotient rule:* f/g is differentiable at a if $g(a) \neq 0$, and

$$(f/g)'(a) = \frac{f'(a)g(a) - f(a)g'(a)}{g(a)^2}.$$

Note that since g is differentiable and hence continuous at a, if $g(a) \neq 0$, there is a neighbourhood U of a such that $g(x) \neq 0$ and $f(x)/g(x)$ is defined for all $x \in U \cap I$.

Proof. We prove (3) and leave the other parts as an exercise. We have

$$\begin{aligned} \frac{f(x)g(x) - f(a)g(a)}{x - a} &= \frac{f(x) - f(a)}{x - a} g(x) + \frac{g(x) - g(a)}{x - a} f(a) \\ &\to f'(a)g(a) + f(a)g'(a) \end{aligned}$$

as $x \to a$, using continuity of g at a. □

Exercise 6.1. Finish the proof of Theorem 6.4.

The next result is analogous to Corollary 5.13.

6.5. Corollary. Rational functions are differentiable. In particular, polynomial functions are differentiable.

6.6. Theorem (chain rule)**.** Let I and J be intervals and $f : I \to \mathbb{R}$ and $g : J \to \mathbb{R}$ be functions such that $f(I) \subset J$, so the composition $g \circ f : I \to \mathbb{R}$ is defined. If f is differentiable at $a \in I$ and g is differentiable at $f(a) \in J$, then $g \circ f$ is differentiable at a and

$$(g \circ f)'(a) = g'(f(a))f'(a).$$

Proof. For $x \in I$, $x \neq a$, let $u(x) = \dfrac{f(x) - f(a)}{x - a} - f'(a)$. Then $u(x) \to 0$ as $x \to a$. Define $u(a) = 0$. Then

$$f(x) - f(a) = (x - a)(f'(a) + u(x))$$

for all $x \in I$. For $y \in J$, $y \neq f(a)$, let $v(y) = \dfrac{g(y) - g(f(a))}{y - f(a)} - g'(f(a))$. Then $v(y) \to 0$ as $y \to f(a)$. Define $v(f(a)) = 0$. Then

$$g(y) - g(f(a)) = (y - f(a))(g'(f(a)) + v(y))$$

for all $y \in J$. Hence, for all $x \in I$,

$$\begin{aligned}(g \circ f)(x) - (g \circ f)(a) &= (f(x) - f(a))\big(g'(f(a)) + v(f(x))\big) \\ &= (x - a)(f'(a) + u(x))\big(g'(f(a)) + v(f(x))\big),\end{aligned}$$

so for $x \neq a$,

$$\frac{(g \circ f)(x) - (g \circ f)(a)}{x - a} = (f'(a) + u(x))\big(g'(f(a)) + v(f(x))\big).$$

As $x \to a$, $f(x) \to f(a)$ by Proposition 6.3, so $v(f(x)) \to 0$, and the right-hand side goes to $f'(a)g'(f(a))$. □

6.7. Theorem (inverse function theorem)**.** Let $I \subset \mathbb{R}$ be an interval and $f : I \to \mathbb{R}$ be a continuous injection with inverse $f^{-1} : f(I) \to I$. If f is differentiable at $a \in I$ with $f'(a) \neq 0$, then f^{-1} is differentiable at $f(a) \in f(I)$ with

$$(f^{-1})'(f(a)) = \frac{1}{f'(a)}.$$

Proof. Let (y_n) be a sequence in $f(I) \setminus \{f(a)\}$ with $y_n \to f(a)$. Let $x_n = f^{-1}(y_n) \in I$. By Theorem 5.26, f^{-1} is continuous, so $x_n \to a$. Then

$$\frac{f^{-1}(y_n) - f^{-1}(f(a))}{y_n - f(a)} = \frac{x_n - a}{f(x_n) - f(a)} \to \frac{1}{f'(a)}$$

as $n \to \infty$. □

Exercise 6.2. In Theorem 6.7, could f^{-1} be differentiable at $f(a)$ if $f'(a)$ was 0?

6.8. Example. Let $n \in \mathbb{N}$ and $I = (0, \infty)$. By Example 6.2, the function $f : I \to I$, $f(x) = x^n$, is differentiable with $f'(x) = nx^{n-1} \neq 0$ for all $x \in I$. Also, f is bijective by Exercise 5.17. Hence, by the inverse function theorem, the n^{th} root function $f^{-1} : I \to I$, $x \mapsto x^{1/n}$, is differentiable with

$$(f^{-1})'(x) = \frac{1}{f'(f^{-1}(x))} = \frac{1}{n(x^{1/n})^{n-1}} = \tfrac{1}{n} x^{\frac{1}{n}-1}$$

for all $x > 0$.

The relevance of the derivative to optimisation problems is expressed by the following result.

6.9. Theorem. Suppose a function $f : (a, b) \to \mathbb{R}$ has a maximum or a minimum at a point $c \in (a, b)$. If f is differentiable at c, then $f'(c) = 0$.

Why is it important that the domain of the function be an *open* interval?

Proof. Say f has a maximum at c, that is, $f(c) \geq f(x)$ for all $x \in (a, b)$ (the case of a minimum is analogous). Take a sequence (x_n) in (a, b) with $x_n \to c$ such that $x_n > c$ for all $n \in \mathbb{N}$. Then $x_n - c > 0$ and $f(x_n) - f(c) \leq 0$, so $\dfrac{f(x_n) - f(c)}{x_n - c} \leq 0$ for all $n \in \mathbb{N}$, and

$$f'(c) = \lim_{n\to\infty} \frac{f(x_n) - f(c)}{x_n - c} \leq 0.$$

If we choose (x_n) such that $x_n < c$ for all $n \in \mathbb{N}$, then we conclude in a similar way that $f'(c) \geq 0$. Therefore $f'(c) = 0$. □

6.10. Definition. A *critical point* of a function f is a point c with $f'(c) = 0$.

6.11. Remark. Let $f : [a, b] \to \mathbb{R}$ be differentiable. Since f is continuous and $[a, b]$ is compact, the extreme value theorem (Theorem 5.17) says that f has a maximum and a minimum on $[a, b]$. Theorem 6.9 drastically narrows the search for the *extreme points* of f, that is, the points at which f assumes its maximum or its minimum. The theorem says that the extreme points lie among the critical points of f and the end points a and b.

We end this section by using Theorem 6.9 to show that although derivatives need not be continuous (Example 6.2 (d)), they satisfy the intermediate value theorem (Theorem 5.21).

6.12. Theorem (Darboux's theorem)**.** If $f : [a, b] \to \mathbb{R}$ is differentiable and s is a real number between $f'(a)$ and $f'(b)$, then there is $c \in [a, b]$ with $f'(c) = s$.

Proof. Say $f'(a) < s < f'(b)$. Define $g : [a, b] \to \mathbb{R}$, $g(x) = sx - f(x)$. Then g is differentiable and $g'(x) = s - f'(x)$. We need a zero of g' in $[a, b]$. Since g is continuous, g has a maximum on $[a, b]$ (Theorem 5.17). It cannot be at a since $g'(a) > 0$ (see the proof of Theorem 6.9) and it cannot be at b since $g'(b) < 0$. Thus g has a maximum at a point $c \in (a, b)$, and then $g'(c) = 0$ by Theorem 6.9. □

6.2. The mean value theorem

6.13. Theorem (Rolle's theorem)**.** Let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). If $f(a) = f(b)$, then there is $c \in (a, b)$ with $f'(c) = 0$.

Proof. By the extreme value theorem, f has a maximum and a minimum on $[a, b]$. If both occur at the end points, then f is constant and c can be any point in (a, b). Otherwise, a maximum or a minimum occurs at an interior point $c \in (a, b)$. Then $f'(c) = 0$ by Theorem 6.9. □

The following result is sometimes called the fundamental theorem of differential calculus. It is, at this point, easy to prove, but it has many important applications.

6.14. Theorem (mean value theorem)**.** Let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). Then there is $c \in (a, b)$ with

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Proof. Apply Rolle's theorem to the function

$$x \mapsto f(x) - \frac{f(b) - f(a)}{b - a}(x - a). \qquad \square$$

6.15. Corollary. Let I be an interval and $f : I \to \mathbb{R}$ be differentiable.

(1) f is increasing on I if and only if $f'(x) \geq 0$ for all $x \in I$.

(2) f is decreasing on I if and only if $f'(x) \leq 0$ for all $x \in I$.

(3) f is constant on I if and only if $f'(x) = 0$ for all $x \in I$.

Proof. We prove (1). The proof of (2) is analogous, and (3) is obtained by combining (1) and (2).

$\Rightarrow$ The proof is similar to the proof of Theorem 6.9.

$\Leftarrow$ Suppose $f'(x) \geq 0$ for all $x \in I$. Let $a, b \in I$, $a < b$. By the mean value theorem applied to f restricted to $[a, b]$, there is $c \in (a, b)$ with $f(b) - f(a) = f'(c)(b - a)$. By assumption, $f'(c) \geq 0$, so $f(b) - f(a) \geq 0$. This shows that f is increasing. □

Exercise 6.3. Show that if I is an interval and $f : I \to \mathbb{R}$ is differentiable with $f'(x) > 0$ for all $x \in I$, then f is strictly increasing. Show that the converse may fail.

6.16. Corollary. If f, g are differentiable functions on an interval I, and $f' = g'$ on I, then f and g differ by a constant.

Proof. Apply Corollary 6.15 (3) to $f - g$. □

6.17. Corollary (generalised mean value theorem)**.** If $f, g : [a, b] \to \mathbb{R}$ are continuous on $[a, b]$ and differentiable on (a, b), then there is $c \in (a, b)$ such that

$$(f(b) - f(a))g'(c) = (g(b) - g(a))f'(c).$$

Proof. Apply the mean value theorem to the function

$$x \mapsto (f(b) - f(a))g(x) - (g(b) - g(a))f(x).$$ □

As an application of the generalised mean value theorem, we prove one version of L'Hôpital's rule.

6.18. Theorem (L'Hôpital's rule)**.** If f and g are continuous on an interval I and differentiable on $I \setminus \{a\}$ for some $a \in I$, and $f(a) = g(a) = 0$, then

$$\lim_{x \to a} \frac{f'(x)}{g'(x)} = L \text{ implies } \lim_{x \to a} \frac{f(x)}{g(x)} = L.$$

Proof. It is implicit in the statement of the theorem that $g'(x) \neq 0$ for all $x \in J \cap I \setminus \{a\}$ for some open interval J containing a. It follows by Rolle's theorem that $g(x) \neq g(a) = 0$ for all $x \in J \cap I \setminus \{a\}$.

Let (x_n) be a sequence in $J \cap I \setminus \{a\}$ with $x_n \to a$. For each $n \in \mathbb{N}$, we apply Corollary 6.17 to f and g restricted to the interval between a and x_n (note that x_n may be smaller or larger than a). We obtain t_n strictly between a and x_n with $(f(x_n) - f(a))g'(t_n) = (g(x_n) - g(a))f'(t_n)$, that is,

$$\frac{f(x_n)}{g(x_n)} = \frac{f'(t_n)}{g'(t_n)}$$

for each $n \in \mathbb{N}$ (note that the denominators are not 0). If $n \to \infty$, then $x_n \to a$, so $t_n \to a$, and by assumption, $f(x_n)/g(x_n) \to L$. □

More exercises

6.4. Prove directly from the definition of the derivative that the derivative of the function $x \mapsto 1/x^2$ at $c \neq 0$ is $-2/c^3$.

6.5. Let $g : \mathbb{R} \to \mathbb{R}$ be a function. Suppose g is differentiable at 0 with $g'(0) > 0$. Show that there is $\delta > 0$ such that if $0 < x < \delta$, then $g(x) > g(0)$.

6.6. Let $a \in \mathbb{R}$ and define $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \begin{cases} a & \text{if } x < 0, \\ x & \text{if } x \geq 0. \end{cases}$ For which values of a is there a differentiable function $g : \mathbb{R} \to \mathbb{R}$ with $g' = f$?

6.7. Let $A \subset \mathbb{R}$ be *symmetric* about 0, that is, $x \in A$ if and only if $-x \in A$. A function $f : A \to \mathbb{R}$ is called *even* if $f(-x) = f(x)$ for all $x \in A$, and *odd* if $f(-x) = -f(x)$ for all $x \in A$.

Suppose A is an interval and f is differentiable. Prove that if f is even, then f' is odd. Prove that if f is odd, then f' is even.

6.8. (a) Consider the function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^3 + x + 1$. Prove that f is injective.

(b) Explain why the inverse function $f^{-1} : f(\mathbb{R}) \to \mathbb{R}$ is differentiable. Calculate $(f^{-1})'(3)$.

6.9. Let $f : \mathbb{R} \to \mathbb{R}$ be a differentiable function whose derivative is bounded, that is, there is $M > 0$ such that $|f'(x)| \leq M$ for all $x \in \mathbb{R}$. Show that f is uniformly continuous.

6.10. Show that the polynomial $x^7 + x^5 + x^3 + x + 1000$ has exactly one root.

6.11. Let $f : \mathbb{R} \to \mathbb{R}$ be differentiable with $f'(x) \geq 0$ for all $x \in \mathbb{R}$. Show that if f is not constant on any nonempty open interval, then f is strictly increasing.

6.12. Let I be an interval and $f : I \to \mathbb{R}$ be differentiable. Show that if f has two distinct fixed points on I, then there is $c \in I$ with $f'(c) = 1$.

6.13. Use Rolle's theorem and induction to show that a polynomial of degree n has at most n roots.

6.14. Define a function $g : \mathbb{R} \to \mathbb{R}$ by the formula

$$g(x) = \begin{cases} \dfrac{x}{2} + x^2 \sin \dfrac{1}{x} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

Show that g is differentiable on $\mathbb{R}$, $g'(0) > 0$, but g is not increasing on any open interval containing 0.

6.15. Prove that if $f : [0, \infty) \to \mathbb{R}$ is differentiable, $f(0) = 0$, and $f'(x) \geq 1$ for every $x > 0$, then $f(x) \geq x$ for every $x > 0$.

6.16. A real-valued function f on an open interval I is said to be *convex* if for all $x, y \in I$ and $t \in (0, 1)$,

$$f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y).$$

This means that the line segment joining any two points on the graph of f lies on or above the graph. We say that f is *concave* if $-f$ is convex.

(a) Suppose f is differentiable. Prove that f is convex if and only if f' is increasing.

(b) Suppose f is twice differentiable. Prove that f is convex if and only if $f''(x) \geq 0$ for all $x \in I$.

(c) Let $u : \mathbb{R} \to \mathbb{R}$ be convex, increasing, and twice differentiable. Show that if f is convex and twice differentiable, then $u \circ f$ is convex.

Integration

7.1. The Riemann integral

7.1. Definition. A *partition* P of a compact interval $[a, b]$, where $a < b$, is a finite subset of $[a, b]$ including the end points, with elements

$$a = x_0 < x_1 < \cdots < x_n = b.$$

A partition Q of $[a, b]$ is a *refinement* of P if $P \subset Q$.

Let $f : [a, b] \to \mathbb{R}$ be a bounded function. For a partition P of $[a, b]$ as above, let

$$m_k = \inf\{f(x) : x \in [x_{k-1}, x_k]\},$$
$$M_k = \sup\{f(x) : x \in [x_{k-1}, x_k]\}$$

for $k = 1, \ldots, n$. The *lower sum* of f with respect to P is

$$L(f, P) = \sum_{k=1}^{n} m_k(x_k - x_{k-1}).$$

The *upper sum* of f with respect to P is

$$U(f, P) = \sum_{k=1}^{n} M_k(x_k - x_{k-1}).$$

7.2. Lemma. Let $f : [a, b] \to \mathbb{R}$ be a bounded function.

(1) For every partition P of $[a, b]$, $L(f, P) \le U(f, P)$.

(2) If Q is a refinement of P, then $L(f, P) \le L(f, Q)$ and $U(f, P) \ge U(f, Q)$.

(3) If P_1 and P_2 are partitions of $[a, b]$, then $L(f, P_1) \le U(f, P_2)$.

Proof. (1) follows immediately from $m_k \le M_k$.

(2) Transform P to Q by adding one point at a time. If a new point is added to P, say y between x_{k-1} and x_k, then the term $m_k(x_k - x_{k-1})$ in $L(f,P)$ is replaced by $m'(y - x_{k-1}) + m''(x_k - y)$, where $m' = \inf_{[x_{k-1},y]} f$ and $m'' = \inf_{[y,x_k]} f$. Note that $m', m'' \ge m_k$ (the infimum of a smaller set is larger). Argue similarly for upper sums.

(3) follows from (1) and (2) using the common refinement $P_1 \cup P_2$ of P_1 and P_2. □

7.3. Definition. A bounded function $f : [a,b] \to \mathbb{R}$ is *integrable* (or *Riemann integrable*) if its *lower integral*

$$L(f) = \sup\{L(f,P) : P \text{ is a partition of } [a,b]\}$$

equals its *upper integral*

$$U(f) = \inf\{U(f,P) : P \text{ is a partition of } [a,b]\}.$$

The common value of $U(f)$ and $L(f)$ is then called the *integral of f over* $[a,b]$, and denoted $\int_a^b f$ or $\int_a^b f(x)dx$.

Roughly speaking, integrability of f means that there is no gap between the lower sums of f and the upper sums of f. The unique number that separates all the lower sums from all the upper sums is the integral of f. In view of Lemma 7.2, the following characterisation is immediate.

7.4. Lemma. A bounded function $f : [a,b] \to \mathbb{R}$ is integrable if and only if for every $\epsilon > 0$, there is a partition P of $[a,b]$ such that $U(f,P) - L(f,P) < \epsilon$.

7.5. Example. (a) Let $f : [a,b] \to \mathbb{R}$ be a constant function, say $f(x) = c$ for all $x \in [a,b]$. For every partition P of $[a,b]$, $m_k = M_k = c$, so $L(f,P) = U(f,P) = c(b-a)$. Hence f is integrable with $\int_a^b f = \int_a^b c\,dx = c(b-a)$.

(b) Consider the function $f : [0,1] \to \mathbb{R}$, $f(x) = x^2$. For $n \in \mathbb{N}$, take the partition $P_n = \{0, \frac{1}{n}, \frac{2}{n}, \dots, 1\}$ of $[0,1]$. Then $M_k = (k/n)^2$, so

$$U(f,P_n) = \sum_{k=1}^{n} \left(\frac{k}{n}\right)^2 \frac{1}{n} = \frac{1}{n^3} \sum_{k=1}^{n} k^2 = \frac{1}{n^3}\frac{1}{6} n(n+1)(2n+1) \to \frac{1}{3}$$

as $n \to \infty$. A similar computation shows that $L(f,P_n) \to \frac{1}{3}$ as $n \to \infty$. This shows that f is integrable with $\int_0^1 f = \int_0^1 x^2\,dx = \frac{1}{3}$.

The next two theorems provide a big supply of integrable functions.

7.6. Theorem. A monotone function $f : [a,b] \to \mathbb{R}$ is integrable.

Proof. First note that since f is monotone, it is bounded: all its values lie between $f(a)$ and $f(b)$. Say f is increasing. Let $\epsilon > 0$ and choose $\delta > 0$ such that $\delta(f(b) - f(a)) < \epsilon$. Let P be a partition of $[a, b]$ fine enough that $x_k - x_{k-1} < \delta$ for $k = 1, \ldots, n$. Then

$$\begin{aligned} U(f, P) - L(f, P) &= \sum_{k=1}^{n} (f(x_k) - f(x_{k-1}))(x_k - x_{k-1}) \\ &\leq \delta \sum_{k=1}^{n} (f(x_k) - f(x_{k-1})) = \delta(f(b) - f(a)) < \epsilon. \end{aligned}$$

By Lemma 7.4, f is integrable. □

7.7. Theorem. A continuous function $f : [a, b] \to \mathbb{R}$ is integrable.

Proof. By the extreme value theorem (Theorem 5.17), since $[a, b]$ is compact and f is continuous, f is bounded. Let $\epsilon > 0$. Since f is uniformly continuous (Theorem 5.20), there is $\delta > 0$ such that if $|x - y| < \delta$, then $|f(x) - f(y)| < \dfrac{\epsilon}{b - a}$. Let P be a partition of $[a, b]$ fine enough that $x_k - x_{k-1} < \delta$ for $k = 1, \ldots, n$. Again by the extreme value theorem, f has a maximum and a minimum on $[x_{k-1}, x_k]$, say $M_k = f(y_k)$, $m_k = f(z_k)$ for some $y_k, z_k \in [x_{k-1}, x_k]$. Then $|y_k - z_k| < \delta$, so $M_k - m_k < \dfrac{\epsilon}{b - a}$. Thus

$$U(f, P) - L(f, P) = \sum_{k=1}^{n} (M_k - m_k)(x_k - x_{k-1}) < \frac{\epsilon}{b - a}(b - a) = \epsilon.$$

By Lemma 7.4, f is integrable. □

7.8. Example. This example shows that a discontinuous function may or may not be integrable.

(a) Let $f : [-1, 1] \to \mathbb{R}$ equal 1 at 0 and equal 0 elsewhere. For $n \in \mathbb{N}$, consider the partition $P_n = \{-1, -\frac{1}{2n}, \frac{1}{2n}, 1\}$ of $[-1, 1]$. Then $L(f, P_n) = 0$ and $U(f, P_n) = \frac{1}{n} \to 0$ as $n \to \infty$. Thus f is integrable with $\int_{-1}^{1} f = 0$, even though f is discontinuous.

(b) Let $f : [0, 1] \to \mathbb{R}$ equal 1 on the rationals in $[0, 1]$ and 0 on the irrationals. By density of $\mathbb{Q}$ and $\mathbb{R} \setminus \mathbb{Q}$ in $\mathbb{R}$, for every partition P of $[0, 1]$, we have $m_k = 0$ and $M_k = 1$ for all k, so $L(f, P) = 0$ and $U(f, P) = 1$. Thus f is not integrable.

7.9. Theorem. Let $f : [a, b] \to \mathbb{R}$ be bounded and $c \in (a, b)$. Then f is integrable on $[a, b]$ if and only if f is integrable on $[a, c]$ and on $[c, b]$, and then

$$\int_a^b f = \int_a^c f + \int_c^b f.$$

Exercise 7.1. Prove Theorem 7.9.

7.10. Remark. If f is integrable on $[a,b]$, we define

$$\int_b^a f = -\int_a^b f.$$

Also, for $c \in [a,b]$, we define $\int_c^c f = 0$.

Then, if I is a compact interval and $f : I \to \mathbb{R}$ is integrable,

$$\int_a^b f + \int_b^c f = \int_a^c f$$

for any three points $a, b, c \in I$. We leave the verification as an exercise.

7.11. Theorem. Suppose f and g are integrable on $[a,b]$. Then:

(1) $f+g$ is integrable on $[a,b]$ with $\int_a^b (f+g) = \int_a^b f + \int_a^b g$.

(2) For every $k \in \mathbb{R}$, kf is integrable on $[a,b]$ with $\int_a^b (kf) = k\int_a^b f$.

(3) If $f \le g$ on $[a,b]$, then $\int_a^b f \le \int_a^b g$.

(4) $|f|$ is integrable on $[a,b]$ and $|\int_a^b f| \le \int_a^b |f|$.

Proof. The tricky parts are (1) and (4). We leave (2) and (3) as exercises.

(1) The proof hinges on the fact that for any $A \subset [a,b]$,

$$\inf_A f + \inf_A g \le \inf_A (f+g), \quad \sup_A (f+g) \le \sup_A f + \sup_A g.$$

Thus, for any partition P of $[a,b]$,

$$L(f,P) + L(g,P) \le L(f+g,P), \quad U(f+g,P) \le U(f,P) + U(g,P).$$

Take $\epsilon > 0$. By definition of the upper integral, there are partitions P_1 and P_2 of $[a,b]$ such that

$$U(f,P_1) \le U(f) + \epsilon/2, \quad U(g,P_2) \le U(g) + \epsilon/2,$$

so

$$\begin{aligned} U(f+g) &\le U(f+g, P_1 \cup P_2) \le U(f, P_1 \cup P_2) + U(g, P_1 \cup P_2) \\ &\le U(f,P_1) + U(g,P_2) \le U(f) + U(g) + \epsilon. \end{aligned}$$

Similarly,

$$L(f+g) \ge L(f) + L(g) - \epsilon,$$

so

$$L(f) + L(g) - \epsilon \le L(f+g) \le U(f+g) \le U(f) + U(g) + \epsilon.$$

Since this holds for every $\epsilon > 0$,

$$L(f) + L(g) \le L(f+g) \le U(f+g) \le U(f) + U(g).$$

Finally, integrability of f and g (which we have not used so far) implies that the smallest and the largest of these four numbers are equal, so all four are equal, and equal to $\int_a^b f + \int_a^b g$.

(4) Let $A \subset [a,b]$. For $x, y \in A$, by the triangle inequality,

$$\begin{aligned} |f(x)| - |f(y)| &\le |f(x) - f(y)| \\ &= f(x) - f(y) \text{ or } f(y) - f(x) \le \sup_A f - \inf_A f. \end{aligned}$$

Hence, for each $y \in A$,

$$|f(x)| \le \sup_A f - \inf_A f + |f(y)| \text{ for all } x \in A,$$

so, taking the supremum over $x \in A$,

$$\sup_A |f| \le \sup_A f - \inf_A f + |f(y)|,$$

and

$$\sup_A f - \inf_A f \ge \sup_A |f| - |f(y)| \ge \sup_A |f| - \inf_A |f|.$$

This shows that if P is a partition of $[a,b]$, then

$$U(|f|, P) - L(|f|, P) \le U(f, P) - L(f, P).$$

By the assumption that f is integrable, for every $\epsilon > 0$, there is P with $U(f,P) - L(f,P) < \epsilon$, so $U(|f|,P) - L(|f|,P) < \epsilon$. Hence $|f|$ is integrable.

Now $-|f| \le f \le |f|$, so $-\int_a^b |f| \le \int_a^b f \le \int_a^b |f|$ by (3), which gives $|\int_a^b f| \le \int_a^b |f|$. □

Exercise 7.2. Finish the proof of Theorem 7.11.

7.2. The fundamental theorem of calculus

This central theorem says that the operations of differentiation and integration are, in a sense, inverse to each other.

7.12. Theorem (fundamental theorem of calculus)**.**

(1) If $f : [a,b] \to \mathbb{R}$ is integrable and $F : [a,b] \to \mathbb{R}$ is differentiable with $F'(x) = f(x)$ for all $x \in [a,b]$, then

$$\int_a^b f = F(b) - F(a).$$

(2) Let $g : [a,b] \to \mathbb{R}$ be integrable and define

$$G(x) = \int_a^x g, \qquad x \in [a,b].$$

Then G is continuous on $[a,b]$. If g is continuous at $c \in [a,b]$, then G is differentiable at c and $G'(c) = g(c)$.

7.13. Definition. In (1) above, F is called an *antiderivative* of f. In (2), G is called an *indefinite integral* of g.

7.14. Remark. We know that not every derivative is continuous (Example 6.2 (d)). Theorem 7.12 says that every continuous function is a derivative.

Proof. (1) Let P be a partition of $[a,b]$. The mean value theorem applied to F on $[x_{k-1},x_k]$ yields $t_k \in (x_{k-1},x_k)$ with

$$F(x_k)-F(x_{k-1}) = F'(t_k)(x_k-x_{k-1}) = f(t_k)(x_k-x_{k-1}).$$

Since $m_k \le f(t_k) \le M_k$, we have

$$L(f,P) \le \sum f(t_k)(x_k-x_{k-1}) \le U(f,P).$$

The sum is a telescoping sum, equal to $F(b)-F(a)$, so $\int_a^b f = F(b)-F(a)$.

(2) Say $|g| \le M$ on $[a,b]$. Then, for $x,y \in [a,b]$,

$$|G(x)-G(y)| = \left|\int_a^x g - \int_a^y g\right| = \left|\int_x^y g\right| \le \left|\int_x^y |g|\right| \le M|x-y|.$$

(The outer vertical bars in $|\int_x^y |g||$ are needed in case $x>y$.) This shows that G is uniformly continuous on $[a,b]$ (given $\epsilon>0$, take $\delta=\epsilon/M$).

Now suppose g is continuous at $c \in [a,b]$. For $x \ne c$,

$$g(c) = \frac{1}{x-c}\int_c^x g(c)\,dt$$

and

$$\frac{G(x)-G(c)}{x-c} = \frac{1}{x-c}\int_c^x g(t)\,dt.$$

Let $\epsilon>0$ and find $\delta>0$ such that $|g(t)-g(c)|<\epsilon$ if $|t-c|<\delta$. Then, if $0<|x-c|<\delta$,

$$\begin{aligned}\left|\frac{G(x)-G(c)}{x-c} - g(c)\right| &= \left|\frac{1}{x-c}\int_c^x \big(g(t)-g(c)\big)\,dt\right| \\ &\le \frac{1}{x-c}\int_c^x |g(t)-g(c)|\,dt \le \epsilon.\end{aligned}$$

This shows that $G'(c)=g(c)$. □

7.15. Remark. Calculating integrals directly from the definition of the integral is almost never possible in practice. The benefit of being able to compute integrals using antiderivatives cannot be overestimated.

7.16. Corollary (mean value theorem for integrals)**.** If $g:[a,b]\to\mathbb{R}$ is continuous, then there is $c\in(a,b)$ such that

$$\int_a^b g = (b-a)g(c).$$

Proof. Apply the mean value theorem to the function $x \mapsto \int_a^x g$ on $[a, b]$, which, by the fundamental theorem of calculus, is an antiderivative for g. □

7.3. The natural logarithm and the exponential function

The fundamental theorem of calculus allows us to conveniently and rigorously define some important functions as indefinite integrals. In this section, we introduce the natural logarithm and its inverse, the exponential function.

7.17. Definition. The *natural logarithm* (or simply *logarithm*) is the function

$$\log : (0, \infty) \to \mathbb{R}, \quad \log x = \int_1^x \frac{dt}{t}.$$

By the fundamental theorem of calculus, log is differentiable with $\log'(x) = 1/x$ for all $x > 0$. In fact, log is the unique antiderivative of the reciprocal function on $(0, \infty)$ that satisfies $\log 1 = 0$. Since $\log'(x) > 0$ for all $x > 0$, log is strictly increasing (Exercise 6.3) and hence injective. For $n \in \mathbb{N}$, $n \geq 2$,

$$\log n = \int_1^n \frac{dt}{t} = \sum_{k=2}^{n} \int_{k-1}^{k} \frac{dt}{t} \geq \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}.$$

Since the harmonic series diverges (Example 3.23), log is unbounded above. Similarly, for $k \in \mathbb{N}$,

$$\int_{\frac{1}{k}}^{\frac{1}{k-1}} \frac{dt}{t} \geq (k-1)\left(\frac{1}{k-1} - \frac{1}{k}\right) = \frac{1}{k},$$

so for $n \in \mathbb{N}$, $n \geq 2$,

$$\log \frac{1}{n} \leq -\frac{1}{2} - \frac{1}{3} - \cdots - \frac{1}{n}.$$

Hence, log is unbounded below as well. Thus, by the intermediate value theorem, the range of log is $\mathbb{R}$, so log is a bijection $(0, \infty) \to \mathbb{R}$.

7.18. Definition. The number e is the unique number with $\log e = 1$.

In the language of group theory, the following result says that the logarithm is a group isomorphism from the multiplicative group of positive real numbers to the additive group of all real numbers.

7.19. Theorem. For all $x, y > 0$, $\log(xy) = \log x + \log y$.

Proof. Fix $y > 0$ and define $f : (0, \infty) \to \mathbb{R}$, $f(x) = \log(xy)$. Then f is differentiable with

$$f'(x) = y \log'(xy) = y \frac{1}{xy} = \frac{1}{x} = \log'(x)$$

for all $x > 0$, so by Corollary 6.16, there is $c \in \mathbb{R}$ with $f = \log + c$. Evaluating at 1 gives $\log y = f(1) = \log 1 + c = c$. □

7.20. Definition. The *exponential function* $\exp : \mathbb{R} \to (0, \infty)$ is the inverse of log.

Since log is strictly increasing, so is exp (Exercise 5.7). Theorem 7.19 immediately yields

$$\exp(x + y) = (\exp x)(\exp y)$$

for all $x, y \in \mathbb{R}$. This, along with the definition of the number $e = \exp 1$, gives $\exp n = e^n$ for all $n \in \mathbb{Z}$. This identity easily extends to rational exponents and can then be taken as the *definition* of e^x for irrational x. Subsequently, for $a > 0$ and x irrational, a^x can be defined to be $e^{x \log a}$.

Exercise 7.3. Let $c \in \mathbb{R}$ and $f : (0, \infty) \to \mathbb{R}$, $f(x) = x^c = e^{c \log x}$. Show that $f'(x) = cx^{c-1}$ for all $x > 0$.

By the inverse function theorem (Theorem 6.7), exp is differentiable on $\mathbb{R}$ with

$$\exp'(x) = \frac{1}{\log'(\exp x)} = \exp x$$

for all $x \in \mathbb{R}$.

7.21. Theorem. The exponential function is the unique differentiable function $f : \mathbb{R} \to \mathbb{R}$ with $f' = f$ and $f(0) = 1$.

Proof. Let f be another such function. Then

$$(f/\exp)' = (f' \exp - f \exp')/\exp^2 = 0,$$

so $f/\exp$ is constant. Evaluating at 0 shows that the constant is 1. □

Finally, let us derive an explicit expression for the number e, which allows us to approximate it as closely as we wish. As $n \to \infty$,

$$\log(1 + \tfrac{1}{n})^n = n \log(1 + \tfrac{1}{n}) = \frac{\log(1 + \frac{1}{n}) - \log 1}{\frac{1}{n}} \to \log'(1) = 1$$

(the first equality follows from Theorem 7.19), so since exp is continuous,

$$(1 + \tfrac{1}{n})^n = \exp \log(1 + \tfrac{1}{n})^n \to \exp 1 = e.$$

Thus

$$e = \lim_{n \to \infty} (1 + \tfrac{1}{n})^n.$$

More exercises

7.4. Prove directly from the definition of the Riemann integral that the function $f : [0,1] \to \mathbb{R}$, $f(x) = 2x + 1$, is integrable with $\int_0^1 f = 2$.

7.5. Let $f(x) = 0$ for $x \leq 0$ and $f(x) = 1$ for $x > 0$. Define $F(x) = \int_0^x f$, $x \in \mathbb{R}$. Find a formula for $F(x)$. Where is F continuous? Where is F differentiable? Where is $F'(x) = f(x)$?

7.6. Prove each of the following statements about a function $f : [a,b] \to \mathbb{R}$ or disprove it by a counterexample.

(a) If $|f|$ is integrable on $[a,b]$, then so is f.

(b) If $\int_a^b f > 0$, then there is an interval $[c,d] \subset [a,b]$, $c < d$, and $\delta > 0$ such that $f > \delta$ on $[c,d]$.

(c) If $f \geq 0$ on $[a,b]$ and $f(c) > 0$ for some $c \in [a,b]$, then $\int_a^b f > 0$.

(d) If f is continuous, $f \geq 0$ on $[a,b]$, and $f(c) > 0$ for some $c \in [a,b]$, then $\int_a^b f > 0$.

7.7. Let $a < c < d < b$. Prove that if $f : [a,b] \to \mathbb{R}$ is integrable, then f is integrable on $[c,d]$.

7.8. Let $f : [a,b] \to \mathbb{R}$ be an integrable function and $g : \mathbb{R} \to \mathbb{R}$ be a continuously differentiable function. Prove that the composition $g \circ f : [a,b] \to \mathbb{R}$ is integrable. *Hint.* Use the mean value theorem (Theorem 6.14) to compare $U(g \circ f, P) - L(g \circ f, P)$ and $U(f,P) - L(f,P)$.

7.9. Let $f, g : [a,b] \to \mathbb{R}$ be functions, not necessarily continuous, such that g is integrable, $\int_a^b g = 0$, and $0 \leq f(x) \leq g(x)$ for all $x \in [a,b]$. Prove that f is integrable with $\int_a^b f = 0$.

7.10. Let $f : [a,b) \to \mathbb{R}$ be a bounded function which is Riemann integrable on $[a,c]$ whenever $a < c < b$. Define the function $F : [a,b) \to \mathbb{R}$ by the formula $F(x) = \int_a^x f$. Prove that F has a limit at b. The integral (or *improper integral*) of f over $[a,b)$ can be defined to equal this limit. *Hint.* Start by showing that if (x_n) is a sequence in $[a,b)$ with $x_n \to b$, then $\left(\int_a^{x_n} f\right)$ is a Cauchy sequence.

7.11. Let $f : [0,1] \to \mathbb{R}$ be continuous and suppose that $\int_0^x f = \int_x^1 f$ for all $x \in [0,1]$. Show that $f(x) = 0$ for all $x \in [0,1]$.

7.12. Let $f : \mathbb{R} \to \mathbb{R}$ be a differentiable function such that $f'(x) > 0$ for all $x \in \mathbb{R}$ and $f(0) = 0$. Show that for all $x \in \mathbb{R}$,

$$\int_0^{f(x)} \frac{dt}{f'(f^{-1}(t))} = x.$$

7.13. Let $f, g : [a, b] \to \mathbb{R}$ be differentiable functions such that the functions $f'g$ and fg' are integrable (this holds in particular if f' and g' are continuous). Prove the formula for *integration by parts:*

$$\int_a^b f'g = f(b)g(b) - f(a)g(a) - \int_a^b fg'.$$

7.14. Prove the following version of the formula for a *change of variables*, also known as *substitution.* Let $f : [c, d] \to \mathbb{R}$ be continuous. Let $\phi : [a, b] \to [c, d]$ be continuously differentiable. Then

$$\int_{\phi(a)}^{\phi(b)} f = \int_a^b (f \circ \phi)\phi'.$$

7.15. We have defined $\log x = \int_1^x dt/t$ for $x > 0$. Use substitution to prove directly from this definition that $\log(xy) = \log x + \log y$ for all $x, y > 0$. (We gave a different proof of this important identity in Section 7.3.)

7.16. For each $n \in \mathbb{N}$, define $\gamma_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n} - \log n$. Prove that the sequence (γ_n) converges. *Hint.* Use the monotone convergence theorem (Theorem 3.16). The limit $\gamma = \lim \gamma_n = 0.5772156649\ldots$ is called *Euler's constant.*

7.17. Let $\lambda : (0, \infty) \to \mathbb{R}$ be a differentiable function such that $\lambda'(1) = 1$ and $\lambda(xy) = \lambda(x) + \lambda(y)$ for all $x, y > 0$. Show that $\lambda = \log$.

7.18. (a) Let $n \in \mathbb{N}$. Prove that $x^n/e^x \to 0$ as $x \to \infty$ (see Exercise 5.22). *Hint.* Start by observing that if $x \geq m \geq 1$, then $\log x = \log m + \int_m^x dt/t \leq \log m + x/m - 1$, so $\log x - x/m$ is bounded above on $[m, \infty)$. Hence x^m/e^x is bounded above on $[m, \infty)$.

(b) Deduce that for every $n \in \mathbb{N}$, $(\log x)^n/x \to 0$ as $x \to \infty$ and $x(\log x)^n \to 0$ as $x \to 0$.

7.19. (a) Let $f : [1, \infty) \to [0, \infty)$ be decreasing and integrable on $[1, n]$ for every $n \in \mathbb{N}$. Prove that the sequence $(\int_1^n f)$ converges if and only if the series $\sum_{n=1}^{\infty} f(n)$ converges. This result is known as the *integral test. Hint.* Observe that for each $n \in \mathbb{N}$, $f(n+1) \leq \int_n^{n+1} f \leq f(n)$. Use the comparison test (Proposition 3.26).

(b) Use the integral test to show that $\sum 1/n$ diverges and $\sum 1/n^2$ converges.

(c) Determine whether the series $\sum_{n=2}^{\infty} \frac{1}{n \log n}$ and $\sum_{n=2}^{\infty} \frac{1}{n(\log n)^2}$ converge.

7.20. Show that the function $h : [0, 1] \to \mathbb{R}$ in Exercise 5.25 is integrable even though it has uncountably many discontinuities. What is $\int_0^1 h$?

Sequences and series of functions

8.1. Pointwise and uniform convergence

We will consider two notions of convergence for sequences and series of functions.

8.1. Definition. Suppose that for each $n \in \mathbb{N}$ we have a function $f_n : A \to \mathbb{R}$. The functions are all defined on the same domain $A \subset \mathbb{R}$. We say that the sequence $(f_n)_{n\in\mathbb{N}}$ *converges pointwise* on A to a function $f : A \to \mathbb{R}$ if, for every $x \in A$, the sequence $(f_n(x))$ of real numbers converges to $f(x)$.

8.2. Example. (a) Let $f_n : [0,1] \to \mathbb{R}$, $f_n(x) = x^n$. For each $x \in [0,1]$,

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

as $n \to \infty$. The pointwise limit function f is not continuous, even though all the functions f_n are.

(b) Let $g_n : \mathbb{R} \to \mathbb{R}$, $g_n(x) = x^{1+\frac{1}{2n-1}} = x\sqrt[2n-1]{x}$. For all $x \in \mathbb{R}$,

$$g_n(x) \to g(x) = \left\{ \begin{array}{cc} x \cdot 1 = x & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ x \cdot (-1) = -x & \text{if } x < 0 \end{array} \right\} = |x|$$

as $n \to \infty$. The pointwise limit function g is not differentiable, even though all the functions g_n are.

(c) Let $h_n : [0,1] \to \mathbb{R}$,

$$h_n(x) = \begin{cases} 4n^2x & \text{if } 0 \le x \le \frac{1}{2n}, \\ -4n^2x + 4n & \text{if } \frac{1}{2n} \le x \le \frac{1}{n}, \\ 0 & \text{if } x \ge \frac{1}{n} \end{cases}$$

(draw the graph!). Then h_n is integrable and $\int_0^1 h_n = 1$ for each $n \in \mathbb{N}$. For each $x \in [0,1]$, $h_n(x) \to 0$ as $n \to \infty$, so $h_n \to 0$ pointwise. Thus $\lim_{n\to\infty} \int_0^1 h_n \neq \int_0^1 \lim_{n\to\infty} h_n$.

As these examples illustrate, pointwise convergence is too weak to interact well with continuity, differentiability, and integration. The following stronger notion of convergence has better properties.

8.3. Definition. Let $A \subset \mathbb{R}$ and $f_n : A \to \mathbb{R}$ for each $n \in \mathbb{N}$. The sequence (f_n) *converges uniformly* on A to $f : A \to \mathbb{R}$ if for every $\epsilon > 0$, there is $N \in \mathbb{N}$ such that $|f_n(x) - f(x)| < \epsilon$ for all $x \in A$ and all $n \ge N$.

Equivalently, for every $\epsilon > 0$, there is $N \in \mathbb{N}$ such that $\sup_A |f_n - f| < \epsilon$ for all $n \ge N$, that is, $\sup_A |f_n - f| \to 0$ as $n \to \infty$. Or, in geometric terms, for every $\epsilon > 0$, there is $N \in \mathbb{N}$ such that the graph of f_n lies within the strip of radius ϵ about the graph of f for all $n \ge N$.

Uniform convergence requires N to depend only on ϵ, whereas pointwise convergence allows N to also depend on the point $x \in A$.

8.4. Example. Consider the sequence of functions $f_n : [0,1] \to \mathbb{R}$, $f_n(x) = x^n$, from Example 8.2 with pointwise limit $f : x \mapsto \begin{cases} 0 & \text{if } x < 1, \\ 1 & \text{if } x = 1. \end{cases}$ The graph of f_n does not lie in the $\frac{1}{2}$-strip about the graph of f for any n, so the convergence of f_n to f is not uniform on $[0,1]$. However, for every $c \in (0,1)$, the convergence is uniform on $[0,c]$ because $\sup_{[0,c]} |f_n - f| = c^n \to 0$ as $n \to \infty$.

Exercise 8.1. Prove that if $f_n \to f$ uniformly on A, and each f_n is bounded on A, then f is bounded on A. Show that this may fail if $f_n \to f$ pointwise.

8.5. Theorem. If $f_n \to f$ uniformly on A, and each f_n is continuous at $c \in A$, then f is continuous at c.

Proof. Let $\epsilon > 0$. By uniform convergence, there is $N \in \mathbb{N}$ such that $|f_N - f| < \epsilon/3$ on A. Since f_N is continuous at c, there is $\delta > 0$ such that $|f_N(x) - f_N(c)| < \epsilon/3$ if $x \in A$ and $|x - c| < \delta$, but then

$$\begin{aligned} |f(x) - f(c)| &\le |f(x) - f_N(x)| + |f_N(x) - f_N(c)| + |f_N(c) - f(c)| \\ &< \epsilon/3 + \epsilon/3 + \epsilon/3 = \epsilon. \end{aligned}$$

□

8.6. Example. (a) Let $f_n : [0,1] \to \mathbb{R}$, $f_n(x) = \dfrac{nx}{1+nx^2}$. Then f_n is continuous, and

$$f_n(x) \to f(x) = \begin{cases} 0 & \text{if } x = 0, \\ 1/x & \text{if } 0 < x \le 1 \end{cases}$$

as $n \to \infty$, so the pointwise limit function f is not continuous. Hence (f_n) does not converge uniformly on $[0,1]$.

(b) Let us modify the preceding example and consider $g_n : [0,1] \to \mathbb{R}$, $g_n(x) = \dfrac{nx}{1+n^2x^2} \ge 0$. Then g_n is continuous and $g_n \to 0$ pointwise on $[0,1]$. The pointwise limit function is continuous, so further investigation is needed to determine whether $g_n \to 0$ uniformly. It is easy to find the maximum of g_n on $[0,1]$. A simple calculation shows that the only zero of the derivative g_n' on $[0,1]$ is $1/n$, and $g_n(1/n) = 1/2$. The end point values are $g_n(0) = 0$ and $g_n(1) = n/(1+n^2)$. The maximum of g_n on $[0,1]$ is the largest of these three values, that is, $1/2$. Thus $\sup_{[0,1]} |g_n| \not\to 0$, so (g_n) does not converge uniformly.

8.7. Theorem. If $f_n \to f$ uniformly on $[a,b]$, and each f_n is integrable on $[a,b]$, then f is integrable on $[a,b]$ and

$$\lim_{n\to\infty} \int_a^b f_n = \int_a^b f.$$

Proof. Each f_n is bounded on $[a,b]$ by assumption (boundedness is part of the definition of integrability), so f is bounded by Exercise 8.1. Let $\epsilon > 0$. There is $N \in \mathbb{N}$ such that $|f_n - f| < \dfrac{\epsilon}{b-a}$ on $[a,b]$ for all $n \ge N$. Since f_N is integrable, there is a partition P of $[a,b]$ such that $U(f_N,P) - L(f_N,P) < \epsilon$. For every $x \in [a,b]$,

$$f_N(x) - \frac{\epsilon}{b-a} < f(x) < f_N(x) + \frac{\epsilon}{b-a},$$

so

$$L(f_N,P) - \epsilon \le L(f,P) \le U(f,P) \le U(f_N,P) + \epsilon$$

and

$$U(f,P) - L(f,P) \le U(f_N,P) - L(f_N,P) + 2\epsilon < 3\epsilon.$$

By Lemma 7.4, this implies that f is integrable. Finally, for $n \ge N$,

$$\left| \int_a^b f_n - \int_a^b f \right| \le \int_a^b |f_n - f| \le \int_a^b \frac{\epsilon}{b-a} = \epsilon.$$

This shows that $\int_a^b f_n \to \int_a^b f$. □

Theorems 8.5 and 8.7 show that uniform convergence preserves continuity and integrability. Differentiability is more subtle. It will be considered in Proposition 8.16.

The notions of pointwise and uniform convergence are easily adapted to series of functions. If f_n, $n \in \mathbb{N}$, and f are functions on $A \subset \mathbb{R}$, we say that the series $\sum f_n$ converges pointwise or uniformly to f on A if the sequence of partial sums $s_n = f_1 + \cdots + f_n$ does. Then we write $\sum f_n = f$ and we must be careful to indicate which type of convergence we mean. By Theorem 8.5, if $\sum f_n = f$ uniformly on A, and each f_n is continuous on A, then the sum f is continuous on A.

We conclude this section with a useful test for the uniform convergence of a series of functions.

8.8. Theorem (Weierstrass M-test)**.** Let $A \subset \mathbb{R}$ and, for each $n \in \mathbb{N}$, let $f_n : A \to \mathbb{R}$ be a bounded function, say $|f_n| \le M_n$ on A. If $\sum M_n$ converges, then $\sum f_n$ converges uniformly on A.

Proof. Write $s_n = f_1 + \cdots + f_n$. Let $\epsilon > 0$. Find $N \in \mathbb{N}$ with $\sum_{j=N}^{\infty} M_j < \epsilon$. Then, for $x \in A$ and $m, n \ge N$, say $m < n$,

$$|s_n(x) - s_m(x)| = |f_{m+1}(x) + \cdots + f_n(x)| \le M_{m+1} + \cdots + M_n < \epsilon.$$

This shows that $(s_n(x))$ is a Cauchy sequence, so it converges to a real number $f(x)$ (Theorem 3.43). Thus we have obtained a pointwise limit function f for (s_n) on A. Finally, to see that $s_n \to f$ uniformly on A, note that for every $x \in A$ and $n \ge N$,

$$|s_n(x) - f(x)| = \Big| \sum_{j=n+1}^{\infty} f_j(x) \Big| \le \sum_{j=N}^{\infty} M_j < \epsilon. \qquad \square$$

8.9. Example. Let $f_n(x) = x^n/n!$ for $n \ge 0$ and $x \in \mathbb{R}$ (recall the convention that $0! = 1$). Let $c > 0$. On $[-c, c]$, $|f_n| \le c^n/n!$, and $\sum c^n/n!$ converges by the ratio test (Theorem 3.29), so by Theorem 8.8, $\sum f_n$ converges uniformly on $[-c, c]$. Since f_n is continuous for each n, we thus obtain a continuous function $\mathbb{R} \to \mathbb{R}$, $x \mapsto \sum_{n=0}^{\infty} \dfrac{x^n}{n!}$. We shall soon see that this function is nothing but the exponential function.

8.2. Power series

We like polynomials because they are so easy to work with. However, most functions are not polynomials. Power series, that is, 'polynomials with infinitely many terms', form a much bigger class that encompasses most (although not all) commonly used functions. Allowing infinitely many terms

raises convergence issues that must be addressed. That is the topic of this section. With a bit of theory under our belts we can work with power series almost as if they were polynomials.

8.10. Definition. A *power series* is a series of the form

$$\sum_{n=0}^{\infty} a_n x^n = a_0 + a_1 x + a_2 x^2 + \cdots ,$$

with *coefficients* $a_0, a_1, a_2, \ldots$ in $\mathbb{R}$.

More generally, one can consider power series of the form $\sum_{n=0}^{\infty} a_n (x-c)^n$. The number c is called the *centre* of the series. To keep the notation simple, we will restrict ourselves to the case $c = 0$. Our results can be easily adapted to the general case.

We will address two fundamental questions about power series.

- For which values of x (besides $x = 0$) does the power series converge? Can we describe the set of such x?
- On the set of points x at which the series converges, what can we say about the sum function? It is continuous or even differentiable?

The key to the first question is the following result.

8.11. Theorem. Suppose the power series $\sum a_n x^n$ converges at $x_0 \neq 0$. Then it converges absolutely at every x with $|x| < |x_0|$, and it converges uniformly on $[-c, c]$ for every c with $0 < c < |x_0|$.

Proof. Since $\sum a_n x_0^n$ converges, $a_n x_0^n \to 0$ (Proposition 3.24); in particular, $(a_n x_0^n)$ is bounded (Theorem 3.8). Find $M > 0$ such that $|a_n x_0^n| \leq M$ for all $n \in \mathbb{N}$. If $|x| < |x_0|$, then

$$|a_n x^n| = |a_n x_0^n| \left|\frac{x}{x_0}\right|^n \leq M \left|\frac{x}{x_0}\right|^n .$$

Since $\sum |x/x_0|^n$ converges, being a geometric series with $|x/x_0| < 1$, so does $\sum |a_n x^n|$ by the comparison test (Proposition 3.26). Also, if $0 < c < |x_0|$, then

$$|a_n x^n| \leq |a_n| c^n = |a_n x_0^n| \left|\frac{c}{x_0}\right|^n \leq M \left|\frac{c}{x_0}\right|^n$$

for all $x \in [-c, c]$. Since $\sum |c/x_0|^n$ converges, $\sum a_n x^n$ converges uniformly on $[-c, c]$ by the Weierstrass M-test (Theorem 8.8). □

The following consequence is immediate.

8.12. Corollary. For a power series $\sum a_n x^n$, precisely one of the following holds.

(i) The series converges for $x = 0$ only.

(ii) The series converges absolutely for all $x \in \mathbb{R}$.

(iii) There is a real number $R > 0$, namely

$$R = \sup\big\{x \in \mathbb{R} : \sum a_n x^n \text{ converges}\big\},$$

such that $\sum a_n x^n$ converges absolutely for $|x| < R$ and diverges for $|x| > R$.

We set $R = 0$ in case (i) and $R = \infty$ in case (ii) and call R the *radius of convergence* of the power series.

Furthermore, in cases (ii) and (iii), the power series converges uniformly to a continuous sum function on every compact subset of $(-R, R)$.

8.13. Remark. It follows from Corollary 8.12 that the set of $x \in \mathbb{R}$ for which a power series $\sum a_n x^n$ converges is an interval. It is called the *interval of convergence* of the power series. In case (i), it is $\{0\}$, and in case (ii) $\mathbb{R}$. In case (iii), it is $(-R, R)$, $[-R, R)$, $(-R, R]$, or $[-R, R]$. We call $(-R, R)$ the *open interval of convergence*.

We have now answered the first fundamental question: the set of points at which a power series converges is an interval. As for the second question, we have seen that the sum function is continuous at least on the open interval of convergence. We now turn to differentiability.

8.14. Theorem. Let the power series $\sum\limits_{n=0}^{\infty} a_n x^n$ have radius of convergence $R \geq 0$. The termwise differentiated series $\sum\limits_{n=1}^{\infty} n a_n x^{n-1}$ has the same radius of convergence. If $R > 0$, then the sum of $\sum a_n x^n$ is a differentiable function on $(-R, R)$ and its derivative is the sum of $\sum n a_n x^{n-1}$.

Before proving the theorem we record a corollary.

8.15. Corollary. (1) On the open interval of convergence, the sum of a power series is an infinitely differentiable function.

(2) Let the power series $\sum\limits_{n=0}^{\infty} a_n x^n$ have radius of convergence $R > 0$. The termwise integrated series $\sum\limits_{n=0}^{\infty} \dfrac{a_n}{n+1} x^{n+1}$ has the same radius of convergence, and its sum on $(-R, R)$ is an antiderivative for the sum of $\sum a_n x^n$.

Proof of Theorem 8.14. To show that the series $\sum a_n x^n$ and $\sum n a_n x^{n-1}$ have the same radius of convergence, it suffices to prove that if one of them converges absolutely for $|x| < r$, then so does the other one. First, suppose

$\sum na_n x^{n-1}$ converges absolutely. Since $|a_n x^n| \le |x||na_n x^{n-1}|$ for $n \ge 1$, $\sum a_n x^n$ also converges absolutely by comparison.

Conversely, suppose $\sum a_n x^n$ converges absolutely for $|x| < r$. Take x with $|x| < r$. Choose $w \in (|x|, r)$. Since $\sum a_n w^n$ converges, there is $M \ge 0$ with $|a_n w^n| \le M$ for all $n \in \mathbb{N}$, so

$$|na_n x^{n-1}| = \left| a_n w^n \frac{n}{w} \left(\frac{x}{w}\right)^{n-1} \right| \le \frac{Mn}{w} \left(\frac{|x|}{w}\right)^{n-1}.$$

Now $\sum n(|x|/w)^{n-1}$ converges by the ratio test, so $\sum na_n x^{n-1}$ converges absolutely by comparison.

Suppose $R > 0$. Let f be the sum of $\sum_{n=0}^{\infty} a_n x^n$ and g be the sum of $\sum_{n=1}^{\infty} na_n x^{n-1}$ on $(-R, R)$. We need to show that f is differentiable and $f' = g$. Let $s_m(x) = \sum_{n=0}^{m} a_n x^n$ and $t_m(x) = \sum_{n=1}^{m} na_n x^{n-1}$. Clearly, $s_m' = t_m$ for all $m \in \mathbb{N}$. Let $0 < c < R$ and let $I = [-c, c]$. By Theorem 8.11, $s_m \to f$ and $t_m \to g$ uniformly on I. The following result completes the proof. □

8.16. Proposition. Let $I \subset \mathbb{R}$ be an interval and $s_n : I \to \mathbb{R}$ be a differentiable function for each $n \in \mathbb{N}$ such that $s_n' : I \to \mathbb{R}$ is continuous. Suppose (s_n) converges pointwise on I to a limit function f, and (s_n') converges uniformly on I to a limit function g. Then f is differentiable on I and $f' = g$.

Proof. Fix $a \in I$. By the fundamental theorem of calculus (Theorem 7.12), part (1), for every $n \in \mathbb{N}$ and $x \in I$, $\int_a^x s_n' = s_n(x) - s_n(a)$. Letting $n \to \infty$, this yields $\int_a^x g = f(x) - f(a)$ by Theorem 8.7. Since g is continuous by Theorem 8.5, f is differentiable and $f' = g$ by the fundamental theorem of calculus, part (2). □

8.17. Example. We know that the power series $\sum x^n/n!$ sums to a continuous function f on all of $\mathbb{R}$ (Example 8.9). By Theorem 8.14, f is differentiable and its derivative is the sum of the power series obtained by differentiating

$$\sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \cdots$$

term by term. The termwise derivative is nothing but the series itself! Thus $f' = f$. Moreover, $f(0) = 1$, so by Theorem 7.21, $f = \exp$, that is,

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} \qquad \text{for all } x \in \mathbb{R}.$$

In particular,

$$e = \lim_{n\to\infty} (1 + \tfrac{1}{n})^n = \sum_{n=0}^{\infty} \frac{1}{n!}.$$

8.18. Example. Consider the function $f : (-1, 1) \to \mathbb{R}$, $f(x) = \log(1 - x)$. It is differentiable with $f'(x) = -\dfrac{1}{1-x}$. We know that $\sum\limits_{n=0}^{\infty} x^n = \dfrac{1}{1-x}$ for $|x| < 1$ (Example 3.25). By Corollary 8.15, the termwise antiderivative of the series $-\sum\limits_{n=0}^{\infty} x^n$ converges on $(-1, 1)$ and its sum is an antiderivative for f', so the sum differs from f by a constant. Hence

$$\log(1 - x) = -\sum_{n=0}^{\infty} \frac{x^{n+1}}{n+1} = -\sum_{n=1}^{\infty} \frac{x^n}{n} \qquad \text{for all } x \in (-1, 1).$$

In particular (just to show you a pretty formula),

$$\log 2 = -\log(1 - \tfrac{1}{2}) = \sum_{n=1}^{\infty} \frac{1}{n\, 2^n}.$$

8.3. Taylor series

We have discussed the properties of the sum function of a given power series. Now we turn the problem around and ask: Given a function, is it the sum of a power series? We start by observing that a power series with a positive radius of convergence is determined by its sum.

8.19. Proposition. Suppose the power series $\sum\limits_{n=0}^{\infty} a_n x^n$ has radius of convergence $R > 0$. Let f be the sum function on $(-R, R)$. Then, for every $n \geq 0$,

$$a_n = \frac{f^{(n)}(0)}{n!}.$$

Here, $f^{(n)}$ denotes the n^{th} derivative of f. By convention, $f^{(0)} = f$.

Proof. Differentiate repeatedly using Theorem 8.14 and set $x = 0$. □

Suppose we have an infinitely differentiable function $f : (-R, R) \to \mathbb{R}$, $R > 0$. We ask: Is there a power series centred at 0 with sum f? In other words, is

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} x^n$$

for all $x \in (-R, R)$? By Proposition 8.19, this is the only power series centred at 0 that can possibly have sum f.

8.20. Definition. This series is called the *Taylor series* of f centred at 0 or the *Maclaurin series* of f.

8.21. Example. (a) We know that the exponential function on $\mathbb{R}$ and the function $x \mapsto \log(1-x)$ on $(-1,1)$ equal the sums of their respective Taylor series centred at 0 (Examples 8.17 and 8.18). The same holds for many other important functions, such as the sine and the cosine (Section 8.4).

(b) Define $g : \mathbb{R} \to \mathbb{R}$ by the formula

$$g(x) = \begin{cases} \exp(-1/x) & \text{if } x > 0, \\ 0 & \text{if } x \le 0. \end{cases}$$

Using Exercise 7.18, you can show that g is infinitely differentiable with $g^{(n)}(0) = 0$ for all $n \ge 0$. Hence the Taylor series of g centred at 0 is identically zero, but g itself is not.

(c) Since $\sum_{n=0}^{\infty} x^n = \dfrac{1}{1-x}$ for $|x| < 1$, we also have

$$\frac{1}{1+x^2} = \sum_{n=0}^{\infty} (-x^2)^n = \sum_{n=0}^{\infty} (-1)^n x^{2n} = 1 - x^2 + x^4 - x^6 + \cdots$$

for $|x| < 1$. The function $x \mapsto \dfrac{1}{1+x^2}$ is infinitely differentiable on all of $\mathbb{R}$, but by Proposition 8.19, its one and only power series expansion centred at 0 is the one we have just written down, and it converges only on $(-1,1)$. Why cannot this seemingly well-behaved function have a power series expansion on a larger set? The answer comes from complex analysis. It is the zeros of the denominator $1+x^2$ at the complex numbers $\pm i$ that prevent the power series expansion from extending farther than distance 1 from 0.

8.22. Definition. Let f be an infinitely differentiable function on $(-R,R)$, $R > 0$. For each $n \ge 0$, the n^{th} *remainder* or *error function* of f is the function

$$E_n(x) = f(x) - \sum_{k=0}^{n} \frac{f^{(k)}(0)}{k!} x^k, \qquad x \in (-R,R).$$

Clearly, $f(x)$ equals the sum of the Maclaurin series of f at x if and only if $E_n(x) \to 0$ as $n \to \infty$. The following theorem gives us a handle on the remainder.

8.23. Theorem (Lagrange's remainder theorem)**.** Let f be an $n+1$ times differentiable function on $(-R,R)$, $R > 0$. For every $x \in (-R,R)$, $x \neq 0$, there is a number c strictly between 0 and x such that

$$E_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!} x^{n+1}.$$

This result can be viewed as a generalisation of the mean value theorem: write down what it says for $n = 0$.

Proof. Fix $x \in (-R, R)$, $x \neq 0$. Define

$$F(t) = f(t) + \sum_{k=1}^{n} \frac{f^{(k)}(t)}{k!}(x-t)^k + A(x-t)^{n+1}, \qquad t \in (-R, R),$$

where the constant A is chosen so that $F(0) = f(x)$. Clearly, $F(x) = f(x)$. By Rolle's theorem (Theorem 6.13) applied to F on the interval between 0 and x, there is c strictly between 0 and x such that (do the computation!)

$$0 = F'(c) = \frac{f^{(n+1)}(c)}{n!}(x-c)^n - (n+1)A(x-c)^n.$$

Thus

$$A = \frac{f^{(n+1)}(c)}{(n+1)!}.$$

Finally,

$$f(x) = F(0) = \sum_{k=0}^{n} \frac{f^{(k)}(0)}{k!}x^k + Ax^{n+1},$$

so $E_n(x) = Ax^{n+1}$. □

8.24. Corollary. Suppose $f : (-R, R) \to \mathbb{R}$, $R > 0$, is an infinitely differentiable function with $M \geq 0$ such that $|f^{(n)}(x)| \leq M$ for all $n \geq 0$ and $x \in (-R, R)$. Then

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!}x^n \qquad \text{for all } x \in (-R, R).$$

Proof. Let $x \in (-R, R)$. By Theorem 8.23, for every $n \geq 0$, there is c_n between 0 and x such that

$$|E_n(x)| = \left| \frac{f^{(n+1)}(c_n)}{(n+1)!}x^{n+1} \right| \leq \frac{M}{(n+1)!}|x|^{n+1},$$

so $E_n(x) \to 0$ as $n \to \infty$ (for every $a \in \mathbb{R}$, $a^n/n! \to 0$ because the series $\sum a^n/n!$ converges). □

8.25. Example. Suppose we have two bounded infinitely differentiable functions $s, c : \mathbb{R} \to \mathbb{R}$ such that $s' = c$, $c' = -s$, $s(0) = 0$, and $c(0) = 1$. Then s and c satisfy the hypotheses of Corollary 8.24, so each equals the sum of its Maclaurin series on all of $\mathbb{R}$, that is,

$$s(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!}x^{2n+1}, \quad c(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!}x^{2n} \quad \text{for all } x \in \mathbb{R}.$$

In particular, s and c are uniquely determined by the above properties.

Conversely, the results of this chapter show that these two power series converge on all of $\mathbb{R}$, and that their sum functions are infinitely differentiable and satisfy $s' = c$, $c' = -s$, $s(0) = 0$, and $c(0) = 1$. Hence $(s^2 + c^2)' = 0$, so $s^2 + c^2$ is a constant function, equal to $s(0)^2 + c(0)^2 = 1$. Consequently, s and c take their values in $[-1, 1]$. This is a starting point for the theory of the trigonometric functions. The final section of the chapter is devoted to a rigorous development of this theory.

8.4. The trigonometric functions

The purpose of this section is to place the basic theory of the trigonometric functions on a firm footing.

8.26. Theorem. (1) The power series $\sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1}$ and $\sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} x^{2n}$ have infinite radius of convergence.

(2) The sum functions $s, c : \mathbb{R} \to \mathbb{R}$,

$$s(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1}, \quad c(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} x^{2n},$$

are infinitely differentiable.

(3) They satisfy the differential equations

$$s' = c, \qquad c' = -s.$$

(4) s is an odd function, that is, $s(-x) = -s(x)$ for all $x \in \mathbb{R}$, and c is an even function, that is, $c(-x) = c(x)$ for all $x \in \mathbb{R}$.

(5) $s^2 + c^2 = 1$.

(6) s and c are bounded and take their values in $[-1, 1]$.

(7) For all $x, y \in \mathbb{R}$,

$$s(x + y) = s(x)c(y) + c(x)s(y).$$

(8) s and c are the unique differentiable functions on $\mathbb{R}$ such that

$$s' = c, \quad c' = -s, \quad s(0) = 0, \quad c(0) = 1.$$

Proof. (1) Apply the ratio test (Theorem 3.29).

(2) Invoke Corollary 8.15 (1).

(3) Differentiate term by term and use Theorem 8.14.

(4) Note that the power series for s only has terms of odd degree and the power series for c only has terms of even degree.

(5) Since $(s^2 + c^2)' = 2sc + 2c(-s) = 0$, the function $s^2 + c^2$ is constant, equal to $s(0)^2 + c(0)^2 = 1$.

(6) This follows directly from (5).

(7) Fix $y \in \mathbb{R}$ and define $f : \mathbb{R} \to \mathbb{R}$ by the formula

$$f(x) = s(x+y) - s(x)c(y) - c(x)s(y).$$

Differentiating twice using (3) gives $f'' = -f$, so $(f^2+(f')^2)' = 2f'(f+f'') = 0$. Also, $f(0) = f'(0) = 0$. Thus $f^2 + (f')^2$ is constant and equal to 0, so $f(x) = 0$ for all $x \in \mathbb{R}$.

(8) Let $\tilde{s}$, $\tilde{c}$ be another such pair. Then

$$\big((s-\tilde{s})^2 + (c-\tilde{c})^2\big)' = 2(s-\tilde{s})(c-\tilde{c}) + 2(c-\tilde{c})(-s+\tilde{s}) = 0,$$

so $(s-\tilde{s})^2 + (c-\tilde{c})^2$ is a constant function and equal to 0 at 0, so it is identically zero. This shows that $\tilde{s} = s$ and $\tilde{c} = c$. □

8.27. Definition. The function s is called the *sine function*, denoted sin. The function c is called the *cosine function*, denoted cos.

Exercise 8.2. Using (3), (4), and (7) in Theorem 8.26, derive the addition formulas for $\sin(x-y)$, $\cos(x+y)$, and $\cos(x-y)$. Also derive the double-angle formulas for $\sin 2x$ and $\cos 2x$.

The proof of Theorem 8.26 was short and easy, given the tools already at our disposal. To establish the periodicity of sine and cosine requires more work and some new ideas. It is by no means obvious from the power series expansions of sine and cosine, or from the differential equations $\sin' = \cos$ and $\cos' = -\sin$, that these functions are periodic.

The addition formula for the sine points us in the right direction. If $y \neq 0$ is a real number with $\cos y = 1$ and $\sin y = 0$, then

$$\sin(x+y) = \sin x \cos y + \cos x \sin y = \sin x$$

for all $x \in \mathbb{R}$, so the sine is periodic with period y. We do not know if such a number actually exists, but this observation suggests that we should investigate the zero set

$$A = \sin^{-1}(0) = \{x \in \mathbb{R} : \sin x = 0\}$$

of the sine. What can we say about A? Let us list some easy facts.

(a) $0 \in A$ because $\sin 0 = 0$.

(b) $A \neq \mathbb{R}$ because sin is not identically zero.

(c) The addition formula shows that if $x, y \in A$, then $x + y \in A$.

(d) Since sin is an odd function, if $x \in A$, then $-x \in A$.

(e) A is a closed subset of $\mathbb{R}$ since sin is continuous (Exercise 5.9).

Here is a definition from group theory that conveniently encapsulates properties (a), (c), and (d) of A.

8.28. Definition. A subset G of $\mathbb{R}$ is called a *subgroup* of $\mathbb{R}$ if:

- $0 \in G$.
- If $x, y \in G$, then $x + y \in G$.
- If $x \in G$, then $-x \in G$.

There are many subgroups of $\mathbb{R}$, for example $\mathbb{R}$ itself, $\{0\}$, $\mathbb{Z}$, the set of even integers, $\mathbb{Q}$, and $\{m + n\sqrt{2} : m, n \in \mathbb{Z}\}$. Subgroups of $\mathbb{R}$ can have a complicated structure and they are hard to understand in general. However, *closed* subgroups of $\mathbb{R}$, such as A, can be very simply described.

8.29. Theorem. If G is a closed subgroup of $\mathbb{R}$, $G \neq \{0\}$, and $G \neq \mathbb{R}$, then there is a unique real number $a > 0$ such that $G = \{an : n \in \mathbb{Z}\}$.

We denote the set $\{an : n \in \mathbb{Z}\}$ of all integral multiples of $a \in \mathbb{R}$ by $a\mathbb{Z}$.

Proof. Let $P = \{x \in G : x > 0\}$. Then $P \neq \varnothing$, because G has an element $x \neq 0$, and then $x \in P$ or $-x \in P$. First we claim that no sequence in P converges to 0. Otherwise, for every $\epsilon > 0$, there is $x \in G$ with $0 < x < \epsilon$. Then $x\mathbb{Z} \subset G$, so every interval of length at least ϵ intersects G. Since $\epsilon > 0$ was arbitrary, it follows that G is dense in $\mathbb{R}$, so $G = \mathbb{R}$ since G is closed, contrary to assumption.

Now let $a = \inf P \geq 0$. Then $a > 0$, for otherwise P contains a sequence converging to 0. Also, $a \in P$, for otherwise there is a strictly decreasing sequence (x_n) in P with $x_n \to a$, but then $(x_n - x_{n+1})$ is a sequence in P converging to $a - a = 0$. Therefore $a\mathbb{Z} \subset G$.

If $x \in G$, let m be the largest integer with $m \leq x/a$. Then $0 \leq x - am < a$ and $x - am \in G$, so $x - am = 0$ by the definition of a, and $x = am \in a\mathbb{Z}$. This shows that $G = a\mathbb{Z}$.

As for uniqueness, if $a, b > 0$ and $a\mathbb{Z} = b\mathbb{Z}$, then a and b are integral multiples of each other, so $a = b$. □

It remains to show that 0 is not the only zero of the sine.

8.30. Proposition. The sine function has a positive zero.

Proof. Suppose sin has no positive zeros. By the double-angle formula $\sin 2x = 2 \sin x \cos x$, cos has no positive zeros either. Since cos is continuous and $\cos 0 = 1$, it follows that $\sin' = \cos > 0$ on $[0, \infty)$, so sin is strictly increasing there. In particular, $\sin 1 > \sin 0 = 0$. Then, for all $x \geq 1$,

$$(x - 1) \sin 1 \leq \int_1^x \sin = \cos 1 - \cos x \leq 2,$$

which is absurd. □

From Theorem 8.29 and Proposition 8.30 we obtain the following result, which serves as a definition of the number π.

8.31. Corollary. There is a unique real number $\pi > 0$ such that

$$\sin^{-1}(0) = \pi\mathbb{Z}.$$

8.32. Lemma. $\cos \pi = -1$.

Proof. Now $\cos^2 \pi = 1 - \sin^2 \pi = 1$, so $\cos \pi = \pm 1$. On $(0, \pi)$, sin has no zeros, so sin does not change sign there. Since $\sin' 0 = \cos 0 = 1 > 0$, sin is positive on $(0, \pi)$, so cos is strictly decreasing there. Thus $\cos \pi = -1$. □

Exercise 8.3. A real number p is called a *period* of a function $f : \mathbb{R} \to \mathbb{R}$ if $f(x+p) = f(x)$ for all $x \in \mathbb{R}$. If f has a nonzero period, then f is said to be *periodic*. Let P be the set of all periods of f. Show that P is a subgroup of $\mathbb{R}$. It is called the *period group* of f. Show that if f is continuous, then P is closed. Conclude by Theorem 8.29 that a nonconstant continuous periodic function has a smallest positive period a, and that its period group is $a\mathbb{Z}$.

8.33. Theorem. The sine and cosine functions are periodic with smallest positive period 2π.

Proof. Note that if p is a period for sin, then $\sin p = \sin 0 = 0$, so $p \in \pi\mathbb{Z}$. Also, $\sin \pi = \sin 2\pi = 0$, and $\cos \pi = -1$ by Lemma 8.32. By the double-angle formula $\cos 2x = \cos^2 x - \sin^2 x$, $\cos 2\pi = 1$. Thus, by the addition formula for sin,

$$\sin(x + \pi) = -\sin x, \qquad \sin(x + 2\pi) = \sin x,$$

for all $x \in \mathbb{R}$. This shows that 2π is a period for sin, but π is not. It follows that the period group of sin is $2\pi\mathbb{Z}$.

Finally, by differentiating $\sin(x + p) = \sin x$ with respect to x, we see that a period for sin is also a period for cos. Similarly, a period for cos is also a period for sin, so sin and cos have the same periods. □

Exercise 8.4. Show that $\sin \frac{\pi}{2} = 1$ and $\cos \frac{\pi}{2} = 0$, and deduce that

$$\sin(x + \tfrac{\pi}{2}) = \cos x, \qquad \cos(x + \tfrac{\pi}{2}) = -\sin x$$

for all $x \in \mathbb{R}$. Conclude that

$$\cos^{-1}(0) = \frac{\pi}{2} + \pi\mathbb{Z}.$$

We can now define each of the other four trigonometric functions

$$\tan = \frac{\sin}{\cos}, \qquad \cot = \frac{\cos}{\sin}, \qquad \sec = \frac{1}{\cos}, \qquad \csc = \frac{1}{\sin}$$

on the complement of the zero set of its denominator.

We can also introduce the inverse trigonometric functions. We end this section by briefly considering the inverse sine. On $(-\frac{\pi}{2}, \frac{\pi}{2})$, $\sin' = \cos > 0$, so sin is strictly increasing. Also, $\sin(-\frac{\pi}{2}) = -1$ and $\sin\frac{\pi}{2} = 1$. By Theorem 5.26, the bijection $[-\frac{\pi}{2}, \frac{\pi}{2}] \to [-1, 1]$, $x \mapsto \sin x$, has a continuous inverse $[-1, 1] \to [-\frac{\pi}{2}, \frac{\pi}{2}]$, called the *arcsine* or *inverse sine* and denoted arcsin or $\sin^{-1}$. By Theorem 6.7, arcsin is differentiable on $(-1, 1)$ with

$$\arcsin' x = \frac{1}{\sin'(\arcsin x)} = \frac{1}{\cos\arcsin x}$$

for all $x \in (-1, 1)$. Since $\arcsin x \in (-\frac{\pi}{2}, \frac{\pi}{2})$, $\cos\arcsin x > 0$, so

$$\cos\arcsin x = \sqrt{1 - \sin^2\arcsin x} = \sqrt{1 - x^2}$$

and

$$\arcsin' x = \frac{1}{\sqrt{1 - x^2}}.$$

More exercises

8.5. For each $n \in \mathbb{N}$, let $f_n : [0, \infty) \to \mathbb{R}$, $f_n(x) = \dfrac{x}{x + n}$. Prove that for every $b > 0$, the sequence (f_n) converges uniformly on $[0, b]$.

8.6. Let $f_n(x) = \dfrac{x}{1 + x^n}$, $x \in [0, \infty)$, $n \in \mathbb{N}$. Find the pointwise limit of (f_n) on $[0, \infty)$. Show that the convergence is not uniform on $[0, \infty)$. Find a smaller set on which the convergence is uniform.

8.7. Let $f_n(x) = \dfrac{x}{1 + nx^2}$, $x \in \mathbb{R}$, $n \in \mathbb{N}$. Find the pointwise limit of (f_n) on $\mathbb{R}$. Is the limit uniform? (*Hint.* Find the maximum and minimum values of f_n.) For which values of x is $(\lim f_n)'(x) = \lim f_n'(x)$?

8.8. Let f_n, $n \in \mathbb{N}$, and f be functions on $A \subset \mathbb{R}$. Complete the following sentence. To say that (f_n) *does not* converge uniformly to f means that there is $\epsilon > 0$ such that for every ....

8.9. Let $A = A_1 \cup A_2 \subset \mathbb{R}$. Let f_n, $n \in \mathbb{N}$, and f be functions on A. Prove that if $f_n \to f$ uniformly on A_1, and $f_n \to f$ uniformly on A_2, then $f_n \to f$ uniformly on A.

8.10. (a) Let f_n, $n \in \mathbb{N}$, and f be functions on $A \subset \mathbb{R}$ such that $f_n \to f$ uniformly and f is continuous. Show that if $x_n \to a$ in A, then $f_n(x_n) \to f(a)$.

(b) What if $f_n \to f$ only pointwise?

8.11. Monotonicity is a powerful tool. For sequences of numbers, it turns boundedness into a sufficient condition for convergence (Theorem 3.16). It can also turn pointwise convergence into uniform convergence.

(a) Let f_n, $n \in \mathbb{N}$, and f be continuous functions $[a,b] \to \mathbb{R}$ such that $f_n \to f$ pointwise and $f_1(x) \le f_2(x) \le \cdots$ for every $x \in [a,b]$. Show that $f_n \to f$ uniformly. This result is known as *Dini's theorem.*

Hint. Fix $\epsilon > 0$ and let $U_n = \{x \in [a,b] : f(x) < f_n(x) + \epsilon\}$, so $U_1 \subset U_2 \subset \cdots$. We want $U_n = [a,b]$ for n large enough. If not, $K_n = [a,b] \setminus U_n \neq \varnothing$ for all n. Apply Theorem 4.13.

(b) What if $[a,b]$ is replaced by (a,b)?

8.12. Use the Weierstrass M-test to prove that the formula $g(x) = \sum_{n=1}^{\infty} \frac{x^n}{n^2}$ defines a continuous function g on the interval $[-1,1]$.

8.13. Let (a_n) be a bounded sequence such that the series $\sum a_n$ diverges. Prove that the radius of convergence of the power series $\sum a_n x^n$ is 1.

8.14. Find the interval of convergence of each of the following power series.

(a) $\sum_{n=1}^{\infty} (\log n)^n x^n$.

(b) $\sum_{n=0}^{\infty} \frac{n^2}{n!} x^n$.

(c) $\sum_{n=1}^{\infty} \frac{(-1)^n}{n3^n} x^n$.

(d) $\sum_{n=0}^{\infty} P(n) x^n$, where P is a nonconstant polynomial.

8.15. Find the radius of convergence of the power series $\sum_{n=0}^{\infty} \frac{n!}{n^n} x^n$.

8.16. (a) Find a power series (centred at 0) that converges conditionally at -1 and converges absolutely at 1, or explain why such a series does not exist.

(b) Prove that a power series can converge conditionally at no more than two points.

8.17. Let $\sum a_n x^n$ and $\sum b_n x^n$ be power series with radius of convergence r and s, respectively. If $r \neq s$, what is the radius of convergence of the power series $\sum (a_n + b_n) x^n$? What if $r = s$?

8.18. Find a power series such that $\sum_{n=0}^{\infty} a_n x^n = \sqrt[3]{x}$ for all $x \in (-1,1)$ or explain why such a series does not exist.

8.19. By the formula for the sum of the geometric series, $\sum_{n=0}^{\infty} x^n = \dfrac{1}{1-x}$ for all $x \in (-1,1)$. Use this to calculate the infinite sum

$$\frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \frac{4}{16} + \frac{5}{32} + \cdots .$$

Hint. Use the theorem about termwise differentiation of a power series.

8.20. Let $s_n = \sum_{k=0}^{n} \dfrac{1}{k!}$, so $s_n \to e$ as $n \to \infty$. Show that $0 < e - s_n < \dfrac{1}{n!\,n}$.

8.21. Prove that e is irrational. *Hint.* Suppose $e = m/n$ with $m, n \in \mathbb{N}$. Then $n!(e - s_n)$ is an integer, but $0 < n!(e - s_n) < 1/n$ by Exercise 8.20.

8.22. Is there an infinitely differentiable function $f : \mathbb{R} \to \mathbb{R}$ such that $f^{(n)}(0) = n^3 - 5n + 2$ for all $n \geq 0$?

8.23. Let $f : \mathbb{R} \setminus \{1\} \to \mathbb{R}$, $f(x) = \dfrac{1}{1-x}$. Then f is infinitely differentiable. Let $c \in \mathbb{R} \setminus \{1\}$ and $r = |c-1|$. Show that there is a power series $\sum a_n(x-c)^n$ with radius of convergence r, such that $f(x) = \sum a_n(x-c)^n$ for all $x \in (c-r, c+r)$. *Hint.* Start by writing

$$\frac{1}{1-x} = \frac{1}{1-c}\,\frac{1}{1 - \dfrac{x-c}{1-c}}.$$

8.24. (a) Compute the Maclaurin series of the function $f : [-1, \infty) \to \mathbb{R}$, $f(x) = \sqrt{1+x}$.

(b) Show that the radius of convergence of the series is 1.

(c) Let $x \in (0,1)$. Use Lagrange's remainder theorem (Theorem 8.23) to show that the sum of the Maclaurin series of f at x equals $f(x)$.

(d) Can you do the same for $x \in (-1,0)$?

(e) Let $s : (-1,1) \to \mathbb{R}$ be the sum function of the Maclaurin series of f. We know that s is infinitely differentiable. Show that s satisfies the differential equation $2(1+x)s'(x) = s(x)$. Conclude that $s(x) = f(x)$ for all $x \in (-1,1)$.

8.25. Let I be an open interval and $f : I \to \mathbb{R}$ be twice differentiable. We say that $c \in I$ is an *inflection point* of f if f'' changes sign at c, that is, for some $\epsilon > 0$, we have $f''(x) < 0$ if $c - \epsilon < x < c$ and $f''(x) > 0$ if $c < x < c + \epsilon$ or the other way around.

Use Lagrange's remainder theorem (Theorem 8.23) to show that a twice-differentiable function cannot have a maximum or a minimum at an inflection point.

Chapter 9

Metric spaces

9.1. Examples of metric spaces

Much of the theory developed in Chapters 3, 4, and 5 can be extended to the vastly more general setting of metric spaces. Even if we were only interested in analysis on the real line, this would still be worthwhile. In the following chapter, we will use the abstract theory of this chapter to prove an existence and uniqueness theorem for solutions of differential equations.

9.1. Definition. A *metric space* is a set X with a function $d : X \times X \to [0, \infty)$, such that:

- $d(x, y) = 0$ if and only if $x = y$.
- $d(x, y) = d(y, x)$ for all $x, y \in X$.
- $d(x, z) \leq d(x, y) + d(y, z)$ for all $x, y, z \in X$ (*triangle inequality*).

We call d a *metric* or a *distance function* on X. We sometimes write (X, d) for the set X with the metric d.

It turns out that all we need in order to develop such notions as convergence, completeness, and continuity is the three simple properties that define a metric. Of the three, the triangle inequality is of course the most substantial.

Examples of metric spaces abound throughout mathematics. In the remainder of this section we will explore a few of them. Be sure to verify the three defining properties of a metric if some of the details have been left out.

9.2. Example. The prototypical example of a metric space is the set $\mathbb{R}$ of real numbers with the metric $d(x, y) = |x - y|$.

9.3. Example. Every set can be made into a metric space by defining the distance between two distinct points to be 1. More explicitly,

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

A set with this metric is called a *discrete space*.

9.4. Example. Euclidean space $\mathbb{R}^n$, for $n \geq 2$, has several different metrics that are commonly used in analysis and geometry. The best known is the *Euclidean metric*

$$d_2(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2},$$

where $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ are points in $\mathbb{R}^n$. It is not obvious that d_2 satisfies the triangle inequality. It is a consequence of the following inequality.

9.5. Theorem (Cauchy-Schwarz inequality)**.** If $a_1, \ldots, a_n, b_1, \ldots, b_n \in \mathbb{R}$, then

$$\Big(\sum_{i=1}^{n} a_i b_i \Big)^2 \leq \Big(\sum_{i=1}^{n} a_i^2 \Big) \Big(\sum_{i=1}^{n} b_i^2 \Big).$$

In other words, for $a, b \in \mathbb{R}^n$,

$$|a \cdot b| \leq \|a\| \, \|b\|.$$

Here, $a \cdot b$ denotes the *inner product* $a_1 b_1 + \cdots + a_n b_n$, and $\|a\|$ denotes the *Euclidean norm* $(a_1^2 + \cdots + a_n^2)^{1/2}$.

Proof. This is clear if $a_1, \ldots, a_n = 0$. Otherwise consider the quadratic polynomial

$$p(x) = \sum_{i=1}^{n} (a_i x + b_i)^2 = Ax^2 + 2Bx + C,$$

where

$$A = \sum_{i=1}^{n} a_i^2 > 0, \quad B = \sum_{i=1}^{n} a_i b_i, \quad C = \sum_{i=1}^{n} b_i^2.$$

Completing the square gives

$$p(x) = \frac{1}{A}(Ax + B)^2 + \frac{D}{A}, \quad \text{where } D = AC - B^2.$$

Clearly, D/A is the smallest value of $p(x)$. Since $p(x)$ is a sum of squares, $p(x) \geq 0$ for all $x \in \mathbb{R}$. Hence $D \geq 0$, that is, $B^2 \leq AC$. □

Now, for $a, b \in \mathbb{R}^n$,

$$\begin{aligned}\|a+b\|^2 &= (a+b)\cdot(a+b) = a\cdot a + 2a\cdot b + b\cdot b \\ &\le \|a\|^2 + 2\|a\|\|b\| + \|b\|^2 = (\|a\| + \|b\|)^2,\end{aligned}$$

so

$$\|a+b\| \le \|a\| + \|b\|.$$

Taking $a = x-y$ and $b = y-z$ gives the triangle inequality for the Euclidean metric d_2.

Among other metrics on $\mathbb{R}^n$ are the L^1 *metric*

$$d_1(x,y) = \sum_{i=1}^{n} |x_i - y_i|,$$

and the L^∞ *metric* or *maximum metric*

$$d_\infty(x,y) = \max_{i=1,\dots,n} |x_i - y_i|$$

(the Euclidean metric d_2 is sometimes called the L^2 *metric*). The three defining properties of a metric are easy to verify for d_1 and d_∞. When $n = 1$, $d_1 = d_2 = d_\infty$. Finally, note that

$$d_\infty \le d_2 \le d_1 \le n d_\infty.$$

We express this by saying that the three metrics are mutually *equivalent*.

Exercise 9.1. For a completely different example of a metric space, let X be the set of all finite strings—call them words—of letters of the alphabet. For distinct words w and w', let $d(w,w') = 2^{-n}$, where n is the first place in which w and w' differ (and let $d(w,w) = 0$ of course). For example, $d(\texttt{car},\texttt{cat}) = 2^{-3}$ and $d(\texttt{car},\texttt{card}) = 2^{-4}$. As usual, the first two defining properties of a metric are obvious. Prove the triangle inequality. Show that d actually satisfies the stronger inequality

$$d(w_1,w_3) \le \max\{d(w_1,w_2), d(w_2,w_3)\}.$$

Such a metric is called an *ultrametric*. Show that if $d(w_1,w_2) \ne d(w_2,w_3)$, we even have

$$d(w_1,w_3) = \max\{d(w_1,w_2), d(w_2,w_3)\}.$$

Exercise 9.2. Here is a more substantial example of an ultrametric space that is important in number theory. Fix a prime number p. If $x \ne 0$ is a rational number, write $x = p^n a/b$ where $n, a, b \in \mathbb{Z}$ and neither a nor b is divisible by p, and set $|x|_p = p^{-n}$. Let $|0|_p = 0$. Show that setting $d_p(x,y) = |x-y|_p$ defines an ultrametric d_p on $\mathbb{Q}$. It is called the *p-adic metric* on $\mathbb{Q}$.

9.6. Example. Let ℓ_∞ be the set of all bounded sequences of real numbers. Note that ℓ_∞ is a vector space with the usual addition and scalar multiplication of sequences. Namely, if (a_n) and (b_n) are bounded sequences of reals, then the sum $(a_n) + (b_n) = (a_n + b_n)$ is also bounded, and so is the scalar multiple $c(a_n) = (ca_n)$ for every $c \in \mathbb{R}$. As a vector space, ℓ_∞ is infinite-dimensional: it has no finite basis.

For $a, b \in \ell_\infty$, let

$$d(a, b) = \sup_{n\in\mathbb{N}} |a_n - b_n|.$$

Note that since the set $\{|a_n - b_n| : n \in \mathbb{N}\}$ is bounded, the supremum exists as a nonnegative real number. We claim that d is a metric on ℓ_∞. It is called the *supremum metric*. First, $d(a, b) = 0$ if and only if $|a_n - b_n| = 0$ for all $n \in \mathbb{N}$, that is, $a = b$. Second, it is clear that $d(a, b) = d(b, a)$. Third, let $a, b, c \in \ell_\infty$. By the triangle inequality for real numbers, for every $n \in \mathbb{N}$,

$$|a_n - c_n| \le |a_n - b_n| + |b_n - c_n| \le d(a, b) + d(b, c),$$

so $d(a, b) + d(b, c)$ is an upper bound for the set $\{|a_n - c_n| : n \in \mathbb{N}\}$. Hence $d(a, b) + d(b, c)$ is no smaller than the least upper bound $d(a, c)$ of this set. Thus d satisfies the triangle inequality.

9.7. Example. Consider a compact interval $I = [a, b] \subset \mathbb{R}$, $a \le b$. Let $\mathscr{C}(I)$ denote the set of all continuous functions $I \to \mathbb{R}$. It is a vector space, which is infinite-dimensional if $a < b$ (can you prove it?).

Let $f, g \in \mathscr{C}(I)$. Then $|f - g|$ is a continuous function on I, so it has a maximum by the extreme value theorem (Theorem 5.17). We define the distance between f and g to be this maximum, that is, we set

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

We claim that d is a metric on $\mathscr{C}(I)$. It is called the *supremum metric* or the *uniform metric*. The first two defining properties of a metric are clear. As for the triangle inequality, let $f, g, h \in \mathscr{C}(I)$. For every $x \in I$,

$$|f(x) - h(x)| \le |f(x) - g(x)| + |g(x) - h(x)| \le d(f, g) + d(g, h),$$

so $d(f, g) + d(g, h)$ is an upper bound for the set $\{|f(x) - h(x)| : x \in I\}$. Hence $d(f, g) + d(g, h)$ is no smaller than the maximum $d(f, h)$ of this set. Thus d satisfies the triangle inequality.

For example, on $I = [0, 1]$, let $f(x) = x$, $g(x) = x^2$, and $h(x) = 1 - x$. Then $d(f, g) = \max\limits_{0\le x\le 1} |x - x^2| = \frac{1}{4}$, $d(f, h) = \max\limits_{0\le x\le 1} |x - (1 - x)| = 1$, and $d(g, h) = \max\limits_{0\le x\le 1} |x^2 - (1 - x)| = 1$, so indeed $d(f, h) = 1 \le \frac{5}{4} = d(f, g) + d(g, h)$.

Finally, here is a very simple way to obtain new metric spaces from old.

9.8. Definition. Let (X, d_X) be a metric space. A *subspace* (Y, d_Y) of (X, d_X) is a subset $Y \subset X$ with the metric d_Y obtained by restricting d_X to $Y \times Y$, that is, $d_Y(y, y') = d_X(y, y')$ for $y, y' \in Y$. We call d_Y the *induced metric*, or more precisely, the metric *induced* on Y from (X, d_X).

Thus we can always consider a subset of a metric space as a metric space in its own right.

9.2. Convergence and completeness in metric spaces

Many of the definitions, theorems, and proofs in Chapters 3, 4, and 5 can be extended from $\mathbb{R}$ to an arbitrary metric space simply by replacing expressions of the form $|x - y|$ by $d(x, y)$. The fundamental definition is the following generalisation of Definition 3.2.

9.9. Definition. Let (X, d) be a metric space. A sequence (a_n) in X *converges* to $b \in X$ if for every $\epsilon > 0$, there is $N \in \mathbb{N}$ such that $d(a_n, b) < \epsilon$ for all $n \geq N$. We call b the *limit* of (a_n) and write $b = \lim_{n\to\infty} a_n$ or $a_n \to b$.

As before, we can show that $a_n \to b$ if and only if $d(a_n, b) \to 0$ as a sequence of real numbers. Also, the limit of a convergent sequence is unique. And we can extend the notion of a neighbourhood to a metric space and reformulate Definition 9.9 as in Remark 3.5.

9.10. Definition. Let (X, d) be a metric space. The *open ball* in X of radius $r > 0$ centred at $a \in X$ is the set

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

A *neighbourhood* of a in X is any subset of X that contains the open ball $B(a, r)$ for some $r > 0$.

9.11. Remark. A sequence (a_n) in a metric space (X, d) converges to $b \in X$ if and only if each neighbourhood of b contains a_n for all but finitely many $n \in \mathbb{N}$.

Let us now investigate what convergence means in the examples of the previous section.

9.12. Example. Let (X, d) be a discrete space (Example 9.3). Say $a_n \to b$ in X. Taking $\epsilon = 1$ in Definition 9.9, we see that there is $N \in \mathbb{N}$ with $d(a_n, b) < 1$ for all $n \geq N$. But $d(a_n, b) < 1$ implies $a_n = b$. So (a_n) is *eventually constant:* there is $b \in X$ and $N \in \mathbb{N}$ such that $a_n = b$ for all $n \geq N$.

In every metric space, an eventually-constant sequence converges. A discrete space has no other convergent sequences.

9.13. Example. Let $(a_k)_{k\in\mathbb{N}}$ be a sequence in $\mathbb{R}^n$ with $a_k = (a_{k1}, \dots, a_{kn})$. We have $a_k \to b$ as $k \to \infty$ with respect to the maximum metric d_∞ if and only if

$$\max_{i=1,\dots,n} |a_{ki} - b_i| \to 0 \quad \text{as } k \to \infty,$$

that is, $|a_{ki} - b_i| \to 0$ as $k \to \infty$ for each $i = 1, \dots, n$. In other words, convergence in $(\mathbb{R}^n, d_\infty)$ is convergence in each coordinate.

Convergence with respect to d_1 and d_2 is exactly the same. As noted in Example 9.4, $d_\infty \le d_2 \le d_1 \le n d_\infty$, so $d_1(a_k, b) \to 0$ if and only if $d_2(a_k, b) \to 0$ if and only if $d_\infty(a_k, b) \to 0$.

9.14. Example. We have $a_k \to b$ in ℓ_∞ if and only if

$$d(a_k, b) = \sup_{n\in\mathbb{N}} |a_{kn} - b_n| \to 0 \quad \text{as } k \to \infty.$$

This implies coordinatewise convergence, that is, for each $n \in \mathbb{N}$,

$$|a_{kn} - b_n| \le d(a_k, b) \to 0 \quad \text{as } k \to \infty,$$

but it is stronger. For example, let $a_1 = (1, 0, 0, \dots)$, $a_2 = (0, 1, 0, \dots)$, $a_3 = (0, 0, 1, \dots), \dots$. Then, for each $n \in \mathbb{N}$, the sequence of n^{th} coordinates $(a_{kn})_{k\in\mathbb{N}}$ goes to 0 (one term of this sequence is 1 and the others are all 0), but $d(a_k, 0) = 1$ for all $k \in \mathbb{N}$, so $a_k \not\to 0$ (where we also denote by 0 the zero vector $(0, 0, 0, \dots)$ in ℓ_∞).

9.15. Example. Let $I \subset \mathbb{R}$ be a compact interval. Let (f_n) be a sequence in $\mathscr{C}(I)$ and $g \in \mathscr{C}(I)$. We have $f_n \to g$ with respect to the supremum metric d on $\mathscr{C}(I)$ (Example 9.7) if and only if

$$d(f_n, g) = \max_{x\in I} |f_n(x) - g(x)| \to 0 \quad \text{as } n \to \infty.$$

In other words, for every $\epsilon > 0$, there is $N \in \mathbb{N}$ such that $|f_n(x) - g(x)| < \epsilon$ for all $x \in I$ and all $n \ge N$. This means precisely that $f_n \to g$ uniformly (Definition 8.3).

Exercise 9.3. What does it mean for a sequence of words to converge with respect to the metric defined in Exercise 9.1?

The theory of metric spaces encompasses convergence of sequences of real numbers, convergence in higher-dimensional Euclidean spaces, uniform convergence of continuous functions on a compact interval, and much more. Any result about convergence in a general metric space will apply to all these different kinds of convergence.

We will now extend our central concept, completeness, to metric spaces. The axiom of completeness invokes the order structure of $\mathbb{R}$ and thus cannot be applied to an arbitrary metric space. The Cauchy criterion, on the other hand, generalises directly to metric spaces. (Recall that, for an ordered

field, the Cauchy criterion is a consequence of the axiom of completeness and, conversely, implies the axiom of completeness with the help of the Archimedean property: see Theorem 3.45.) Metric spaces satisfying the Cauchy criterion turn out to be particularly useful in mathematics and they are said to be complete.

9.16. Definition. A sequence (a_n) in a metric space (X, d) is a *Cauchy sequence* if for every $\epsilon > 0$, there is $N \in \mathbb{N}$ such that if $m, n \geq N$, then $d(a_m, a_n) < \epsilon$.

As before, we can show that a convergent sequence is Cauchy (this was the easy half of Theorem 3.43).

9.17. Proposition. A convergent sequence in a metric space is a Cauchy sequence.

Proof. Suppose $a_n \to b$ in a metric space (X, d). Let $\epsilon > 0$. There is $N \in \mathbb{N}$ such that $d(a_n, b) < \epsilon/2$ for all $n \geq N$. Then, if $m, n \geq N$,

$$d(a_m, a_n) \leq d(a_m, b) + d(a_n, b) < \epsilon/2 + \epsilon/2 = \epsilon.$$

This shows that (a_n) is Cauchy. □

We now turn the Cauchy criterion into the definition of what it means for a metric space to be complete.

9.18. Definition. A metric space X is *complete* if every Cauchy sequence in X converges in X.

9.19. Example. By Theorem 3.43, $\mathbb{R}$ is complete as a metric space with the usual metric. The subspace $(0, 1)$ of $\mathbb{R}$ is not complete because there are Cauchy sequences in $(0, 1)$ with no limit in $(0, 1)$, for example the sequence $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots$. The subspace $[0, 1]$ of $\mathbb{R}$, on the other hand, is complete. Namely, if (a_n) is a Cauchy sequence in $[0, 1]$, then (a_n) is a Cauchy sequence in $\mathbb{R}$, so (a_n) converges to a limit, say b, in $\mathbb{R}$. Since $0 \leq a_n \leq 1$ for all $n \in \mathbb{N}$, we also have $0 \leq b \leq 1$ by Theorem 3.13, so (a_n) converges in $[0, 1]$. In fact, by Theorem 9.28, a subspace of $\mathbb{R}$ is complete if and only if it is closed.

9.20. Example. Let (X, d) be a discrete space (Example 9.3). Say (a_n) is Cauchy in X. Taking $\epsilon = 1$ in Definition 9.16, we see that there is $N \in \mathbb{N}$ with $d(a_m, a_n) < 1$ for all $m, n \geq N$. But $d(a_m, a_n) < 1$ implies $a_m = a_n$. Thus (a_n) is eventually constant and hence convergent. This shows that every discrete space is complete.

9.21. Example. Let $(a_k)_{k \in \mathbb{N}}$ be a Cauchy sequence in $\mathbb{R}^n$ with the maximum metric d_∞. For each $i = 1, \ldots, n$,

$$|a_{ji} - a_{ki}| \leq \max_{i=1,\ldots,n} |a_j - a_k|,$$

so the sequence of i^{th} coordinates $(a_{ki})_{k\in\mathbb{N}}$ is Cauchy in $\mathbb{R}$, and hence convergent with limit $b_i \in \mathbb{R}$ by the completeness of $\mathbb{R}$. Then $a_k \to b$ coordinatewise and hence with respect to d_1, d_2, and d_∞ (Example 9.13). Thus $\mathbb{R}^n$ is complete with respect to each of the three metrics.

9.22. Example. Let $I \subset \mathbb{R}$ be a compact interval. Let (f_n) be a Cauchy sequence in $\mathscr{C}(I)$. For each $x \in I$,

$$|f_m(x) - f_n(x)| \le \max_I |f_m - f_n| = d(f_m, f_n),$$

so the sequence $(f_n(x))$ is Cauchy in $\mathbb{R}$, and hence convergent with a limit that we shall call $f(x)$. This defines a function $f : I \to \mathbb{R}$ such that $f_n \to f$ pointwise. We want to show that $f \in \mathscr{C}(I)$ and that $f_n \to f$ uniformly.

Let $\epsilon > 0$ and find $N \in \mathbb{N}$ such that $d(f_m, f_n) < \epsilon$ for all $m, n \ge N$. Then, for each $x \in I$ and $m, n \ge N$,

$$|f_m(x) - f_n(x)| \le d(f_m, f_n) < \epsilon.$$

Letting $m \to \infty$ gives $|f_n(x) - f(x)| \le \epsilon$ for all $x \in I$ and $n \ge N$. Hence $f_n \to f$ uniformly on I. Therefore f is continuous (Theorem 8.5), so $f \in \mathscr{C}(I)$ and $f_n \to f$ in $\mathscr{C}(I)$.

This shows that $\mathscr{C}(I)$ is complete with respect to the supremum metric.

Exercise 9.4. Show that ℓ_∞ with the supremum metric is complete. *Hint.* Follow the approach of Example 9.22.

Exercise 9.5. Is the ultrametric space in Exercise 9.1 complete?

We conclude this section by determining when a subspace of a complete metric space is itself complete. For this, we need to generalise Section 4.1.

9.23. Definition. A subset U of a metric space (X, d) is *open* if it is a neighbourhood of each of its points. That is, for every $a \in U$, there is $\epsilon > 0$ such that $B(a, \epsilon) \subset U$.

The following proposition generalises Proposition 4.2. We leave the proof as an exercise.

9.24. Proposition. (1) X and $\varnothing$ are open. An open ball is open.

(2) The union of an arbitrary collection of open sets is open.

(3) The intersection of finitely many open sets is open.

Exercise 9.6. (a) Show that for every $a \in \mathbb{R}^n$ and $r > 0$,

$$B_\infty(a, r/n) \subset B_1(a, r) \subset B_2(a, r) \subset B_\infty(a, r),$$

where the subscripts refer to one of the metrics d_1, d_2, d_∞.

(b) Show that the three metrics define the same notion of a subset of $\mathbb{R}^n$ being open.

9.25. Definition. A subset A of a metric space (X, d) is *closed* if its complement $X \setminus A$ is open.

The next proposition is dual to Proposition 9.24.

9.26. Proposition. (1) X and $\varnothing$ are closed.

(2) The intersection of an arbitrary collection of closed sets is closed.

(3) The union of finitely many closed sets is closed.

The following result generalises Theorem 4.6.

9.27. Theorem. A subset A of a metric space (X, d) is closed if and only if whenever $a_n \in A$ for all $n \in \mathbb{N}$, and $a_n \to c$ in X, we have $c \in A$.

The proof is virtually identical to the proof of Theorem 4.6.

Proof. $\Rightarrow$ Say $a_n \in A$, $n \in \mathbb{N}$, and $a_n \to c$ in X. If $c \notin A$, then, since $X \setminus A$ is open, $X \setminus A$ is a neighbourhood of c, so $a_n \in X \setminus A$ for all but finitely many n, which is absurd.

$\Leftarrow$ We prove the contrapositive. Suppose A is not closed, that is, $X \setminus A$ is not open. This means that there is $c \in X \setminus A$ such that $X \setminus A$ is not a neighbourhood of c. Thus, for each $n \in \mathbb{N}$, $X \setminus A$ does not contain the open ball $B(c, \frac{1}{n})$, so there is $a_n \in A$ with $d(a_n, c) < \frac{1}{n}$. Then $a_n \to c$. □

9.28. Theorem. A subset of a complete metric space is complete (as a metric space with the induced metric) if and only if it is closed.

Proof. Let (X, d_X) be a complete metric space. Endow $Y \subset X$ with the induced metric d_Y. Recall that $d_Y(y, y') = d_X(y, y')$ for all $y, y' \in Y$.

Suppose (Y, d_Y) is complete. Let $a_n \in Y$, $n \in \mathbb{N}$, such that $a_n \to b$ in (X, d_X). Since (a_n) converges in (X, d_X), (a_n) is Cauchy in (X, d_X) (Proposition 9.17) and hence in (Y, d_Y). Since (Y, d_Y) is complete, (a_n) converges to a limit c in (Y, d_Y). Then also $a_n \to c$ in (X, d_X). Finally, by the uniqueness of limits, $b = c \in Y$. This shows that Y is closed in X.

Conversely, suppose Y is closed and let (a_n) be Cauchy in (Y, d_Y). Then (a_n) is Cauchy in (X, d_X), so since (X, d_X) is complete, (a_n) converges in (X, d_X) to a limit $b \in X$. Since Y is closed, $b \in Y$, so $a_n \to b$ in (Y, d_Y). This shows that (Y, d_Y) is complete. □

More exercises

9.7. Let (X, d) be a metric space. Let $a \in X$ and $r > 0$. Show that the open ball $B(a, r)$ is an open subset of X. *Hint.* You need to show that for each $y \in B(a, r)$, there is $s > 0$ such that $B(y, s) \subset B(a, r)$.

9.8. Let (X, d) be a metric space. Let $a \in X$ and $r \geq 0$. Show that the 'closed ball' $B = \{x \in X : d(x, a) \leq r\}$ with centre a and radius r is indeed a closed subset of X. *Hint.* You need to show that the complement $X \setminus B$ is open, that is, for every $y \in X \setminus B$ there is $\epsilon > 0$ such that the open ball $B(y, \epsilon)$ is contained in $X \setminus B$. Draw a picture!

9.9. Determine whether the following subsets of $\mathscr{C}([0, 1])$ are open or closed with respect to the uniform metric.

(a) $\{f : f(\frac{1}{2}) = 3\}$.

(b) $\{f : f(\frac{1}{2}) \neq 3\}$.

(c) $\{f : f(x) > 0 \text{ for all } x \in [0, 1]\}$.

(d) $\{f : |f(x)| \leq 2 \text{ for all } x \in [0, 1]\}$.

(e) $\{f : f \text{ is a polynomial function}\}$.

(f) $\{f : f \text{ is increasing}\}$.

(g) $\{f : f \text{ is differentiable}\}$.

9.10. Prove that a sequence (a_n) in a metric space X has a subsequence with limit $b \in X$ if and only if every neighbourhood of b contains a_n for infinitely many n.

9.11. (a) Let $I = [a, b]$, $a < b$. Show that the function d_1 on $\mathscr{C}(I) \times \mathscr{C}(I)$, defined by the formula $d_1(f, g) = \int_a^b |f - g|$, is a metric on $\mathscr{C}(I)$.

(b) Show that the metric space $(\mathscr{C}(I), d_1)$ is not complete.

9.12. In this exercise, we compare the supremum metric d on $\mathscr{C}(I)$, where $I = [a, b]$, $a < b$, and the metric d_1 defined in Exercise 9.11.

(a) Show that if a subset of $\mathscr{C}(I)$ is open with respect to d_1, then it is also open with respect to d.

(b) Find a subset of $\mathscr{C}(I)$ that is open with respect to d, but not with respect to d_1.

9.13. (a) Find an incomplete metric on $\mathbb{R}$.

(b) Does every set have a complete metric on it? Does every set have an incomplete metric on it?

9.14. Let p be a prime number. Show that the sequence $1, 1+p, 1+p+p^2, \ldots$ converges to $\dfrac{1}{1-p}$ in $(\mathbb{Q}, d_p)$ (see Exercise 9.2). Conclude that $\mathbb{Z}$ is not closed in $\mathbb{Q}$ with respect to the p-adic metric d_p if $p \geq 3$.

9.15. Let (X, d) be a metric space. Let E be a subset of X. Prove that the following are equivalent.

(1) There are $a \in X$ and $r > 0$ such that $E \subset B(a, r)$.

(2) For every $a \in X$, there is $r > 0$ such that $E \subset B(a, r)$.

(3) There is $R > 0$ such that $d(x, y) < R$ for all $x, y \in E$.

If these conditions hold, then we say that E is *bounded*.

9.16. (a) Show that if a function $f : \mathbb{R} \to \mathbb{R}$ is continuous, then its graph $\{(x, f(x)) : x \in \mathbb{R}\}$ is a closed subset of $\mathbb{R}^2$.

(b) Show that the converse fails in general, but holds if f is bounded.

9.17. Compactness for metric spaces is defined in exactly the same way as for subsets of $\mathbb{R}$ (Definition 4.8). A metric space X is said to be *compact* if every sequence in X has a subsequence that converges in X.

A subset K of a metric space X is said to be *compact* if it is compact when viewed as a metric space in its own right with the induced metric. This means that every sequence in K has a subsequence that converges, as a sequence in X, to a limit in K.

(a) Prove that a compact subset of a metric space is closed and bounded. (This is one half of the Heine-Borel theorem.)

(b) Prove that the union of finitely many compact subsets of a metric space is compact. *Hint.* You need to prove this directly from the definition of compactness. You cannot follow the proof of Corollary 4.11, because you do not have the other half of the Heine-Borel theorem (see Exercise 9.20).

9.18. (a) Prove that the three metrics d_1, d_2, d_∞ on $\mathbb{R}^n$ define the same notion of a subset of $\mathbb{R}^n$ being compact. *Hint.* Use Example 9.13.

(b) Prove that a subset of $\mathbb{R}^n$ of the form

$$\{x \in \mathbb{R}^n : a_i \le x_i \le b_i \text{ for } i = 1, \dots, n\},$$

where $a_i \le b_i$ for $i = 1, \dots, n$, is compact.

9.19. Prove that a compact metric space is complete.

9.20. Let $E = \{f \in \mathscr{C}([0, 1]) : 0 \le f(x) \le 1 \text{ for all } x \in [0, 1]\}$. Show that E is closed and bounded but not compact with respect to the uniform metric.

This shows that the characterisation of compact sets in $\mathbb{R}$ given by the Heine-Borel theorem (Theorem 4.10) fails for metric spaces in general.

9.21. An *open cover* of a metric space X is a collection of open subsets of X whose union is X. A subcollection whose union is X is called a *subcover*.

Suppose the metric space X is not compact. Prove that there is an open cover of X with no finite subcover. *Hint.* Use Exercise 9.10.

This shows that if a metric space X has the property that every open cover of X has a finite subcover, then X is compact. In fact, this property is equivalent to compactness, but the missing implication is not easy to prove.

The contraction principle

10.1. The contraction principle

The definitions of continuity from Chapter 5 (Definition 5.9) readily extend to the setting of metric spaces and they are still equivalent.

10.1. Definition. Let (X, d_X) and (Y, d_Y) be metric spaces. A map $f : X \to Y$ is *continuous* at $c \in X$ if the following equivalent conditions hold.

(i) For every $\epsilon > 0$, there is $\delta > 0$ such that if $x \in X$ and $d_X(x, c) < \delta$, then $d_Y(f(x), f(c)) < \epsilon$.

(ii) For every neighbourhood V of $f(c)$ in Y, there is a neighbourhood U of c in X such that $f(U) \subset V$.

(iii) If (x_n) is a sequence with $x_n \to c$ in X, then $f(x_n) \to f(c)$ in Y.

We say that f is *continuous* if f is continuous at each point of X.

Exercise 10.1. Show that definitions (i), (ii), and (iii) are equivalent.

We are particularly interested in continuous maps of a special kind.

10.2. Definition. Let (X, d_X) and (Y, d_Y) be metric spaces. A map $f : X \to Y$ is called a *contraction* if there is $\alpha \in [0, 1)$ such that

$$d_Y(f(x), f(x')) \leq \alpha\, d_X(x, x') \quad \text{for all } x, x' \in X.$$

A contraction is evidently continuous: given $\epsilon > 0$, just choose $\delta > 0$ so that $\alpha\delta < \epsilon$.

Now we state and prove the main theorem of this chapter, the contraction principle. It is also known as the Banach fixed point theorem.

10.3. Theorem (contraction principle)**.** Let (X, d) be a nonempty complete metric space and $f : X \to X$ be a contraction. Then f has a unique fixed point, that is, there is a unique point $p \in X$ such that $f(p) = p$.

Proof. By assumption, there is $\alpha \in [0, 1)$ such that $d(f(x), f(y)) \le \alpha\, d(x, y)$ for all $x, y \in X$. Note first that f has at most one fixed point. Namely, if p and q are fixed points of f, then

$$d(p, q) = d(f(p), f(q)) \le \alpha\, d(p, q),$$

so $d(p, q) = 0$ since $\alpha < 1$, and $p = q$.

Now choose any $x_1 \in X$ and recursively define a sequence (x_n) in X by the formula $x_{n+1} = f(x_n)$ for all $n \in \mathbb{N}$. Then, for every $n \in \mathbb{N}$,

$$d(x_{n+1}, x_n) \le \alpha\, d(x_n, x_{n-1}) \le \cdots \le \alpha^{n-1} d(x_2, x_1),$$

so for $n > m \ge 1$,

$$\begin{aligned} d(x_n, x_m) &\le d(x_n, x_{n-1}) + d(x_{n-1}, x_{n-2}) + \cdots + d(x_{m+1}, x_m) \\ &\le (\alpha^{n-2} + \alpha^{n-3} + \cdots + \alpha^{m-1}) d(x_2, x_1) \\ &= \alpha^{m-1}(1 + \alpha + \cdots + \alpha^{n-m-1}) d(x_2, x_1) \\ &\le \alpha^{m-1}\left(\sum_{k=0}^{\infty} \alpha^k\right) d(x_2, x_1) = \frac{\alpha^{m-1}}{1-\alpha}\, d(x_2, x_1). \end{aligned}$$

Since $\alpha \in [0, 1)$, $\dfrac{\alpha^{m-1}}{1-\alpha}\, d(x_2, x_1)$ can be made arbitrarily small by taking m large enough. Hence (x_n) is Cauchy, so since X is complete, (x_n) converges to a limit $p \in X$. Finally, since $x_n \to p$ and f is continuous, $x_{n+1} = f(x_n) \to f(p)$, so by the uniqueness of limits, $f(p) = p$. □

Note that the proof of the contraction principle is quite constructive. It shows that for any choice of a point $c \in X$, the sequence

$$c, f(c), f(f(c)), f(f(f(c))), \ldots$$

converges to the fixed point of f. In many cases, we can compute as many of these values as we please, and thus approximate the fixed point.

The contraction principle is a powerful tool for solving a wide variety of equations. Any equation $g(x) = h(x)$ whatsoever can be formulated as a fixed point problem: just write it as $f(x) = x$ with $f(x) = g(x) - h(x) + x$. If x can be interpreted as a point in a complete metric space and f as a contraction of that space, then the contraction principle shows that the equation has a unique solution. We shall now consider several examples that illustrate the contraction principle.

10.4. Example. As a first, very simple example, let $I = [-a, a]$ with $a \in (0, \frac{1}{2})$ and consider the map $f : I \to I$, $f(x) = x^2$. Since I is a closed subset of the complete metric space $\mathbb{R}$, I is complete with the induced metric. Also, for $x, y \in I$, $|f(x) - f(y)| = |x + y||x - y| \le 2a|x - y|$, so f is a contraction since $2a < 1$. Indeed, f has a unique fixed point, namely 0, and for every $c \in I$, the sequence $c, f(c) = c^2, f(f(c)) = c^4, \dots$ converges to 0.

By simply removing 0 from I, we get an incomplete metric space $X = I \setminus \{0\}$ and a contraction $f : X \to X$ without a fixed point.

10.5. Example. Note that if $1 \le x \le 5$, then $1 < \sqrt{5} \le \sqrt{3x+2} \le \sqrt{17} < 5$, so we have a map $f : [1, 5] \to [1, 5]$, $f(x) = \sqrt{3x+2}$. We claim that f is a contraction of the metric space $[1, 5]$, which is complete as a closed subspace of the complete space $\mathbb{R}$. Namely, f is differentiable with $f'(x) = \frac{3}{2}(3x+2)^{-1/2} \ge 0$, and the maximum of $f'(x)$ is $\alpha = f'(1) = \frac{3}{2\sqrt{5}} < 1$. Hence, if $x, y \in [1, 5]$, by the mean value theorem, there is c between x and y such that

$$|f(x) - f(y)| = |f'(c)||x - y| \le \alpha\,|x - y|.$$

Therefore, by the contraction principle, f has a unique fixed point $p \in [1, 5]$. Choosing 3 as an initial point and applying f a few times, we obtain the following sequence:

3

3.3166...

3.4568...

3.5171...

3.5428...

3.5536...

3.5582...

3.5601...

The function f in this example is simple enough that we can use the quadratic formula to solve the equation $f(p) = p$. We get $p = \frac{1}{2}(3 + \sqrt{17}) = 3.5615\dots$ (the other solution, $\frac{1}{2}(3 - \sqrt{17})$, lies outside $[1, 5]$).

Exercise 10.2. Define a map $g : \ell_\infty \to \ell_\infty$ by the formula

$$(x_1, x_2, x_3, \dots) \mapsto (1 + \tfrac{1}{2}x_1 + \tfrac{1}{3}x_2, 1 + \tfrac{1}{2}x_2 + \tfrac{1}{3}x_3, 1 + \tfrac{1}{2}x_3 + \tfrac{1}{3}x_4, \dots).$$

Show that $d(g(x), g(y)) \le \frac{5}{6}d(x, y)$ for all $x, y \in \ell_\infty$, so g is a contraction. Assuming that ℓ_∞ is complete (Exercise 9.4), conclude that g has a unique fixed point. Find the fixed point.

Let S be the set of all sequences of real numbers, bounded and unbounded. Then g extends to a map $G : S \to S$ defined by the same formula.

Show that G has infinitely many fixed points. Conclude that G is not a contraction with respect to any complete metric on S.

10.6. Example. Let $A = (a_{ij})$ be an $n \times n$ matrix with real entries. Let $b \in \mathbb{R}^n$. We want to solve the system of linear equations $Ax = b$ using the contraction principle, under a suitable condition on A. First, we turn $Ax = b$ into a fixed point equation, writing it as $Bx+b = x$ with $B = (b_{ij}) = I - A$, where $I = (\delta_{ij})$ is the $n \times n$ identity matrix. Define $f : \mathbb{R}^n \to \mathbb{R}^n$, $f(x) = Bx + b$. If f is a contraction with respect to any of the metrics d_1, d_2, or d_∞, then the contraction principle implies that $Ax = b$ has a unique solution that can be found as the limit of the sequence $c, f(c), f(f(c)), \dots$ for any $c \in \mathbb{R}^n$. Let us use d_∞. For all $x, y \in \mathbb{R}^n$,

$$d_\infty(f(x), f(y)) = d_\infty(Bx + b, By + b) = \max_{i=1,\dots,n} \left| \sum_{j=1}^{n} b_{ij}(x_j - y_j) \right|$$
$$\leq \max_{i=1,\dots,n} \sum_{j=1}^{n} |b_{ij}|\, |x_j - y_j| \leq \max_{i=1,\dots,n} \sum_{j=1}^{n} |b_{ij}|\, d_\infty(x, y).$$

Thus f is a contraction with respect to d_∞ if

$$\sum_{j=1}^{n} |b_{ij}| = \sum_{j=1}^{n} |a_{ij} - \delta_{ij}| < 1 \quad \text{for } i = 1, \dots, n.$$

This sufficient condition for f to be a contraction (saying, roughly speaking, that A is close to I) is an explicit condition on the matrix A that is very easy to check if the entries of A are known.

10.7. Example. Let $h \in (0, 1)$ and $I = [-h, h]$. Let $\mathscr{C}(I)$ be the metric space of all continuous functions $I \to \mathbb{R}$ with the uniform metric d (Example 9.7). Define a map $f : \mathscr{C}(I) \to \mathscr{C}(I)$ by letting $f(\phi)$ for $\phi \in \mathscr{C}(I)$ be the function $x \mapsto 1 + \int_0^x \phi$. Then f is well defined: $f(\phi)$ is not only continuous on I, so $f(\phi) \in \mathscr{C}(I)$, but in fact differentiable with $f(\phi)' = \phi$ by the fundamental theorem of calculus. We claim that f is a contraction. Namely, for $\phi, \psi \in \mathscr{C}(I)$,

$$d(f(\phi), f(\psi)) = \max_{x \in [-h,h]} |f(\phi)(x) - f(\psi)(x)| = \max_{x \in [-h,h]} \left| \int_0^x \phi - \int_0^x \psi \right|$$
$$\leq \max_{x \in [-h,h]} \left| \int_0^x |\phi - \psi| \right| \leq \max_{x \in [-h,h]} \left| \int_0^x d(\phi, \psi) \right| = h d(\phi, \psi).$$

Since $\mathscr{C}(I)$ is complete (Example 9.22), f has a unique fixed point. In other words, there is a unique continuous function $\phi : [-h, h] \to \mathbb{R}$ such that $\phi(x) = 1 + \int_0^x \phi$ for all $x \in [-h, h]$. We can find ϕ as the limit of the sequence $\phi_1, f(\phi_1), f(f(\phi_1)), \dots$ for any $\phi_1 \in \mathscr{C}(I)$. As a simple choice, take $\phi_1 = 1$. We get:

$\phi_1 = 1,$
$\phi_2 = f(\phi_1) = 1 + \int_0^x 1 = 1 + x,$
$\phi_3 = f(\phi_2) = 1 + \int_0^x (1 + t)dt = 1 + x + \frac{1}{2}x^2,$
$\phi_4 = f(\phi_3) = 1 + \int_0^x (1 + t + \frac{1}{2}t^2)dt = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3, \ldots.$

It is quite clear what ϕ is, isn't it?

10.2. Picard's theorem

Our final goal is to prove a fundamental theorem on the existence and uniqueness of solutions to differential equations of a very general kind. Our strategy is to express the differential equation as a fixed point problem in the space $\mathscr{C}(I)$ of continuous functions on a compact interval I, and then apply completeness of $\mathscr{C}(I)$ and the contraction principle to conclude that the fixed point problem has a unique solution.

Here is the set-up. Let D be an open subset of $\mathbb{R}^2$ and $f : D \to \mathbb{R}$ be a continuous function. We want to solve the *initial value problem*

$$\frac{dy}{dx} = f(x, y), \quad y(x_0) = y_0, \tag{1}$$

where (x_0, y_0) is a given point in D. This means finding a continuously differentiable function ϕ on some interval I containing x_0, such that:

- the graph of ϕ lies in D,
- $\phi'(x) = f(x, \phi(x))$ for all $x \in I$,
- $\phi(x_0) = y_0$.

Key observation. If ϕ is a solution of (1), then, by the fundamental theorem of calculus,

$$\phi(x) = \phi(x_0) + \int_{x_0}^{x} \phi'(t)\, dt = y_0 + \int_{x_0}^{x} f(t, \phi(t))\, dt$$

for all $x \in I$. Conversely, if $\phi : I \to \mathbb{R}$ is a continuous function on an interval I containing x_0, such that the graph of ϕ lies in D and

$$\phi(x) = y_0 + \int_{x_0}^{x} f(t, \phi(t))\, dt \quad \text{for all } x \in I, \tag{2}$$

then ϕ is a solution of (1), again by the fundamental theorem of calculus (see Exercise 10.5). Thus solving the initial value problem (1) is equivalent to finding a continuous solution ϕ of the integral equation (2).

Note that (2) is a fixed point problem: it says that the function ϕ is a fixed point of the map that takes ϕ to the function $x \mapsto y_0 + \int_{x_0}^{x} f(t, \phi(t))\, dt$. Thus the initial value problem (1) has been reformulated as a fixed point problem in the space of continuous functions on I.

We need to impose a mild additional condition on f, a so-called *Lipschitz condition* (see Remark 10.11). We assume that there is a rectangle

$$R = \{(x, y) \in \mathbb{R}^2 : |x - x_0| \le a, |y - y_0| \le b\} \subset D$$

with $a, b > 0$, such that there is a constant $K > 0$ with

$$|f(x, y) - f(x, y')| \le K|y - y'|$$

for all $(x, y), (x, y') \in R$.

If the partial derivative $\partial f/\partial y$ exists and is continuous on D, then the Lipschitz condition is satisfied for every rectangle $R \subset D$. Namely, since R is compact (Exercise 9.18), $\partial f/\partial y$ is bounded on R (Exercises 10.8 and 10.9), so there is $K > 0$ with $|\partial f/\partial y| \le K$ on R. The Lipschitz condition then follows from the mean value theorem (Theorem 6.14) applied to $f(x, y)$ as a function of y with x fixed.

Since f is continuous and R is compact, f is bounded on R, so there is a constant $M > 0$ with $|f| \le M$ on R. Let h be any positive number with

$$h \le a, \quad h \le b/M, \quad h < 1/K.$$

We will show that (1) has a unique solution on $I = [x_0 - h, x_0 + h]$.

Let A be the set of those $\phi \in \mathscr{C}(I)$ whose graph lies in R, that is, for which $|\phi - y_0| \le b$ on I.

Exercise 10.3. Show that if $\phi_n \in A$, $n \in \mathbb{N}$, and $\phi_n \to \phi$ uniformly on I, then $\phi \in A$. Hence A is closed in $\mathscr{C}(I)$ (Theorem 9.27), so A is complete with respect to the uniform metric d (Theorem 9.28).

We note that if ϕ is a solution of (1), then $\phi \in A$. Otherwise, there is $x \in I$, say $x > x_0$, with $|\phi(x) - y_0| > b$. Let $w \in (x_0, x_0 + h)$ be the infimum of such x. Then $|\phi(w) - y_0| = b$. By the mean value theorem applied to ϕ on $[x_0, w]$, there is $c \in (x_0, w)$ with $\phi(w) - \phi(x_0) = \phi'(c)(w - x_0)$. Then $(c, \phi(c)) \in R$ and

$$b = |\phi(w) - y_0| = |f(c, \phi(c))|\,(w - x_0) < Mh \le b,$$

which is absurd.

Now define $F : A \to A$ to be the map that takes $\phi \in A$ to the function

$$F(\phi) : I \to \mathbb{R}, \quad x \mapsto y_0 + \int_{x_0}^{x} f(t, \phi(t))\,dt.$$

Since the graph of ϕ lies in $R \subset D$, the integrand $t \mapsto f(t, \phi(t))$ is well defined and continuous on I (Exercise 10.5), so by the fundamental theorem of calculus, $F(\phi)$ is not only continuous but even differentiable. Moreover, $F(\phi) \in A$, because for $x \in I$,

$$|F(\phi)(x) - y_0| = \left| \int_{x_0}^{x} f(t, \phi(t))\,dt \right| \le \left| \int_{x_0}^{x} M\,dt \right| \le Mh \le b.$$

We claim that F is a contraction. Namely, for $\phi, \psi \in A$ and $x \in I$,

$$\begin{aligned} |F(\phi)(x) - F(\psi)(x)| &= \left| \int_{x_0}^{x} \big(f(t,\phi(t)) - f(t,\psi(t))\big)\, dt \right| \\ &\le \left| \int_{x_0}^{x} \big|f(t,\phi(t)) - f(t,\psi(t))\big| dt \right| \\ &\le K \left| \int_{x_0}^{x} |\phi(t) - \psi(t)|\, dt \right| \\ &\le K|x - x_0| \max_{t\in I} |\phi(t) - \psi(t)| \\ &\le Kh\, d(\phi,\psi), \end{aligned}$$

so

$$d(F(\phi), F(\psi)) \le Kh\, d(\phi,\psi).$$

Since $Kh < 1$, F is a contraction. As A is complete, we conclude that F has a unique fixed point. In other words, there is a unique continuously differentiable function ϕ on I, whose graph lies in D, and which solves the equivalent problems (1) and (2).

Let us summarise what we have proved.

10.8. Theorem (Picard's theorem)**.** Let D be an open subset of $\mathbb{R}^2$ and $f : D \to \mathbb{R}$ be a continuous function. Let $(x_0, y_0) \in D$ and

$$R = \{(x,y) \in \mathbb{R}^2 : |x - x_0| \le a, |y - y_0| \le b\} \subset D$$

with $a, b > 0$, such that there is a constant $K > 0$ with

$$|f(x,y) - f(x,y')| \le K|y - y'|$$

for all $(x,y), (x,y') \in R$. Take $M > 0$ with $|f| \le M$ on R. Let h be any positive number with

$$h \le a, \quad h \le b/M, \quad h < 1/K.$$

Then there is a unique continuously differentiable function ϕ on the interval $I = [x_0 - h, x_0 + h]$, such that the graph of ϕ lies in D and ϕ solves the initial value problem

$$\phi'(x) = f(x, \phi(x)) \text{ for all } x \in I, \qquad \phi(x_0) = y_0.$$

In fact, the graph of ϕ lies in R.

10.9. Example. As a first example, let $(x_0, y_0) = (0,1) \in D = \mathbb{R}^2$ and $f(x,y) = y$, so we can let $K = 1$. With R as in Theorem 10.8, we can take $M = \max_R |f| = b + 1$, so to get a unique solution to the initial value problem $y' = y$, $y(0) = 1$, on $[-h, h]$, the number $h > 0$ must satisfy $h \le a$, $h \le b/M = b/(b+1)$, and $h < 1/K = 1$. Every $h < 1$ satisfies these inequalities if b is chosen large enough and, say, $a = 1$. Thus, for every $h \in (0,1)$, the initial value problem has a unique solution on $[-h, h]$. The

solution is the limit of the sequence $\phi_1, F(\phi_1), F(F(\phi_1)), \ldots$, where F takes ϕ to $x \mapsto 1 + \int_0^x \phi$, and ϕ_1 is any function in $\mathscr{C}([-h,h])$. With $\phi_1 = 1$, this iteration was carried out in Example 10.7. The solution, of course, is the exponential function.

10.10. Example. There is $\epsilon > 0$ and a continuously differentiable function $g : (-\epsilon, \epsilon) \to \mathbb{R}$ such that $g'(x) = \dfrac{x^5 \log(3 + g(x)^4)}{2e^{g(x)} - \cos(x^3 g(x)) + 1}$ for all $x \in (-\epsilon, \epsilon)$ and $g(0) = -7$. This follows from Picard's theorem simply because the continuous function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x,y) = \dfrac{x^5 \log(3 + y^4)}{2e^y - \cos(x^3 y) + 1}$, is differentiable with respect to y, and $\partial f/\partial y$ is continuous on $\mathbb{R}^2$.

10.11. Remark. If we omit the Lipschitz condition in Picard's theorem, the uniqueness of the solution may fail. For example, the initial value problem $y' = y^{1/3}$, $y(0) = 0$, has as a solution not only the function that is identically zero, but also the continuously differentiable function ϕ with $\phi(x) = 0$ for $x \le 0$ and $\phi(x) = (\frac{2}{3}x)^{3/2}$ for $x \ge 0$.

A solution still exists on a small enough interval without the Lipschitz condition (that is, with f merely continuous), but a different and more difficult method of proof is required.

The following example illustrates the importance of the uniqueness part of Picard's theorem. Uniqueness can help us extend solutions to larger intervals, and it may provide additional information about solutions.

10.12. Example. Let us determine the largest interval on which Picard's theorem guarantees a solution to the initial value problem

$$y' = 1 + y^2, \qquad y(0) = 0.$$

Here, $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x,y) = 1 + y^2$. Let $R = \{(x,y) \in \mathbb{R}^2 : |x| \le a, |y| \le b\}$, with $a, b > 0$. If we set

$$h(a,b) = \min\left\{a, \frac{b}{\max\limits_R |f|}, \frac{1}{\max\limits_R |\partial f/\partial y|}\right\},$$

then by Picard's theorem, there is a unique solution on $[-r, r]$ for every $r < h(a,b)$. By uniqueness, if $r < s < h(a,b)$, the solutions on $[-r,r]$ and $[-s,s]$ must agree on $[-r,r]$. Hence there is in fact a unique solution on $(-h(a,b), h(a,b))$. We need to maximise $h(a,b)$.

Now $\max_R |f| = 1 + b^2$. Also, $\partial f/\partial y = 2y$, so $\max_R |\partial f/\partial y| = 2b$. It is an easy exercise to show that the maximum of $\frac{b}{1+b^2}$ for $b > 0$ is $\frac{1}{2}$, taken at $b = 1$. There, $\frac{1}{2b}$ also equals $\frac{1}{2}$, so the largest h can be is $\frac{1}{2}$. Thus the largest interval on which Picard's theorem guarantees a solution is $I = (-\frac{1}{2}, \frac{1}{2})$.

Let ϕ be the unique solution on I. Let $\psi : I \to \mathbb{R}$, $\psi(x) = -\phi(-x)$. Then $\psi'(x) = \phi'(-x) = 1 + \phi(-x)^2 = 1 + \psi(x)^2$ and $\psi(0) = 0$, so ψ is also a solution on I. By uniqueness, $\psi = \phi$, that is, $\phi(-x) = -\phi(x)$ for all $x \in I$. Thus uniqueness implies that ϕ is an odd function.

There is in fact a solution on a larger interval, namely the tangent on $(-\frac{\pi}{2}, \frac{\pi}{2})$. Picard's theorem applies to a very large class of equations, and yet its proof is relatively easy. The trade-off is that the theorem cannot be expected to produce the largest interval on which a solution exists.

More exercises

10.4. Let $f : X \to Y$ and $g : Y \to Z$ be maps of metric spaces. Show that if f and g are continuous, then the composition $g \circ f : X \to Z$ is also continuous.

10.5. Let D be an open subset of $\mathbb{R}^2$ and $f : D \to \mathbb{R}$ be continuous. Let $I \subset \mathbb{R}$ be an interval and $\phi : I \to \mathbb{R}$ be continuous. Suppose the graph of ϕ lies in D, so that the composition $g : I \to \mathbb{R}$, $g(t) = f(t, \phi(t))$, is defined. Prove that g is continuous. *Hint.* Use sequences and recall Example 9.13.

10.6. Let (X, d_X) and (Y, d_Y) be metric spaces and $f : X \to Y$ be a map. Show that the following are equivalent.

(i) f is continuous, meaning that for every $a \in X$ and $\epsilon > 0$, there is $\delta > 0$ such that if $d_X(x, a) < \delta$, then $d_Y(f(x), f(a)) < \epsilon$.

(ii) For every open subset V of Y, the preimage $f^{-1}(V)$ is open in X.

10.7. Let $X = \mathscr{C}([a, b])$ be the set of continuous functions $[a, b] \to \mathbb{R}$ with the uniform metric. Let $f \in X$. Show that the map $F : X \to X$ with $F(g) = fg$ (meaning f times g) is continuous.

10.8. Prove the following generalisation of Theorem 5.16. If X and Y are metric spaces, $f : X \to Y$ is a continuous map, and $K \subset X$ is compact, then the image $f(K)$ is compact.

10.9. Prove the following generalisation of the extreme value theorem (Theorem 5.17). A continuous real-valued function on a nonempty compact metric space has a maximum and a minimum value.

10.10. Let X be a nonempty discrete metric space. Explicitly describe the contractions $X \to X$. Does every contraction $X \to X$ have a unique fixed point?

10.11. Show that a differentiable function $f : \mathbb{R} \to \mathbb{R}$ with $|f'(x)| \leq \frac{1}{2}$ for all $x \in \mathbb{R}$ has a unique fixed point.

10.12. Show that the map $f : [0, 2] \to [0, 2]$, $f(x) = \sqrt[3]{2x + 1}$, is a contraction. *Hint.* Use the method of Example 10.5.

10.13. Use the contraction principle to show that there is a unique real number $a \leq -1$ such that $e^a + a = -1$.

10.14. Use the contraction principle to show that there is a unique real number $a \geq 2$ such that $\log a = a - 2$.

10.15. Let $X = \mathscr{C}([3,5])$ be the set of continuous functions $[3,5] \to \mathbb{R}$ with the supremum metric. Let the map $f : X \to X$ take $\phi \in X$ to the function $f(\phi)$ with

$$f(\phi)(x) = 2\int_4^x \frac{\phi(t)}{t}\,dt + 4.$$

This is a well-defined map because by the fundamental theorem of calculus, $f(\phi)$ is not only continuous, so $f(\phi) \in X$, but even differentiable on $[3,5]$.

(a) Show that f is a contraction on X.

(b) Find the unique fixed point of f. You may use any method you can think of, as long as you verify that the function you come up with is indeed a fixed point for f.

10.16. Define a map $f : \mathscr{C}([0,1]) \to \mathscr{C}([0,1])$ by setting

$$f(\phi)(x) = x + \int_0^x t\phi(t)\,dt.$$

Show that f is a contraction with respect to the uniform metric on $\mathscr{C}([0,1])$. Show that its fixed point is a solution of the differential equation $y' = xy+1$. (You do not have to find the solution.)

10.17. Use Picard's theorem to show that the initial value problem $y' = y^2$, $y(0) = 1$, has a unique solution on $(-\frac{1}{4}, \frac{1}{4})$. Solve the equation by separation of variables and find the largest interval to which the solution extends.

10.18. Use Picard's theorem to show that the initial value problem $y' = x^2 + y^2$, $y(0) = 0$, has a unique solution on $(-\sqrt{2}/2, \sqrt{2}/2)$. Show that the solution is an odd function.

Index

absolute value, 5
absolutely convergent series, 30
addition, 2
additive
  identity, 2
  inverse, 2
algebraic limit theorem, 25, 47
algebraic number, 22
alternating harmonic series, 31
alternating series test, 31
antiderivative, 68
Archimedean property, 17
arcsine, 87
associativity, 2
axiom of completeness, 16

ball
  closed, 100
  open, 95
Banach fixed point theorem, 104
bijection, 10
bijective function, 10
binary expansion, 37
Bolzano-Weierstrass theorem, 33
bound
  greatest lower, 16
  least upper, 16
  lower, 15
  upper, 15
boundary, 43
bounded
  function, 53
  sequence, 25
  set, 15, 101

cancellation law, 12
  for functions, 13
Cantor set, 43
cardinality, 19
Cauchy criterion, 33
Cauchy sequence, 33, 97
Cauchy-Schwarz inequality, 92
centre, 77
chain rule, 57
change of variables, 72
closed
  ball, 100
  interval, 4
  set, 40, 99
closure, 42
codomain, 9
coefficient, 77
commutativity, 2
compact
  metric space, 101
  set, 41, 101
comparison test, 29
complement, 7
complete metric space, 97
completeness, axiom of, 16
composition, 9
concave function, 61
conditionally convergent series, 30
continuous function, 47, 103
  at a point, 47, 103
continuously differentiable function, 55
contraction, 103
contraction principle, 104
contrapositive, 2
convergent
  sequence, 23, 95
  series, 28, 76

converse, 2
convex function, 61
cosine, 84
countable set, 20
countably infinite set, 20
cover, open, 101
critical point, 58

Darboux's theorem, 58
De Morgan's laws, 7
decimal expansion, 37
decreasing
  function, 51
  sequence, 27
degenerate interval, 4
degree, 48
dense set, 18
derivative, 55
differentiable function, 55
  at a point, 55
Dini's theorem, 88
discrete space, 92
disjoint sets, 7
distance function, 91
distributivity, 2
divergent
  sequence, 23
  series, 28
domain, 9

e, 69, 70, 80
element, 6
empty set $\varnothing$, 7
equinumerous sets, 19
equivalent metrics, 93
error function, 81
Euclidean metric, 92
Euclidean norm, 92
Euler's constant γ, 72
even function, 61
eventually constant sequence, 95
expansion
  binary, 37
  decimal, 37
  to a base, 37
exponential function, 70
extreme point, 58
extreme value theorem, 49

family of sets, 8
fibre, 9
field, 2
  ordered, 3
function, 9
  bijective, 10
  bounded, 53
    above, 53
    below, 53
    locally, 54
  concave, 61
  continuous, 47, 103
    at a point, 47, 103
    uniformly, 49
  convex, 61
  decreasing, 51
    strictly, 51
  differentiable, 55
    at a point, 55
    continuously, 55
  even, 61
  exponential, 70
  identity, 9
  increasing, 51
    strictly, 51
  injective, 10
  integrable, 64
  inverse, 11
  monotone, 51
    strictly, 51
  odd, 61
  one-to-one, 10
  onto, 10
  periodic, 86
  polynomial, 48
  rational, 48
  Riemann integrable, 64
  surjective, 10
  trigonometric, 33, 83–87
fundamental theorem of calculus, 67

geometric series, 29
graph, 11
greatest lower bound, 16

harmonic series, 29
Heine-Borel theorem, 41

identity
  additive, 2
  multiplicative, 2
identity function, 9
image
  inverse, 9
  of a function, 9
  of a subset, 9
  of an element, 9
improper integral, 71
increasing
  function, 51
  sequence, 27
indefinite integral, 68
index set, 8
induced metric, 95
induction, 1

inductive set, 22
inductively defined sequence, 27
inequality
  Cauchy-Schwarz, 92
  triangle, 5, 91
infimum, 16
inflection point, 89
initial value problem, 107
injection, 10
injective function, 10
inner product, 92
integer, 1
integrable function, 64
integral, 64
  improper, 71
  indefinite, 68
  lower, 64
  upper, 64
integral test, 72
integration by parts, 72
interior, 43
intermediate value theorem, 50
intersection, 7, 8
interval, 4
  closed, 4
  degenerate, 4
  nondegenerate, 4
  of convergence, 78
    open, 78
  open, 4
inverse
  additive, 2
  multiplicative, 2
inverse function, 11
inverse function theorem, 57
inverse image, 9
inverse sine, 87
isolated point, 47

L^1 metric, 93
L^2 metric, 93
L^∞ metric, 93
Lagrange's remainder theorem, 81
least upper bound, 16
L'Hôpital's rule, 60
limit
  inferior, 38
  of a function, 45, 54
  of a sequence, 23, 95
  superior, 37
limit comparison test, 30
limit point, 45
Lipschitz condition, 108
locally bounded function, 54
logarithm (natural), 69
lower
  bound, 15
  integral, 64
  sum, 63

Maclaurin series, 81
map, mapping, 9, *see also* function
maximum, 5, 16
maximum metric, 93
mean value theorem, 59
  for integrals, 68
  generalised, 60
metric, 91
  equivalent, 93
  Euclidean, 92
  induced, 95
  L^1, 93
  L^2, 93
  L^∞, 93
  maximum, 93
  p-adic, 93
  supremum, 94
  uniform, 94
metric space, 91
  compact, 101
  complete, 97
  discrete, 92
  ultrametric, 93
minimum, 5, 16
monotone
  function, 51
  sequence, 27
monotone convergence theorem, 27
multiplication, 2
multiplicative
  identity, 2
  inverse, 2

natural logarithm, 69
natural number, 1, 22
negative number, 4
neighbourhood, 24, 95
nested interval property, 19
nondegenerate interval, 4
norm, Euclidean, 92
number
  algebraic, 22
  integer, 1
  natural, 1, 22
  negative, 4
  positive, 4
  rational, 1
  transcendental, 22

odd function, 61
one-to-one function, 10
onto function, 10
open
  ball, 95

cover, 101
interval, 4
set, 39, 98
or (conjunction), 7
order limit theorem, 26
ordered field, 3
ordered pair, 8

p-adic metric, 93
pair, ordered, 8
partial sum, 28
partition, 63
period, 86
period group, 86
periodic function, 86
π, 86
Picard's theorem, 109
pointwise convergent
sequence, 73
series, 76
polynomial function, 48
positive number, 4
power series, 77
preimage, 9
product of sets, 8
product rule, 56
proper subset, 6

quotient rule, 56

radius of convergence, 78
range, 9
ratio test, 30
rational function, 48
rational number, 1
rearrangement, 31
recursion formula, 27
recursively defined sequence, 27
refinement, 63
reflexive relation, 20
remainder, 81
Riemann integrable function, 64
Rolle's theorem, 59
root, 18, 48, 53, 58
root test, 31
rule, 9, 11

sequence, 23
bounded, 25
above, 25
below, 25
Cauchy, 33, 97
convergent, 23, 95
pointwise, 73
uniformly, 74
decreasing, 27
strictly, 27
divergent, 23
eventually constant, 95
increasing, 27
strictly, 27
inductively defined, 27
monotone, 27
strictly, 27
recursively defined, 27
series, 28
alternating harmonic, 31
convergent, 28, 76
absolutely, 30
conditionally, 30
pointwise, 76
uniformly, 76
divergent, 28
geometric, 29
harmonic, 29
Maclaurin, 81
power, 77
Taylor, 81
set, 6
bounded, 15, 101
above, 15
below, 15
closed, 40, 99
compact, 41, 101
countable, 20
countably infinite, 20
dense, 18
inductive, 22
open, 39, 98
symmetric, 61
uncountable, 20
sine, 84
inverse, 87
source, 9
squeeze theorem, 25, 46
strictly decreasing
function, 51
sequence, 27
strictly increasing
function, 51
sequence, 27
strictly monotone
function, 51
sequence, 27
subcover, 101
subgroup, 85
subsequence, 32
subset, 6
proper, 6
subspace, 95
substitution, 72
sum
lower, 63
of a series, 28

upper, 63
supremum, 16
supremum metric, 94
surjection, 10
surjective function, 10
symmetric relation, 20
symmetric set, 61

target, 9
Taylor series, 81
term, 23
test
alternating series, 31
comparison, 29
integral, 72
limit comparison, 30
ratio, 30
root, 31
Weierstrass M-, 76
theorem
algebraic limit, 25, 47
Banach fixed point, 104
Bolzano-Weierstrass, 33
Darboux's, 58
Dini's, 88
extreme value, 49
fundamental, of calculus, 67
Heine-Borel, 41
intermediate value, 50
inverse function, 57
Lagrange's remainder, 81
mean value, 59
for integrals, 68
generalised, 60
monotone convergence, 27
order limit, 26
Picard's, 109
Rolle's, 59
squeeze, 25, 46
transcendental number, 22
transitive relation, 4, 20
triangle inequality, 5, 91
trigonometric function, 33, 83–87

ultrametric, 93
uncountable set, 20
uniform metric, 94
uniformly continuous function, 49
uniformly convergent
sequence, 74
series, 76
union, 6, 8
upper
bound, 15
integral, 64
sum, 63

value, 9

Weierstrass M-test, 76
well-ordering property, 1, 22

CHAPTER (1)

ERRORS

1. Introduction

In numerical analysis solving a problem is only a part of the process. Another part is to know how far the results are accurate. This is a very important part and is often more difficult than achieving the results themselves. In this part, we take in consideration the errors that arise whether from rounding errors in arithmetic operations or from some other source. Throughout this book, as we look at the numerical solution of various problems, we will simultaneously consider the errors involved in whatever computational procedure is being used.

2. Absolute error and relative error

The error in a computed quantity is defined as

$$\text{Error} = \text{true value} - \text{approximate value}$$

The relative error is a measure of the error in relation to the size of the true value:

$$\text{Relative error} = \frac{\text{error}}{\text{true value}}$$

To simplify the notation when working with these numbers, we will usually denote the true and approximate values of a number x by x_T and x_A, respectively. Then we write

$$\text{Error}(x_A) = x_T - x_A$$

$$\text{Rel}(x_A) = \frac{x_T - x_A}{x_T}$$

As an illustration, consider the well-known approximation

$$\pi = \frac{22}{7}$$

Here $x_T = \pi = 3.14159265\ldots$ and $x_A = 22/7 = 3.1428571\ldots$,

$$\text{Error}\left(\frac{22}{7}\right)=\pi-\frac{22}{7}=-0.00126$$

$$\text{Rel}\left(\frac{22}{7}\right)=\frac{\pi-(22/7)}{\pi}=-0.000402$$

An idea related to relative error is that of significant digits. For a number x_A, the number of its leading digits that are correct relative to the corresponding digits in the true value x_T is called the number of significant digits in x_A. For a more precise definition, assuming the numbers are written in decimal, calculate the magnitude of the error $|x_T - x_A|$. If this error is less than or equal to five units in the $(m+1)$st digit of x_T, counting rightward from the first nonzero digit, then we say x_A has, at least, m significant digits of accuracy relative to x_T. In other words, we say that x_A has m significant digits with respect to x_T if

$$\left|\frac{x_T - x_A}{x_T}\right| \le 0.5\times 10^{-m} \qquad (1.1)$$

Example (1.1)

(a) $x_A = 0.222$ has three digits of accuracy relative to $x_T = 2/9$.

(b) $x_A = 23.496$ has four digits of accuracy relative to $x_T = 23.494$.

(c) $x_A = 0.02138$ has just two digits of accuracy relative to $x_T = 0.02144$.

(d) $x_A = 22/7$ has three digits of accuracy relative to $x_T = \pi$.

Most people find it easier to measure relative error than significant digits; and in some textbooks, satisfaction of (1.1) is used as the definition of x_A having m significant digits of accuracy.

3. Functional error

If e is the error in the approximated value x_A to the true value x_T so that $x_T = x_A + e$. If e_f denote the error when a function f is evaluated at x_A instead of at x_A, we have

$$f(x_T) = f(x_A) + e_f$$

Therefore

$$e_f = f(x_T) - f(x_A)$$
$$= f(x_A + e) - f(x_A)$$

Expanding $f(x_A + e)$ in a Taylor series, we have

$$e_f = f(x_A + e) - f(x_A)$$
$$= f(x_A) + ef'(x_A) + \frac{1}{2}e^2 f''(x_A) + \ldots - f(x_A)$$

Therefore,

$$e_f = ef'(x_A) + \frac{1}{2}e^2 f''(x_A) + \cdots$$

Hence if e is small (and the second and higher derivatives of f evaluated at x_A are not excessively large) we see that

$$e_f \simeq ef'(x_A)$$

Thus

$$|e_f| \approx |e||f'(x_A)|$$

and if x_A has m significant digits of accuracy, then

$$|e_f| \leq 0.5 \times 10^{-m} |f'(x_A)|$$

4. Sources of errors

Imagine solving a scientific-mathematical problem, and suppose this involves a computational procedure. Errors will usually be involved in this process, often of several different kinds. We will give a simple classification of the kinds of error that might occur.

A. Round-off error

When carrying out numerical calculations, digital computers have precision limit on their ability to represent numbers.

The difference between the result produced by a given algorithm using exact arithmetic and the result produced by the same algorithm using finite-precision, rounded arithmetic is called the round-off error. For example,

$$\text{Exact number } \frac{4}{3} = 1.333\ldots$$

$$\text{Rounded number to four significant digits } \frac{4}{3} = 1.333$$

$$\text{Hence Round-off error} = 1.333\ldots - 1.333 = 0.00033\ldots$$

$$\text{Exact number } \frac{5}{3} = 1.666\ldots$$

$$\text{Rounded number to four significant digits } \frac{5}{3} = 1.667$$

$$\text{Hence, Round-off error=1.666\ldots-1.667=-0.000333\ldots}$$

The following table shows the result of rounding exact numbers to N significant digits:

Number	N	Round number	Round-of error
23.764462	5	23.764	0.000462
0.0092746	3	0.00927	0.0000046
1.650045	3	1.6500	0.000045
0.0003786	3	0.000379	−0.0000004
0.57386	3	0.574	−0.00014

The following table shows the result of rounding exact numbers to N decimal places:

Number	N	Round number	Round-of error
23.764462	5	23.76446	0.000002
0.0092746	3	0.009	0.0002746
0.0003786	3	0.000	0.0003786
1.650045	5	1.65004	0.000005
0.57386	3	0.574	–0.00014

B. Imprecision of the given data

If the data are obtained experimentally, then they are known within the limits of experimental error (which can normally be estimated), and this will limit the accuracy of the results of any subsequent calculations. This is obvious fact that the accuracy of results is limited by the accuracy of any initial data.

C. Mistakes

Mistakes are errors which are created by the person performing the calculations. A common mistake is to invert the order of two digits occurring in a number. For example, it is very easy to use the number 62381 instead of the number 63281.

When doing calculations, as many checks as possible should be incorporated in the method itself so that any mistakes come quickly to light.

D. Mathematical approximation error (Truncation error)

Mathematical approximation errors are due to replacing an exact quantity by an approximation one. For example, we introduce an error if we use only a finite number of terms from an infinite series expansion. This error is called a truncation error, that is, the error due to truncating the series somewhere. For example $\sin x$ can be expressed as the infinite series expansion

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \frac{x^{11}}{11!} + \ldots$$

and when x is small the sum of the first three terms, namely

$$x - \frac{x^3}{3!} + \frac{x^5}{5!}$$

will give a good approximation to $\sin x$. The truncation error is then the sum of the remaining terms of the infinite series expansion namely

$$-\frac{x^7}{7!} + \frac{x^5}{9!} - \frac{x^{11}}{11!} + \ldots$$

In general if $f(x)$ is approximated using Taylor series about x_0, where $x = x_0 + h$, then

$$f(x) = f(x_0) + \frac{h}{1!} f'(x_0) + \frac{h^2}{2!} f''(x_0) + \ldots + \frac{h^n}{h!} f^{(n)}(x_0) + R_n$$

where R_n is the truncation error. This error can be calculated as

$$R_n = \frac{h^{n+1}}{(n+1)!} f^{(n+1)}(x_0 + \theta h),\ 0 \le \theta \le 1$$

CHAPTER (2)

FINITE DIFFERENCES

1. Introduction

The calculus of finite differences plays an important role in Numerical methods. It deals with the variations in a function when the independent variable changes by finite jumps which may be equal or unequal. In this chapter, we shall study the variations in a function due to the changes in the independent variable by equal intervals.

2. Finite differences

Let $y = f(x)$ be a discrete function. If $x_0,\ x_0 + h,\ x_0 + 2h, \cdots$ $x_0 + nh$ are the successive values of x , where two consecutive values differ by a quantity h, then the corresponding values of y are $y_0, y_1, y_2, \ldots, y_n$. The value of the independent variable x is usually called the arguments and the corresponding functional value is known as the entry. The arguments and entries can be shown in a tabular form as follows:

Argument x	x_0	$x_1 = x_0 + h$	$x_2 = x_0 + 2h$	…	$x_n = x_0 + nh$
Entry $y = f(x)$	$y_0 = f(x_0)$	$y_1 = f(x_0 + h)$	$y_2 = f(x_0 + 2h)$	…	$y_n = f(x_0 + nh)$

To determine the values of $f(x)$ or $f'(x)$ etc., for some intermediate arguments, the following three types of differences are found useful:

(i) Forward differences
(ii) Backward differences
(iii) Central differences

3. Forward differences

If we subtract from each value of y (except y_0) the preceding value of y we get $y_1 - y_0, y_2 - y_1, \ldots, y_n - y_{n-1}$ respectively, known as the first differences of y . These results which may be denoted $\Delta y_0, \Delta y_1, \ldots, \Delta y_n$

i.e.

$$\Delta y_0 = y_1 - y_0 \ , \Delta y_1 = y_2 - y_1 \ , \ldots, \ \Delta y_{n-1} = y_n - y_{n-1}$$

where Δ is a symbol representing an operation of forward difference, are called first forward differences. Thus, the first forward differences are given by

$$\Delta y_i = y_{i+1} - y_i \ , i = 0,1,2,\ldots,n.$$

Now, the second forward differences are defined as the differences of the first differences, that is,

$$\begin{aligned}
\Delta^2 y_0 &= \Delta(\Delta y_0) = \Delta(y_1 - y_0) = \Delta y_1 - \Delta y_0 \\
&= (y_2 - y_1) - (y_1 - y_0) = y_2 - 2y_1 + y_0 \\
\Delta^2 y_1 &= \Delta(\Delta y_1) = \Delta y_2 - \Delta y_1 = y_3 - 2y_2 + y_1 \\
&\ldots \qquad\qquad \ldots \\
\Delta^2 y_n &= \Delta y_{n+1} - \Delta y_n = y_{n+2} - 2y_{n+1} + y_n
\end{aligned}$$

Here, Δ^2 is called second forward difference operator.

Similarly, the third forward differences are:

$$\begin{aligned}
\Delta^3 y_0 &= \Delta(\Delta^2 y_0) = \Delta^2 y_1 - \Delta^2 y_0 = \Delta(\Delta y_1) - \Delta(\Delta y_0) \\
&= \Delta(y_2 - y_1) - \Delta(y_1 - y_0) = \Delta y_2 - 2\Delta y_1 + \Delta y_0 \\
&= (y_3 - y_2) - 2(y_2 - y_2) + y_1 - y_0 \\
&= y_3 - 3y_2 + 3y_1 - y_0 \\
\Delta^3 y_1 &= \Delta^2 y_2 - \Delta^2 y_1 = y_4 - 3y_3 + 3y_2 - y_1 \\
&\ldots \qquad \ldots \qquad \ldots \\
\Delta^3 y_n &= \Delta^2 y_{n+1} - \Delta^2 y_n = y_{n+1} - 3y_{n+2} + 3y_{n+1} - y_n
\end{aligned}$$

In general, the n th forward differences are defined as

$$\Delta^n y_k = \Delta^{n-1} y_{k+1} - \Delta^{n-1} y_k$$

In function notation, the forward differences are as written below:

$$\Delta f(x) = f(x+h) - f(x)$$

$$\Delta^2 f(x) = f(x+2h) - 2f(x+h) + f(x)$$

$$\Delta^3 f(x) = f(x+3h) - 3f(x+2h) + 3f(x+h) - f(x)$$

and so on, where h is step size.

The forward differences are usually arranged in a tabular form in the following manner:

x argument	$y = f(x)$ entry	1st difference	2nd difference	3rd difference	4th difference	5th difference
x_0	$y_0 = f(x_0)$					
		Δy_0				
$x_1 = x_0 + h$	$y_1 = f(x_1)$		$\Delta^2 y_0$			
		Δy_1		$\Delta^3 y_0$		
$x_2 = x_0 + 2h$	$y_2 = f(x_2)$		$\Delta^2 y_1$		$\Delta^4 y_0$	
		Δy_2		$\Delta^3 y_1$		$\Delta^5 y_0$
$x_3 = x_0 + 3h$	$y_3 = f(x_3)$		$\Delta^2 y_2$		$\Delta^4 y_1$	
		Δy_3		$\Delta^3 y_2$		
$x_4 = x_0 + 4h$	$y_4 = f(x_4)$		$\Delta^2 y_3$			
		Δy_4				
$x_5 = x_0 + 5h$	$y_5 = f(x_5)$					

The first term in the table y_0 is called the leading term and the differences $\Delta y_0, \Delta^2 y_0, \Delta^3 y_0, \ldots$ are called leading differences. It can be seen that the differences $\Delta^k y_i$ with a subscript 'i' lie along the diagonal sloping downwards, that is, forward with respect to the direction of x. The above difference table is known as Forward difference table or Diagonal difference table.

- **Properties of Δ**

  The operator "Δ" satisfies the following properties:

  (i) $\Delta[f(x) \pm g(x)] = \Delta f(x) \pm \Delta g(x)$, i.e. Δ is linear.

  (ii) $\Delta[\alpha f(x)] = \alpha \Delta f(x)$, α is a constant.

  (iii) $\Delta^m \Delta^n f(x) = \Delta^{m+n} f(x) = \Delta^n \Delta^m f(x)$, where m and n are positive integers.

  (iv) $\Delta[f(x) \cdot g(x)] \neq f(x) \cdot \Delta g(x)$.

- **Observation 1**

  We can express any higher order forward difference of y_0 in terms of the entries $y_0, y_1, y_2, \ldots, y_n$. From

$$\Delta y_0 = y_1 - y_0$$
$$\Delta^2 y_0 = y_2 - 2y_1 + y_0$$
$$\Delta^3 y_0 = y_3 - 3y_2 + 3y_1 - y_0$$

  and so on, we can see that the coefficients of the entries on the RHS are binomial coefficients. Therefore, in general,

$$\Delta^n y_0 = y_n - C_1^n y_{n-1} + C_2^n y_{n-2} - \cdots + (-1)^n y_0$$

➢ **Observation 2**

We can express any value of y in terms of leading entry y_0

We know that $\Delta y_0 = y_1 - y_0$

$$\therefore \quad y_1 = y_0 + \Delta y_0 = (1+\Delta)y_0$$

Now,

$$y_2 = y_1 + \Delta y_1 = (1+\Delta)y_1 = (1+\Delta)^2 y_0$$

Similarly, $y_3 = (1+\Delta)^3 y_0$ and so on.

In general,

$$y_n = (1+\Delta)^n y_0 = y_0 + C_1^n \Delta y_0 + C_2^n \Delta^2 y_0 + \cdots + \Delta^n y_0$$

4. Backward differences

The differences $y_1 - y_0, y_2 - y_1, \ldots, y_n - y_{n-1}$ when denoted by $\nabla y_1, \nabla y_2, \ldots, \nabla y_n$ respectively, are called the first backward differences, where ∇ is the backward difference operator called nabla operator.

$$\therefore \quad \nabla y_1 = y_1 - y_0, \quad \nabla y_2 = y_2 - y_1, \ldots, \nabla y_n = y_n - y_{n-1}$$

Now the second backward differences are defined as the differences of the first backward differences, i.e.

$$\nabla^2 y_2 = \nabla(\nabla y_2) = \nabla(y_2 - y_1) = \nabla y_2 - \nabla y_1 = (y_2 - y_1) - (y_1 - y_0)$$
$$= y_2 - 2y_1 + y_0$$

$\nabla^2 y_3 = \nabla y_3 - \nabla y_2 = y_3 - 2y_2 + y_1$ and so on.

In general,

$$\nabla^n y_k = \nabla^{n-1} y_k - \nabla^{n-1} y_{k-1}$$

In function notation, these differences are written as

$$\nabla f(x) = f(x) - f(x-h)$$
$$\nabla f(x+h) = f(x+h) - f(x)$$
$$\nabla^2 f(x+2h) = f(x+2h) - 2f(x+h) + f(x)$$
$$\nabla^3 f(x+3h) = f(x+3h) - 3f(x+2h) + 3f(x+h) - f(x)$$

and so on, where h is step size.

These backward differences are arranged in a tabular form in the following manner. In this table, the difference $\nabla^k y_i$ with a fixed subscript 'i' lies along the diagonal sloping upwards; that is, backwards with respect to the direction of increasing argument x.

x argument	$y = f(x)$ entry	1st difference	2nd difference	3rd difference	4th difference	5th difference
x_0	$y_0 = f(x_0)$					
		∇y_1				
$x_1 = x_0 + h$	$y_1 = f(x_1)$		$\nabla^2 y_2$			
		∇y_2		$\nabla^3 y_3$		
$x_2 = x_0 + 2h$	$y_2 = f(x_2)$		$\nabla^2 y_3$		$\nabla^4 y_4$	
		∇y_3		$\nabla^3 y_4$		$\nabla^5 y_5$
$x_3 = x_0 + 3h$	$y_3 = f(x_3)$		$\nabla^2 y_4$		$\nabla^4 y_5$	
		∇y_4		$\nabla^3 y_5$		
$x_4 = x_0 + 4h$	$y_4 = f(x_4)$		$\nabla^2 y_5$			
		∇y_5				
$x_5 = x_0 + 5h$	$y_5 = f(x_5)$					

- **Properties of ∇**
  - (i) $\nabla[f(x) \pm g(x)] = \nabla f(x) \pm \nabla g(x)$, i.e. ∇ is a linear operator.
  - (ii) $\nabla[\alpha f(x)] = \alpha \nabla f(x)$, α is a constant.
  - (iii) $\nabla^m \nabla^n f(x) = \nabla^{m+n} f(x)$, m and n are positive integers.
  - (iv) $\nabla[f(x)g(x)] \neq [\nabla f(x)].g(x)$.
- **Observation 1**

  We can express any higher order backward difference of y_n in terms of the entries $y_0, y_1, y_2, \ldots, y_n$. From

  $$\nabla y_n = y_n - y_{n-1}$$
  $$\nabla^2 y_n = y_n - 2y_{n-1} + y_{n-2}$$
  $$\nabla^3 y_n = y_n - 3y_{n-1} + 3y_{n-2} - y_{n-3}$$

  and so on, we can see that the coefficients of the entries on the RHS are binomial coefficients. Therefore, in general,

  $$\nabla^n y_n = y_n - C_1^n y_{n-1} + C_2^n y_{n-2} - \cdots + (-1)^n y_0$$
- **Observation 2**

We can express any value of y in terms of y_n and the backward differences ∇y_n , $\nabla^2 y_n$, etc. By definition,

$$\nabla y_n = y_n - y_{n-1}$$

or

$$y_{n-1} = y_n - \nabla y_n = (1-\nabla) y_n$$

Now,

$$y_{n-2} = y_{n-1} - \nabla y_{n-1} = (1-\nabla) y_{n-1} = (1-\nabla)^2 y_n$$

Similarly,

$$y_{n-3} = (1-\nabla)^3 y_n$$

and so on. In general,

$$y_{n-k} = (1-\nabla)^k y_n$$

$$\therefore \quad y_{n-k} = y_n - C_1^k \nabla y_n + C_2^k \nabla^2 y_n - \cdots + (-1)^k \nabla^k y_n$$

5. Central differences

Sometimes, it is more convenient to employ another system of differences known as central differences. In this system the symbol δ is used instead of Δ and is known as central difference operator. The subscript of δy for any difference is the average of the subscripts of the two entries.

$$\therefore \quad \delta y_{1/2} = y_1 - y_0, \delta y_{3/2} = y_2 - y_1, \delta y_{5/2} = y_3 - y_2, \ldots$$

For higher order differences, we have

$$\delta^2 y_1 = \delta y_{3/2} - \delta y_{1/2}, \ \delta^2 y_2 = \delta y_{3/2} - \delta y_{12}, \ldots, \ \delta^2 y_{3/2} = \delta^2 y_2 - \delta^2 y_1,$$

and so on. The central differences are tabulated below.

x argument	$y = f(x)$ entry	1st difference	2nd difference	3rd difference	4th difference	5th difference
x_0	$y_0 = f(x_0)$					
		$\delta y_{1/2}$				
$x_1 = x_0 + h$	$y_1 = f(x_1)$		$\delta^2 y_1$			
		$\delta y_{3/2}$		$\delta^3 y_{3/2}$		
$x_2 = x_0 + 2h$	$y_2 = f(x_2)$		$\delta^2 y_2$		$\delta^4 y_2$	
		$\delta y_{5/2}$		$\delta^3 y_{5/2}$		$\delta^5 y_{5/2}$
$x_3 = x_0 + 3h$	$y_3 = f(x_3)$		$\delta^2 y_3$		$\delta^4 y_3$	
		$\delta y_{7/2}$		$\delta^3 y_{7/2}$		
$x_4 = x_0 + 4h$	$y_4 = f(x_4)$		$\delta^2 y_4$			
		$\delta y_{9/2}$				
$x_5 = x_0 + 5h$	$y_5 = f(x_5)$					

We can see from the table that central differences on the same horizontal line have the same subscript. Also, all odd differences have a fractional subscript, and the even differences have integer subscript.

- **Note**

  From all the three tables, we can see that only the notation changes, not the differences. For examples,

$$y_1 - y_0 = \Delta y_0 = \nabla y_1 = \delta y_{1/2}$$

6. Other differences operators

So far we have studied the operators Δ, ∇ and δ. Now we shall introduce other operators like E, μ, D etc. which also play a vital role in numerical methods.

- **Shift operator** E

  If h is step size for the argument x then the operator E is defined as

$$E f(x) = f(x + h).$$

  It is also called translation operator due to the reason that it results the next value of the function. The higher orders of shift operator are defined as

$$E^2 f(x) = E[Ef(x)] = Ef(x + h) = f(x + 2h)$$

  Similarly,

$$E^3 f(x) = f(x + 3h),$$

$$E^4 f(x) = f(x+4h)$$

In general

$$E^n f(x) = f(x+nh) \text{ for any real } n$$

The inverse shift operator E^{-1} is defined as

$$E^{-1} f(x) = f(x-h)$$

Similarly

$$E^{-n} f(x) = f(x-nh) \text{ for any real } n$$

If y_k, is the function $f(x)$ then $Ey_k = y_{k+1}$ and

$$E^n y_k = y_{k+n}$$

- **Average operator μ**

The average operator μ is defined by

$$\mu f(x) = \frac{1}{2}[f(x+h/2) + f(x-h/2)]$$

i.e. $$\mu y(x) = \frac{1}{2}\left[y(x+h/2) + y(x-h/2)\right]$$

- **Differential operator D**

The differential operator D is defined as $Df(x) = \frac{d}{dx} f(x)$

In general,

$$D^n f(x) = \frac{d^n}{dx^n} f(x)$$

- **Note**

All the above operators are linear and obey index laws.

7. Relation between different differences operators

- **Relation between Δ and E**

$$\begin{aligned}\Delta f(x) &= f(x+h) - f(x) \\ &= Ef(x) - f(x) \\ &= (E-1)f(x)\end{aligned}$$

Thus $\Delta = E - 1$ or $E = 1 + \Delta$

- **Relation between E and ∇**

$$\nabla f(x) = f(x) - f(x-h)$$
$$= f(x) - E^{-1} f(x) = \left(1 - E^{-1}\right) f(x)$$
$$\therefore \nabla = 1 - E^{-1} \text{ or } E^{-1} = 1 - \nabla$$
$$\therefore E = (1-\nabla)^{-1} \quad \left[\because \left(E^{-1}\right)^{-1} = E\right]$$

- **Relation between E and δ**

$$\delta f(x) = f(x + h/2) - f(x - h/2)$$
$$= E^{1/2} f(x) - E^{-1/2} f(x)$$
$$= \left(E^{1/2} - E^{-1/2}\right) f(x)$$
$$\therefore \quad \delta = E^{1/2} - E^{-1/2}$$

Also,

$$\delta = E^{1/2}\left(1 - E^{-1}\right) = E^{1/2}\nabla$$
$$\delta = E^{-1/2}(E - 1) = E^{-1/2}\Delta$$

Hence

$$\delta = E^{1/2}\nabla = E^{-1/2}\Delta$$

- **Relation between E and μ**

$$\mu f(x) = \frac{1}{2}[f(x + h/2) + f(x - h/2)]$$
$$= \frac{1}{2}\left[E^{1/2} f(x) + E^{-1/2} f(x)\right]$$
$$= \frac{1}{2}\left(E^{1/2} + E^{-1/2}\right) f(x)$$
$$\therefore \mu = \frac{1}{2}\left(E^{1/2} + E^{-1/2}\right)$$

- **Relation of D with other Operators**

We know that $Df(x) = \frac{d}{dx} f(x) = f'(x)$ etc.

By Taylor's series

$$f(x+h) = f(x) + \frac{h}{1!} f'(x) + \frac{h^2}{2!} f''(x) + \frac{h^3}{3!} f'''(x) + \ldots$$

Or

$$Ef(x) = f(x) + hDf(x) + \frac{h^2}{2!}D^2 f(x) + \frac{h^3}{3!}D^3 f(x) + \ldots$$

$$= \left[1 + hD + \frac{h^2 D^2}{2!} + \frac{h^3 D^3}{3!} + \ldots\right] f(x) = e^{hD} f(x)$$

Thus

$$E = e^{hD}$$

Taking logarithms on both sides, we get

$$hD = \ln E = \ln(1 + \Delta)$$

$$D = \frac{1}{h}\left[\Delta - \frac{\Delta^2}{2} + \frac{\Delta^3}{3} - \frac{\Delta^4}{4} + \ldots\right]$$

Also,

$$\nabla = 1 - E^{-1}$$

Thus

$$E^{-1} = 1 - \nabla = e^{-hD}$$

Taking logarithm on both sides,

$$-hD = \ln(1 - \nabla)$$

$$D = -\frac{1}{h}\left[-\nabla - \frac{\nabla^2}{2} - \frac{\nabla^3}{3} - \frac{\nabla^4}{4} - \ldots\right]$$

$$= \frac{1}{h}\left[\nabla + \frac{\nabla^2}{2} + \frac{\nabla^3}{3} + \frac{\nabla^4}{4} + \ldots\right]$$

$$\therefore \quad \sinh(hD) = \frac{e^{hD} - e^{-hD}}{2} = \frac{E - E^{-1}}{2}$$

$$= \frac{1}{2}\left[E^{1/2} + E^{-1/2}\right]\left[E^{1/2} - E^{-1/2}\right] = \mu\delta$$

$$\therefore hD = \sinh^{-1}(\mu\delta)$$

Example (2.1)

Construct the forward difference table from the following data:

x	0	1	2	3	4
y	1	1.5	2.2	3.1	4.6

.

Then evaluate $\Delta^3 y_1, y_n$ and y_5.

Solution

The forward differences table is as given below:

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
0	1				
		0.5			
1	1.5		0.2		
		0.7		0	
2	2.2		0.2		0.4
		0.9		0.4	
3	3.1		0.6		
		1.5			
4	4.6				

Now

$$\Delta^3 y_1 = (E-1)^3 y_1 = \left(E^3 - 3E^2 + 3E - 1\right) y_1$$
$$= y_4 - 3y_3 + 3y_2 - y_1 = 4.6 - 3(3.1) + 3(2.2) - 1.5 = 0.4$$

Again from observation 2 of section, we have

$$y_n = y_0 + C_1^n \Delta y_0 + C_2^n \Delta^2 y_0 + C_3^n \Delta^3 y_0 + C_4^n \Delta^4 y_0$$
$$= 1 + n(0.5) + \frac{1}{2} n(n-1)(0.2) + \frac{1}{3!} n(n-1)(n-2)(0)$$
$$+ \frac{1}{4!} n(n-1)(n-2)(n-3)(0.4)$$
$$= 1 + \frac{1}{2} n + \frac{1}{10}\left(n^2 - n\right) + \frac{1}{60}\left(n^4 - 6n^3 + 11n^2 - 6n\right)$$
$$\therefore y_5 = \frac{1}{60}\left[5^4 - 6(5)^3 + 17(5)^2 + 18(5) + 60\right] = 7.5$$

Example (2.2)

Evaluate

(i) $\Delta \cos x$ (ii) $\Delta \ln f(x)$ (iii) $\Delta^2 \sin(px+q)$ (1v) $\Delta \tan^{-1} x$ (v) $\Delta^n e^{ax+b}$

Solution

Let h be the interval step size.

(i)

$$\Delta\cos x = \cos(x+h) - \cos x = -2\sin\left(x + \frac{h}{2}\right)\sin\frac{h}{2}$$

(ii)

$$\begin{aligned}\Delta\ln f\,(x) &= \ln f\,(x+h) - \ln f\,(x)\\ &= \ln\left[\frac{f\,(x+h)}{f\,(x)}\right] = \ln\left[\frac{f\,(x)+\Delta f\,(x)}{f\,(x)}\right]\\ &= \ln\left[1+\frac{\Delta f\,(x)}{f\,(x)}\right]\end{aligned}$$

(iii)

$$\begin{aligned}\Delta\sin(px+q) &= \sin[p(x+h)+q] - \sin(px+q)\\ &= 2\cos\left(px+q+\frac{ph}{2}\right)\sin\frac{ph}{2}\\ &= 2\sin\frac{ph}{2}\sin\left(\frac{\pi}{2}+px+q+\frac{ph}{2}\right)\\ &= 2\sin\frac{ph}{2}\sin\left(px+q+\frac{1}{2}(\pi+ph)\right)\end{aligned}$$

Hence

$$\begin{aligned}\Delta^2\sin(px+q) &= 2\sin\frac{ph}{2}\Delta\left[\sin\left(px+q+\frac{1}{2}(\pi+ph)\right)\right]\\ &= \left(2\sin\frac{ph}{2}\right)^2\sin\left(px+q+2\cdot\frac{1}{2}(\pi+ph)\right)\end{aligned}$$

(iv)

$$\begin{aligned}\Delta\tan^{-1}x &= \tan^{-1}(x+h) - \tan^{-1}x\\ &= \tan^{-1}\left[\frac{x+h-x}{1+x(x+h)}\right]\\ &= \tan^{-1}\frac{h}{1+x(x+h)}\end{aligned}$$

(v)

$$\begin{aligned}\Delta e^{ax+b} &= e^{a(x+h)+b} - e^{ax+b} \\ &= e^{ax+b}(e^{ah}-1) \\ \Delta^2 e^{ax+b} &= \Delta\left[\Delta e^{ax+b}\right] = \Delta\left[(e^{ah}-1)e^{ax+b}\right] \\ &= (e^{ah}-1)^2 e^{ax+b}, \qquad \left[(e^{ah}-1) \text{ is constant}\right]\end{aligned}$$

Proceeding on, we get,

$$\Delta^n\left(e^{ax+b}\right) = \left(e^{ah}-1\right)^n e^{ax+b}$$

Example (2.3)

Prove the following results:

(i) $\Delta\nabla = \nabla\Delta = \Delta - \nabla = \delta^2$

(ii) $\Delta + \nabla = \dfrac{\Delta}{\nabla} - \dfrac{\nabla}{\Delta}$

(iii) $\left(E^{1/2} + E^{-1/2}\right)(1+\Delta)^{1/2} = 2 + \Delta$

(iv) $1 + \mu^2\delta^2 = \left(1 + \dfrac{\delta^2}{2}\right)^2$

(v) $\Delta = \dfrac{\delta^2}{2} + \delta\sqrt{\left(1 + \dfrac{\delta^2}{4}\right)}$

(vi) $\mu^{-1} = 1 - \dfrac{1}{8}\delta^2 + \dfrac{3}{128}\delta^4 - \dfrac{5}{1024}\delta^6 + \ldots$

Solution

(i) We have,

$$\begin{aligned}\Delta\nabla f(x) &= \Delta[\nabla f(x)] = \Delta[f(x) - f(x-h)] \\ &= \Delta f(x) - \Delta f(x-h) \\ &= [f(x+h) - f(x)] - [f(x) - f(x-h)] \\ &= \Delta f(x) - \nabla f(x) = (\Delta - \nabla) f(x) \\ \therefore \Delta\nabla &= \Delta - \nabla\end{aligned}$$

Similarly,

$$\begin{aligned}\nabla\Delta f(x) &= \nabla[\Delta f(x)] = \nabla[f(x+h) - f(x)] \\ &= \nabla f(x+h) - \nabla f(x) \\ &= [f(x+h) - f(x)] - [f(x) - f(x-h)] \\ &= \Delta f(x) - \nabla f(x) = (\Delta - \nabla) f(x) \\ \therefore \nabla\Delta &= \Delta - \nabla\end{aligned}$$

Again

$$\begin{aligned}\delta^2 f(x) &= \left[E^{1/2} - E^{-1/2}\right]^2 f(x) \\ &= \left(E + E^{-1} - 2\right) f(x) \\ &= f(x+h) + f(x-h) - 2f(x) \\ &= f(x+h) - f(x)] - [f(x) - f(x-h)] \\ &= \Delta f(x) - \nabla f(x) = (\Delta - \nabla) f(x) \\ \delta^2 &= \Delta - \nabla\end{aligned}$$

Hence

$$\Delta\nabla = \nabla\Delta = \Delta - \nabla = \delta^2$$

(ii)

$$\begin{aligned}\text{R.H.S.} &= \frac{\Delta}{\nabla} - \frac{\nabla}{\Delta} = \frac{\Delta^2 - \nabla^2}{\nabla\Delta} \\ &= \frac{(\Delta + \nabla)(\Delta - \nabla)}{(\Delta - \nabla)} \\ &= \Delta + \nabla = \text{ L.H.S.}\end{aligned}$$

(iii)

$$\left(E^{1/2}+E^{-1/2}\right)(1+\Delta)^{1/2}=\left(E^{1/2}+E^{-1/2}\right)E^{1/2}$$
$$=E+1=1+\Delta+1=2+\Delta$$

(iv)

$$1+\mu^2\delta^2=1+\left[\frac{E^{1/2}+E^{-1/2}}{2}\right]^2\left[E^{1/2}-E^{-1/2}\right]^2$$
$$=1+\left[\frac{E-E^{-1}}{2}\right]^2=\frac{4+\left(E-E^{-1}\right)^2}{4}$$
$$=\left[\frac{E+E^{-1}}{2}\right]^2 \qquad (2.1)$$

Now,

$$\left[1+\frac{1}{2}\delta^2\right]^2=\left[1+\frac{1}{2}\left[E^{1/2}-E^{-1/2}\right]^2\right]^2$$
$$=\left[1+\frac{1}{2}\left[E+E^{-1}-2\right]\right]^2=\left[\frac{E+E^{-1}}{2}\right]^2 \qquad (2.2)$$

Hence, from Eqns. (2.1) and (2.2), we have

$$1+\mu^2\delta^2=\left[1+\frac{1}{2}\delta^2\right]^2$$

(v)

$$\text{RHS}=\frac{1}{2}\delta^2+\delta\sqrt{1+\frac{\delta^2}{4}}=\frac{1}{2}\delta\left[\delta+\sqrt{4+\delta^2}\right]$$
$$=\frac{1}{2}\delta\left[\left(E^{1/2}-E^{-1/2}\right)+\sqrt{4+\left(E^{1/2}-E^{-1/2}\right)^2}\right]$$
$$=\frac{1}{2}\delta\left[\left(E^{1/2}-E^{-1/2}\right)+\left(E^{1/2}+E^{-1/2}\right)\right]$$
$$=\frac{1}{2}\left(E^{1/2}-E^{-1/2}\right)\left(2E^{1/2}\right)=E-1=\Delta=\text{LHS}$$

(vi) By definition, we have

$$\mu^2 = \left[\frac{1}{2}\left(E^{1/2} + E^{-1/2}\right)\right]^2$$

$$= \frac{1}{4}\left[\left(E^{1/2} - E^{-1/2}\right)^2 + 4\right]$$

$$= \frac{1}{4}\left(\delta^2 + 4\right) = \frac{\delta^2}{4} + 1$$

$$\therefore \mu = \left[1 + \frac{\delta^2}{4}\right]^{1/2}$$

or

$$\mu^{-1} = \left[1 + \frac{\delta^2}{4}\right]^{-1/2}$$

$$= 1 - \frac{1}{1!}\left(\frac{1}{2}\right)\frac{\delta^2}{4} + \frac{1}{2!}\left(\frac{1}{2}\right)\left(\frac{1}{2}+1\right)\left[\frac{\delta^2}{4}\right]^2 - \frac{1}{3!}\left(\frac{1}{2}\right)\left(\frac{1}{2}+1\right)\left(\frac{1}{2}+2\right)\left[\frac{\delta^2}{4}\right]^3 + \cdots$$

$$= 1 - \frac{1}{8}\delta^2 + \frac{3}{128}\delta^4 - \frac{5}{1024}\delta^6 + \cdots$$

CHAPTER (3)

INTERPOLATION

1. Introduction

Interpolation is a technique of obtaining the value of a function for any intermediate values of the independent variable, i.e. argument within an interval, when the values of the arguments are given. Suppose that the following values of $y = f(x)$ for a set of values of x are given:

x (argument)	x_0	x_1	x_2	$\cdots$	x_n
$y(x)$	y_0	y_1	y_2	$\cdots$	y_n

.

Then the process of finding the value of y corresponding to any value of $x = x_i$ between x_0 and x_n is called in interpolation.

The process of finding the value of a function outside the given range of arguments is called extrapolation

If the form of the function $f(x)$ is known we can find $f(x)$ for any value of x by simple substitution. But in most practical problems that occur in engineering and science the form of the function $f(x)$ is unknown and it is very difficult to determine its exact form which is the help of tabulated set of values in such cases we replace $f(x)$ by simple function $\varphi(x)$ is called interpolating function which assumes the same values as those of $f(x)$ and from which others are values may be computed to the desired degree of accuracy.

If $\varphi(x)$ is a polynomial then it is called interpolating polynomial and the process is known as polynomial interpolation. If $\varphi(x)$ is a finite trigonometric series the process is called trigonometric interpolation. Usually, polynomial interpolation is preferred due to the reason that they are free from singularities is and the easy to differentiate and integrate. Even though there are other methods like graphical method and method of curve fitting, in this chapter we will study polynomial interpolation using the calculus of finite differences by driving two important interpolation formulae which are used often in all fields by means of forward and the backward differences of a function.

2. Newton forward interpolation formula

Let $y = f(x)$ be a function which takes the values $y_0, y_1, \cdots, y_n$ for $(n+1)$ values $x_0, x_1, \cdots, x_n$ of the dependent variable. Let these values be equidistant $x_i = x_0 + ih$, $i = 0,1,2,\cdots,n$ and let $P(x)$ be a polynomial of n degree such as

$$P(x_i) = f(x_i) = y_i \ , \ i = 0,1,2,\cdots,n .$$

$$\begin{aligned} P(x) &= a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) \\ &\quad + a_3(x - x_0)(x - x_1)(x - x_2) + \cdots \\ &\quad + a_n(x - x_0)(x - x_1)\cdots(x - x_{n-1}) \end{aligned} \qquad (3.1)$$

Putting $x = x_0, x_1, \cdots, x_n$ successfully in equation (3.1), we get

$$\begin{aligned} y_0 &= a_0 \ , \ y_1 = a_0 + a_1(x_1 - x_0) \ , \\ y_2 &= a_0 + a_1(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1) \\ &\vdots \qquad \vdots \qquad \vdots \\ y_n &= a_0 + a_1(x_n - x_0) + a_2(x_n - x_0)(x_n - x_1) + \cdots \\ &\quad + a_n(x_n - x_0)(x_n - x_1)\cdots(x_n - x_{n-1}) \end{aligned}$$

from these

$$a_0 = y_0 \ , \ a_1 = \frac{y_1 - a_0}{x_1 - x_0} = \frac{y_1 - y_0}{h} = \frac{\Delta y_0}{h} \ ,$$

$$\begin{aligned} a_2 &= \frac{y_2 - a_0 - a_1(x_2 - x_0)}{(x_2 - x_0)(x_2 - x_1)} = \frac{y_2 - y_0 - 2y_1 + 2y_0}{2h^2} \\ &= \frac{y_2 - 2y_1 + y_0}{2!h^2} = \frac{\Delta^2 y_0}{2!h^2}, \\ &\vdots \qquad \vdots \qquad \vdots \\ a_n &= \frac{\Delta^n y_0}{n!h^n} \end{aligned}$$

Putting these values in equation (3.1) we get

$$\begin{aligned} P(x) &= y_0 + \frac{\Delta y_0}{h}(x - x_0) + \frac{\Delta^2 y_0}{2!h^2}(x - x_0)(x - x_1) \\ &\quad + \frac{\Delta^3 y_0}{3!h^3}(x - x_0)(x - x_1)(x - x_2) + \cdots \\ &\quad + \frac{\Delta^n y_0}{n!h^n}(x - x_0)(x - x_1)\cdots(x - x_{n-1}) \end{aligned} \qquad (3.2)$$

Putting

$$\frac{x - x_0}{h} = q$$

$$x = x_0 + qh$$

$$\Downarrow \quad \Downarrow \quad \Downarrow$$

$$x - x_i = x - x_0 + x_0 - x_i = qh - ih = (q - i)h \ , \ i = 1,2,\cdots,n$$

where $0 < q < 1$ is real number, Eq. (3.2) takes the form

$$P(x) = y_0 + q\Delta y_0 + \frac{q(q-1)}{2!}\Delta^2 y_0 + \frac{q(q-1)(q-2)}{3!}\Delta^3 y_0 + \cdots + \frac{q(q-1)(q-2)\cdots(q-n+1)}{n!}\Delta^n y_0 \qquad (3.3)$$

Equation (3.3) is known as Newton forward interpolation formula

- **Note**

  Formula (3.3) is called Newton forward interpolation formula due the fact that this formula contains values of the tabulated function from y_0 onward to right and none to the left of this value. This formula is used mainly to interpolating the values of y near the beginning of a set of tabulated values and to extrapolating y a little to the left of y_0. The first two terms of the equation will give a linear interpolation while the first three terms a quadratic interpolation and so on.

3. Newton backward interpolation formula

Let $y = f(x)$ be a function which takes the values $y_0, y_1, \cdots, y_n$ for $(n+1)$ values $x_0, x_1, \cdots, x_n$ of the dependent variable. Let these values be equidistant $x_i = x_0 + ih$, $i = 0,1,2,\cdots,n$ and let $P(x)$ be a polynomial of n degree such as

$$P(x_i) = f(x_i) = y_i \ , \ i = 0,1,2,\cdots,n .$$

Suppose that it is required to evaluate $y(x)$ near the end of the table values then we can assume that

$$P(x) = a_0 + a_1(x - x_n) + a_2(x - x_n)(x - x_{n-1}) + a_3(x - x_n)(x - x_{n-1})(x - x_{n-2}) + \cdots + a_n(x - x_n)(x - x_{n-1})\cdots(x - x_1) \qquad (3.4)$$

Putting $x = x_0, x_1, \cdots, x_n$ successfully in Eq. (3.4), we get

$$a_0 = y_n ,$$
$$y_{n-1} = a_0 + a_1(x_{n-1} - x_n)$$
$$y_{n-2} = a_0 + a_1(x_{n-2} - x_n) + a_2(x_{n-2} - x_n)(x_{n-2} - x_{n-1})$$
$$\vdots \qquad \vdots \qquad \vdots$$
$$y_0 = a_0 + a_1(x_0 - x_n) + a_2(x_0 - x_n)(x_0 - x_{n-1})$$
$$+ a_3(x_0 - x_n)(x_0 - x_{n-1})(x_0 - x_{n-2}) + \cdots$$
$$+ a_n(x_0 - x_n)(x_0 - x_{n-1})\cdots(x_0 - x_1)$$

These equations give

$$a_0 = y_n \ , \ a_1 = \frac{y_{n-1} - a_0}{x_{n-1} - x_n} = \frac{y_{n-1} - y_n}{x_{n-1} - x_n} = \frac{y_n - y_{n-1}}{x_n - x_{n-1}} = \frac{\nabla y_n}{h} \ ,$$
$$a_2 = \frac{y_{n-2} - a_0 - a_1(x_{n-2} - x_n)}{(x_{n-2} - x_n)(x_{n-2} - x_{n-1})} = \frac{y_{n-2} - y_n - 2y_{n-1} + 2y_n}{-2h^2}$$
$$= \frac{y_n - 2y_{n-1} + y_{n-2}}{2!h^2} = \frac{\nabla^2 y_n}{2!h^2},$$
$$\vdots \qquad \vdots \qquad \vdots$$
$$a_n = \frac{\nabla^n y_n}{n!h^n}$$

Putting these values in Eq.(3.4) we get

$$P(x) = y_n + \frac{\nabla y_n}{h}(x - x_n) + \frac{\nabla^2 y_n}{2!h^2}(x - x_n)(x - x_{n-1})$$
$$+ \frac{\nabla^3 y_n}{3!h^3}(x - x_n)(x - x_{n-1})(x - x_{n-2}) + \cdots$$
$$+ \frac{\nabla^n y_n}{n!h^n}(x - x_n)(x - x_{n-1})\cdots(x - x_1) \qquad (3.5)$$

Let

$$\frac{x - x_n}{h} = q$$
$$x = x_n + qh$$
$$\Downarrow \quad \Downarrow \quad \Downarrow$$
$$x - x_i = x - x_n + x_n - x_i = qh + (n - i)h = (q + n - i)h \ , \ i = 1, 2, \cdots, n$$

Where q is real number. Then Eq.(3.5) takes the form

$$P(x) = y_n + q\nabla y_n + \frac{q(q+1)}{2!}\nabla^2 y_n + \frac{q(q+1)(q+2)}{3!}\nabla^3 y_n + \cdots + \frac{q(q+1)(q+2)\cdots(q+n-1)}{n!}\nabla^n y_n \quad (3.6)$$

Eq.(36) is known as Newton backward interpolation formula

- **Note**

  Since the formula (3.6) involves the backward differences it is called backward interpolation formula and it is used to interpolate the values of y near to the end of a set of tabular values. This may also be used to extrapolate the values of y a little to the right of y_n

Example (3.1)

Find a polynomial which takes the following values

x	0	1	2	3	4	5
$y(x)$	5.2	8.0	10.4	12.4	14.0	15.2

Solution

We take

$$x_0 = 0,\ h = x_1 - x_0 = 1,\ q = \frac{x - x_0}{h} = \frac{x - 0}{1} = x$$

The forward differences table is as follows:

x	$y(x)$	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
0	5.2				
		2.8			
1	8.0		–0.4		
		2.4		0	
2	10.4		–0.4		0
		2.0		0	
3	12.4		–0.4		0
		1.6		0	
4	14.0		–0.4		
		1.2			
5	15.2				

Using Newton forward interpolation formula, we get

$$P(x) = y_0 + q\Delta y_0 + \frac{q(q-1)}{2!}\Delta^2 y_0 + \frac{q(q-1)(q-2)}{3!}\Delta^3 y$$

$$= 5.2 + 2.8x - \frac{0.4}{2}(x)(x-1)$$

$$= 5.2 + 2.6x - 0.2x^2$$

Example (3.2)

Find a polynomial which takes the following values:

x	1	1.5	2.0	2.5
$y(x)$	4.0	18.25	44.0	84.25

and hence compute $y(1.25)$.

Solution

Take

$$x_0 = 1.0\,,\ h = x_1 - x_0 \ \Rightarrow\ h = 1.5 - 1.0 = 0.5$$

$$x = x_0 + qh \ \Rightarrow\ q = \frac{x - x_0}{0.5} = \frac{x - 1.0}{0.5} = 2(x - 1)$$

The forward differences table is as follows:

x	$y(x)$	Δy	$\Delta^2 y$	$\Delta^3 y$
1.0	4.0			
		14.25		
1.5	18.25		11.5	
		25.75		3.0
2.0	44.0		14.5	
		40.25		
2.5	84.25			

Thus

$$y_0 = 4.0\,,\ \Delta y_0 = 14.25\ \,,\ \Delta^2 y_0 = 11.5\,,\ \Delta^3 y_0 = 3.0$$

Using Newton forward interpolation formula, we get

$$P(x) = y_0 + q\Delta y_0 + \frac{q(q-1)}{2!}\Delta^2 y_0 + \frac{q(q-1)(q-2)}{3!}\Delta^3 y$$

$$= 4 + 2(x-1)(14.25) + \frac{1}{2!}\big[2(x-1)\big]\big[2(x-1)-1\big](11.5)$$

$$+ \frac{1}{3!}\big[2(x-1)\big]\big[2(x-1)-1\big]\big[2(x-1)-2\big](3)$$

Now

$$y(1.25) = 4.0 + (0.5)(14.25) + \frac{(0.5)(-0.5)}{2!}(11.5)$$

$$+ \frac{(0.5)(-0.5)(-1.5)}{3!}(3) = 9.875$$

Example (3.3)

Find a polynomial which takes the following values

x	1	3	5	7	9
y	3	14	19	21	23

and hence compute $y(2), y(10)$.

Solution

The differences table as follows:

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
1	3				
		11			
3	14		–6		
		5		3	
5	19		–3		0
		2		3	
7	21		0		
		2			
9	23				

Take $x_0 = 1, y_0 = 3, h = 2, q = \dfrac{x-1}{2}$

Using Newton forward interpolation formula, we get

$$y(x) = y_0 + q\Delta y_0 + \frac{q(q-1)}{2!}\Delta^2 y_0 + \frac{q(q-1)(q-2)}{3!}\Delta^3 y_0$$

$$= 3 + \frac{x-1}{2}(11) + \frac{1}{2!}\frac{x-1}{2}\left[\frac{x-1}{2} - 1\right](-6)$$

$$+ \frac{1}{3!}\frac{x-1}{2}\left[\frac{x-1}{2} - 1\right]\left[\frac{x-1}{2} - 2\right](3)$$

$$= \frac{1}{16}\left(x^3 - 21x^2 + 159x - 91\right).$$

Again $x_n = 9, y_n = 23, h = 2, q = \dfrac{x-9}{2}$

Using Newton backward interpolation formula, we get

$$\begin{aligned} y(x) &= y_n + q\nabla y_n + \frac{q(q+1)}{2!}\nabla^2 y_n + \frac{q(q+1)(q+2)}{3!}\nabla^3 y_n \\ &= 23 + \frac{x-9}{2}(2) + \frac{1}{2!}\frac{x-9}{2}\left[\frac{x-9}{2}+1\right](0) \\ &+ \frac{1}{3!}\frac{x-9}{2}\left[\frac{x-9}{2}+1\right]\left[\frac{x-9}{2}+2\right](3) \\ &= 23 + (x-9) + \frac{1}{16}(x-9)(x-7)(x-5). \end{aligned}$$

Then

$$y(2) = \frac{1}{16}\left(2^3 - 212^2 + 159(2) - 91\right) = 9.4375$$

and

$$y(10) = 23 + (10-9) + \frac{1}{16}(10-9)(10-7)(10-5)$$

Example (3.4)

The amount A of a substance remaining in a reacting system after a time t in a certain chemical experiment is tabulated below

t	2	5	8	11
A	94.8	87.9	81.3	75.1

Obtain the value of A when $t = 9$ using Newton backward interpolation formula.

Solution

Since the value $t = 9$ is near the end of the table, to get the corresponding value of t we use Newton backward interpolation formula.

The backward differences are calculated and tabulated below:

t	A	∇A	$\nabla^2 A$	$\nabla^3 A$
2.0	94.8			
		–6.9		
5.0	87.9		0.3	
		–6.6		0.1
8.0	81.3		0.4	
		–6.2		
11.0	75.1			

Here

$$h = t_1 - t_0 \Rightarrow h = 5 - 2 = 3,\ t_n = 11.0$$

Hence the interpolation polynomial is

$$A(t) = A_n + q\nabla A_n + \frac{q(q+1)}{2!}\nabla^2 A_n + \frac{q(q+1)(q+2)}{3!}\nabla^3 A_n.$$

If $t = 9$, we have

$$t = t_n + qh \Rightarrow q = \frac{t - t_n}{h} = \frac{9 - 11.0}{3} = -\frac{2}{3}$$

Therefore

$$A(9) = 75.1 + \left(-\frac{2}{3}\right)(-6.2) + \frac{1}{2!}\left(-\frac{2}{3}\right)\left(-\frac{2}{3}+1\right)(0.4)$$
$$+ \frac{1}{3!}\left(-\frac{2}{3}\right)\left(-\frac{2}{3}+1\right)\left(-\frac{2}{3}+2\right)(0.1) = 79.183951$$

Example (3.5)

Find the missing value in the following table

x	16	18	20	22	24	26
y	43	89	–	155	268	388

Solution

Since five values are given, it is possible to express y as a polynomial of fourth degree. Hence the fifth differences of y are zeros. Taking the origin for x at 16 , from the given data we have:

$$y_0 = 43,\ y_1 = 89, y_3 = 155, y_4 = 268, y_5 = 388,$$

and we have to find y_2 . We know that $\Delta^5 y_0 = 0$

$$\Delta^5 y_0 = (E - 1)^5 y_0 = 0$$

i.e.

$$\left(E^5 - C_1^5 E^4 + C_2^5 E^3 - C_3^5 E^2 + C_4^5 E - 1\right) y_0 = 0$$

$$\left(E^5 - 5E^4 + 10E^3 - 10E^2 + 5E - 1\right) y_0 = 0,$$
$$E^5 y_0 - 5E^4 y_0 + 10E^3 y_0 - 10E^2 y_0 + 5Ey_0 - y_0 = 0,$$
$$y_5 - 5y_4 + 10y_3 - 10y_2 + 5y_1 - y_0 = 0$$

Substituting the given values, we have

$$388 - 5(268) + 10(155) - 10y_2 + 5(89) - 43 = 0$$
$$\Downarrow$$
$$y_2 = 100$$

CHAPTER (4)

NUMERICAL DIFFERENTIATION

1. Introduction

This chapter deals with numerical approximations of derivatives. The first question that comes up to mind is: why do we need to approximate derivatives at all? After all, we know how to analytically differentiate every function. Nevertheless, there are several reasons as of why we still need to approximate derivatives:

- Even if there exists an underlying function that we need to differentiate, we might know its values only at a sampled data set without knowing the function itself.
- There are some cases where it may not be obvious that an underlying function exists and all that we have is a discrete data set. We may still be interested in studying changes in the data, which are related, of course, to derivatives.
- There are times in which exact formulas are available but they are very complicated to the point that an exact computation of the derivative requires a lot of function evaluations. It might be significantly simpler to approximate the derivative numerically instead of computing its exact value.
- When approximating solutions to ordinary (or partial) differential equations, we typically represent the solution as a discrete approximation that is defined on a grid. Since we then have to evaluate derivatives at the grid points, we need to be able to come up with methods for approximating the derivatives at these points, and again, this will typically be done using only values that are defined on a lattice. The underlying function itself (which in this case is the solution of the equation) is unknown.

Consider a set of values (x_i, y_i) of a function $y = f(x)$. The process of computing the derivative or a derivative of the function at some value x from the given set of values is called numerical differentiation. This may be done by first approximating the function by a suitable interpolation formula and then differentiating it as many times as desired.

2. Derivatives using Newton forward interpolation formula

If the values of x are equispaced and the derivative is required near the beginning of the table, we employ Newton forward interpolation formula.

Newton forward interpolation formula is

$$y(x) = y_0 + q\Delta y_0 + \frac{q(q-1)}{2!}\Delta^2 y_0 + \frac{q(q-1)(q-2)}{3!}\Delta^3 y_0 + \cdots + \frac{q(q-1)(q-2)\cdots(q-n+1)}{n!}\Delta^n y_0, \quad (4.1)$$

where $q = \frac{x - x_0}{h}$.

Differentiating both sides of equation (4.1) with respect to q , we have

$$\frac{dy}{dq} = \Delta y_0 + \frac{2q-1}{2!}\Delta^2 y_0 + \frac{3q^2 - 6q + 2}{3!}\Delta^3 y_0 + \frac{4q^3 - 18q^2 + 22q - 6}{4!}\Delta^4 y_0 + \cdots$$

Now

$$\frac{dy}{dx} = \frac{dy}{dq}\cdot\frac{dq}{dx} = \frac{1}{h}\frac{dy}{dq}, \quad \left(\frac{dq}{dx} = \frac{1}{h}\right)$$

$$\frac{dy}{dx} = \frac{1}{h}\left[\Delta y_0 + \frac{2q-1}{2!}\Delta^2 y_0 + \frac{3q^2 - 6q + 2}{3!}\Delta^3 y_0 + \frac{4q^3 - 18q^2 + 22q - 6}{4!}\Delta^4 y_0 + \cdots\right] \quad (4.2)$$

At $x = x_0 \Rightarrow q = 0$. Hence putting $q = 0$ in equation, we get

$$\left.\frac{dy}{dx}\right|_{x=x_0} = \frac{1}{h}\left[\Delta y_0 - \frac{1}{2}\Delta^2 y_0 + \frac{1}{3}\Delta^3 y_0 - \frac{1}{4}\Delta^4 y_0 + \cdots\right]$$

Differentiating Eq.(4.2) with respect to x , we get

$$\frac{d^2y}{dx^2} = \frac{d}{dq}\left(\frac{dy}{dx}\right)\frac{dq}{dx} = \frac{1}{h}\cdot\frac{d}{dq}\left(\frac{dy}{dx}\right)$$

$$= \frac{1}{h^2}\left[\Delta^2 y_0 + (q-1)\Delta^3 y_0 + \frac{6q^2 - 18q + 11}{12}\Delta^4 y_0 + \cdots\right] \quad (4.3)$$

Putting $q = 0$ in equation, we get

$$\left.\frac{d^2y}{dx^2}\right|_{x=x_0} = \frac{1}{h^2}\left[\Delta^2 y_0 - \Delta^3 y_0 + \frac{11}{12}\Delta^4 y_0 - \cdots\right]$$

Similarly

$$\left.\frac{d^3y}{dx^3}\right|_{x=x_0} = \frac{1}{h^3}\left[\Delta^3 y_0 - \frac{2}{3}\Delta^4 y_0 + \cdots\right]$$

And so on.

3. Derivatives using Newton backward interpolation formula

If the derivative is required near the end of the table, we use the backward interpolation formula.

Newton backward interpolation formula

$$y(x) = y_n + q\nabla y_n + \frac{q(q+1)}{2!}\nabla^2 y_n + \frac{q(q+1)(q+2)}{3!}\nabla^3 y_n + \cdots + \frac{q(q+1)(q+2)\cdots(q+n-1)}{n!}\nabla^n y_n, \quad (4.4)$$

where $q = \dfrac{x - x_n}{h}$.

Differentiating both sides of Eq. (4.4) with respect to q , we have

$$\frac{dy}{dq} = \nabla y_n + \frac{2q+1}{2!}\nabla^2 y_n + \frac{3q^2+6q+2}{3!}\nabla^3 y_n + \frac{4q^3+18q^2+22q+6}{4!}\nabla^4 y_n + \cdots$$

Now

$$\frac{dy}{dx} = \frac{dy}{dq}\cdot\frac{dq}{dx} = \frac{1}{h}\frac{dy}{dq}, \quad \left(\frac{dq}{dx} = \frac{1}{h}\right)$$

$$\frac{dy}{dx} = \frac{1}{h}\left[\nabla y_n + \frac{2q+1}{2!}\nabla^2 y_n + \frac{3q^2+6q+2}{3!}\nabla^3 y_n + \frac{4q^3+18q^2+22q+6}{4!}\nabla^4 y_n + \cdots\right] \quad (4.5)$$

At $x = x_n \Rightarrow q = 0$. Hence, putting $q = 0$ in equation, we get

$$\left.\frac{dy}{dx}\right|_{x=x_n} = \frac{1}{h}\left[\nabla y_n + \frac{1}{2}\nabla^2 y_n + \frac{1}{3}\nabla^3 y_n + \frac{1}{4}\nabla^4 y_n + \cdots\right]$$

Again differentiating Eq. (4.5) with respect to x we get

$$\frac{d^2y}{dx^2} = \frac{d}{dq}\left(\frac{dy}{dx}\right)\frac{dq}{dx} = \frac{1}{h}\cdot\frac{d}{dq}\left(\frac{dy}{dx}\right)$$

$$= \frac{1}{h^2}\left[\nabla^2 y_n + (q+1)\nabla^3 y_n + \frac{6q^2+18q+11}{12}\nabla^4 y_n + \cdots\right] \tag{4.6}$$

Putting $q = 0$ in Eq. (4.6) , we get

$$\left.\frac{d^2y}{dx^2}\right|_{x=x_n} = \frac{1}{h^2}\left[\nabla^2 y_n + \nabla^3 y_n + \frac{11}{12}\nabla^4 y_n - \cdots\right]$$

Similarly

$$\left.\frac{d^3y}{dx^3}\right|_{x=x_n} = \frac{1}{h^3}\left[\nabla^3 y_n + \frac{2}{3}\nabla^4 y_n + \cdots\right]$$

and so on.

Example (4.1)

Find the first, second and third derivatives of $y(x)$ at $x = 1.5$ if

x	1.5	2.0	2.5	3.0	3.5	4.0
$y(x)$	3.375	7.000	13.625	24.000	38.875	59.000

Solution

We have to find the derivative at the point $x = 1.5$ which is at the beginning of the given data. Therefore we use here the derivative of Newton forward interpolation formula. The forward differences table as follows

x	$y(x)$	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
1.5	3.375				
		3.625			
2.0	7.000		3.000		
		6.625		0.750	
2.5	13.625		3.750		0
		10.375		0.750	
3.0	24.000		4.500		0
		14.875		0.750	
3.5	38.875		5.250		
		20.125			
4.0	59.000				

Here $x_0 = 1.5$, $h = x_1 - x_0 = 0.5$, from Eq. (4.2) we have

$$\left.\frac{dy}{dx}\right|_{x=x_0} = \frac{1}{h}\left[\Delta y_0 - \frac{1}{2}\Delta^2 y_0 + \frac{1}{3}\Delta^3 y_0 - \frac{1}{4}\Delta^4 y_0 + \cdots\right]$$

Thus

$$y'(1.5) = \frac{1}{0.5}\left[3.625 - \frac{1}{2}(3) + \frac{1}{3}(0.75)\right] = 4.75$$

from Eq.(4.3) we have

$$\left.\frac{d^2 y}{dx^2}\right|_{x=x_0} = \frac{1}{h^2}\left[\Delta^2 y_0 - \Delta^3 y_0 + \frac{11}{12}\Delta^4 y_0 - \cdots\right]$$

Hence

$$y''(1.5) = \frac{1}{(0.5)^2}\left[3 - 0.75\right] = 9$$

Again from Eq.(4.4) we have

$$\left.\frac{d^3y}{dx^3}\right|_{x=x_0} = \frac{1}{h^3}\left[\Delta^3 y_0 - \frac{2}{3}\Delta^4 y_0 + \cdots\right]$$

Thus

$$y'''(1.5) = \frac{1}{(0.5)^3}[0.75] = 6$$

Example (4.2)

The population of a certain town is shown in the following table

x	1951	1961	1971	1981	1991
y	19.96	36.65	58.81	77.21	94.61

Find the rate of growth of the population in the year 1981.

Solution

Here we have to find the derivative at 1981 which is near the end of the table. Hence we use derivative of Newton backward difference formula. The table of differences is as follows

x	y	∇y	$\nabla^2 y$	$\nabla^3 y$	$\nabla^4 y$
1951	19.96				
		16.69			
1961	36.65		5.47		
		22.16		–9.23	
1971	58.81		–3.76		11.9
		18.40		2.76	
1981	77.21		–1		
		17.40			
1991	94.61				

Hence

$$h = 10,\ x_n = 1991,\ q = \frac{x - x_n}{h} = \frac{1981 - 1991}{10} = -1$$

we know from Eq.(4.5) that

$$\left.\frac{dy}{dx}\right|_{x=x_n} = \frac{1}{h}\left[\nabla y_n + \frac{2q+1}{2!}\nabla^2 y_n + \frac{3q^2+6q+2}{3!}\nabla^3 y_n + \frac{4q^3+18q^2+22q+6}{4!}\nabla^4 y_n + \cdots\right]$$

Now we have to find out the rate of growth of the population in the year 1981

$$y'(1981) = \frac{1}{10}\left[17.4 + \frac{2(-1)+1}{2!}(-1) + \frac{3(-1)^2+6(-1)+2}{3!}(2.76) + \frac{4(-1)^3+18(-1)^2+22(-1)+6}{4!}(11.99)\right] = 1.6440833$$

The rate of growth of the population in year 1981 is 1.6440833

Example (4.3)

Find the first and second derivative of the function tabulated below at the point $x = 1.9$.

x	1.0	1.2	1.4	1.6	1.8	2.0
$y(x)$	0.000	0.128	0.544	1.296	2.432	4.00

Solution

We have to find the derivative at the point $x = 1.9$ which is near the end of the given data. Therefore we use the derivative of Newton backward interpolation formula. The backward differences table as follows

x	$y(x)$	∇y	$\nabla^2 y$	$\nabla^3 y$	$\nabla^4 y$
1.0	0.000				
		0.128			
1.2	0.128		0.288		
		0.416		0.048	
1.4	0.544		0.336		0
		0.752		0.048	
1.6	1.296		0.384		0
		1.136		0.048	
1.8	2.432		0.432		
		1.568			
2.0	4.000				

Here

$$x_n = 2\ ,\ h = x_1 - x_0 = 0.2\ ,\ q = \frac{x - x_n}{h} = \frac{1.9 - 2.0}{0.2} = -0.5$$

we know from Eq.(4.5) that

$$\left.\frac{dy}{dx}\right|_{x=x_n} = \frac{1}{h}\left[\nabla y_n + \frac{2q+1}{2!}\nabla^2 y_n + \frac{3q^2+6q+2}{3!}\nabla^3 y_n + \frac{4q^3+18q^2+22q+6}{4!}\nabla^4 y_n + \cdots\right]$$

Thus

$$y'(1.9) = \frac{1}{0.2}\left[1.568 + \frac{2(-0.5)+1}{2!}(0.432) + \frac{3(-0.5)^2+6(-0.5)+2}{3!}(0.048)\right] = 7.83$$

we know from Eq.(4.6) that

$$\frac{d^2y}{dx^2} = \frac{1}{h^2}\left[\nabla^2 y_n + (q+1)\nabla^3 y_n + \frac{6q^2+18q+11}{12}\nabla^4 y_n + \cdots\right]$$

Hence

$$y''(1.9)=\frac{1}{(0.2)^2}\left[0.432+(-0.5+1)(0.048)\right]=11.4\,.$$

4. Two points first derivative approximation

I. First derivative forward differences approximation

The Taylor expansion of $f(x_i+h)$ about x_i is given by:

$$f(x_i+h)=f(x_i)+\frac{h}{1!}f'(x_i)+\frac{h^2}{2!}f''(x_i)+\frac{h^3}{3!}f'''(x_i)+\cdots$$
$$+\frac{h^n}{n!}f^{(n)}(x_i)+\frac{h^{n+1}}{(n+1)!}f^{(n+1)}(\xi),\quad \xi\in(x_i,x_i+h)$$

For such expansion to be valid, we assume that $f(x)$ has $(n+1)$th continuous derivatives at the point $x=x_i$. Neglecting terms of degree higher than two, we obtain

$$f(x_i+h)=f(x_i)+\frac{h}{1!}f'(x_i)+\frac{h^2}{2!}f''(\xi),\quad \xi\in(x_i,x_i+h)$$

which turns into

$$f'(x_i)=\frac{f(x_i+h)-f(x_i)}{h}-\frac{h}{2!}f''(\xi),\quad \xi\in(x_i,x_i+h) \qquad (4.7)$$

Eq. (4.7) can be written as

$$f'(x_i)=F+E_F,$$

where

$$F=\frac{f(x_i+h)-f(x_i)}{h}\ ,\ E_F=-\frac{h}{2!}f''(\xi),\quad \xi\in(x_i,x_i+h)$$

F is called forward differences formula for approximating $f'(x_i)$ and E_F is the error.

II. First derivative backward differences approximation

The Taylor expansion of $f(x_i-h)$ about x_i is given by:

$$f(x_i-h)=f(x_i)-\frac{h}{1!}f'(x_i)+\frac{h^2}{2!}f''(x_i)-\frac{h^3}{3!}f'''(x_i)+\cdots$$
$$+(-1)^n\frac{h^n}{n!}f^{(n)}(x_i)+(-1)^{n+1}\frac{h^{n+1}}{(n+1)!}f^{(n+1)}(\xi),\quad \xi\in(x_i-h,x_i)$$

For such expansion to be valid, we assume that $f(x)$ has (n+1)*th* continuous derivatives at the point $x = x_i$. Neglecting terms of degree higher than two, we obtain

$$f(x_i - h) = f(x_i) - \frac{h}{1!}f'(x_i) + \frac{h^2}{2!}f''(\xi), \quad \xi \in (x_i - h, x_i)$$

which turns into

$$f'(x_i) = \frac{f(x_i) - f(x_i - h)}{h} + \frac{h}{2!}f''(\xi), \quad \xi \in (x_i - h, x_i) \tag{4.8}$$

Eq. (4.8) can be written as

$$f'(x_i) = B + E_B,$$

where

$$B = \frac{f(x_i) - f(x_i - h)}{h}, \; E_B = \frac{h}{2!}f''(\xi), \quad \xi \in (x_i - h, x_i)$$

B is called backward differences formula for approximating $f'(x_i)$ and E_B is the error.

Example (4.4)

Find the first derivative approximation of the function $f(x) = \cos(\pi x)$ at $x = \frac{\pi}{4}$ using forward differences approximation formula (take $h = 0.01$)

Solution

The forward differences approximation formula of the first derivative defined as

$$f'(x_i) = \frac{f(x_i + h) - f(x_i)}{h},$$

then we have

$$f'(\pi/4) = \frac{f(\pi/4 + 0.01) - f(\pi/4)}{0.01}$$

$$= \frac{0.700000476 - 0.707106781}{0.01} = 0.71063051$$

Example (4.5)

Find the first derivative of the function tabulated below at the point $x = 0.2$ using both forward differences and backward differences approximation formulae

x	0.1	0.2	0.3	0.4	0.5
y	0.0001	0.0016	0.0081	0.0256	0.0625

Solution

The forward differences approximation formula of the first derivative defined as

$$f'(x_i) = \frac{f(x_i + h) - f(x_i)}{h},$$

then we have

$$f'(0.2) = \frac{f(0.3) - f(0.2)}{0.1} = \frac{0.0081 - 0.0016}{0.1} = 0.065$$

The backward differences approximation formula of the first derivative defined as

$$f'(x_i) = \frac{f(x_i) - f(x_i - h)}{h},$$

then we have

$$f'(0.2) = \frac{f(0.2) - f(0.1)}{0.1} = \frac{0.0016 - 0.0001}{0.1} = 0.015$$

5. Three points first derivative approximation

The Taylor expansion of $f(x_i + h)$ about x_i is given by:

$$f(x_i + h) = f(x_i) + \frac{h}{1!}f'(x_i) + \frac{h^2}{2!}f''(x_i) + \frac{h^3}{3!}f'''(x_i) + \cdots + \frac{h^n}{n!}f^{(n)}(x_i) + \frac{h^{n+1}}{(n+1)!}f^{(n+1)}(\xi), \quad \xi \in (x_i, x_i + h). \tag{4.9}$$

While, the Taylor expansion of $f(x_i - h)$ about x_i is given by:

$$f(x_i - h) = f(x_i) - \frac{h}{1!}f'(x_i) + \frac{h^2}{2!}f''(x_i) - \frac{h^3}{3!}f'''(x_i) + \cdots$$
$$+ (-1)^n \frac{h^n}{n!}f^{(n)}(x_i) + (-1)^{n+1}\frac{h^{n+1}}{(n+1)!}f^{(n+1)}(\xi), \quad \xi \in (x_i - h, x_i). \tag{4.10}$$

Subtracting Eq. (4.10) from Eq. (4.9) and neglecting terms of degree higher than three, we obtain

$$f(x_i + h) - f(x_i - h) = 2hf'(x_i) + \frac{h^3}{3!}\left[f'''(\xi_1) + f'''(\xi_2)\right]$$

If the third-order derivative $f'''(x)$ is a continuous function in the interval $[x_i - h, x_i + h]$, then the intermediate value theorem implies that there exists a point $\xi \in (x_i - h, x_i + h)$ such that

$$f'''(\xi) = \frac{1}{2}\left[f'''(\xi_1) + f'''(\xi_2)\right]$$

Hence

$$f'(x_i) = \frac{f(x_i + h) - f(x_i - h)}{2h} - \frac{h^2}{6}f'''(\xi) \tag{4.11}$$

Eq. (4.11) can be written as

$f'(x_i) = C + E_C$,

where

$$C = \frac{f(x_i + h) - f(x_i - h)}{h}, \quad E_C = \frac{h}{2!}f''(\xi), \quad \xi \in (x_i - h, x_i + h)$$

C is called central differences formula for approximating $f'(x_i)$ and E_C is the error.

Example (4.6)

Find the first derivative of the function tabulated below at the point $x = 0.2$ using central differences approximation formula

x	0.1	0.2	0.3	0.4	0.5
y	0.0001	0.0016	0.0081	0.0256	0.0625

Solution

The central differences approximation formula of the first derivative defined as:

$$f'(x_i)=\frac{f(x_i+h)-f(x_i-h)}{2h},$$

so we have

$$f'(0.2)=\frac{f(0.3)-f(0.1)}{2(0.1)}=\frac{0.0081-0.0001}{2(0.1)}=0.04$$

6. Three points second derivative approximation

For the second derivative approximation, we add Eq. (4.9) and Eq. (4.10) and neglecting terms of degree higher than four to obtain

$$f(x_i+h)+f(x_i-h)=2f(x_i)+h^2f''(x_i)+\frac{2h^4}{4!}f''''(\xi),\ \xi\in(x_i-h,x_i+h)$$

So, we have

$$f''(x_i)=\frac{f(x_i+h)-2f(x_i)+f(x_i-h)}{h^2}-\frac{h^2}{12}f''''(\xi),\ \xi\in(x_i-h,x_i+h) \quad (4.12)$$

Eq. (4.12) can be written as

$$f''(x_i)=S+E_S,$$

where

$$S=\frac{f(x_i+h)-2f(x_i)+f(x_i-h)}{h^2},\ E_S=-\frac{h^2}{12}f''''(\xi),\quad \xi\in(x_i-h,x_i+h)$$

S is called differences approximation formula of $f''(x_i)$ and E_S is the error.

Example (4.7)

Find the second derivative of the function tabulated below at the point $x=0.2$ using differences approximation formula

x	0.1	0.2	0.3	0.4	0.5
y	0.0001	0.0016	0.0081	0.0256	0.0625

Solution

The differences approximation formula of the second derivative defined as

$$f''(x_i) = \frac{f(x_i + h) - 2f(x_i) + f(x_i - h)}{h^2}.$$

So, we have

$$f''(0.2) = \frac{f(0.3) - 2f(0.2) + f(0.1)}{(0.1)^2}$$

$$= \frac{0.0081 - 2(0.0016) + 0.0001}{(0.1)^2} = 0.5$$

CHAPTER (5)

NUMERICAL INTEGRATION

1. Introduction

The process of computing $\int_a^b y(x)dx$ where $y = f(x)$ is given by a set of tabulated values $[x_i, y_i]$, $i = 0,1,2,\cdots,n$, $a = x_0$, $b = x_n$ is called numerical integration. Like that of numerical differentiation, here we also replace $y = f(x)$ by an interpolation formula and integrate it between the given limits. In this way we can derive a quadrature formula for approximate integration of a function defined by a set of numerical values.

2. General quadrature formula

In this section we will derive a general quadrature formula for equidistant mesh points.

Let

$$I = \int_a^b y\,dx \text{ , where } y = f(x),$$

takes the values $y_0, y_1, \ldots, y_n$ for $x_0, x_1, \ldots, x_n$. Let us divide the interval (a,b) into n equal parts of width h, so that

$$a = x_0, x_1 = x_0 + h, x_2 = x_0 + 2h, \ldots, x_n = x_0 + nh = b .$$

Then,

$$I = \int_{x_0}^{x_0+nh} f(x)dx$$

Putting, $x = x_0 + qh$, so that $dx = hdq$ in above, we get,

$$I = h\int_0^n f(x_0 + qh)dq = h\int_0^n y(x)dq .$$

Now replacing $y(x)$ by Newton forward interpolation formula we get,

$$I = h\int_0^n \left[y_0 + q\Delta y_0 + \frac{q(q-1)}{2!}\Delta^2 y_0 + \frac{q(q-1)(q-2)}{3!}\Delta^3 y_0 + \frac{q(q-1)(q-2)(q-3)}{4!}\Delta^4 y_0 + \frac{q(q-1)(q-2)(q-3)(q-4)}{5!}\Delta^5 y_0 + \frac{q(q-1)(q-2)(q-3)(q-4)(q-5)}{6!}\Delta^6 y_0 + \cdots \right] dq$$

Now integrating a term by term we get after substituting the limits as

$$I = h\left[ny_0 + \frac{n^2}{2}\Delta y_0 + \frac{1}{2}\left\{\frac{n^3}{3} - \frac{n^2}{2}\right\}\Delta^2 y_0 + \frac{1}{3!}\left\{\frac{n^4}{4} - n^3 + n^2\right\}\Delta^3 y_0 + \frac{1}{4!}\left\{\frac{n^5}{5} - \frac{3n^4}{2} + \frac{11n^3}{3} - 3n^2\right\}\Delta^4 y_0 + \frac{1}{5!}\left\{\frac{n^6}{6} - 2n^5 + \frac{35n^4}{4} - \frac{50n^3}{3} + 12n^2\right\}\Delta^5 y_0 + \frac{1}{6!}\left\{\frac{n^7}{7} - \frac{15n^6}{6} + 17n^5 - \frac{225n^4}{4} + \frac{274n^3}{3} - 60n^2\right\}\Delta^6 y_0 \right] \quad (5.1)$$

Eq.(5.1) is known as Newton-Cote's quadrature formula which is general quadratic formula for equidistant mesh points. In the following sections we deduce important quadrature formula for this equation taking $n = 1, 2, 3$.

3. Trapezoidal rule

Putting $n = 1$ in Eq. (5.1) and neglecting second and higher order differences we get

$$\int_{x_0}^{x_0+h} y(x)dx = h\int_0^1 y(x)dq = h\left[y_0 + \frac{1}{2}\Delta y_0 \right]$$

$$= h\left[y_0 + \frac{1}{2}(y_1 - y_0) \right] = \frac{h}{2}\left[y_0 + y_1 \right]$$

Similarly

$$\int_{x_0+h}^{x_0+2h} y(x)dx = \frac{h}{2}[y_1 + y_2]$$
$$\vdots \qquad \vdots \qquad \vdots$$
$$\int_{x_0+(n-1)h}^{x_0+nh} y(x)dx = \frac{h}{2}[y_{n-1} + y_n]$$

Adding these n integrals, we get,

$$I = \int_{x_0}^{x_0+nh} y(x)dx = \frac{h}{2}\left[(y_0 + y_n) + 2(y_1 + y_2 + \ldots + y_{n-1})\right] \tag{5.2}$$

Eq.(5.2) is known as trapezoidal rule.

4. Simpson's $1/3$ rule

Here, taking $n = 2$ in Eq.(5.1) and neglecting third and higher-order differences, we get

$$\int_{x_0}^{x_0+2h} y(x)dx = h\int_0^2 y(x)dq = h\left[2y_0 + 2\Delta y_0 + \frac{1}{2}\left(\frac{8}{3} - 2\right)\Delta^2 y_0\right]$$
$$= h\left[2y_0 + 2(y_1 - y_0) + \frac{1}{3}(y_2 - 2y_1 + y_0)\right]$$
$$= \frac{h}{3}[y_0 + 4y_1 + y_2]$$

Similarly

$$\int_{x_0+2h}^{x_0+4h} y(x)dx = \frac{h}{3}[y_2 + 4y_3 + y_4]$$
$$\vdots \qquad \vdots \qquad \vdots$$
$$\int_{x_0+(n-2)h}^{x_0+nh} y(x)dx = \frac{h}{3}[y_{n-2} + 4y_{n-1} + y_n],$$

where n is even. Adding all these integrals, we get

$$I = \int_{x_0}^{x_0+nh} y(x)dx = \frac{h}{3}\left[(y_0 + y_n) + 4(y_1 + y_3 + \ldots + y_{n-1}) + 2(y_2 + y_4 + \ldots + y_{n-2})\right] \tag{5.3}$$

Eq.(5.3) is known as Simpson's $1/3$ rule.

5. Simpson's $3/8$ rule

Putting $n=3$ in Eq.(5.1) and neglecting all differences above the third order, we get

$$\int_{x_0}^{x_0+3h} y(x)dx = h\int_0^3 y(x)dq$$

$$= h\left[3y_0+\frac{9}{2}\Delta y_0+\frac{1}{2}\left(\frac{27}{3}-\frac{9}{2}\right)\Delta^2 y_0+\frac{1}{3!}\left(\frac{81}{4}-27+9\right)\Delta^3 y_0\right]$$

$$= h\left[3y_0+\frac{9}{2}(y_1-y_0)+\frac{9}{4}(y_2-2y_1+y_0)+\frac{3}{8}(y_3-3y_2+3y_1-y_0)\right]$$

$$= \frac{3h}{8}\left[y_0+3y_1+3y_2+y_3\right]$$

Similarly

$$\int_{x_0+3h}^{x_0+6h} y(x)dx = \frac{3h}{8}\left[y_3+3y_4+3y_5+y_6\right]$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$\int_{x_0+(n-3)h}^{x_0+nh} y(x)dx = \frac{3h}{8}\left[y_{n-3}+3y_{n-2}+3y_{n-1}+y_n\right]$$

Adding all these integrals, where n is a multiple of 3, we get

$$I = \int_{x_0}^{x_0+nh} y(x)dx = \frac{3h}{8}\Big[\left(y_0+y_n\right)$$

$$+3\left(y_1+y_2+y_4+y_5+y_7+y_8+\ldots+y_{n-2}+y_{n-1}\right)$$

$$+2\left(y_3+y_6+\ldots+y_{n-3}\right)\Big] \qquad (5.4)$$

Eq. (5.4) known as Simpson's $3/8$ rule.

Note

- The trapezoidal rule $f(x)$ is linear function of x i.e. of the form $f(x)=ax+b$. It is the simplest rule but least accurate.
- In Simpson's $1/3$ rule, $f(x)$ is a polynomial of second degree, i.e. $f(x)=ax^2+bx+c$. To apply this rule, the number of intervals n must be even.
- In Simpson's $3/8$ rule $f(x)$ is a polynomial of third degree, i.e. $f(x)=ax^3+bx^2+cx+d$. To apply this rule the number of intervals n must be a multiple of 3.

Example (5.1)

Evaluate

$$I = \int_0^{10} \frac{dx}{1+x^2},$$

by using

1. Trapezoidal rule
2. Simpson's $1/3$ rule. Compare the results with the actual value.

Solution

Taking $n = 10$, divide the whole range of the integration into ten equal parts. The value of the integrand function for each point of sub-division are given below:

x	y	y_n
0	1.00000	y_0
1	0.50000	y_1
2	0.200000	y_2
3	0.100000	y_3
4	0.0588235	y_4
5	0.0384615	y_5
6	0.027027	y_6
7	0.0200000	y_7
8	0.0153846	y_8
9	0.0121951	y_9
10	9.9009901×10^{-3}	y_{10}
$\sum$		

1. By Trapezoidal rule

$$I=\int_0^{10}\frac{dx}{1+x^2}=\frac{h}{2}\left[(y_0+y_{10})+2(y_1+y_2+y_3+y_4+y_5+y_6+y_7+y_8+y_9)\right]$$

$$=\frac{1}{2}\left[(1+9.9009901\times10^{-3})+2(0.5+0.2+0.1+0.0588235+0.0384615\right.$$

$$\left.+0.027027+0.02+0.0153846+0.0121951)\right]=1.4768422$$

2. By Simpson's $1/3$ rule

$$I=\int_0^{10}\frac{dx}{1+x^2}=\frac{h}{3}\left[(y_0+y_{10})+4(y_1+y_3+y_5+y_7+y_9)+2(y_2+y_4+y_6+y_8)\right]$$

$$=\frac{1}{3}\left[(1+9.9009901\times10^{-3})+4(0.5+0.1+0.0384615+0.02+0.0121951)\right.$$

$$\left.+2(0.2+0.0588235+0.027027+0.0153846)\right]=1.4316659$$

Example (5.2)

The velocity v of a particle at a distance x from a point on its path is given by the following table:

$x\ (ft)$	0	10	20	30	40	50	60
$v\ (ft/s)$	47	58	64	65	61	52	38

Estimate the time taken to travel to $60ft$ using Simpson's $1/3$ rule. Compare the result with Simpson's $3/8$ rule.

Solution

We know that the rate of displacement is velocity , i.e. $v=\frac{dx}{dt}$. Therefore the time taken to travel $60ft$ is given by

$$t=\int_0^{60}\frac{1}{v}dx=\int_0^{60}y\,dx$$

where $y=1/v$. The table is as given below.

x	$y = 1/v$	y_n
0	0.0212765	y_0
10	0.0172413	y_1
20	0.015625	y_2
30	0.0153846	y_3
40	0.0163934	y_4
50	0.0192307	y_5
60	0.0263157	y_6

By Simpson's $1/3$ rule

$$I = \int_0^{60} y\,dx = \frac{h}{3}\left[(y_0 + y_6) + 4(y_1 + y_3 + y_5) + 2(y_2 + y_4)\right]$$

$$= \frac{10}{3}\left[(0.0212765 + 0.0263157) + 4(0.0172413 + 0.0153846 + 0.0192307)\right.$$

$$\left. +2(0.015625 + 0.0163934)\right] = 1.063518$$

Hence the time taken to travel $60ft$ is 1.064s.

By Simpson's $3/8$ rule

$$I = \int_0^{60} y\,dx = \frac{3h}{8}\left[(y_0 + y_6) + 3(y_1 + y_2 + y_4 + y_5) + 2y_3\right]$$

$$= \frac{30}{8}\left[(0.0212765 + 0.0263157) + 3(0.0172413 + 0.015625 + 0.0163934 + 0.0192307)\right.$$

$$\left. +2(0.0153846)\right] = 1.0643723$$

By this method also the time taken to travel $60ft$ is 1.064s.

Example (5.3)

Find the following integral by

(i) Trapezoidal rule (ii) Simpson's $1/3$ rule (iii) Simpson's $3/8$ rule

$$I = \int_4^{5.2} \ln x\,dx$$

Solution

Taking $n = 6$, divide the whole range of the integration into six equal parts. The value of the integrand function for each point of sub-division are given below:

x	4	4.2	4.4	4.6	4.8	5	5.2
$f(x) = \ln x$	1.386	1.435	1.482	1.526	1.569	1.609	1.649

1. By Trapezoidal rule

$$I = \int_4^{5.2} \ln x \, dx = \frac{h}{2}\left[(y_0 + y_6) + 2(y_1 + y_2 + y_3 + y_4 + y_5)\right]$$

$$= \frac{0.2}{2}\left[(1.386 + 1.649) + 2(1.435 + 1.482 + 1.526 + 1.569 + 1.609)\right] = 1.8277$$

2. By Simpson's $1/3$ rule

$$I = \int_4^{5.2} \ln x \, dx = \frac{h}{3}\left[(y_0 + y_6) + 4(y_1 + y_3 + y_5) + 2(y_2 + y_4)\right]$$

$$= \frac{0.2}{3}\left[(1.386 + 1.649) + 4(1.435 + 1.526 + 1.609) + 2(1.482 + 1.569)\right] = 1.8278$$

3. By Simpson's $3/8$ rule

$$I = \int_4^{5.2} \ln x \, dx = \frac{3h}{8}\left[(y_0 + y_6) + 3(y_1 + y_2 + y_4 + y_5) + 2y_3\right]$$

$$= \frac{0.6}{8}\left[(1.386 + 1.649) + 3(1.435 + 1.482 + 1.569 + 1.609) + 2(1.526)\right] = 1.8279$$

Example (5.4)

A rocket is launched from the ground . Its acceleration is registered during the 90 seconds and are given in the table below. Using Simpson's $3/8$ rule, find the velocity of the rocket at $t = 90$.

t (s)	0	10	20	30	40	50	60	70	80	90
a (m/ s^2)	30	31.63	33.64	35.47	37.75	40.33	43.25	46.69	50.67	54.87

Solution

We know that the rate of velocity is acceleration , i.e. $a = \frac{dv}{dt}$ Therefore the velocity of the rocket at $t = 90$ is given by

$$v = \int_0^{90} a\, dt\ .$$

By Simpson's $3/8$ rule

$$I = \int_0^{90} a\, dt = \frac{3h}{8}\left[(y_0 + y_9) + 3(y_1 + y_2 + y_4 + y_5 + y_7 + y_8) + 2(y_3 + y_6)\right]$$

$$= \frac{30}{8}\left[(30 + 54.87) + 3(31.63 + 33.64 + 37.75 + 40.33 + 46.69 + 50.67) + 2(35.47 + 43.25)\right]$$

$$= 3616.65$$

CHAPTER (6)

SOLUTIONS OF ALGEBRIAC AND TRANSCENDENTAL EQUATIONS

1. Introduction

We have seen that an expression of the form

$$f(x) = a_0 x^n + a_1 x^{n-1} + a_2 x^{n-2} + \cdots + a_n ,$$

where $a's$ are constants $(a_0 \neq 0)$ and n is positive integer, is called a polynomial in x of degree n and the equation $f(x) = 0$ is called an algebraic equation of degree n . If $f(x)$ contains some other functions like exponential, trigonometric, logarithmic , then $f(x) = 0$ is called transcendental equation. For example

$x^3 - 3x + 6 = 0$, $x^5 - 7x^4 + 3x^2 + 36x - 7 = 0$

are algebraic equations. Whereas

$x^2 - 3\cos x + 1 = 0$, $xe^x - 2 = 0, x \log x = 1.2$

are transcendental equations.

In this chapter we will solve algebraic and the transcendental equations. For equation s of degree two or three or four, methods are available to solve them. But the need often arises to solve higher degree or transcendental equation for which no direct method exists. Such equations can be solved by approximate methods. Before we proceed to solve such equations let us recall the fundamental theorem on roots of $f(x) = 0$ in $a \leq x \leq b$.

Theorem 6.1

If $f(x) = 0$ is continuous function in a closed interval $[a,b]$ and $f(a), f(b)$ are of opposite signs, then the equation $f(x) = 0$ will have at least one real root between a and b .

2. Bisection method

Let the function $f(x)$ be continuous between a and b . For definiteness let $f(a)$ be negative and $f(b)$ be positive, then there is a root of $f(x) = 0$ lying between a and b . Let the first the approximation be $x_1 = \dfrac{a+b}{2}$ (the average of the ends of the range).

Now if $f(x_1)=0$, then x_1 is a root of $f(x)=0$. Otherwise, the root will lie between a and x_1 or x_1 and b depending upon whether $f(x_1)$ is positive or negative.

Then, as before we bisect the interval and continue the process till the root is found to the desired accuracy. If $f(x_1)$ is positive, therefore the root lies between a and x_1. The second approximation to the root now is $x_2=\frac{a+x_1}{2}$. If $f(x_2)$ is negative, then the root lies between x_1 and x_2 then, the third approximation to the root is $x_3=\frac{x_1+x_2}{2}$ and so on. This method is simple but slowly convergent.

Example (6.1)

Find a root of the equation

$$x^3-x-11=0,$$

correct to four decimal places using bisection method.

Solution

Let

$$f(x)=x^3-x-11.$$

Since $f(2)=-5<0$ and $f(3)=13>0$, then there exist a real root lies between 2 and 3. Hence, the first approximation to the root is

$$x_1=\frac{2+3}{2}=2.5.$$

Now

$$f(2.5)=(2.5)^3-2.5-11=2.125>0.$$

Therefore the second approximation lies between 2 and 2.5. Thus the second approximation to the root is

$$x_2 = \frac{2+2.5}{2} = 2.25.$$

Now

$$f\ (2.25) = (2.25)^3 - 2.25 - 11 = -1.859375 < 0.$$

Therefore the third approximation lies between 2.5 and 2.25. Thus the third approximation to the root is

$$x_3 = \frac{x_1 + x_2}{2} = \frac{2.5+2.25}{2} = 2.375.$$

Now

$$f\ (2.375) = (2.375)^3 - 2.375 - 11 = 0.0214843 > 0.$$

Therefore the fourth approximation lies between 2.25 and 2.375. Thus the fourth approximation to the root is

$$x_4 = \frac{x_2 + x_3}{2} = \frac{2.25+2.375}{2} = 2.3125.$$

Now

$$f\ (2.3125) = (2.3125)^3 - 2.3125 - 11 = -0.9460449 < 0.$$

Therefore the fifth approximation lies between 2.375 and 2.3125. Thus the fifth approximation to the root is

$$x_5 = \frac{x_3 + x_4}{2} = \frac{2.375+2.3125}{2} = 2.34375.$$

Now

$$f\ (2.34375) = (2.34375)^3 - 2.34375 - 11 = -0.4691467 < 0.$$

Therefore the sixth approximation lies between 2.375 and 2.34375. Thus the sixth approximation to the root is

$$x_6 = \frac{x_3 + x_5}{2} = \frac{2.375+2.34375}{2} = 2.359375.$$

Now

$$f(2.359375) = (2.359375)^3 - 2.359375 - 11 = -0.2255592 < 0.$$

Therefore the seventh approximation lies between 2.375 and 2.359375. Thus the seventh approximation to the root is

$$x_7 = \frac{x_3 + x_6}{2} = \frac{2.375 + 2.359375}{2} = 2.3671875.$$

Now

$$f(2.3671875) = (2.3671875)^3 - 2.3671875 - 11 = -0.1024708 < 0.$$

Which means that the eighth approximation lies between 2.375 and 2.3671875. Thus the eighth approximation to the root is

$$x_8 = \frac{x_3 + x_7}{2} = \frac{2.375 + 2.3671875}{2} = 2.3710938.$$

Now

$$f(2.3710938) = (2.3710938)^3 - 2.3710938 - 11 = -0.040601 < 0.$$

Which means that the ninth approximation lies between 2.375 and 2.3710938. Thus the ninth approximation to the root is

$$x_9 = \frac{x_3 + x_8}{2} = \frac{2.375 + 2.3710938}{2} = 2.3730469.$$

Now

$$f(2.3730469) = (2.3730469)^3 - 2.3730469 - 11 = -9.585864 \times 10^{-3} < 0.$$

Therefore the tenth approximation lies between 2.375 and 2.3730469. Thus the tenth approximation to the root is

$$x_{10} = \frac{x_3 + x_9}{2} = \frac{2.375 + 2.3730469}{2} = 2.3740235.$$

Now

$$f(2.3740235) = (2.3740235)^3 - 2.3740235 - 11 = 5.942463 \times 10^{-3} > 0.$$

Therefore the eleventh approximation lies between 2.3730469 and 2.3740235. Thus the eleventh approximation to the root is

$$x_{11} = \frac{x_9 + x_{10}}{2} = \frac{2.3730469 + 2.3740235}{2} = 2.3735352.$$

Now

$$f\ (2.3135352) = (2.3135352)^3 - 2.3135352 - 11 = -1.823398 \times 10^{-3} < 0$$

Therefore the twelfth approximation lies between 2.3740235 and 2.3735352. Thus the twelfth approximation to the root is

$$x_{12} = \frac{x_{10} + x_{11}}{2} = \frac{2.3740235 + 2.3735352}{2} = 2.3737793.$$

Now

$$f\ (2.3737793) = (2.3737793)^3 - 2.3737793 - 11 = 2.059107 \times 10^{-3} > 0.$$

Therefore the thirteenth approximation lies between 2.3735352 and 2.3737793. Thus the thirteenth approximation to the root is

$$x_{13} = \frac{x_{11} + x_{12}}{2} = \frac{2.3735352 + 2.3737793}{2} = 2.3736572.$$

Now

$$f\ (2.3736572) = (2.3736572)^3 - 2.3736572 - 11 = 1.17748 \times 10^{-4} > 0.$$

Therefore the fourteenth approximation lies between 2.3735352 and 2.3736572. Thus the fourteenth approximation to the root is

$$x_{14} = \frac{x_{11} + x_{13}}{2} = \frac{2.3735352 + 2.3736572}{2} = 2.3735962.$$

Now

$$f\ (2.3735962) = (2.3735962)^3 - 2.3735962 - 11 = -8.52851 \times 10^{-4} < 0.$$

Therefore the fifteenth approximation lies between 2.3736572 and 2.3735962. Thus the fifteenth approximation to the root is

$$x_{15} = \frac{x_{13} + x_{14}}{2} = \frac{2.3736572 + 2.3735962}{2} = 2.3736267.$$

Now

$f\ (2.3736267) = (2.3736267)^3 - 2.3736267 - 11 = -3.67558 \times 10^{-4} < 0.$

Therefore from x_{14} and x_{15} we can see that $f\ (x_{14})$ and $f\ (x_{15})$ are nearly equal to 0. Hence the root is correct to 4 decimal places is 2.37362.

Example (6.2)

Using bisection method, find the negative root of

$$x^3 - x + 11 = 0$$

Solution

Let

$$f\ (x) = x^3 - x + 11.$$

Hence

$$f\ (-x) = -x^3 + x + 11.$$

The negative root of $f\ (x) = 0$ is the positive root of $f\ (-x) = 0$. Therefore we will find the positive root of $f\ (-x) = 0$,

i.e.

$$x^3 - x - 11 = 0.$$

Proceeding as explained in example (1), we get $x = 2.37362$ and hence the negative root is $x = -2.37362$.

3. Iteration method

Let $f\ (x) = 0$ by the given equation whose roots are to be determined this equation can be written in the form

$$x = \phi(x). \qquad (6.1)$$

Let $x = x_0$ an initial approximation to the actual root say α of Eq. (6.1). Then the first approximation is $x_1 = \phi(x_0)$ and successive approximations are $x_2 = \phi(x_1),\ x_3 = \phi(x_2),\ x_4 = \phi(x_3)\ , \ldots ,\ x_n = \phi(x_{n-1})$. If the sequence of approximate roots $x_0, x_1, x_2, \ldots, x_n$ converges to α, then the value x_n it is taking as the root of the equation

$f(x)=0$. For the convergence purpose the function $\phi(x)$ have to be chosen carefully. The choice of $\phi(x)$ is determined according to the following theorem.

Theorem 6.2

If α is a root of $f(x)=0$ which is equivalent to $x=\phi(x)$. Let I be an interval contains the point $x=\alpha$.Then the sequence of approximations $x_0, x_1, x_2, \ldots, x_n$ will converge to the root α , if

$$|\phi'(x)|<1 \;\forall\; x \in I \ .$$

- **Note**

The smaller values of $\phi'(x)$ the more rapid convergence

Example (6.3)

Find a real root of the equation

$$x^3+x^2-1=0.$$

By iteration method.

Solution

Let $f(x)=x^3+x^2-1$. Now $f(0)=-1$ and $f(1)=1$. Hence a real root lies between 0 and 1 . Rewrite $x^3+x^2-1=0$ as

$$x=\frac{1}{\sqrt{1+x}}=\phi(x).$$

Now

$$\phi'(x)=-\frac{1}{2(1+x)^{3/2}}.$$

It is clear that

$$|\phi'(x)|<1 \;\forall\; x \in [0,1].$$

Hence the iteration method can be applied. Let $x_0=0.65$ be the initial approximation to the desired root, then

$$x_0 = 0.65,$$

$$x_1 = \phi(x_0) = \frac{1}{\sqrt{1+x_0}} = \frac{1}{\sqrt{1.65}} = 0.7784989,$$

$$x_2 = \frac{1}{\sqrt{1+x_1}} = \frac{1}{\sqrt{1.7784989}} = 0.7498479,$$

$$x_3 = \frac{1}{\sqrt{1+x_2}} = \frac{1}{\sqrt{1.7498479}} = 0.7559617,$$

$$x_4 = \frac{1}{\sqrt{1+x_3}} = \frac{1}{\sqrt{1.7559617}} = 0.7546446,$$

$$x_5 = \frac{1}{\sqrt{1+x_4}} = \frac{1}{\sqrt{1.7546446}} = 0.7549278,$$

$$x_6 = \frac{1}{\sqrt{1+x_5}} = \frac{1}{\sqrt{1.7549278}} = 0.7548668,$$

$$x_7 = \frac{1}{\sqrt{1+x_6}} = \frac{1}{\sqrt{1.7548668}} = 0.7548799,$$

$$x_8 = \frac{1}{\sqrt{1+x_7}} = \frac{1}{\sqrt{1.7548799}} = 0.7548771,$$

$$x_9 = \frac{1}{\sqrt{1+x_8}} = \frac{1}{\sqrt{1.7548771}} = 0.7548777,$$

$$x_{10} = \frac{1}{\sqrt{1+x_9}} = \frac{1}{\sqrt{1.7548777}} = 0.7548776,$$

$$x_{11} = \frac{1}{\sqrt{1+x_{10}}} = \frac{1}{\sqrt{1.7548776}} = 0.7548776,$$

Hence the root is 0.7548776.

Example (6.4)

Find a real root of the equation $\cos x - 3x + 1 = 0$ correct to seven decimal places.

Solution

Let $f(x)=\cos x-3x+1$. Now $f(0)=2>0$ and $f(\pi/2)=-\frac{3\pi}{2}+1<0$. Therefore there exist a real root lies between 0 and $\pi/2$. Rewrite $\cos x-3x+1=0$ as

$$x=\frac{1}{3}(\cos x+1)=\phi(x).$$

Now

$$\phi'(x)=-\frac{\sin x}{3}.$$

It is clear that

$$|\phi'(x)|=\left|-\frac{\sin x}{3}\right|<\frac{1}{3}\ \forall\ x\ .$$

Hence the iteration method can be applied. Let $x_0=0.5$ be the initial approximation to the desired root, then

$$x_1=\phi(x_0)=\frac{1}{3}(\cos x+1)=0.6258608,$$

$$x_2=\frac{1}{3}(\cos(0.6258608)+1)=0.6034863,$$

$$x_3=\frac{1}{3}(\cos(0.6034863)+1)=0.6077873,$$

$$x_4=\frac{1}{3}(\cos(0.6077873)+1)=0.6069711,$$

$$x_5=\frac{1}{3}(\cos(0.6069711)+1)=0.6071264,$$

$$x_6=\frac{1}{3}(\cos(0.6071264)+1)=0.6070969,$$

$$x_7=\frac{1}{3}(\cos(0.6070969)+1)=0.6071025,$$

$$x_8=\frac{1}{3}(\cos(0.6071025)+1)=0.6071014,$$

$$x_9=\frac{1}{3}(\cos(0.6071014)+1)=0.6071016,$$

$$x_{10} = \frac{1}{3}\left(\cos(0.6071016) + 1\right) = 0.6071016,$$

Hence the root is 0.6071016.

4. Newton-Raphson method

This method, is a particular form of the iteration method discussed in section 3. When an approximate value of a root of an equation is given, a better and closer approximation to the root can be found using this method. It can be derived as follows:

Let x_0 be an approximation of a root of the given equation $f(x) = 0$, which may be algebraic or transcendental. Let $x_0 + h$ be the exact value or the better approximation of the corresponding root, h being a small quantity. Then $f(x_0 + h) = 0$. Expanding $f(x_0 + h) = 0$ by Taylor's theorem, we get

$$f(x_0 + h) = f(x_0) + hf'(x_0) + \frac{h^2}{2!}f''(x_0) + \ldots = 0.$$

Since h is small, we can neglect second, third and higher degree terms in h and thus we get,

$$f(x_0) + hf'(x_0) = 0.$$

Or

$$h = -\frac{f(x_0)}{f'(x_0)}; \quad f'(x_0) \neq 0.$$

Hence,

$$x_1 = x_0 + h = x_0 - \frac{f(x_0)}{f'(x_0)}$$

Now substituting x_1 for x_0 and x_2 for x_1, then the next better approximations are given by

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)},$$

and

$$x_3 = x_2 - \frac{f(x_n)}{f'(x_n)}.$$

Proceeding in the same way n times, we get the general formula

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \quad \text{for } n = 0,1,2,\ldots, \tag{6.2}$$

which is known as Newton-Raphson formula.

Example (6.5)

Find an iterative formula to find $\sqrt{N}$, where N is a positive number and hence, find $\sqrt{12}$ correct to four decimal places.

Solution

Let

$$x = \sqrt{N} \;\Rightarrow\; x^2 - N = 0.$$

Assume

$$f(x) = x^2 - N\ .$$

Then,

$$f'(x) = 2x$$

Now, from Newton-Raphson formula,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^2 - N}{2x_n}$$
$$= \frac{1}{2}\left[x_n + \left(\frac{N}{x_n}\right)\right] \tag{6.3}$$

Eq. (6.3) is the required iterative formula. Putting $N = 12$ in $f(x)$, we have $f(x) = x^2 - 12$.

Now, $f(3)<0$ and $f(4)>0$. Therefore, the root lies in between 3 and 4. Let the initial approximation x_0 be 3.1. Then, from Eq. (6.3) the first approximation to the root

$$x_1=\frac{1}{2}\left[x_0+\frac{12}{x_0}\right]=\frac{1}{2}\left[3.1+\frac{12}{3.1}\right]=3.4854839.$$

The second approximation is

$$x_2=\frac{1}{2}\left[x_1+\frac{12}{x_1}\right]=\frac{1}{2}\left[3.4854839+\frac{12}{3.4854839}\right]=3.4641672.$$

The third approximation is

$$x_3=\frac{1}{2}\left[3.4641672+\frac{12}{3.4641672}\right]=3.4641016.$$

The fourth approximation is

$$x_4=\frac{1}{2}\left[3.4641016+\frac{12}{3.4641016}\right]=3.4641016.$$

Thus, the value of $\sqrt{12}$ correct to four decimals is 3.4641.

Example (6.6)

Solve $x^3+2x^2+10x-20=0$ by Newton-Raphson method.

Solution

Let

$$f(x)=x^3+2x^2+10x-20.$$

Therefore

$$f'(x)=3x^2+4x+10.$$

From Eq. (6.2)

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$
$$= x_n - \left[\frac{x_n^3 + 2x_n^2 + 10x_n - 20}{3x_n^2 + 4x_n + 10}\right]$$
$$= \frac{2\left(x_n^3 + x_n^2 + 10\right)}{3x_n^2 + 4x_n + 10}. \qquad (6.4)$$

Now we can see that $f(1) = -7 < 0$ and $f(2) = 16 > 0$. Therefore, the root lies in between 1 and 2. Let $x_0 = 1.2$ be the initial approximation $(\because f(1.2) < 0)$.

Putting $n = 0$ in Eq. (6.4), first approximation x_1 is given by

$$x_1 = \frac{2\left(x_0^3 + x_0^7 + 10\right)}{3x_0^2 + 4x_0 + 10} = \frac{2\left[(1.2)^2 + (1.2)^2 + 10\right]}{3(1.2)^2 + 4(1.2) + 10}$$
$$= \frac{26.336}{19.12} = 1.3774059.$$

The second approximation x_2 is

$$x_2 = \frac{2\left(x_1^3 + x_1^2 + 10\right)}{3x_1^2 + 4x_1 + 10} = \frac{2\left[(1.3774059)^3 + (1.3774059)^2 + 10\right]}{3(1.3774059)^2 + 4(1.3774059) + 10}$$
$$= \frac{29.021052}{21.201364} = 1.3688295.$$

The third approximation x_3 is given by

$$x_3 = \frac{2\left(x_2^3 + x_2^2 + 10\right)}{3x_2^2 + 4x_2 + 10} = \frac{2\left[(1.3688295)^3 + (1.3688295)^2 + 10\right]}{3(1.3688295)^2 + 4(1.3688295) + 10}$$
$$= \frac{28.876924}{210064} = 1.3688081.$$

The fourth approximation x_4 (to the root) is given by

$$x_4 = \frac{2\left(x_3^3 + x_3^2 + 10\right)}{3x_3^2 + 4x_3 + 10} = \frac{2\left[(1.3688081)^3 + (1.3688081)^2 + 10\right]}{3(1.3688081)^2 + 4(13688081) + 10}$$
$$= \frac{28.876567}{21.09614} = 1.3688081.$$

Hence the root is 1.3688081.

Example (6.7)

Using Newton-Raphson method, find the root of the equation

$$x \ln x = 1.2 \ .$$

Solution

Let

$$f(x) = x \ln x - 1.2 \ \Rightarrow \ f'(x) = \ln x + 1.$$

From Newton-Raphson formula,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n \ln x_n - 1.2}{\ln x_n + 1}.$$

Therefore

$$x_{n+1} = \frac{x_n + 1.2}{\ln x_n + 1}. \qquad (6.5)$$

Now $f(2.5) = -0.2051499 < 0$ and $f(3) = 0.2313637 > 0$. Therefore, the real root of $f(x)$ lies in $(2.5,3)$. Let $x_0 = 2.7$ be the initial approximation. Putting $n = 0$ in Eq. (6.5), the first approximation x_1 is given by

$$x_1 = \frac{x_0 + 1.2}{\ln x_0 + 1} = \frac{2.7 + 1.2}{\ln 2.7 + 1} = 1.9566.$$

The second approximation x_2 is

$$x_2 = \frac{x_1 + 1.2}{\ln x_1 + 1} = \frac{1.9566 + 1.2}{\ln(1.9566) + 1} = 1.8888.$$

Similarly, the third approximation is

$$x_3 = \frac{x_2 + 1.2}{\ln x_2 + 1} = \frac{1.8888 + 1.2}{\ln(1.8888) + 1} = 1.88809.$$

Hence, the root is 1.88809.

Example (6.8)

Solve $\sin x = 1 + x^3$ using Newton-Raphson method.

Solution

Let

$$f(x) = \sin x - 1 - x^3 \quad \Rightarrow \quad f'(x) = \cos x - 3x^2.$$

Then, from Newton-Raphson formula,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{\sin x_n - 1 - x_n^3}{\cos x_n - 3x_n^2} \quad .$$

Hence

$$x_{n+1} = \frac{x_n \cos x_n - \sin x_n - 2x_n^3 + 1}{\cos x_n - 3x_n^2}. \qquad (6.6)$$

Now

$$f(-1) = \sin(-1) - 1 - (-1)^3 = -0.8414709 < 0,$$

and

$$f(-2) = \sin(-2) - 1 - (-2)^3 = 6.0907026 > 0,$$

which means that the root lies in between -1 and -2. Let $x_0 = -1.1$ be the initial approximation. Then, by putting

$n = 0,1,2,\ldots$ in Eq. (6.6), we obtain the successive approximations as

$$x_1 = \frac{x_0 \cos x_0 - \sin x_0 - 2x_0^3 + 1}{\cos x_0 - 3x_0^2} = \frac{4.0542516}{-3.1764039} = -1.2763653$$

$$x_2 = \frac{5.7452469}{-4.5971297} = -1.2497465$$

$$x_3 = \frac{5.4584049}{-4.370036} = -1.2490526$$

$$x_4 = \frac{5.4510835}{-4.364176} = -1.2490522$$

$$x_5 = \frac{5.4510786}{-4.3641722} = -1.2490521$$

$$x_6 = \frac{5.4510785}{-4.3641721} = -1.2490522$$

Hence the approximated root is x_6, i.e. -1.2490522 .