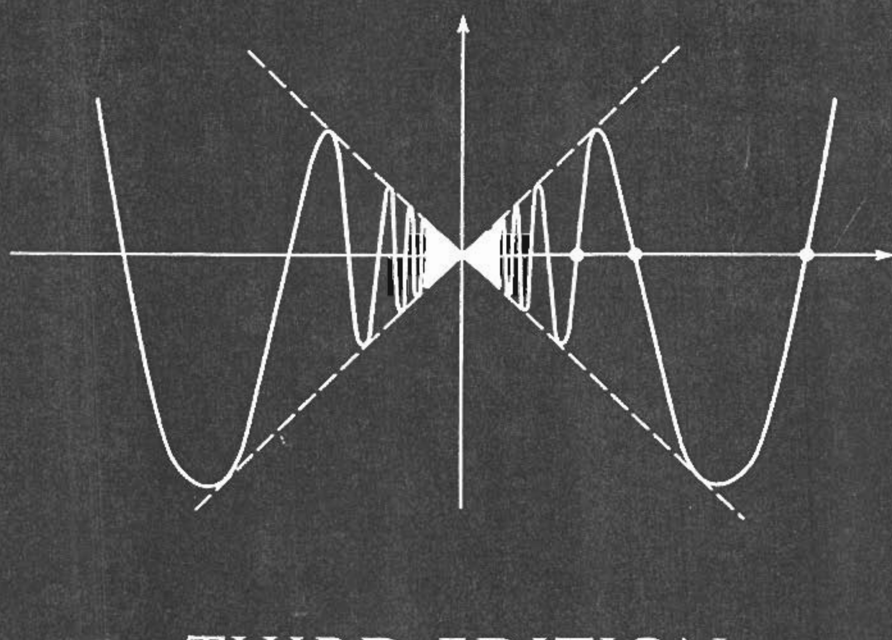
BARTLE / SHERBERT

ROBERT G. BARTLE
DONALD R. SHERBERT

INTRODUCTION TO REAL ANALYSIS



THIRD EDITION

Introduction to Real Analysis

Jo

ACQUISITION EDITOR	Barbara Holland
ASSOCIATE EDITOR	Sharon Prendergast
PRODUCTION EDITOR	Ken Santor
PHOTO EDITOR	Nicole Horlacher
ILLUSTRATION COORDINATOR	Gene Aiello

This book was set in Times Roman by Eigentype Compositors, and printed and bound by Hamilton Printing Company. The cover was printed by Phoenix Color Corporation.

This book is printed on acid-free paper. ♾

The paper in this book was manufactured by a mill whose forest management programs include sustained yield harvesting of its timberlands. Sustained yield harvesting principles ensure that the numbers of trees cut each year does not exceed the amount of new growth.

Library of Congress Cataloging in Publication Data:

Bartle, Robert Gardner, 1927–
Introduction to real analysis / Robert G. Bartle, Donald R., Sherbert. – 3rd ed.
p. cm.
Includes bibliographical references and index.
ISBN 0-471-32148-6 (alk. paper)
1. Mathematical analysis. 2. Functions of real variables.
I. Sherbert, Donald R., 1935– . II. Title.
QA300.B294 2000
515–dc21 99-13829
CIP

A. M. S. Classification 26-01

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

PREFACE

The study of real analysis is indispensible for a prospective graduate student of pure or applied mathematics. It also has great value for any undergraduate student who wishes to go beyond the routine manipulations of formulas to solve standard problems, because it develops the ability to think deductively, analyze mathematical situations, and extend ideas to a new context. In recent years, mathematics has become valuable in many areas, including economics and management science as well as the physical sciences, engineering, and computer science. Our goal is to provide an accessible, reasonably paced textbook in the fundamental concepts and techniques of real analysis for students in these areas. This book is designed for students who have studied calculus as it is traditionally presented in the United States. While students find this book challenging, our experience is that serious students at this level are fully capable of mastering the material presented here.

The first two editions of this book were very well received, and we have taken pains to maintain the same spirit and user-friendly approach. In preparing this edition, we have examined every section and set of exercises, streamlined some arguments, provided a few new examples, moved certain topics to new locations, and made revisions. Except for the new Chapter 10, which deals with the generalized Riemann integral, we have not added much new material. While there is more material than can be covered in one semester, instructors may wish to use certain topics as honors projects or extra credit assignments.

It is desirable that the student have had some exposure to proofs, but we do not assume that to be the case. To provide some help for students in analyzing proofs of theorems, we include an appendix on "Logic and Proofs" that discusses topics such as implications, quantifiers, negations, contrapositives, and different types of proofs. We have kept the discussion informal to avoid becoming mired in the technical details of formal logic. We feel that it is a more useful experience to learn how to construct proofs by first watching and then doing than by reading about techniques of proof.

We have adopted a medium level of generality consistently throughout the book: we present results that are general enough to cover cases that actually arise, but we do not strive for maximum generality. In the main, we proceed from the particular to the general. Thus we consider continuous functions on open and closed intervals in detail, but we are careful to present proofs that can readily be adapted to a more general situation. (In Chapter 11 we take particular advantage of the approach.) We believe that it is important to provide the student with many examples to aid them in their understanding, and we have compiled rather extensive lists of exercises to challenge them. While we do leave routine proofs as exercises, we do not try to attain brevity by relegating difficult proofs to the exercises. However, in some of the later sections, we do break down a moderately difficult exercise into a sequence of steps.

In Chapter 1 we present a brief summary of the notions and notations for sets and functions that we use. A discussion of Mathematical Induction is also given, since inductive proofs arise frequently. We also include a short section on finite, countable and infinite sets. We recommend that this chapter be covered quickly, or used as background material, returning later as necessary.

Chapter 2 presents the properties of the real number system $\mathbb{R}$. The first two sections deal with the Algebraic and Order Properties and provide some practice in writing proofs of elementary results. The crucial Completeness Property is given in Section 2.3 as the Supremum Property, and its ramifications are discussed throughout the remainder of this chapter.

In Chapter 3 we give a thorough treatment of sequences in $\mathbb{R}$ and the associated limit concepts. The material is of the greatest importance; fortunately, students find it rather natural although it takes some time for them to become fully accustomed to the use of ε. In the new Section 3.7, we give a brief introduction to infinite series, so that this important topic will not be omitted due to a shortage of time.

Chapter 4 on limits of functions and Chapter 5 on continuous functions constitute the heart of the book. Our discussion of limits and continuity relies heavily on the use of sequences, and the closely parallel approach of these chapters reinforces the understanding of these essential topics. The fundamental properties of continuous functions (on intervals) are discussed in Section 5.3 and 5.4. The notion of a "gauge" is introduced in Section 5.5 and used to give alternative proofs of these properties. Monotone functions are discussed in Section 5.6.

The basic theory of the derivative is given in the first part of Chapter 6. This important material is standard, except that we have used a result of Carathéodory to give simpler proofs of the Chain Rule and the Inversion Theorem. The remainder of this chapter consists of applications of the Mean Value Theorem and may be explored as time permits.

Chapter 7, dealing with the Riemann integral, has been completely revised in this edition. Rather than introducing upper and lower integrals (as we did in the previous editions), we here define the integral as a limit of Riemann sums. This has the advantage that it is consistent with the students' first exposure to the integral in calculus and in applications; since it is not dependent on order properties, it permits immediate generalization to complex- and vector-valued functions that students may encounter in later courses. Contrary to popular opinion, this limit approach is no more difficult than the order approach. It also is consistent with the generalized Riemann integral that is discussed in detail in Chapter 10. Section 7.4 gives a brief discussion of the familiar numerical methods of calculating the integral of continuous functions.

Sequences of functions and uniform convergence are discussed in the first two sections of Chapter 8, and the basic transcendental functions are put on a firm foundation in Section 8.3 and 8.4 by using uniform convergence. Chapter 9 completes our discussion of infinite series. Chapters 8 and 9 are intrinsically important, and they also show how the material in the earlier chapters can be applied.

Chapter 10 is completely new; it is a presentation of the generalized Riemann integral (sometimes called the "Henstock-Kurzweil" or the "gauge" integral). It will be new to many readers, and we think they will be amazed that such an apparently minor modification of the definition of the Riemann integral can lead to an integral that is more general than the Lebesgue integral. We believe that this relatively new approach to integration theory is both accessible and exciting to anyone who has studied the basic Riemann integral.

The final Chapter 11 deals with topological concepts. Earlier proofs given for intervals are extended to a more abstract setting. For example, the concept of compactness is given proper emphasis and metric spaces are introduced. This chapter will be very useful for students continuing to graduate courses in mathematics.

Throughout the book we have paid more attention to topics from numerical analysis and approximation theory than is usual. We have done so because of the importance of these areas, and to show that real analysis is not merely an exercise in abstract thought.

We have provided rather lengthy lists of exercises, some easy and some challenging. We have provided "hints" for many of these exercises, to help students get started toward a solution or to check their "answer". More complete solutions of almost every exercise are given in a separate Instructor's Manual, which is available to teachers upon request to the publisher.

It is a satisfying experience to see how the mathematical maturity of the students increases and how the students gradually learn to work comfortably with concepts that initially seemed so mysterious. But there is no doubt that a lot of hard work is required on the part of both the students and the teachers.

In order to enrich the historical perspective of the book, we include brief biographical sketches of some famous mathematicians who contributed to this area. We are particularly indebted to Dr. Patrick Muldowney for providing us with his photograph of Professors Henstock and Kurzweil. We also thank John Wiley & Sons for obtaining photographs of the other mathematicians.

We have received many helpful comments from colleagues at a wide variety of institutions who have taught from earlier editions and liked the book enough to express their opinions about how to improve it. We appreciate their remarks and suggestions, even though we did not always follow their advice. We thank them for communicating with us and wish them well in their endeavors to impart the challenge and excitement of learning real analysis and "real" mathematics. It is our hope that they will find this new edition even more helpful than the earlier ones.

February 24, 1999
Ypsilanti and Urbana

Robert G. Bartle
Donald R. Sherbert

THE GREEK ALPHABET

A	α	Alpha	N	ν	Nu
B	β	Beta	Ξ	ξ	Xi
Γ	γ	Gamma	O	o	Omicron
Δ	δ	Delta	Π	π	Pi
E	ε	Epsilon	P	ρ	Rho
Z	ζ	Zeta	Σ	σ	Sigma
H	η	Eta	T	τ	Tau
Θ	θ	Theta	Y	υ	Upsilon
I	ι	Iota	Φ	φ	Phi
K	κ	Kappa	X	χ	Chi
Λ	λ	Lambda	Ψ	ψ	Psi
M	μ	Mu	Ω	ω	Omega

To our wives, Carolyn and Janice,
with our appreciation for their
patience, support, and love.

CONTENTS

CHAPTER 1

PRELIMINARIES

In this initial chapter we will present the background needed for the study of real analysis. Section 1.1 consists of a brief survey of set operations and functions, two vital tools for all of mathematics. In it we establish the notation and state the basic definitions and properties that will be used throughout the book. We will regard the word "set" as synonymous with the words "class", "collection", and "family", and we will not define these terms or give a list of axioms for set theory. This approach, often referred to as "naive" set theory, is quite adequate for working with sets in the context of real analysis.

Section 1.2 is concerned with a special method of proof called Mathematical Induction. It is related to the fundamental properties of the natural number system and, though it is restricted to proving particular types of statements, it is important and used frequently. An informal discussion of the different types of proofs that are used in mathematics, such as contrapositives and proofs by contradiction, can be found in Appendix A.

In Section 1.3 we apply some of the tools presented in the first two sections of this chapter to a discussion of what it means for a set to be finite or infinite. Careful definitions are given and some basic consequences of these definitions are derived. The important result that the set of rational numbers is countably infinite is established.

In addition to introducing basic concepts and establishing terminology and notation, this chapter also provides the reader with some initial experience in working with precise definitions and writing proofs. The careful study of real analysis unavoidably entails the reading and writing of proofs, and like any skill, it is necessary to practice. This chapter is a starting point.

Section 1.1 Sets and Functions

To the reader: In this section we give a brief review of the terminology and notation that will be used in this text. We suggest that you look through quickly and come back later when you need to recall the meaning of a term or a symbol.

If an element x is in a set A, we write

$$x \in A$$

and say that x is a **member** of A, or that x **belongs** to A. If x is *not* in A, we write

$$x \notin A.$$

If every element of a set A also belongs to a set B, we say that A is a **subset** of B and write

$$A \subseteq B \qquad \text{or} \qquad B \supseteq A$$

We say that a set A is a **proper subset** of a set B if $A \subseteq B$, but there is at least one element of B that is not in A. In this case we sometimes write

$$A \subset B.$$

1.1.1 Definition Two sets A and B are said to be **equal**, and we write $A = B$, if they contain the same elements.

Thus, to prove that the sets A and B are equal, we must show that

$$A \subseteq B \qquad \text{and} \qquad B \subseteq A.$$

A set is normally defined by either listing its elements explicitly, or by specifying a property that determines the elements of the set. If P denotes a property that is meaningful and unambiguous for elements of a set S, then we write

$$\{x \in S : P(x)\}$$

for the set of all elements x in S for which the property P is true. If the set S is understood from the context, then it is often omitted in this notation.

Several special sets are used throughout this book, and they are denoted by standard symbols. (We will use the symbol $:=$ to mean that the symbol on the left is being *defined* by the symbol on the right.)

- The set of **natural numbers** $\mathbb{N} := \{1, 2, 3, \cdots\}$,
- The set of **integers** $\mathbb{Z} := \{0, 1, -1, 2, -2, \cdots\}$,
- The set of **rational numbers** $\mathbb{Q} := \{m/n : m, n \in \mathbb{Z} \text{ and } n \neq 0\}$,
- The set of **real numbers** $\mathbb{R}$.

The set $\mathbb{R}$ of real numbers is of fundamental importance for us and will be discussed at length in Chapter 2.

1.1.2 Examples (a) The set

$$\{x \in \mathbb{N} : x^2 - 3x + 2 = 0\}$$

consists of those natural numbers satisfying the stated equation. Since the only solutions of this quadratic equation are $x = 1$ and $x = 2$, we can denote this set more simply by $\{1, 2\}$.

(b) A natural number n is **even** if it has the form $n = 2k$ for some $k \in \mathbb{N}$. The set of even natural numbers can be written

$$\{2k : k \in \mathbb{N}\},$$

which is less cumbersome than $\{n \in \mathbb{N} : n = 2k, \ k \in \mathbb{N}\}$. Similarly, the set of **odd** natural numbers can be written

$$\{2k - 1 : k \in \mathbb{N}\}. \qquad \square$$

Set Operations

We now define the methods of obtaining new sets from given ones. Note that these set operations are based on the meaning of the words "or", "and", and "not". For the union, it is important to be aware of the fact that the word "or" is used in the *inclusive sense*, allowing the possibility that x may belong to both sets. In legal terminology, this inclusive sense is sometimes indicated by "and/or".

1.1.3 Definition **(a)** The **union** of sets A and B is the set

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

(b) The **intersection** of the sets A and B is the set

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

(c) The **complement of** B **relative to** A is the set

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

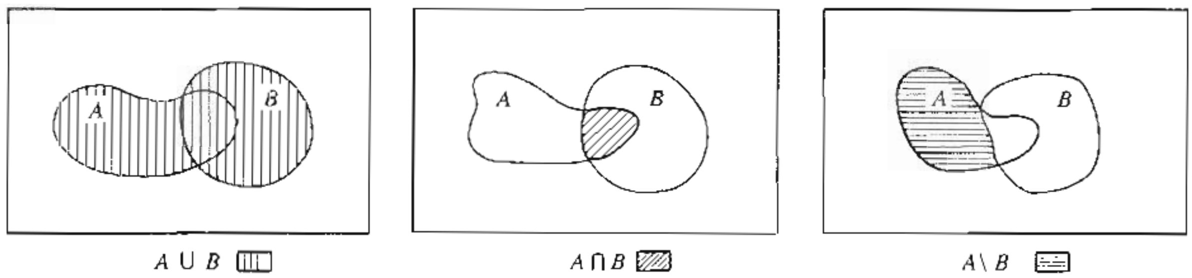


Figure 1.1.1 (a) $A \cup B$ (b) $A \cap B$ (c) $A \setminus B$

The set that has no elements is called the **empty set** and is denoted by the symbol $\emptyset$. Two sets A and B are said to be **disjoint** if they have no elements in common; this can be expressed by writing $A \cap B = \emptyset$.

To illustrate the method of proving set equalities, we will next establish one of the *DeMorgan laws* for three sets. The proof of the other one is left as an exercise.

1.1.4 Theorem *If A, B, C are sets, then*

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

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

Proof. To prove (a), we will show that every element in $A \setminus (B \cup C)$ is contained in both $(A \setminus B)$ and $(A \setminus C)$, and conversely.

If x is in $A \setminus (B \cup C)$, then x is in A, but x is not in $B \cup C$. Hence x is in A, but x is neither in B nor in C. Therefore, x is in A but not B, and x is in A but not C. Thus, $x \in A \setminus B$ and $x \in A \setminus C$, which shows that $x \in (A \setminus B) \cap (A \setminus C)$.

Conversely, if $x \in (A \setminus B) \cap (A \setminus C)$, then $x \in (A \setminus B)$ and $x \in (A \setminus C)$. Hence $x \in A$ and both $x \notin B$ and $x \notin C$. Therefore, $x \in A$ and $x \notin (B \cup C)$, so that $x \in A \setminus (B \cup C)$.

Since the sets $(A \setminus B) \cap (A \setminus C)$ and $A \setminus (B \cup C)$ contain the same elements, they are equal by Definition 1.1.1. Q.E.D.

There are times when it is desirable to form unions and intersections of more than two sets. For a finite collection of sets $\{A_1, A_2, \cdots, A_n\}$, their union is the set A consisting of all elements that belong to *at least one* of the sets A_k, and their intersection consists of all elements that belong to *all* of the sets A_k.

This is extended to an infinite collection of sets $\{A_1, A_2, \cdots, A_n, \cdots\}$ as follows. Their **union** is the set of elements that belong to *at least one* of the sets A_n. In this case we write

$$\bigcup_{n=1}^{\infty} A_n := \{x : x \in A_n \text{ for some } n \in \mathbb{N}\}.$$

Similarly, their **intersection** is the set of elements that belong to *all* of these sets A_n. In this case we write

$$\bigcap_{n=1}^{\infty} A_n := \{x : x \in A_n \text{ for all } n \in \mathbb{N}\}.$$

Cartesian Products

In order to discuss functions, we define the Cartesian product of two sets.

1.1.5 Definition If A and B are nonempty sets, then the **Cartesian product** $A \times B$ of A and B is the set of all ordered pairs (a, b) with $a \in A$ and $b \in B$. That is,

$$A \times B := \{(a, b) : a \in A, b \in B\}.$$

Thus if $A = \{1, 2, 3\}$ and $B = \{1, 5\}$, then the set $A \times B$ is the set whose elements are the ordered pairs

$$(1, 1), \quad (1, 5), \quad (2, 1), \quad (2, 5), \quad (3, 1), \quad (3, 5).$$

We may visualize the set $A \times B$ as the set of six points in the plane with the coordinates that we have just listed.

We often draw a diagram (such as Figure 1.1.2) to indicate the Cartesian product of two sets A and B. However, it should be realized that this diagram may be a simplification. For example, if $A := \{x \in \mathbb{R} : 1 \le x \le 2\}$ and $B := \{y \in \mathbb{R} : 0 \le y \le 1 \text{ or } 2 \le y \le 3\}$, then instead of a rectangle, we should have a drawing such as Figure 1.1.3.

We will now discuss the fundamental notion of a *function* or a *mapping*.

To the mathematician of the early nineteenth century, the word "function" meant a definite formula, such as $f(x) := x^2 + 3x - 5$, which associates to each real number x another number $f(x)$. (Here, $f(0) = -5$, $f(1) = -1$, $f(5) = 35$.) This understanding excluded the case of different formulas on different intervals, so that functions could not be defined "in pieces".

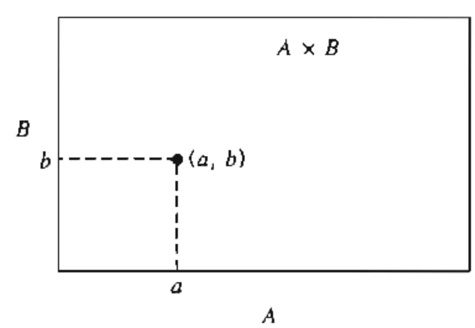


Figure 1.1.2

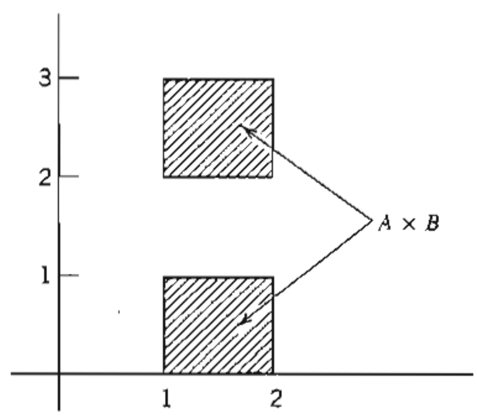


Figure 1.1.3

As mathematics developed, it became clear that a more general definition of "function" would be useful. It also became evident that it is important to make a clear distinction between the function itself and the values of the function. A revised definition might be:

> A function f from a set A into a set B is a rule of correspondence that assigns to each element x in A a uniquely determined element $f(x)$ in B.

But however suggestive this revised definition might be, there is the difficulty of interpreting the phrase "rule of correspondence". In order to clarify this, we will express the definition entirely in terms of sets; in effect, we will define a function to be its **graph**. While this has the disadvantage of being somewhat artificial, it has the advantage of being unambiguous and clearer.

1.1.6 Definition Let A and B be sets. Then a **function** from A to B is a set f of ordered pairs in $A \times B$ such that for each $a \in A$ there exists a unique $b \in B$ with $(a, b) \in f$. (In other words, if $(a, b) \in f$ and $(a, b') \in f$, then $b = b'$.)

The set A of first elements of a function f is called the **domain** of f and is often denoted by $D(f)$. The set of all second elements in f is called the **range** of f and is often denoted by $R(f)$. Note that, although $D(f) = A$, we only have $R(f) \subseteq B$. (See Figure 1.1.4.)

The essential condition that:

$$(a, b) \in f \qquad \text{and} \qquad (a, b') \in f \qquad \text{implies that} \qquad b = b'$$

is sometimes called the *vertical line test*. In geometrical terms it says every vertical line $x = a$ with $a \in A$ intersects the graph of f exactly once.

The notation

$$f : A \to B$$

is often used to indicate that f is a function from A into B. We will also say that f is a **mapping** of A into B, or that f **maps** A into B. If (a, b) is an element in f, it is customary to write

$$b = f(a) \qquad \text{or sometimes} \qquad a \mapsto b.$$

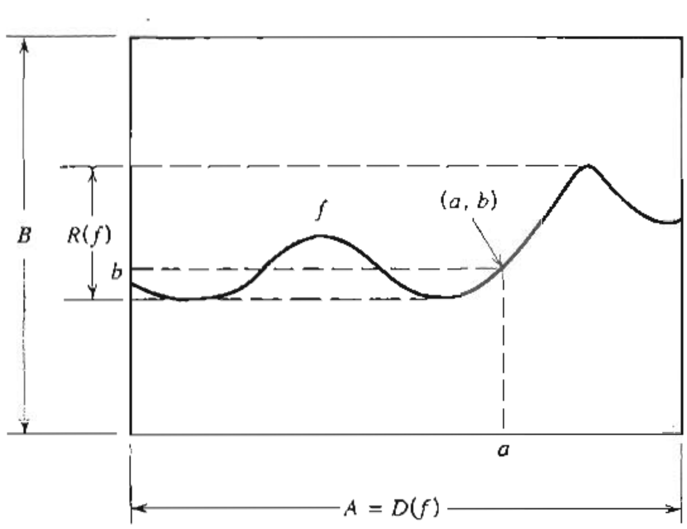


Figure 1.1.4 A function as a graph

If $b = f(a)$, we often refer to b as the **value** of f at a, or as the **image** of a under f.

Transformations and Machines

Aside from using graphs, we can visualize a function as a *transformation* of the set $D(f) = A$ into the set $R(f) \subseteq B$. In this phraseology, when $(a, b) \in f$, we think of f as taking the element a from A and "transforming" or "mapping" it into an element $b = f(a)$ in $R(f) \subseteq B$. We often draw a diagram, such as Figure 1.1.5, even when the sets A and B are not subsets of the plane.

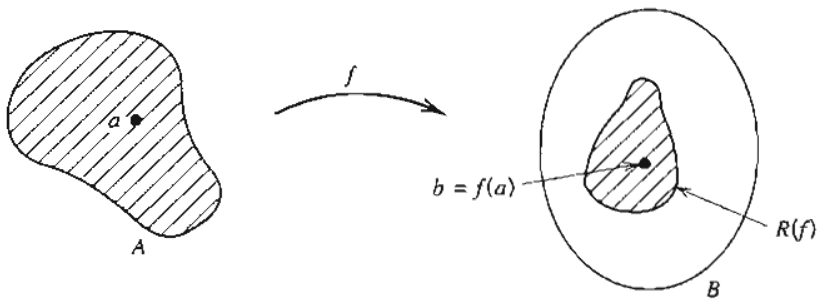


Figure 1.1.5 A function as a transformation

There is another way of visualizing a function: namely, as a *machine* that accepts elements of $D(f) = A$ as *inputs* and produces corresponding elements of $R(f) \subseteq B$ as *outputs*. If we take an element $x \in D(f)$ and put it into f, then out comes the corresponding value $f(x)$. If we put a different element $y \in D(f)$ into f, then out comes $f(y)$ which may or may not differ from $f(x)$. If we try to insert something that does not belong to $D(f)$ into f, we find that it is not accepted, for f can operate only on elements from $D(f)$. (See Figure 1.1.6.)

This last visualization makes clear the distinction between f and $f(x)$: the first is the machine itself, and the second is the output of the machine f when x is the input. Whereas no one is likely to confuse a meat grinder with ground meat, enough people have confused functions with their values that it is worth distinguishing between them notationally.

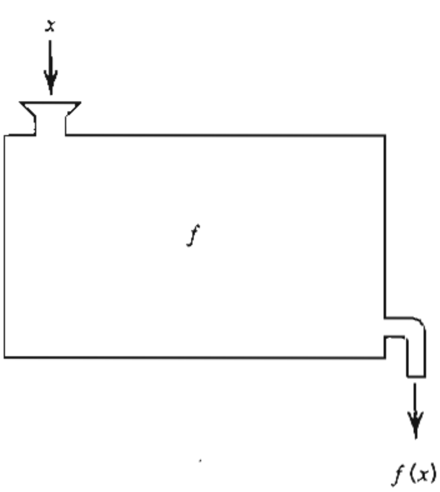


Figure 1.1.6 A function as a machine

Direct and Inverse Images

Let $f : A \to B$ be a function with domain $D(f) = A$ and range $R(f) \subseteq B$.

1.1.7 Definition If E is a subset of A, then the **direct image** of E under f is the subset $f(E)$ of B given by

$$f(E) := \{f(x) : x \in E\}.$$

If H is a subset of B, then the **inverse image** of H under f is the subset $f^{-1}(H)$ of A given by

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

Remark The notation $f^{-1}(H)$ used in this connection has its disadvantages. However, we will use it since it is the standard notation.

Thus, if we are given a set $E \subseteq A$, then a point $y_1 \in B$ is in the direct image $f(E)$ if and only if there exists at least one point $x_1 \in E$ such that $y_1 = f(x_1)$. Similarly, given a set $H \subseteq B$, then a point x_2 is in the inverse image $f^{-1}(H)$ if and only if $y_2 := f(x_2)$ belongs to H. (See Figure 1.1.7.)

1.1.8 Examples (a) Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) := x^2$. Then the direct image of the set $E := \{x : 0 \le x \le 2\}$ is the set $f(E) = \{y : 0 \le y \le 4\}$.

If $G := \{y : 0 \le y \le 4\}$, then the inverse image of G is the set $f^{-1}(G) = \{x : -2 \le x \le 2\}$. Thus, in this case, we see that $f^{-1}(f(E)) \neq E$.

On the other hand, we have $f\left(f^{-1}(G)\right) = G$. But if $H := \{y : -1 \le y \le 1\}$, then we have $f\left(f^{-1}(H)\right) = \{y : 0 \le y \le 1\} \neq H$.

A sketch of the graph of f may help to visualize these sets.

(b) Let $f : A \to B$, and let G, H be subsets of B. We will show that

$$f^{-1}(G \cap H) \subseteq f^{-1}(G) \cap f^{-1}(H).$$

For, if $x \in f^{-1}(G \cap H)$, then $f(x) \in G \cap H$, so that $f(x) \in G$ and $f(x) \in H$. But this implies that $x \in f^{-1}(G)$ and $x \in f^{-1}(H)$, whence $x \in f^{-1}(G) \cap f^{-1}(H)$. Thus the stated implication is proved. [The opposite inclusion is also true, so that we actually have set equality between these sets; see Exercise 13.] □

Further facts about direct and inverse images are given in the exercises.

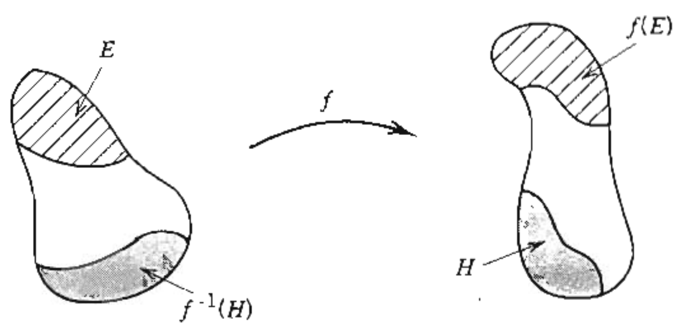


Figure 1.1.7 Direct and inverse images

Special Types of Functions

The following definitions identify some very important types of functions.

1.1.9 Definition Let $f : A \to B$ be a function from A to B.

(a) The function f is said to be **injective** (or to be **one-one**) if whenever $x_1 \neq x_2$, then $f(x_1) \neq f(x_2)$. If f is an injective function, we also say that f is an **injection**.

(b) The function f is said to be **surjective** (or to map A **onto** B) if $f(A) = B$; that is, if the range $R(f) = B$. If f is a surjective function, we also say that f is a **surjection**.

(c) If f is both injective and surjective, then f is said to be **bijective**. If f is bijective, we also say that f is a **bijection**.

- In order to prove that a function f is injective, we must establish that:

  $$\text{for all } x_1, x_2 \text{ in } A, \text{ if } f(x_1) = f(x_2), \text{ then } x_1 = x_2.$$

  To do this we assume that $f(x_1) = f(x_2)$ and show that $x_1 = x_2$.
  [In other words, the graph of f satisfies the *first horizontal line test:* Every horizontal line $y = b$ with $b \in B$ intersects the graph f in *at most* one point.]
- To prove that a function f is surjective, we must show that for any $b \in B$ there exists at least one $x \in A$ such that $f(x) = b$.
  [In other words, the graph of f satisfies the *second horizontal line test:* Every horizontal line $y = b$ with $b \in B$ intersects the graph f in *at least* one point.]

1.1.10 Example Let $A := \{x \in \mathbb{R} : x \neq 1\}$ and define $f(x) := 2x/(x-1)$ for all $x \in A$. To show that f is injective, we take x_1 and x_2 in A and assume that $f(x_1) = f(x_2)$. Thus we have

$$\frac{2x_1}{x_1 - 1} = \frac{2x_2}{x_2 - 1},$$

which implies that $x_1(x_2 - 1) = x_2(x_1 - 1)$, and hence $x_1 = x_2$. Therefore f is injective.

To determine the range of f, we solve the equation $y = 2x/(x-1)$ for x in terms of y. We obtain $x = y/(y-2)$, which is meaningful for $y \neq 2$. Thus the range of f is the set $B := \{y \in \mathbb{R} : y \neq 2\}$. Thus, f is a bijection of A onto B. □

Inverse Functions

If f is a function from A into B, then f is a special subset of $A \times B$ (namely, one passing the *vertical line test.*) The set of ordered pairs in $B \times A$ obtained by interchanging the members of ordered pairs in f is not generally a function. (That is, the set f may not pass *both* of the *horizontal line tests.*) However, if f is a bijection, then this interchange does lead to a function, called the "inverse function" of f.

1.1.11 Definition If $f : A \to B$ is a bijection of A onto B, then

$$g := \{(b, a) \in B \times A : (a, b) \in f\}$$

is a function on B into A. This function is called the **inverse function** of f, and is denoted by f^{-1}. The function f^{-1} is also called the **inverse** of f.

We can also express the connection between f and its inverse f^{-1} by noting that $D(f) = R(f^{-1})$ and $R(f) = D(f^{-1})$ and that

$$b = f(a) \qquad \text{if and only if} \qquad a = f^{-1}(b).$$

For example, we saw in Example 1.1.10 that the function

$$f(x) := \frac{2x}{x-1}$$

is a bijection of $A := \{x \in \mathbb{R} : x \neq 1\}$ onto the set $B := \{y \in \mathbb{R} : y \neq 2\}$. The function inverse to f is given by

$$f^{-1}(y) := \frac{y}{y-2} \qquad \text{for} \quad y \in B.$$

Remark We introduced the notation $f^{-1}(H)$ in Definition 1.1.7. It makes sense even if f does not have an inverse function. However, if the inverse function f^{-1} does exist, then $f^{-1}(H)$ is the direct image of the set $H \subseteq B$ under f^{-1}.

Composition of Functions

It often happens that we want to "compose" two functions f, g by first finding $f(x)$ and then applying g to get $g(f(x))$; however, this is possible only when $f(x)$ belongs to the domain of g. In order to be able to do this for *all* $f(x)$, we must assume that the range of f is contained in the domain of g. (See Figure 1.1.8.)

1.1.12 Definition If $f : A \to B$ and $g : B \to C$, and if $R(f) \subseteq D(g) = B$, then the **composite function** $g \circ f$ (note the order!) is the function from A into C defined by

$$(g \circ f)(x) := g(f(x)) \qquad \text{for all} \quad x \in A.$$

1.1.13 Examples (a) The order of the composition must be carefully noted. For, let f and g be the functions whose values at $x \in \mathbb{R}$ are given by

$$f(x) := 2x \qquad \text{and} \qquad g(x) := 3x^2 - 1.$$

Since $D(g) = \mathbb{R}$ and $R(f) \subseteq \mathbb{R} = D(g)$, then the domain $D(g \circ f)$ is also equal to $\mathbb{R}$, and the composite function $g \circ f$ is given by

$$(g \circ f)(x) = 3(2x)^2 - 1 = 12x^2 - 1.$$

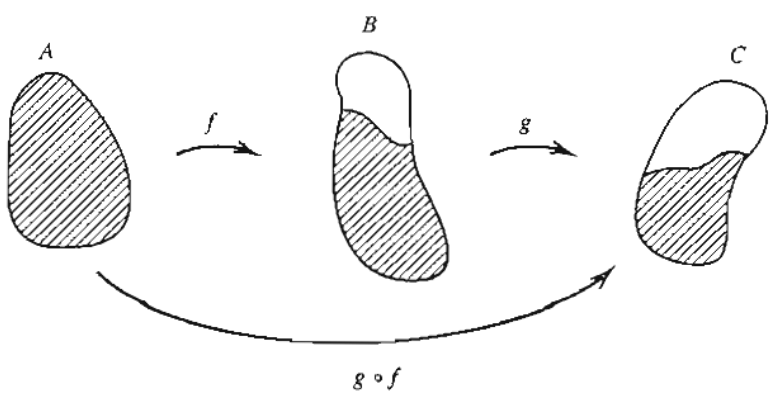


Figure 1.1.8 The composition of f and g

On the other hand, the domain of the composite function $f \circ g$ is also $\mathbb{R}$, but

$$(f \circ g)(x) = 2(3x^2 - 1) = 6x^2 - 2.$$

Thus, in this case, we have $g \circ f \neq f \circ g$.

(b) In considering $g \circ f$, some care must be exercised to be sure that the range of f is contained in the domain of g. For example, if

$$f(x) := 1 - x^2 \qquad \text{and} \qquad g(x) := \sqrt{x},$$

then, since $D(g) = \{x : x \geq 0\}$, the composite function $g \circ f$ is given by the formula

$$(g \circ f)(x) = \sqrt{1 - x^2}$$

only for $x \in D(f)$ that satisfy $f(x) \geq 0$; that is, for x satisfying $-1 \leq x \leq 1$.

We note that if we reverse the order, then the composition $f \circ g$ is given by the formula

$$(f \circ g)(x) = 1 - x,$$

but only for those x in the domain $D(g) = \{x : x \geq 0\}$. □

We now give the relationship between composite functions and inverse images. The proof is left as an instructive exercise.

1.1.14 Theorem *Let $f : A \to B$ and $g : B \to C$ be functions and let H be a subset of C. Then we have*

$$(g \circ f)^{-1}(H) = f^{-1}(g^{-1}(H)).$$

Note the *reversal* in the order of the functions.

Restrictions of Functions

If $f : A \to B$ is a function and if $A_1 \subset A$, we can define a function $f_1 : A_1 \to B$ by

$$f_1(x) := f(x) \qquad \text{for} \quad x \in A_1.$$

The function f_1 is called the **restriction of f to A_1**. Sometimes it is denoted by $f_1 = f|A_1$.

It may seem strange to the reader that one would ever choose to throw away a part of a function, but there are some good reasons for doing so. For example, if $f : \mathbb{R} \to \mathbb{R}$ is the **squaring function**:

$$f(x) := x^2 \qquad \text{for} \quad x \in \mathbb{R},$$

then f is not injective, so it cannot have an inverse function. However, if we restrict f to the set $A_1 := \{x : x \geq 0\}$, then the restriction $f|A_1$ is a bijection of A_1 onto A_1. Therefore, this restriction has an inverse function, which is the **positive square root function**. (Sketch a graph.)

Similarly, the trigonometric functions $S(x) := \sin x$ and $C(x) := \cos x$ are not injective on all of $\mathbb{R}$. However, by making suitable restrictions of these functions, one can obtain the **inverse sine** and the **inverse cosine** functions that the reader has undoubtedly already encountered.

Exercises for Section 1.1

1. If A and B are sets, show that $A \subseteq B$ if and only if $A \cap B = A$.
2. Prove the second De Morgan Law [Theorem 1.1.4(b)].
3. Prove the Distributive Laws:
   (a) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$,
   (b) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.
4. The **symmetric difference** of two sets A and B is the set D of all elements that belong to either A or B but not both. Represent D with a diagram.
   (a) Show that $D = (A \backslash B) \cup (B \backslash A)$.
   (b) Show that D is also given by $D = (A \cup B) \backslash (A \cap B)$.
5. For each $n \in \mathbb{N}$, let $A_n = \{(n+1)k : k \in \mathbb{N}\}$.
   (a) What is $A_1 \cap A_2$?
   (b) Determine the sets $\bigcup\{A_n : n \in \mathbb{N}\}$ and $\bigcap\{A_n : n \in \mathbb{N}\}$.
6. Draw diagrams in the plane of the Cartesian products $A \times B$ for the given sets A and B.
   (a) $A = \{x \in \mathbb{R} : 1 \le x \le 2 \text{ or } 3 \le x \le 4\}$, $B = \{x \in \mathbb{R} : x = 1 \text{ or } x = 2\}$.
   (b) $A = \{1, 2, 3\}$, $B = \{x \in \mathbb{R} : 1 \le x \le 3\}$.
7. Let $A := B := \{x \in \mathbb{R} : -1 \le x \le 1\}$ and consider the subset $C := \{(x, y) : x^2 + y^2 = 1\}$ of $A \times B$. Is this set a function? Explain.
8. Let $f(x) := 1/x^2$, $x \neq 0$, $x \in \mathbb{R}$.
   (a) Determine the direct image $f(E)$ where $E := \{x \in \mathbb{R} : 1 \le x \le 2\}$.
   (b) Determine the inverse image $f^{-1}(G)$ where $G := \{x \in \mathbb{R} : 1 \le x \le 4\}$.
9. Let $g(x) := x^2$ and $f(x) := x + 2$ for $x \in \mathbb{R}$, and let h be the composite function $h := g \circ f$.
   (a) Find the direct image $h(E)$ of $E := \{x \in \mathbb{R} : 0 \le x \le 1\}$.
   (b) Find the inverse image $h^{-1}(G)$ of $G := \{x \in \mathbb{R} : 0 \le x \le 4\}$.
10. Let $f(x) := x^2$ for $x \in \mathbb{R}$, and let $E := \{x \in \mathbb{R} : -1 \le x \le 0\}$ and $F := \{x \in \mathbb{R} : 0 \le x \le 1\}$. Show that $E \cap F = \{0\}$ and $f(E \cap F) = \{0\}$, while $f(E) = f(F) = \{y \in \mathbb{R} : 0 \le y \le 1\}$. Hence $f(E \cap F)$ is a proper subset of $f(E) \cap f(F)$. What happens if 0 is deleted from the sets E and F?
11. Let f and E, F be as in Exercise 10. Find the sets $E \backslash F$ and $f(E) \backslash f(F)$ and show that it is *not* true that $f(E \backslash F) \subseteq f(E) \backslash f(F)$.
12. Show that if $f : A \to B$ and E, F are subsets of A, then $f(E \cup F) = f(E) \cup f(F)$ and $f(E \cap F) \subseteq f(E) \cap f(F)$.
13. Show that if $f : A \to B$ and G, H are subsets of B, then $f^{-1}(G \cup H) = f^{-1}(G) \cup f^{-1}(H)$ and $f^{-1}(G \cap H) = f^{-1}(G) \cap f^{-1}(H)$.
14. Show that the function f defined by $f(x) := x/\sqrt{x^2+1}$, $x \in \mathbb{R}$, is a bijection of $\mathbb{R}$ onto $\{y : -1 < y < 1\}$.
15. For $a, b \in \mathbb{R}$ with $a < b$, find an explicit bijection of $A := \{x : a < x < b\}$ onto $B := \{y : 0 < y < 1\}$.
16. Give an example of two functions f, g on $\mathbb{R}$ to $\mathbb{R}$ such that $f \neq g$, but such that $f \circ g = g \circ f$.
17. (a) Show that if $f : A \to B$ is injective and $E \subseteq A$, then $f^{-1}(f(E)) = E$. Give an example to show that equality need not hold if f is not injective.
    (b) Show that if $f : A \to B$ is surjective and $H \subseteq B$, then $f(f^{-1}(H)) = H$. Give an example to show that equality need not hold if f is not surjective.
18. (a) Suppose that f is an injection. Show that $f^{-1} \circ f(x) = x$ for all $x \in D(f)$ and that $f \circ f^{-1}(y) = y$ for all $y \in R(f)$.
    (b) If f is a bijection of A onto B, show that f^{-1} is a bijection of B onto A.

19. Prove that if $f : A \to B$ is bijective and $g : B \to C$ is bijective, then the composite $g \circ f$ is a bijective map of A onto C.
20. Let $f : A \to B$ and $g : B \to C$ be functions.
    (a) Show that if $g \circ f$ is injective, then f is injective.
    (b) Show that if $g \circ f$ is surjective, then g is surjective.
21. Prove Theorem 1.1.14.
22. Let f, g be functions such that $(g \circ f)(x) = x$ for all $x \in D(f)$ and $(f \circ g)(y) = y$ for all $y \in D(g)$. Prove that $g = f^{-1}$.

Section 1.2 Mathematical Induction

Mathematical Induction is a powerful method of proof that is frequently used to establish the validity of statements that are given in terms of the natural numbers. Although its utility is restricted to this rather special context, Mathematical Induction is an indispensable tool in all branches of mathematics. Since many induction proofs follow the same formal lines of argument, we will often state only that a result follows from Mathematical Induction and leave it to the reader to provide the necessary details. In this section, we will state the principle and give several examples to illustrate how inductive proofs proceed.

We shall assume familiarity with the set of natural numbers:

$$\mathbb{N} := \{1, 2, 3, \cdots\},$$

with the usual arithmetic operations of addition and multiplication, and with the meaning of a natural number being less than another one. We will also assume the following fundamental property of $\mathbb{N}$.

1.2.1 Well-Ordering Property of $\mathbb{N}$ *Every nonempty subset of $\mathbb{N}$ has a least element.*

A more detailed statement of this property is as follows: If S is a subset of $\mathbb{N}$ and if $S \neq \emptyset$, then there exists $m \in S$ such that $m \leq k$ for all $k \in S$.

On the basis of the Well-Ordering Property, we shall derive a version of the Principle of Mathematical Induction that is expressed in terms of subsets of $\mathbb{N}$.

1.2.2 Principle of Mathematical Induction *Let S be a subset of $\mathbb{N}$ that possesses the two properties:*

(1) The number $1 \in S$.

(2) For every $k \in \mathbb{N}$, if $k \in S$, then $k + 1 \in S$.

Then we have $S = \mathbb{N}$.

Proof. Suppose to the contrary that $S \neq \mathbb{N}$. Then the set $\mathbb{N} \backslash S$ is not empty, so by the Well-Ordering Principle it has a least element m. Since $1 \in S$ by hypothesis (1), we know that $m > 1$. But this implies that $m - 1$ is also a natural number. Since $m - 1 < m$ and since m is the least element in $\mathbb{N}$ such that $m \notin S$, we conclude that $m - 1 \in S$.

We now apply hypothesis (2) to the element $k := m - 1$ in S, to infer that $k + 1 = (m - 1) + 1 = m$ belongs to S. But this statement contradicts the fact that $m \notin S$. Since m was obtained from the assumption that $\mathbb{N} \backslash S$ is not empty, we have obtained a contradiction. Therefore we must have $S = \mathbb{N}$. Q.E.D.

The Principle of Mathematical Induction is often set forth in the framework of properties or statements about natural numbers. If $P(n)$ is a meaningful statement about $n \in \mathbb{N}$, then $P(n)$ may be true for some values of n and false for others. For example, if $P_1(n)$ is the statement: "$n^2 = n$", then $P_1(1)$ is true while $P_1(n)$ is false for all $n > 1, n \in \mathbb{N}$. On the other hand, if $P_2(n)$ is the statement: "$n^2 > 1$", then $P_2(1)$ is false, while $P_2(n)$ is true for all $n > 1, n \in \mathbb{N}$.

In this context, the Principle of Mathematical Induction can be formulated as follows.

For each $n \in \mathbb{N}$, let $P(n)$ be a statement about n. Suppose that:

(1′) *$P(1)$ is true.*
(2′) *For every $k \in \mathbb{N}$, if $P(k)$ is true, then $P(k+1)$ is true.*

Then $P(n)$ is true for all $n \in \mathbb{N}$.

The connection with the preceding version of Mathematical Induction, given in 1.2.2, is made by letting $S := \{n \in \mathbb{N} : P(n) \text{ is true}\}$. Then the conditions (1) and (2) of 1.2.2 correspond exactly to the conditions (1′) and (2′), respectively. The conclusion that $S = \mathbb{N}$ in 1.2.2 corresponds to the conclusion that $P(n)$ is true for all $n \in \mathbb{N}$.

In (2′) the assumption "if $P(k)$ is true" is called the **induction hypothesis**. In establishing (2′), we are not concerned with the actual truth or falsity of $P(k)$, but only with the validity of the implication "if $P(k)$, then $P(k+1)$". For example, if we consider the statements $P(n)$: "$n = n + 5$", then (2′) is logically correct, for we can simply add 1 to both sides of $P(k)$ to obtain $P(k+1)$. However, since the statement $P(1)$: "$1 = 6$" is false, we cannot use Mathematical Induction to conclude that $n = n + 5$ for all $n \in \mathbb{N}$.

It may happen that statements $P(n)$ are false for certain natural numbers but then are true for all $n \geq n_0$ for some particular n_0. The Principle of Mathematical Induction can be modified to deal with this situation. We will formulate the modified principle, but leave its verification as an exercise. (See Exercise 12.)

1.2.3 Principle of Mathematical Induction (second version) *Let $n_0 \in \mathbb{N}$ and let $P(n)$ be a statement for each natural number $n \geq n_0$. Suppose that:*

(1) *The statement $P(n_0)$ is true.*
(2) *For all $k \geq n_0$, the truth of $P(k)$ implies the truth of $P(k+1)$.*

Then $P(n)$ is true for all $n \geq n_0$.

Sometimes the number n_0 in (1) is called the **base**, since it serves as the starting point, and the implication in (2), which can be written $P(k) \Rightarrow P(k+1)$, is called the **bridge**, since it connects the case k to the case $k+1$.

The following examples illustrate how Mathematical Induction is used to prove assertions about natural numbers.

1.2.4 Examples (a) For each $n \in \mathbb{N}$, the sum of the first n natural numbers is given by

$$1 + 2 + \cdots + n = \tfrac{1}{2}n(n+1).$$

To prove this formula, we let S be the set of all $n \in \mathbb{N}$ for which the formula is true. We must verify that conditions (1) and (2) of 1.2.2 are satisfied. If $n = 1$, then we have $1 = \frac{1}{2} \cdot 1 \cdot (1+1)$ so that $1 \in S$, and (1) is satisfied. Next, we *assume* that $k \in S$ and wish to infer from this assumption that $k + 1 \in S$. Indeed, if $k \in S$, then

$$1 + 2 + \cdots + k = \tfrac{1}{2}k(k+1).$$

If we add $k + 1$ to both sides of the assumed equality, we obtain

$$\begin{aligned} 1 + 2 + \cdots + k + (k + 1) &= \tfrac{1}{2}k(k + 1) + (k + 1) \\ &= \tfrac{1}{2}(k + 1)(k + 2). \end{aligned}$$

Since this is the stated formula for $n = k + 1$, we conclude that $k + 1 \in S$. Therefore, condition (2) of 1.2.2 is satisfied. Consequently, by the Principle of Mathematical Induction, we infer that $S = \mathbb{N}$, so the formula holds for all $n \in \mathbb{N}$.

(b) For each $n \in \mathbb{N}$, the sum of the squares of the first n natural numbers is given by

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

To establish this formula, we note that it is true for $n = 1$, since $1^2 = \frac{1}{6} \cdot 1 \cdot 2 \cdot 3$. If we assume it is true for k, then adding $(k + 1)^2$ to both sides of the assumed formula gives

$$\begin{aligned} 1^2 + 2^2 + \cdots + k^2 + (k + 1)^2 &= \tfrac{1}{6}k(k + 1)(2k + 1) + (k + 1)^2 \\ &= \tfrac{1}{6}(k + 1)(2k^2 + k + 6k + 6) \\ &= \tfrac{1}{6}(k + 1)(k + 2)(2k + 3). \end{aligned}$$

Consequently, the formula is valid for all $n \in \mathbb{N}$.

(c) Given two real numbers a and b, we will prove that $a - b$ is a factor of $a^n - b^n$ for all $n \in \mathbb{N}$.

First we see that the statement is clearly true for $n = 1$. If we now assume that $a - b$ is a factor of $a^k - b^k$, then

$$\begin{aligned} a^{k+1} - b^{k+1} &= a^{k+1} - ab^k + ab^k - b^{k+1} \\ &= a(a^k - b^k) + b^k(a - b). \end{aligned}$$

By the induction hypothesis, $a - b$ is a factor of $a(a^k - b^k)$ and it is plainly a factor of $b^k(a - b)$. Therefore, $a - b$ is a factor of $a^{k+1} - b^{k+1}$, and it follows from Mathematical Induction that $a - b$ is a factor of $a^n - b^n$ for all $n \in \mathbb{N}$.

A variety of divisibility results can be derived from this fact. For example, since $11 - 7 = 4$, we see that $11^n - 7^n$ is divisible by 4 for all $n \in \mathbb{N}$.

(d) The inequality $2^n > 2n + 1$ is false for $n = 1, 2$, but it is true for $n = 3$. If we assume that $2^k > 2k + 1$, then multiplication by 2 gives, when $2k + 2 > 3$, the inequality

$$2^{k+1} > 2(2k + 1) = 4k + 2 = 2k + (2k + 2) > 2k + 3 = 2(k + 1) + 1.$$

Since $2k + 2 > 3$ for all $k \geq 1$, the bridge is valid for all $k \geq 1$ (even though the statement is false for $k = 1, 2$). Hence, with the base $n_0 = 3$, we can apply Mathematical Induction to conclude that the inequality holds for all $n \geq 3$.

(e) The inequality $2^n \leq (n + 1)!$ can be established by Mathematical Induction.

We first observe that it is true for $n = 1$, since $2^1 = 2 = 1 + 1$. If we assume that $2^k \leq (k + 1)!$, it follows from the fact that $2 \leq k + 2$ that

$$2^{k+1} = 2 \cdot 2^k \leq 2(k + 1)! \leq (k + 2)(k + 1)! = (k + 2)!.$$

Thus, if the inequality holds for k, then it also holds for $k + 1$. Therefore, Mathematical Induction implies that the inequality is true for all $n \in \mathbb{N}$.

(f) If $r \in \mathbb{R}$, $r \neq 1$, and $n \in \mathbb{N}$, then

$$1 + r + r^2 + \cdots + r^n = \frac{1 - r^{n+1}}{1 - r}.$$

This is the formula for the sum of the terms in a "geometric progression". It can be established using Mathematical Induction as follows. First, if $n = 1$, then $1 + r = (1 - r^2)/(1 - r)$. If we assume the truth of the formula for $n = k$ and add the term r^{k+1} to both sides, we get (after a little algebra)

$$1 + r + r^k + \cdots + r^{k+1} = \frac{1 - r^{k+1}}{1 - r} + r^{k+1} = \frac{1 - r^{k+2}}{1 - r},$$

which is the formula for $n = k + 1$. Therefore, Mathematical Induction implies the validity of the formula for all $n \in \mathbb{N}$.

[This result can also be proved without using Mathematical Induction. If we let $s_n := 1 + r + r^2 + \cdots + r^n$, then $rs_n = r + r^2 + \cdots + r^{n+1}$, so that

$$(1 - r)s_n = s_n - rs_n = 1 - r^{n+1}.$$

If we divide by $1 - r$, we obtain the stated formula.]

(g) Careless use of the Principle of Mathematical Induction can lead to obviously absurd conclusions. The reader is invited to find the error in the "proof" of the following assertion.

Claim: If $n \in \mathbb{N}$ and if the maximum of the natural numbers p and q is n, then $p = q$.

"Proof." Let S be the subset of $\mathbb{N}$ for which the claim is true. Evidently, $1 \in S$ since if $p, q \in \mathbb{N}$ and their maximum is 1, then both equal 1 and so $p = q$. Now assume that $k \in S$ and that the maximum of p and q is $k + 1$. Then the maximum of $p - 1$ and $q - 1$ is k. But since $k \in S$, then $p - 1 = q - 1$ and therefore $p = q$. Thus, $k + 1 \in S$, and we conclude that the assertion is true for all $n \in \mathbb{N}$.

(h) There are statements that are true for *many* natural numbers but that are not true for *all* of them.

For example, the formula $p(n) := n^2 - n + 41$ gives a prime number for $n = 1, 2, \cdots, 40$. However, $p(41)$ is obviously divisible by 41, so it is not a prime number. □

Another version of the Principle of Mathematical Induction is sometimes quite useful. It is called the "Principle of Strong Induction", even though it is in fact equivalent to 1.2.2.

1.2.5 Principle of Strong Induction *Let S be a subset of $\mathbb{N}$ such that*

(1″) $1 \in S$.

(2″) *For every $k \in \mathbb{N}$, if $\{1, 2, \cdots, k\} \subseteq S$, then $k + 1 \in S$.*

Then $S = \mathbb{N}$.

We will leave it to the reader to establish the equivalence of 1.2.2 and 1.2.5.

Exercises for Section 1.2

1. Prove that $1/1 \cdot 2 + 1/2 \cdot 3 + \cdots + 1/n(n + 1) = n/(n + 1)$ for all $n \in \mathbb{N}$.
2. Prove that $1^3 + 2^3 + \cdots + n^3 = \left[\frac{1}{2}n(n + 1)\right]^2$ for all $n \in \mathbb{N}$.
3. Prove that $3 + 11 + \cdots + (8n - 5) = 4n^2 - n$ for all $n \in \mathbb{N}$.
4. Prove that $1^2 + 3^2 + \cdots + (2n - 1)^2 = (4n^3 - n)/3$ for all $n \in \mathbb{N}$.
5. Prove that $1^2 - 2^2 + 3^2 + \cdots + (-1)^{n+1}n^2 = (-1)^{n+1}n(n + 1)/2$ for all $n \in \mathbb{N}$.

6. Prove that $n^3 + 5n$ is divisible by 6 for all $n \in \mathbb{N}$.
7. Prove that $5^{2n} - 1$ is divisible by 8 for all $n \in \mathbb{N}$.
8. Prove that $5^n - 4n - 1$ is divisible by 16 for all $n \in \mathbb{N}$.
9. Prove that $n^3 + (n+1)^3 + (n+2)^3$ is divisible by 9 for all $n \in \mathbb{N}$.
10. Conjecture a formula for the sum $1/1 \cdot 3 + 1/3 \cdot 5 + \cdots + 1/(2n-1)(2n+1)$, and prove your conjecture by using Mathematical Induction.
11. Conjecture a formula for the sum of the first n odd natural numbers $1 + 3 + \cdots + (2n-1)$, and prove your formula by using Mathematical Induction.
12. Prove the Principle of Mathematical Induction 1.2.3 (second version).
13. Prove that $n < 2^n$ for all $n \in \mathbb{N}$.
14. Prove that $2^n < n!$ for all $n \geq 4$, $n \in \mathbb{N}$.
15. Prove that $2n - 3 \leq 2^{n-2}$ for all $n \geq 5$, $n \in \mathbb{N}$.
16. Find all natural numbers n such that $n^2 < 2^n$. Prove your assertion.
17. Find the largest natural number m such that $n^3 - n$ is divisible by m for all $n \in \mathbb{N}$. Prove your assertion.
18. Prove that $1/\sqrt{1} + 1/\sqrt{2} + \cdots + 1/\sqrt{n} > \sqrt{n}$ for all $n \in \mathbb{N}$.
19. Let S be a subset of $\mathbb{N}$ such that (a) $2^k \in S$ for all $k \in \mathbb{N}$, and (b) if $k \in S$ and $k \geq 2$, then $k - 1 \in S$. Prove that $S = \mathbb{N}$.
20. Let the numbers x_n be defined as follows: $x_1 := 1$, $x_2 := 2$, and $x_{n+2} := \frac{1}{2}(x_{n+1} + x_n)$ for all $n \in \mathbb{N}$. Use the Principle of Strong Induction (1.2.5) to show that $1 \leq x_n \leq 2$ for all $n \in \mathbb{N}$.

Section 1.3 Finite and Infinite Sets

When we count the elements in a set, we say "one, two, three, $\cdots$", stopping when we have exhausted the set. From a mathematical perspective, what we are doing is defining a bijective mapping between the set and a portion of the set of natural numbers. If the set is such that the counting does not terminate, such as the set of natural numbers itself, then we describe the set as being infinite.

The notions of "finite" and "infinite" are extremely primitive, and it is very likely that the reader has never examined these notions very carefully. In this section we will define these terms precisely and establish a few basic results and state some other important results that seem obvious but whose proofs are a bit tricky. These proofs can be found in Appendix B and can be read later.

1.3.1 Definition (a) The empty set $\emptyset$ is said to have 0 **elements**.

(b) If $n \in \mathbb{N}$, a set S is said to have n **elements** if there exists a bijection from the set $\mathbb{N}_n := \{1, 2, \cdots, n\}$ onto S.

(c) A set S is said to be **finite** if it is either empty or it has n elements for some $n \in \mathbb{N}$.

(d) A set S is said to be **infinite** if it is not finite.

Since the inverse of a bijection is a bijection, it is easy to see that a set S has n elements if and only if there is a bijection from S onto the set $\{1, 2, \cdots, n\}$. Also, since the composition of two bijections is a bijection, we see that a set S_1 has n elements if and only if there is a bijection from S_1 onto another set S_2 that has n elements. Further, a set T_1 is finite if and only if there is a bijection from T_1 onto another set T_2 that is finite.

It is now necessary to establish some basic properties of finite sets to be sure that the definitions do not lead to conclusions that conflict with our experience of counting. From the definitions, it is not entirely clear that a finite set might not have n elements for *more than one* value of n. Also it is conceivably possible that the set $\mathbb{N} := \{1, 2, 3, \cdots\}$ might be a finite set according to this definition. The reader will be relieved that these possibilities do not occur, as the next two theorems state. The proofs of these assertions, which use the fundamental properties of $\mathbb{N}$ described in Section 1.2, are given in Appendix B.

1.3.2 Uniqueness Theorem *If S is a finite set, then the number of elements in S is a unique number in $\mathbb{N}$.*

1.3.3 Theorem *The set $\mathbb{N}$ of natural numbers is an infinite set.*

The next result gives some elementary properties of finite and infinite sets.

1.3.4 Theorem **(a)** *If A is a set with m elements and B is a set with n elements and if $A \cap B = \emptyset$, then $A \cup B$ has $m + n$ elements.*

(b) *If A is a set with $m \in \mathbb{N}$ elements and $C \subseteq A$ is a set with 1 element, then $A \backslash C$ is a set with $m - 1$ elements.*

(c) *If C is an infinite set and B is a finite set, then $C \backslash B$ is an infinite set.*

Proof. (a) Let f be a bijection of $\mathbb{N}_m$ onto A, and let g be a bijection of $\mathbb{N}_n$ onto B. We define h on $\mathbb{N}_{m+n}$ by $h(i) := f(i)$ for $i = 1, \cdots, m$ and $h(i) := g(i - m)$ for $i = m + 1, \cdots, m + n$. We leave it as an exercise to show that h is a bijection from $\mathbb{N}_{m+n}$ onto $A \cup B$.

The proofs of parts (b) and (c) are left to the reader, see Exercise 2. Q.E.D.

It may seem "obvious" that a subset of a finite set is also finite, but the assertion must be deduced from the definitions. This and the corresponding statement for infinite sets are established next.

1.3.5 Theorem *Suppose that S and T are sets and that $T \subseteq S$.*

(a) *If S is a finite set, then T is a finite set.*

(b) *If T is an infinite set, then S is an infinite set.*

Proof. (a) If $T = \emptyset$, we already know that T is a finite set. Thus we may suppose that $T \neq \emptyset$. The proof is by induction on the number of elements in S.

If S has 1 element, then the only nonempty subset T of S must coincide with S, so T is a finite set.

Suppose that every nonempty subset of a set with k elements is finite. Now let S be a set having $k + 1$ elements (so there exists a bijection f of $\mathbb{N}_{k+1}$ onto S), and let $T \subseteq S$. If $f(k+1) \notin T$, we can consider T to be a subset of $S_1 := S \backslash \{f(k+1)\}$, which has k elements by Theorem 1.3.4(b). Hence, by the induction hypothesis, T is a finite set.

On the other hand, if $f(k+1) \in T$, then $T_1 := T \backslash \{f(k+1)\}$ is a subset of S_1. Since S_1 has k elements, the induction hypothesis implies that T_1 is a finite set. But this implies that $T = T_1 \cup \{f(k+1)\}$ is also a finite set.

(b) This assertion is the contrapositive of the assertion in (a). (See Appendix A for a discussion of the contrapositive.) Q.E.D.

Countable Sets

We now introduce an important type of infinite set.

1.3.6 Definition **(a)** A set S is said to be **denumerable** (or **countably infinite**) if there exists a bijection of $\mathbb{N}$ onto S.

(b) A set S is said to be **countable** if it is either finite or denumerable.

(c) A set S is said to be **uncountable** if it is not countable.

From the properties of bijections, it is clear that S is denumerable if and only if there exists a bijection of S onto $\mathbb{N}$. Also a set S_1 is denumerable if and only if there exists a bijection from S_1 onto a set S_2 that is denumerable. Further, a set T_1 is countable if and only if there exists a bijection from T_1 onto a set T_2 that is countable. Finally, an infinite countable set is denumerable.

1.3.7 Examples **(a)** The set $E := \{2n : n \in \mathbb{N}\}$ of *even* natural numbers is denumerable, since the mapping $f : \mathbb{N} \to E$ defined by $f(n) := 2n$ for $n \in \mathbb{N}$, is a bijection of $\mathbb{N}$ onto E.

Similarly, the set $O := \{2n - 1 : n \in \mathbb{N}\}$ of *odd* natural numbers is denumerable.

(b) The set $\mathbb{Z}$ of *all* integers is denumerable.

To construct a bijection of $\mathbb{N}$ onto $\mathbb{Z}$, we map 1 onto 0, we map the set of even natural numbers onto the set $\mathbb{N}$ of positive integers, and we map the set of odd natural numbers onto the negative integers. This mapping can be displayed by the enumeration:

$$\mathbb{Z} = \{0, 1, -1, 2, -2, 3, -3, \cdots\}.$$

(c) The union of two disjoint denumerable sets is denumerable.

Indeed, if $A = \{a_1, a_2, a_3, \cdots\}$ and $B = \{b_1, b_2, b_3, \cdots\}$, we can enumerate the elements of $A \cup B$ as:

$$a_1, b_1, a_2, b_2, a_3, b_3, \cdots. \qquad \square$$

1.3.8 Theorem *The set* $\mathbb{N} \times \mathbb{N}$ *is denumerable.*

Informal Proof. Recall that $\mathbb{N} \times \mathbb{N}$ consists of all ordered pairs (m, n), where $m, n \in \mathbb{N}$. We can enumerate these pairs as:

$$(1, 1), \quad (1, 2), \quad (2, 1), \quad (1, 3), \quad (2, 2), \quad (3, 1), \quad (1, 4), \cdots,$$

according to increasing sum $m + n$, and increasing m. (See Figure 1.3.1.) Q.E.D.

The enumeration just described is an instance of a "diagonal procedure", since we move along diagonals that each contain finitely many terms as illustrated in Figure 1.3.1. While this argument is satisfying in that it shows exactly what the bijection of $\mathbb{N} \times \mathbb{N} \to \mathbb{N}$ should do, it is not a "formal proof", since it doesn't define this bijection precisely. (See Appendix B for a more formal proof.)

As we have remarked, the construction of an explicit bijection between sets is often complicated. The next two results are useful in establishing the countability of sets, since they do not involve showing that certain mappings are bijections. The first result may seem intuitively clear, but its proof is rather technical; it will be given in Appendix B.

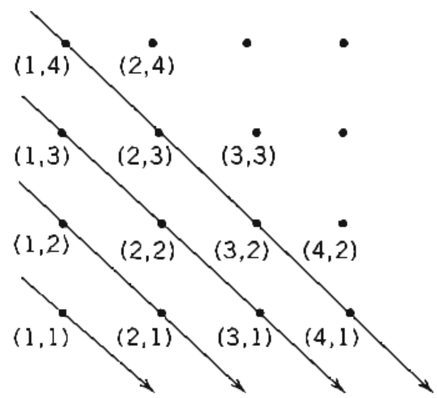


Figure 1.3.1 The set $\mathbb{N} \times \mathbb{N}$

1.3.9 Theorem *Suppose that S and T are sets and that $T \subseteq S$.*

(a) *If S is a countable set, then T is a countable set.*

(b) *If T is an uncountable set, then S is an uncountable set.*

1.3.10 Theorem *The following statements are equivalent:*

(a) *S is a countable set.*

(b) *There exists a surjection of $\mathbb{N}$ onto S.*

(c) *There exists an injection of S into $\mathbb{N}$.*

Proof. (a) $\Rightarrow$ (b) If S is finite, there exists a bijection h of some set $\mathbb{N}_n$ onto S and we define H on $\mathbb{N}$ by

$$H(k) := \begin{cases} h(k) & \text{for} \quad k = 1, \cdots, n, \\ h(n) & \text{for} \quad k > n. \end{cases}$$

Then H is a surjection of $\mathbb{N}$ onto S.

If S is denumerable, there exists a bijection H of $\mathbb{N}$ onto S, which is also a surjection of $\mathbb{N}$ onto S.

(b) $\Rightarrow$ (c) If H is a surjection of $\mathbb{N}$ onto S, we define $H_1 : S \to \mathbb{N}$ by letting $H_1(s)$ be the least element in the set $H^{-1}(s) := \{n \in \mathbb{N} : H(n) = s\}$. To see that H_1 is an injection of S into $\mathbb{N}$, note that if $s, t \in S$ and $n_{st} := H_1(s) = H_1(t)$, then $s = H(n_{st}) = t$.

(c) $\Rightarrow$ (a) If H_1 is an injection of S into $\mathbb{N}$, then it is a bijection of S onto $H_1(S) \subseteq \mathbb{N}$. By Theorem 1.3.9(a), $H_1(S)$ is countable, whence the set S is countable. Q.E.D.

1.3.11 Theorem *The set $\mathbb{Q}$ of all rational numbers is denumerable.*

Proof. The idea of the proof is to observe that the set $\mathbb{Q}^+$ of positive rational numbers is contained in the enumeration:

$$\frac{1}{1}, \frac{1}{2}, \frac{2}{1}, \frac{1}{3}, \frac{2}{2}, \frac{3}{1}, \frac{1}{4}, \cdots,$$

which is another "diagonal mapping" (see Figure 1.3.2). However, this mapping is not an injection, since the different fractions $\frac{1}{2}$ and $\frac{2}{4}$ represent the same rational number.

To proceed more formally, note that since $\mathbb{N} \times \mathbb{N}$ is countable (by Theorem 1.3.8), it follows from Theorem 1.3.10(b) that there exists a surjection f of $\mathbb{N}$ onto $\mathbb{N} \times \mathbb{N}$. If

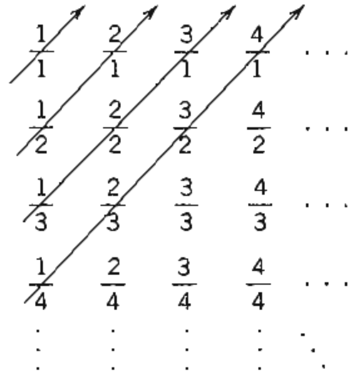


Figure 1.3.2 The set $\mathbb{Q}^+$

$g : \mathbb{N} \times \mathbb{N} \to \mathbb{Q}^+$ is the mapping that sends the ordered pair (m, n) into the rational number having a representation m/n, then g is a surjection onto $\mathbb{Q}^+$. Therefore, the composition $g \circ f$ is a surjection of $\mathbb{N}$ onto $\mathbb{Q}^+$, and Theorem 1.3.10 implies that $\mathbb{Q}^+$ is a countable set.

Similarly, the set $\mathbb{Q}^-$ of all negative rational numbers is countable. It follows as in Example 1.3.7(b) that the set $\mathbb{Q} = \mathbb{Q}^- \cup \{0\} \cup \mathbb{Q}^+$ is countable. Since $\mathbb{Q}$ contains $\mathbb{N}$, it must be a denumerable set. Q.E.D.

The next result is concerned with unions of sets. In view of Theorem 1.3.10, we need not be worried about possible overlapping of the sets. Also, we do not have to construct a bijection.

1.3.12 Theorem *If A_m is a countable set for each $m \in \mathbb{N}$, then the union $A := \bigcup_{m=1}^{\infty} A_m$ is countable.*

Proof. For each $m \in \mathbb{N}$, let φ_m be a surjection of $\mathbb{N}$ onto A_m. We define $\psi : \mathbb{N} \times \mathbb{N} \to A$ by

$$\psi(m, n) := \varphi_m(n).$$

We claim that ψ is a surjection. Indeed, if $a \in A$, then there exists a least $m \in \mathbb{N}$ such that $a \in A_m$, whence there exists a least $n \in \mathbb{N}$ such that $a = \varphi_m(n)$. Therefore, $a = \psi(m, n)$.

Since $\mathbb{N} \times \mathbb{N}$ is countable, it follows from Theorem 1.3.10 that there exists a surjection $f : \mathbb{N} \to \mathbb{N} \times \mathbb{N}$ whence $\psi \circ f$ is a surjection of $\mathbb{N}$ onto A. Now apply Theorem 1.3.10 again to conclude that A is countable. Q.E.D.

Remark A less formal (but more intuitive) way to see the truth of Theorem 1.3.12 is to enumerate the elements of A_m, $m \in \mathbb{N}$, as:

$$\begin{aligned} A_1 &= \{a_{11}, a_{12}, a_{13}, \cdots\}, \\ A_2 &= \{a_{21}, a_{22}, a_{23}, \cdots\}, \\ A_3 &= \{a_{31}, a_{32}, a_{33}, \cdots\}, \\ &\cdots \quad \cdots \quad \cdots. \end{aligned}$$

We then enumerate this array using the "diagonal procedure":

$$a_{11}, a_{12}, a_{21}, a_{13}, a_{22}, a_{31}, a_{14}, \cdots,$$

as was displayed in Figure 1.3.1.

The argument that the set $\mathbb{Q}$ of rational numbers is countable was first given in 1874 by Georg Cantor (1845–1918). He was the first mathematician to examine the concept of infinite set in rigorous detail. In contrast to the countability of $\mathbb{Q}$, he also proved the set $\mathbb{R}$ of real numbers is an uncountable set. (This result will be established in Section 2.5.)

In a series of important papers, Cantor developed an extensive theory of infinite sets and transfinite arithmetic. Some of his results were quite surprising and generated considerable controversy among mathematicians of that era. In a 1877 letter to his colleague Richard Dedekind, he wrote, after proving an unexpected theorem, "I see it, but I do not believe it".

We close this section with one of Cantor's more remarkable theorems.

1.3.13 Cantor's Theorem *If A is any set, then there is no surjection of A onto the set $\mathcal{P}(A)$ of all subsets of A.*

Proof. Suppose that $\varphi : A \to \mathcal{P}(A)$ is a surjection. Since $\varphi(a)$ is a subset of A, either a belongs to $\varphi(a)$ or it does not belong to this set. We let

$$D := \{a \in A : a \notin \varphi(a)\}.$$

Since D is a subset of A, if φ is a surjection, then $D = \varphi(a_0)$ for some $a_0 \in A$.

We must have either $a_0 \in D$ or $a_0 \notin D$. If $a_0 \in D$, then since $D = \varphi(a_0)$, we must have $a_0 \in \varphi(a_0)$, contrary to the definition of D. Similarly, if $a_0 \notin D$, then $a_0 \notin \varphi(a_0)$ so that $a_0 \in D$, which is also a contradiction.

Therefore, φ cannot be a surjection. Q.E.D.

Cantor's Theorem implies that there is an unending progression of larger and larger sets. In particular, it implies that the collection $\mathcal{P}(\mathbb{N})$ of all subsets of the natural numbers $\mathbb{N}$ is uncountable.

Exercises for Section 1.3

1. Prove that a nonempty set T_1 is finite if and only if there is a bijection from T_1 onto a finite set T_2.
2. Prove parts (b) and (c) of Theorem 1.3.4.
3. Let $S := \{1, 2\}$ and $T := \{a, b, c\}$.
   (a) Determine the number of different injections from S into T.
   (b) Determine the number of different surjections from T onto S. 6
4. Exhibit a bijection between $\mathbb{N}$ and the set of all odd integers greater than 13.
5. Give an explicit definition of the bijection f from $\mathbb{N}$ onto $\mathbb{Z}$ described in Example 1.3.7(b).
6. Exhibit a bijection between $\mathbb{N}$ and a proper subset of itself.
7. Prove that a set T_1 is denumerable if and only if there is a bijection from T_1 onto a denumerable set T_2.
8. Give an example of a countable collection of finite sets whose union is not finite.
9. Prove in detail that if S and T are denumerable, then $S \cup T$ is denumerable.
10. Determine the number of elements in $\mathcal{P}(S)$, the collection of all subsets of S, for each of the following sets:
    (a) $S := \{1, 2\}$,
    (b) $S := \{1, 2, 3\}$,
    (c) $S := \{1, 2, 3, 4\}$.
    Be sure to include the empty set and the set S itself in $\mathcal{P}(S)$.
11. Use Mathematical Induction to prove that if the set S has n elements, then $\mathcal{P}(S)$ has 2^n elements.
12. Prove that the collection $\mathcal{F}(\mathbb{N})$ of all *finite* subsets of $\mathbb{N}$ is countable.

CHAPTER 2

THE REAL NUMBERS

In this chapter we will discuss the essential properties of the real number system $\mathbb{R}$. Although it is possible to give a formal construction of this system on the basis of a more primitive set (such as the set $\mathbb{N}$ of natural numbers or the set $\mathbb{Q}$ of rational numbers), we have chosen not to do so. Instead, we exhibit a list of fundamental properties associated with the real numbers and show how further properties can be deduced from them. This kind of activity is much more useful in learning the tools of analysis than examining the logical difficulties of constructing a model for $\mathbb{R}$.

The real number system can be described as a "complete ordered field", and we will discuss that description in considerable detail. In Section 2.1, we first introduce the "algebraic" properties—often called the "field" properties in abstract algebra—that are based on the two operations of addition and multiplication. We continue the section with the introduction of the "order" properties of $\mathbb{R}$ and we derive some consequences of these properties and illustrate their use in working with inequalities. The notion of absolute value, which is based on the order properties, is discussed in Section 2.2.

In Section 2.3, we make the final step by adding the crucial "completeness" property to the algebraic and order properties of $\mathbb{R}$. It is this property, which was not fully understood until the late nineteenth century, that underlies the theory of limits and continuity and essentially all that follows in this book. The rigorous development of real analysis would not be possible without this essential property.

In Section 2.4, we apply the Completeness Property to derive several fundamental results concerning $\mathbb{R}$, including the Archimedean Property, the existence of square roots, and the density of rational numbers in $\mathbb{R}$. We establish, in Section 2.5, the Nested Interval Property and use it to prove the uncountability of $\mathbb{R}$. We also discuss its relation to binary and decimal representations of real numbers.

Part of the purpose of Sections 2.1 and 2.2 is to provide examples of proofs of elementary theorems from explicitly stated assumptions. Students can thus gain experience in writing formal proofs before encountering the more subtle and complicated arguments related to the Completeness Property and its consequences. However, students who have previously studied the axiomatic method and the technique of proofs (perhaps in a course on abstract algebra) can move to Section 2.3 after a cursory look at the earlier sections. A brief discussion of logic and types of proofs can be found in Appendix A at the back of the book.

Section 2.1 The Algebraic and Order Properties of $\mathbb{R}$

We begin with a brief discussion of the "algebraic structure" of the real number system. We will give a short list of basic properties of addition and multiplication from which all other algebraic properties can be derived as theorems. In the terminology of abstract algebra, the system of real numbers is a "field" with respect to addition and multiplication. The basic properties listed in 2.1.1 are known as the *field axioms.* A *binary operation* associates with each pair (a, b) a unique element $B(a, b)$, but we will use the conventional notations of $a + b$ and $a \cdot b$ when discussing the properties of addition and multiplication.

2.1.1 Algebraic Properties of $\mathbb{R}$ On the set $\mathbb{R}$ of real numbers there are two binary operations, denoted by $+$ and $\cdot$ and called **addition** and **multiplication**, respectively. These operations satisfy the following properties:

(A1) $a + b = b + a$ for all a, b in $\mathbb{R}$ (*commutative property of addition*);
(A2) $(a + b) + c = a + (b + c)$ for all a, b, c in $\mathbb{R}$ (*associative property of addition*);
(A3) there exists an element 0 in $\mathbb{R}$ such that $0 + a = a$ and $a + 0 = a$ for all a in $\mathbb{R}$ (*existence of a zero element*);
(A4) for each a in $\mathbb{R}$ there exists an element $-a$ in $\mathbb{R}$ such that $a + (-a) = 0$ and $(-a) + a = 0$ (*existence of negative elements*);
(M1) $a \cdot b = b \cdot a$ for all a, b in $\mathbb{R}$ (*commutative property of multiplication*);
(M2) $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ for all a, b, c in $\mathbb{R}$ (*associative property of multiplication*);
(M3) there exists an element 1 in $\mathbb{R}$ *distinct from* 0 such that $1 \cdot a = a$ and $a \cdot 1 = a$ for all a in $\mathbb{R}$ (*existence of a unit element*);
(M4) for each $a \neq 0$ in $\mathbb{R}$ there exists an element $1/a$ in $\mathbb{R}$ such that $a \cdot (1/a) = 1$ and $(1/a) \cdot a = 1$ (*existence of reciprocals*);
(D) $a \cdot (b + c) = (a \cdot b) + (a \cdot c)$ and $(b + c) \cdot a = (b \cdot a) + (c \cdot a)$ for all a, b, c in $\mathbb{R}$ (*distributive property of multiplication over addition*).

These properties should be familiar to the reader. The first four are concerned with addition, the next four with multiplication, and the last one connects the two operations. The point of the list is that all the familiar techniques of algebra can be derived from these nine properties, in much the same spirit that the theorems of Euclidean geometry can be deduced from the five basic axioms stated by Euclid in his *Elements.* Since this task more properly belongs to a course in abstract algebra, we will not carry it out here. However, to exhibit the spirit of the endeavor, we will sample a few results and their proofs.

We first establish the basic fact that the elements 0 and 1, whose existence were asserted in (A3) and (M3), are in fact unique. We also show that multiplication by 0 always results in 0.

2.1.2 Theorem **(a)** *If z and a are elements in $\mathbb{R}$ with $z + a = a$, then $z = 0$.*
(b) *If u and $b \neq 0$ are elements in $\mathbb{R}$ with $u \cdot b = b$, then $u = 1$.*
(c) *If $a \in \mathbb{R}$, then $a \cdot 0 = 0$.*

Proof. (a) Using (A3), (A4), (A2), the hypothesis $z + a = a$, and (A4), we get

$$z = z + 0 = z + (a + (-a)) = (z + a) + (-a) = a + (-a) = 0.$$

(b) Using (M3), (M4), (M2), the assumed equality $u \cdot b = b$, and (M4) again, we get

$$u = u \cdot 1 = u \cdot (b \cdot (1/b)) = (u \cdot b) \cdot (1/b) = b \cdot (1/b) = 1.$$

(c) We have (why?)

$$a + a \cdot 0 = a \cdot 1 + a \cdot 0 = a \cdot (1 + 0) = a \cdot 1 = a.$$

Therefore, we conclude from (a) that $a \cdot 0 = 0$. Q.E.D.

We next establish two important properties of multiplication: the uniqueness of reciprocals and the fact that a product of two numbers is zero only when one of the factors is zero.

2.1.3 Theorem (a) *If $a \neq 0$ and b in $\mathbb{R}$ are such that $a \cdot b = 1$, then $b = 1/a$.*
(b) *If $a \cdot b = 0$, then either $a = 0$ or $b = 0$.*

Proof. (a) Using (M3), (M4), (M2), the hypothesis $a \cdot b = 1$, and (M3), we have

$$b = 1 \cdot b = ((1/a) \cdot a) \cdot b = (1/a) \cdot (a \cdot b) = (1/a) \cdot 1 = 1/a.$$

(b) It suffices to assume $a \neq 0$ and prove that $b = 0$. (Why?) We multiply $a \cdot b$ by $1/a$ and apply (M2), (M4) and (M3) to get

$$(1/a) \cdot (a \cdot b) = ((1/a) \cdot a) \cdot b = 1 \cdot b = b.$$

Since $a \cdot b = 0$, by 2.1.2(c) this also equals

$$(1/a) \cdot (a \cdot b) = (1/a) \cdot 0 = 0.$$

Thus we have $b = 0$. Q.E.D.

These theorems represent a small sample of the algebraic properties of the real number system. Some additional consequences of the field properties are given in the exercises.

The operation of **subtraction** is defined by $a - b := a + (-b)$ for a, b in $\mathbb{R}$. Similarly, **division** is defined for a, b in $\mathbb{R}$ with $b \neq 0$ by $a/b := a \cdot (1/b)$. In the following, we will use this customary notation for subtraction and division, and we will use all the familiar properties of these operations. We will ordinarily drop the use of the dot to indicate multiplication and write ab for $a \cdot b$. Similarly, we will use the usual notation for exponents and write a^2 for aa, a^3 for $(a^2)a$; and, in general, we define $a^{n+1} := (a^n)a$ for $n \in \mathbb{N}$. We agree to adopt the convention that $a^1 = a$. Further, if $a \neq 0$, we write $a^0 = 1$ and a^{-1} for $1/a$, and if $n \in \mathbb{N}$, we will write a^{-n} for $(1/a)^n$, when it is convenient to do so. In general, we will freely apply all the usual techniques of algebra without further elaboration.

Rational and Irrational Numbers

We regard the set $\mathbb{N}$ of natural numbers as a subset of $\mathbb{R}$, by identifying the natural number $n \in \mathbb{N}$ with the n-fold sum of the unit element $1 \in \mathbb{R}$. Similarly, we identify $0 \in \mathbb{Z}$ with the zero element of $0 \in \mathbb{R}$, and we identify the n-fold sum of -1 with the integer $-n$. Thus, we consider $\mathbb{N}$ and $\mathbb{Z}$ to be subsets of $\mathbb{R}$.

Elements of $\mathbb{R}$ that can be written in the form b/a where $a, b \in \mathbb{Z}$ and $a \neq 0$ are called **rational numbers**. The set of all rational numbers in $\mathbb{R}$ will be denoted by the standard notation $\mathbb{Q}$. The sum and product of two rational numbers is again a rational number (prove this), and moreover, the field properties listed at the beginning of this section can be shown to hold for $\mathbb{Q}$.

The fact that there are elements in $\mathbb{R}$ that are not in $\mathbb{Q}$ is not immediately apparent. In the sixth century B.C. the ancient Greek society of Pythagoreans discovered that the diagonal of a square with unit sides could not be expressed as a ratio of integers. In view of the Pythagorean Theorem for right triangles, this implies that the square of no rational number can equal 2. This discovery had a profound impact on the development of Greek mathematics. One consequence is that elements of $\mathbb{R}$ that are not in $\mathbb{Q}$ became known as **irrational numbers**, meaning that they are not ratios of integers. Although the word "irrational" in modern English usage has a quite different meaning, we shall adopt the standard mathematical usage of this term.

We will now prove that there does not exist a rational number whose square is 2. In the proof we use the notions of even and odd numbers. Recall that a natural number is **even** if it has the form $2n$ for some $n \in \mathbb{N}$, and it is **odd** if it has the form $2n - 1$ for some $n \in \mathbb{N}$. Every natural number is either even or odd, and no natural number is both even and odd.

2.1.4 Theorem *There does not exist a rational number r such that $r^2 = 2$.*

Proof. Suppose, on the contrary, that p and q are integers such that $(p/q)^2 = 2$. We may assume that p and q are positive and have no common integer factors other than 1. (Why?) Since $p^2 = 2q^2$, we see that p^2 is even. This implies that p is also even (because if $p = 2n - 1$ is odd, then its square $p^2 = 2(2n^2 - 2n + 1) - 1$ is also odd). Therefore, since p and q do not have 2 as a common factor, then q must be an odd natural number.

Since p is even, then $p = 2m$ for some $m \in \mathbb{N}$, and hence $4m^2 = 2q^2$, so that $2m^2 = q^2$. Therefore, q^2 is even, and it follows from the argument in the preceding paragraph that q is an even natural number.

Since the hypothesis that $(p/q)^2 = 2$ leads to the contradictory conclusion that q is both even and odd, it must be false. Q.E.D.

The Order Properties of $\mathbb{R}$

The "order properties" of $\mathbb{R}$ refer to the notions of positivity and inequalities between real numbers. As with the algebraic structure of the system of real numbers, we proceed by isolating three basic properties from which all other order properties and calculations with inequalities can be deduced. The simplest way to do this is to identify a special subset of $\mathbb{R}$ by using the notion of "positivity".

2.1.5 The Order Properties of $\mathbb{R}$ There is a nonempty subset $\mathbb{P}$ of $\mathbb{R}$, called the set of **positive real numbers**, that satisfies the following properties:

(i) If a, b belong to $\mathbb{P}$, then $a + b$ belongs to $\mathbb{P}$.

(ii) If a, b belong to $\mathbb{P}$, then ab belongs to $\mathbb{P}$.

(iii) If a belongs to $\mathbb{R}$, then exactly one of the following holds:

$$a \in \mathbb{P}, \qquad a = 0, \qquad -a \in \mathbb{P}.$$

The first two conditions ensure the compatibility of order with the operations of addition and multiplication, respectively. Condition 2.1.5(iii) is usually called the **Trichotomy Property**, since it divides $\mathbb{R}$ into three distinct types of elements. It states that the set $\{-a : a \in P\}$ of **negative** real numbers has no elements in common with the set $\mathbb{P}$ of positive real numbers, and, moreover, the set $\mathbb{R}$ is the union of three disjoint sets.

If $a \in \mathbb{P}$, we write $a > 0$ and say that a is a **positive** (or a **strictly positive**) real number. If $a \in \mathbb{P} \cup \{0\}$, we write $a \geq 0$ and say that a is a **nonnegative** real number. Similarly, if $-a \in \mathbb{P}$, we write $a < 0$ and say that a is a **negative** (or a **strictly negative**) real number. If $-a \in \mathbb{P} \cup \{0\}$, we write $a \leq 0$ and say that a is a **nonpositive** real number.

The notion of inequality between two real numbers will now be defined in terms of the set $\mathbb{P}$ of positive elements.

2.1.6 Definition Let a, b be elements of $\mathbb{R}$.

(a) If $a - b \in \mathbb{P}$, then we write $a > b$ or $b < a$.

(b) If $a - b \in \mathbb{P} \cup \{0\}$, then we write $a \geq b$ or $b \leq a$.

The Trichotomy Property 2.1.5(iii) implies that for $a, b \in \mathbb{R}$ exactly one of the following will hold:

$$a > b, \qquad a = b, \qquad a < b.$$

Therefore, if both $a \leq b$ and $b \leq a$, then $a = b$.

For notational convenience, we will write

$$a < b < c$$

to mean that both $a < b$ and $b < c$ are satisfied. The other "double" inequalities $a \leq b < c$, $a \leq b \leq c$, and $a < b \leq c$ are defined in a similar manner.

To illustrate how the basic Order Properties are used to derive the "rules of inequalities", we will now establish several results that the reader has used in earlier mathematics courses.

2.1.7 Theorem *Let a, b, c be any elements of $\mathbb{R}$.*

(a) *If $a > b$ and $b > c$, then $a > c$.*

(b) *If $a > b$, then $a + c > b + c$.*

(c) *If $a > b$ and $c > 0$, then $ca > cb$.*
If $a > b$ and $c < 0$, then $ca < cb$.

Proof. (a) If $a - b \in \mathbb{P}$ and $b - c \in \mathbb{P}$, then 2.1.5(i) implies that $(a - b) + (b - c) = a - c$ belongs to $\mathbb{P}$. Hence $a > c$.

(b) If $a - b \in \mathbb{P}$, then $(a + c) - (b + c) = a - b$ is in $\mathbb{P}$. Thus $a + c > b + c$.

(c) If $a - b \in \mathbb{P}$ and $c \in \mathbb{P}$, then $ca - cb = c(a - b)$ is in $\mathbb{P}$ by 2.1.5(ii). Thus $ca > cb$ when $c > 0$.

On the other hand, if $c < 0$, then $-c \in \mathbb{P}$, so that $cb - ca = (-c)(a - b)$ is in $\mathbb{P}$. Thus $cb > ca$ when $c < 0$. Q.E.D.

It is natural to expect that the natural numbers are positive real numbers. This property is derived from the basic properties of order. The key observation is that the square of any nonzero real number is positive.

2.1.8 Theorem **(a)** *If $a \in \mathbb{R}$ and $a \neq 0$, then $a^2 > 0$.*

(b) $1 > 0$.

(c) *If $n \in \mathbb{N}$, then $n > 0$.*

Proof. (a) By the Trichotomy Property, if $a \neq 0$, then either $a \in \mathbb{P}$ or $-a \in \mathbb{P}$. If $a \in \mathbb{P}$, then by 2.1.5(ii), $a^2 = a \cdot a \in \mathbb{P}$. Also, if $-a \in \mathbb{P}$, then $a^2 = (-a)(-a) \in \mathbb{P}$. We conclude that if $a \neq 0$, then $a^2 > 0$.

(b) Since $1 = 1^2$, it follows from (a) that $1 > 0$.

(c) We use Mathematical Induction. The assertion for $n = 1$ is true by (b). If we suppose the assertion is true for the natural number k, then $k \in \mathbb{P}$, and since $1 \in \mathbb{P}$, we have $k + 1 \in \mathbb{P}$ by 2.1.5(i). Therefore, the assertion is true for all natural numbers. Q.E.D.

It is worth noting that *no smallest positive real number can exist*. This follows by observing that if $a > 0$, then since $\frac{1}{2} > 0$ (why?), we have that

$$0 < \tfrac{1}{2}a < a.$$

Thus if it is claimed that a is the smallest positive real number, we can exhibit a smaller positive number $\frac{1}{2}a$.

This observation leads to the next result, which will be used frequently as a method of proof. For instance, to prove that a number $a \geq 0$ is actually equal to zero, we see that it suffices to show that a is smaller than an arbitrary positive number.

2.1.9 Theorem *If $a \in \mathbb{R}$ is such that $0 \leq a < \varepsilon$ for every $\varepsilon > 0$, then $a = 0$.*

Proof. Suppose to the contrary that $a > 0$. Then if we take $\varepsilon_0 := \frac{1}{2}a$, we have $0 < \varepsilon_0 < a$. Therefore, it is false that $a < \varepsilon$ for every $\varepsilon > 0$ and we conclude that $a = 0$. Q.E.D.

Remark It is an exercise to show that if $a \in \mathbb{R}$ is such that $0 \leq a \leq \varepsilon$ for every $\varepsilon > 0$, then $a = 0$.

The product of two positive numbers is positive. However, the positivity of a product of two numbers does not imply that each factor is positive. The correct conclusion is given in the next theorem. It is an important tool in working with inequalities.

2.1.10 Theorem *If $ab > 0$, then either*

(i) *$a > 0$ and $b > 0$, or*

(ii) *$a < 0$ and $b < 0$.*

Proof. First we note that $ab > 0$ implies that $a \neq 0$ and $b \neq 0$. (Why?) From the Trichotomy Property, either $a > 0$ or $a < 0$. If $a > 0$, then $1/a > 0$ (why?), and therefore $b = (1/a)(ab) > 0$. Similarly, if $a < 0$, then $1/a < 0$, so that $b = (1/a)(ab) < 0$. Q.E.D.

2.1.11 Corollary *If $ab < 0$, then either*

(i) *$a < 0$ and $b > 0$, or*

(ii) *$a > 0$ and $b < 0$.*

Inequalities

We now show how the Order Properties presented in this section can be used to "solve" certain inequalities. The reader should justify each of the steps.

2.1.12 Examples (a) Determine the set A of all real numbers x such that $2x + 3 \leq 6$.

We note that we have†

$$x \in A \quad \iff \quad 2x + 3 \leq 6 \quad \iff \quad 2x \leq 3 \quad \iff \quad x \leq \tfrac{3}{2}.$$

Therefore $A = \left\{x \in \mathbb{R} : x \leq \frac{3}{2}\right\}$.

(b) Determine the set $B := \left\{x \in \mathbb{R} : x^2 + x > 2\right\}$.

We rewrite the inequality so that Theorem 2.1.10 can be applied. Note that

$$x \in B \quad \iff \quad x^2 + x - 2 > 0 \quad \iff \quad (x - 1)(x + 2) > 0.$$

Therefore, we either have (i) $x - 1 > 0$ and $x + 2 > 0$, or we have (ii) $x - 1 < 0$ and $x + 2 < 0$. In case (i) we must have both $x > 1$ and $x > -2$, which is satisfied if and only

†The symbol $\iff$ should be read "if and only if".

if $x > 1$. In case (ii) we must have both $x < 1$ and $x < -2$, which is satisfied if and only if $x < -2$.

We conclude that $B = \{x \in \mathbb{R} : x > 1\} \cup \{x \in \mathbb{R} : x < -2\}$.

(c) Determine the set

$$C := \left\{x \in \mathbb{R} : \frac{2x+1}{x+2} < 1\right\}.$$

We note that

$$x \in C \iff \frac{2x+1}{x+2} - 1 < 0 \iff \frac{x-1}{x+2} < 0.$$

Therefore we have either (i) $x - 1 < 0$ and $x + 2 > 0$, or (ii) $x - 1 > 0$ and $x + 2 < 0$. (Why?) In case (i) we must have both $x < 1$ and $x > -2$, which is satisfied if and only if $-2 < x < 1$. In case (ii), we must have both $x > 1$ and $x < -2$, which is never satisfied.

We conclude that $C = \{x \in \mathbb{R} : -2 < x < 1\}$. □

The following examples illustrate the use of the Order Properties of $\mathbb{R}$ in establishing certain inequalities. The reader should verify the steps in the arguments by identifying the properties that are employed.

It should be noted that the existence of square roots of positive numbers has not yet been established; however, we assume the existence of these roots for the purpose of these examples. (The existence of square roots will be discussed in Section 2.4.)

2.1.13 Examples (a) Let $a \geq 0$ and $b \geq 0$. Then

(1) $$a < b \iff a^2 < b^2 \iff \sqrt{a} < \sqrt{b}$$

We consider the case where $a > 0$ and $b > 0$, leaving the case $a = 0$ to the reader. It follows from 2.1.5(i) that $a + b > 0$. Since $b^2 - a^2 = (b - a)(b + a)$, it follows from 2.1.7(c) that $b - a > 0$ implies that $b^2 - a^2 > 0$. Also, it follows from 2.1.10 that $b^2 - a^2 > 0$ implies that $b - a > 0$.

If $a > 0$ and $b > 0$, then $\sqrt{a} > 0$ and $\sqrt{b} > 0$. Since $a = (\sqrt{a})^2$ and $b = (\sqrt{b})^2$, the second implication is a consequence of the first one when a and b are replaced by $\sqrt{a}$ and $\sqrt{b}$, respectively.

We also leave it to the reader to show that if $a \geq 0$ and $b \geq 0$, then

(1′) $$a \leq b \iff a^2 \leq b^2 \iff \sqrt{a} \leq \sqrt{b}$$

(b) If a and b are positive real numbers, then their **arithmetic mean** is $\frac{1}{2}(a + b)$ and their **geometric mean** is $\sqrt{ab}$. The **Arithmetic-Geometric Mean Inequality** for a, b is

(2) $$\sqrt{ab} \leq \tfrac{1}{2}(a + b)$$

with equality occurring if and only if $a = b$.

To prove this, note that if $a > 0$, $b > 0$, and $a \neq b$, then $\sqrt{a} > 0$, $\sqrt{b} > 0$ and $\sqrt{a} \neq \sqrt{b}$. (Why?) Therefore it follows from 2.1.8(a) that $(\sqrt{a} - \sqrt{b})^2 > 0$. Expanding this square, we obtain

$$a - 2\sqrt{ab} + b > 0,$$

whence it follows that

$$\sqrt{ab} < \tfrac{1}{2}(a + b).$$

Therefore (2) holds (with strict inequality) when $a \neq b$. Moreover, if $a = b(> 0)$, then both sides of (2) equal a, so (2) becomes an equality. This proves that (2) holds for $a > 0$, $b > 0$.

On the other hand, suppose that $a > 0$, $b > 0$ and that $\sqrt{ab} = \frac{1}{2}(a + b)$. Then, squaring both sides and multiplying by 4, we obtain

$$4ab = (a + b)^2 = a^2 + 2ab + b^2,$$

whence it follows that

$$0 = a^2 - 2ab + b^2 = (a - b)^2.$$

But this equality implies that $a = b$. (Why?) Thus, equality in (2) implies that $a = b$.

Remark The general Arithmetic-Geometric Mean Inequality for the positive real numbers $a_1, a_2, \cdots, a_n$ is

$$(3) \qquad (a_1 a_2 \cdots a_n)^{1/n} \leq \frac{a_1 + a_2 + \cdots + a_n}{n}$$

with equality occurring if and only if $a_1 = a_2 = \cdots = a_n$. It is possible to prove this more general statement using Mathematical Induction, but the proof is somewhat intricate. A more elegant proof that uses properties of the exponential function is indicated in Exercise 8.3.9 in Chapter 8.

(c) **Bernoulli's Inequality**. If $x > -1$, then

$$(4) \qquad (1 + x)^n \geq 1 + nx \qquad \text{for all} \quad n \in \mathbb{N}$$

The proof uses Mathematical Induction. The case $n = 1$ yields equality, so the assertion is valid in this case. Next, we assume the validity of the inequality (4) for $k \in \mathbb{N}$ and will deduce it for $k + 1$. Indeed, the assumptions that $(1 + x)^k \geq 1 + kx$ and that $1 + x > 0$ imply (why?) that

$$\begin{aligned}(1 + x)^{k+1} &= (1 + x)^k \cdot (1 + x) \\ &\geq (1 + kx) \cdot (1 + x) = 1 + (k + 1)x + kx^2 \\ &\geq 1 + (k + 1)x.\end{aligned}$$

Thus, inequality (4) holds for $n = k + 1$. Therefore, (4) holds for all $n \in \mathbb{N}$. □

Exercises for Section 2.1

1. If $a, b \in \mathbb{R}$, prove the following.
   (a) If $a + b = 0$, then $b = -a$, (b) $-(-a) = a$,
   (c) $(-1)a = -a$, (d) $(-1)(-1) = 1$. From b & c
2. Prove that if $a, b \in \mathbb{R}$, then
   (a) $-(a + b) = (-a) + (-b)$, (b) $(-a) \cdot (-b) = a \cdot b$,
   (c) $1/(-a) = -(1/a)$, (d) $-(a/b) = (-a)/b$ if $b \neq 0$.
3. Solve the following equations, justifying each step by referring to an appropriate property or theorem.
   (a) $2x + 5 = 8$, (b) $x^2 = 2x$,
   (c) $x^2 - 1 = 3$, (d) $(x - 1)(x + 2) = 0$.
4. If $a \in \mathbb{R}$ satisfies $a \cdot a = a$, prove that either $a = 0$ or $a = 1$.
5. If $a \neq 0$ and $b \neq 0$, show that $1/(ab) = (1/a)(1/b)$.

6. Use the argument in the proof of Theorem 2.1.4 to show that there does not exist a rational number s such that $s^2 = 6$.

7. Modify the proof of Theorem 2.1.4 to show that there does not exist a rational number t such that $t^2 = 3$.

8. (a) Show that if x, y are rational numbers, then $x + y$ and xy are rational numbers.
   (b) Prove that if x is a rational number and y is an irrational number, then $x + y$ is an irrational number. If, in addition, $x \neq 0$, then show that xy is an irrational number.

9. Let $K := \{s + t\sqrt{2} : s, t \in \mathbb{Q}\}$. Show that K satisfies the following:
   (a) If $x_1, x_2 \in K$, then $x_1 + x_2 \in K$ and $x_1 x_2 \in K$.
   (b) If $x \neq 0$ and $x \in K$, then $1/x \in K$.
   (Thus the set K is a *subfield* of $\mathbb{R}$. With the order inherited from $\mathbb{R}$, the set K is an ordered field that lies between $\mathbb{Q}$ and $\mathbb{R}$).

10. (a) If $a < b$ and $c \leq d$, prove that $a + c < b + d$.
    (b) If $0 < a < b$ and $0 \leq c \leq d$, prove that $0 \leq ac \leq bd$.

11. (a) Show that if $a > 0$, then $1/a > 0$ and $1/(1/a) = a$.
    (b) Show that if $a < b$, then $a < \frac{1}{2}(a + b) < b$.

12. Let a, b, c, d be numbers satisfying $0 < a < b$ and $c < d < 0$. Give an example where $ac < bd$, and one where $bd < ac$.

13. If $a, b \in \mathbb{R}$, show that $a^2 + b^2 = 0$ if and only if $a = 0$ and $b = 0$.

14. If $0 \leq a < b$, show that $a^2 \leq ab < b^2$. Show by example that it does *not* follow that $a^2 < ab < b^2$.

15. If $0 < a < b$, show that (a) $a < \sqrt{ab} < b$, and (b) $1/b < 1/a$.

16. Find all real numbers x that satisfy the following inequalities.
    (a) $x^2 > 3x + 4$, (b) $1 < x^2 < 4$,
    (c) $1/x < x$, (d) $1/x < x^2$.

17. Prove the following form of Theorem 2.1.9: If $a \in \mathbb{R}$ is such that $0 \leq a \leq \varepsilon$ for every $\varepsilon > 0$, then $a = 0$.

18. Let $a, b \in \mathbb{R}$, and suppose that for every $\varepsilon > 0$ we have $a \leq b + \varepsilon$. Show that $a \leq b$.

19. Prove that $\left[\frac{1}{2}(a + b)\right]^2 \leq \frac{1}{2}\left(a^2 + b^2\right)$ for all $a, b \in \mathbb{R}$. Show that equality holds if and only if $a = b$.

20. (a) If $0 < c < 1$, show that $0 < c^2 < c < 1$.
    (b) If $1 < c$, show that $1 < c < c^2$.

21. (a) Prove there is no $n \in \mathbb{N}$ such that $0 < n < 1$. (Use the Well-Ordering Property of $\mathbb{N}$.)
    (b) Prove that no natural number can be both even and odd.

22. (a) If $c > 1$, show that $c^n \geq c$ for all $n \in \mathbb{N}$, and that $c^n > c$ for $n > 1$.
    (b) If $0 < c < 1$, show that $c^n \leq c$ for all $n \in \mathbb{N}$, and that $c^n < c$ for $n > 1$.

23. If $a > 0, b > 0$ and $n \in \mathbb{N}$, show that $a < b$ if and only if $a^n < b^n$. [Hint: Use Mathematical Induction].

24. (a) If $c > 1$ and $m, n \in \mathbb{N}$, show that $c^m > c^n$ if and only if $m > n$.
    (b) If $0 < c < 1$ and $m, n \in \mathbb{N}$, show that $c^m < c^n$ if and only if $m > n$.

25. Assuming the existence of roots, show that if $c > 1$, then $c^{1/m} < c^{1/n}$ if and only if $m > n$.

26. Use Mathematical Induction to show that if $a \in \mathbb{R}$ and $m, n, \in \mathbb{N}$, then $a^{m+n} = a^m a^n$ and $(a^m)^n = a^{mn}$.

Section 2.2 Absolute Value and the Real Line

From the Trichotomy Property 2.1.5(iii), we are assured that if $a \in \mathbb{R}$ and $a \neq 0$, then exactly one of the numbers a and $-a$ is positive. The absolute value of $a \neq 0$ is defined to be the positive one of these two numbers. The absolute value of 0 is defined to be 0.

2.2.1 Definition The **absolute value** of a real number a, denoted by $|a|$, is defined by

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

For example, $|5| = 5$ and $|-8| = 8$. We see from the definition that $|a| \geq 0$ for all $a \in \mathbb{R}$, and that $|a| = 0$ if and only if $a = 0$. Also $|-a| = |a|$ for all $a \in \mathbb{R}$. Some additional properties are as follows.

2.2.2 Theorem **(a)** *$|ab| = |a||b|$ for all $a, b \in \mathbb{R}$.*

(b) *$|a|^2 = a^2$ for all $a \in \mathbb{R}$.*

(c) *If $c \geq 0$, then $|a| \leq c$ if and only if $-c \leq a \leq c$.*

(d) *$-|a| \leq a \leq |a|$ for all $a \in \mathbb{R}$.*

Proof. (a) If either a or b is 0, then both sides are equal to 0. There are four other cases to consider. If $a > 0, b > 0$, then $ab > 0$, so that $|ab| = ab = |a||b|$. If $a > 0, b < 0$, then $ab < 0$, so that $|ab| = -ab = a(-b) = |a||b|$. The remaining cases are treated similarly.

(b) Since $a^2 \geq 0$, we have $a^2 = |a^2| = |aa| = |a||a| = |a|^2$.

(c) If $|a| \leq c$, then we have both $a \leq c$ and $-a \leq c$ (why?), which is equivalent to $-c \leq a \leq c$. Conversely, if $-c \leq a \leq c$, then we have both $a \leq c$ and $-a \leq c$ (why?), so that $|a| \leq c$.

(d) Take $c = |a|$ in part (c). Q.E.D.

The following important inequality will be used frequently.

2.2.3 Triangle Inequality *If $a, b \in \mathbb{R}$, then $|a + b| \leq |a| + |b|$.*

Proof. From 2.2.2(d), we have $-|a| \leq a \leq |a|$ and $-|b| \leq b \leq |b|$. On adding these inequalities, we obtain

$$-(|a| + |b|) \leq a + b \leq |a| + |b|.$$

Hence, by 2.2.2(c) we have $|a + b| \leq |a| + |b|$. Q.E.D.

It can be shown that equality occurs in the Triangle Inequality if and only if $ab > 0$, which is equivalent to saying that a and b have the same sign. (See Exercise 2.)

There are many useful variations of the Triangle Inequality. Here are two.

2.2.4 Corollary *If $a, b \in \mathbb{R}$, then*

(a) $\big||a| - |b|\big| \leq |a - b|$,

(b) $|a - b| \leq |a| + |b|$.

Proof. (a) We write $a = a - b + b$ and then apply the Triangle Inequality to get $|a| = |(a - b) + b| \leq |a - b| + |b|$. Now subtract $|b|$ to get $|a| - |b| \leq |a - b|$. Similarly, from

$|b| = |b - a + a| \le |b - a| + |a|$, we obtain $-|a - b| = -|b - a| \le |a| - |b|$. If we combine these two inequalities, using 2.2.2(c), we get the inequality in (a).

(b) Replace b in the Triangle Inequality by $-b$ to get $|a - b| \le |a| + |-b|$. Since $|-b| = |b|$ we obtain the inequality in (b). Q.E.D.

A straightforward application of Mathematical Induction extends the Triangle Inequality to any finite number of elements of $\mathbb{R}$.

2.2.5 Corollary *If $a_1, a_2, \cdots, a_n$ are any real numbers, then*

$$|a_1 + a_2 + \cdots + a_n| \le |a_1| + |a_2| + \cdots + |a_n|.$$

The following examples illustrate how the properties of absolute value can be used.

2.2.6 Examples **(a)** Determine the set A of $x \in \mathbb{R}$ such that $|2x + 3| < 7$.

From a modification of 2.2.2(c) for the case of strict inequality, we see that $x \in A$ if and only if $-7 < 2x + 3 < 7$, which is satisfied if and only if $-10 < 2x < 4$. Dividing by 2, we conclude that $A = \{x \in \mathbb{R} : -5 < x < 2\}$.

(b) Determine the set $B := \{x \in \mathbb{R} : |x - 1| < |x|\}$.

One method is to consider cases so that the absolute value symbols can be removed. Here we take the cases

$$\text{(i) } x \ge 1, \qquad \text{(ii) } 0 \le x < 1, \qquad \text{(iii) } x < 0.$$

(Why did we choose these three cases?) In case (i) the inequality becomes $x - 1 < x$, which is satisfied without further restriction. Therefore all x such that $x \ge 1$ belong to the set B. In case (ii), the inequality becomes $-(x - 1) < x$, which requires that $x > \frac{1}{2}$. Thus, this case contributes all x such that $\frac{1}{2} < x < 1$ to the set B. In case (iii), the inequality becomes $-(x - 1) < -x$, which is equivalent to $1 < 0$. Since this statement is false, no value of x from case (iii) satisfies the inequality. Forming the union of the three cases, we conclude that $B = \{x \in \mathbb{R} : x > \frac{1}{2}\}$.

There is a second method of determining the set B based on the fact that $a < b$ if and only if $a^2 < b^2$ when both $a \ge 0$ and $b \ge 0$. (See 2.1.13(a).) Thus, the inequality $|x - 1| < |x|$ is equivalent to the inequality $|x - 1|^2 < |x|^2$. Since $|a|^2 = a^2$ for any a by 2.2.2(b), we can expand the square to obtain $x^2 - 2x + 1 < x^2$, which simplifies to $x > \frac{1}{2}$. Thus, we again find that $B = \{x \in \mathbb{R} : x > \frac{1}{2}\}$. This method of squaring can sometimes be used to advantage, but often a case analysis cannot be avoided when dealing with absolute values.

(c) Let the function f be defined by $f(x) := (2x^2 + 3x + 1)/(2x - 1)$ for $2 \le x \le 3$. Find a constant M such that $|f(x)| \le M$ for all x satisfying $2 \le x \le 3$.

We consider separately the numerator and denominator of

$$|f(x)| = \frac{|2x^2 + 3x + 1|}{|2x - 1|}.$$

From the Triangle Inequality, we obtain

$$|2x^2 + 3x + 1| \le 2|x|^2 + 3|x| + 1 \le 2 \cdot 3^2 + 3 \cdot 3 + 1 = 28$$

since $|x| \le 3$ for the x under consideration. Also, $|2x - 1| \ge 2|x| - 1 \ge 2 \cdot 2 - 1 = 3$ since $|x| \ge 2$ for the x under consideration. Thus, $1/|2x - 1| \le 1/3$ for $x \ge 2$. (Why?) Therefore, for $2 \le x \le 3$ we have $|f(x)| \le 28/3$. Hence we can take $M = 28/3$. (Note that we have found one such constant M; evidently any number $H > 28/3$ will also satisfy $|f(x)| \leq H$. It is also possible that $28/3$ is not the smallest possible choice for M.) □

The Real Line

A convenient and familiar geometric interpretation of the real number system is the real line. In this interpretation, the absolute value $|a|$ of an element a in $\mathbb{R}$ is regarded as the distance from a to the origin 0. More generally, the **distance** between elements a and b in $\mathbb{R}$ is $|a - b|$. (See Figure 2.2.1.)

We will later need precise language to discuss the notion of one real number being "close to" another. If a is a given real number, then saying that a real number x is "close to" a should mean that the distance $|x - a|$ between them is "small". A context in which this idea can be discussed is provided by the terminology of neighborhoods, which we now define.

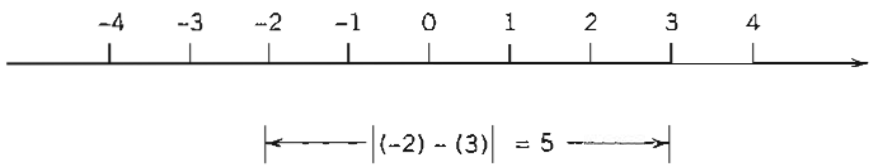


Figure 2.2.1 The distance between $a = -2$ and $b = 3$

2.2.7 Definition Let $a \in \mathbb{R}$ and $\varepsilon > 0$. Then the ε-**neighborhood** of a is the set $V_\varepsilon(a) := \{x \in \mathbb{R}: |x - a| < \varepsilon\}$.

For $a \in \mathbb{R}$, the statement that x belongs to $V_\varepsilon(a)$ is equivalent to either of the statements (see Figure 2.2.2)

$$-\varepsilon < x - a < \varepsilon \quad \iff \quad a - \varepsilon < x < a + \varepsilon.$$

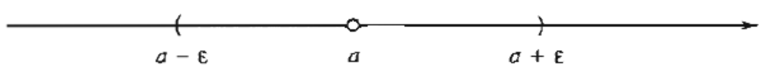


Figure 2.2.2 An ε-neighborhood of a

2.2.8 Theorem *Let $a \in \mathbb{R}$. If x belongs to the neighborhood $V_\varepsilon(a)$ for every $\varepsilon > 0$, then $x = a$.*

Proof. If a particular x satisfies $|x - a| < \varepsilon$ for every $\varepsilon > 0$, then it follows from 2.1.9 that $|x - a| = 0$, and hence $x = a$. Q.E.D.

2.2.9 Examples **(a)** Let $U := \{x : 0 < x < 1\}$. If $a \in U$, then let ε be the smaller of the two numbers a and $1 - a$. Then it is an exercise to show that $V_\varepsilon(a)$ is contained in U. Thus each element of U has some ε-neighborhood of it contained in U.

(b) If $I := \{x : 0 \leq x \leq 1\}$, then for any $\varepsilon > 0$, the ε-neighborhood $V_\varepsilon(0)$ of 0 contains points not in I, and so $V_\varepsilon(0)$ is not contained in I. For example, the number $x_\varepsilon := -\varepsilon/2$ is in $V_\varepsilon(0)$ but not in I.

(c) If $|x - a| < \varepsilon$ and $|y - b| < \varepsilon$, then the Triangle Inequality implies that

$$\begin{aligned} |(x + y) - (a + b)| &= |(x - a) + (y - b)| \\ &\leq |x - a| + |y - b| < 2\varepsilon. \end{aligned}$$

Thus if x, y belong to the ε-neighborhoods of a, b, respectively, then $x + y$ belongs to the 2ε-neighborhood of $a + b$ (but not necessarily to the ε-neighborhood of $a + b$). □

Exercises for Section 2.2

1. If $a, b \in \mathbb{R}$ and $b \neq 0$, show that:
   (a) $|a| = \sqrt{a^2}$, (b) $|a/b| = |a|/|b|$.
2. If $a, b \in \mathbb{R}$, show that $|a + b| = |a| + |b|$ if and only if $ab \geq 0$.
3. If $x, y, z \in \mathbb{R}$ and $x \leq z$, show that $x \leq y \leq z$ if and only if $|x - y| + |y - z| = |x - z|$. Interpret this geometrically.
4. Show that $|x - a| < \varepsilon$ if and only if $a - \varepsilon < x < a + \varepsilon$.
5. If $a < x < b$ and $a < y < b$, show that $|x - y| < b - a$. Interpret this geometrically.
6. Find all $x \in \mathbb{R}$ that satisfy the following inequalities:
   (a) $|4x - 5| \leq 13$, (b) $|x^2 - 1| \leq 3$.
7. Find all $x \in \mathbb{R}$ that satisfy the equation $|x + 1| + |x - 2| = 7$.
8. Find all $x \in \mathbb{R}$ that satisfy the following inequalities.
   (a) $|x - 1| > |x + 1|$, (b) $|x| + |x + 1| < 2$.
9. Sketch the graph of the equation $y = |x| - |x - 1|$.
10. Find all $x \in \mathbb{R}$ that satisfy the inequality $4 < |x + 2| + |x - 1| < 5$.
11. Find all $x \in \mathbb{R}$ that satisfy both $|2x - 3| < 5$ and $|x + 1| > 2$ simultaneously.
12. Determine and sketch the set of pairs (x, y) in $\mathbb{R} \times \mathbb{R}$ that satisfy:
    (a) $|x| = |y|$, (b) $|x| + |y| = 1$,
    (c) $|xy| = 2$, (d) $|x| - |y| = 2$.
13. Determine and sketch the set of pairs (x, y) in $\mathbb{R} \times \mathbb{R}$ that satisfy:
    (a) $|x| \leq |y|$, (b) $|x| + |y| \leq 1$,
    (c) $|xy| \leq 2$, (d) $|x| - |y| \geq 2$.
14. Let $\varepsilon > 0$ and $\delta > 0$, and $a \in \mathbb{R}$. Show that $V_\varepsilon(a) \cap V_\delta(a)$ and $V_\varepsilon(a) \cup V_\delta(a)$ are γ-neighborhoods of a for appropriate values of γ.
15. Show that if $a, b \in \mathbb{R}$, and $a \neq b$, then there exist ε-neighborhoods U of a and V of b such that $U \cap V = \emptyset$.
16. Show that if $a, b \in \mathbb{R}$ then
    (a) $\max\{a, b\} = \frac{1}{2}(a + b + |a - b|)$ and $\min\{a, b\} = \frac{1}{2}(a + b - |a - b|)$.
    (b) $\min\{a, b, c\} = \min\{\min\{a, b\}, c\}$.
17. Show that if $a, b, c \in \mathbb{R}$, then the "middle number" is $\text{mid}\{a, b, c\} = \min\{\max\{a, b\}, \max\{b, c\}, \max\{c, a\}\}$.

Section 2.3 The Completeness Property of $\mathbb{R}$

Thus far, we have discussed the algebraic properties and the order properties of the real number system $\mathbb{R}$. In this section we shall present one more property of $\mathbb{R}$ that is often called the "Completeness Property". The system $\mathbb{Q}$ of rational numbers also has the algebraic and order properties described in the preceding sections, but we have seen that $\sqrt{2}$ *cannot* be represented as a rational number; therefore $\sqrt{2}$ does not belong to $\mathbb{Q}$. This observation shows the necessity of an additional property to characterize the real number system. This additional property, the Completeness (or the Supremum) Property, is an essential property of $\mathbb{R}$, and we will say that $\mathbb{R}$ is a *complete ordered field*. It is this special property that permits us to define and develop the various limiting procedures that will be discussed in the chapters that follow.

There are several different ways to describe the Completeness Property. We choose to give what is probably the most efficient approach by assuming that each nonempty bounded subset of $\mathbb{R}$ has a supremum.

Suprema and Infima

We now introduce the notions of upper bound and lower bound for a set of real numbers. These ideas will be of utmost importance in later sections.

2.3.1 Definition Let S be a nonempty subset of $\mathbb{R}$.

(a) The set S is said to be **bounded above** if there exists a number $u \in \mathbb{R}$ such that $s \le u$ for all $s \in S$. Each such number u is called an **upper bound** of S.

(b) The set S is said to be **bounded below** if there exists a number $w \in \mathbb{R}$ such that $w \le s$ for all $s \in S$. Each such number w is called a **lower bound** of S.

(c) A set is said to be **bounded** if it is both bounded above and bounded below. A set is said to be **unbounded** if it is not bounded.

For example, the set $S := \{x \in \mathbb{R} : x < 2\}$ is bounded above; the number 2 and any number larger than 2 is an upper bound of S. This set has no lower bounds, so that the set is not bounded below. Thus it is unbounded (even though it is bounded above).

If a set has one upper bound, then it has infinitely many upper bounds, because if u is an upper bound of S, then the numbers $u+1, u+2, \cdots$ are also upper bounds of S. (A similar observation is valid for lower bounds.)

In the set of upper bounds of S and the set of lower bounds of S, we single out their least and greatest elements, respectively, for special attention in the following definition. (See Figure 2.3.1.)

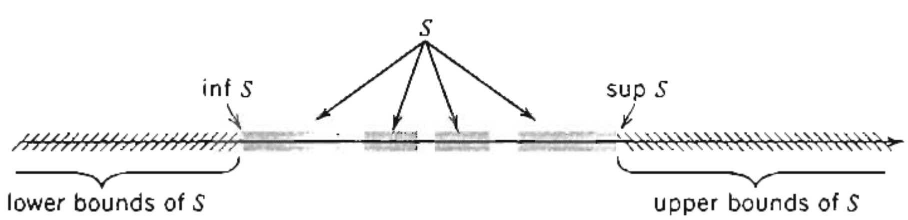


Figure 2.3.1 inf S and sup S

2.3.2 Definition Let S be a nonempty subset of $\mathbb{R}$.

(a) If S is bounded above, then a number u is said to be a **supremum** (or a **least upper bound**) of S if it satisfies the conditions:

 (1) u is an upper bound of S, and

 (2) if v is any upper bound of S, then $u \le v$.

(b) If S is bounded below, then a number w is said to be an **infimum** (or a **greatest lower bound**) of S if it satisfies the conditions:

(1′) w is a lower bound of S, and

(2′) if t is any lower bound of S, then $t \leq w$.

It is not difficult to see that *there can be only one supremum of a given subset S of $\mathbb{R}$.* (Then we can refer to *the* supremum of a set instead of *a* supremum.) For, suppose that u_1 and u_2 are both suprema of S. If $u_1 < u_2$, then the hypothesis that u_2 is a supremum implies that u_1 cannot be an upper bound of S. Similarly, we see that $u_2 < u_1$ is not possible. Therefore, we must have $u_1 = u_2$. A similar argument can be given to show that the infimum of a set is uniquely determined.

If the supremum or the infimum of a set S exists, we will denote them by

$$\sup S \quad \text{and} \quad \inf S.$$

We also observe that if u' is an arbitrary upper bound of a nonempty set S, then $\sup S \leq u'$. This is because $\sup S$ is the least of the upper bounds of S.

First of all, it needs to be emphasized that in order for a nonempty set S in $\mathbb{R}$ to have a supremum, it must have an upper bound. Thus, not every subset of $\mathbb{R}$ has a supremum; similarly, not every subset of $\mathbb{R}$ has an infimum. Indeed, there are four possibilities for a nonempty subset S of $\mathbb{R}$: it can

(i) have both a supremum and an infimum,
(ii) have a supremum but no infimum,
(iii) have a infimum but no supremum,
(iv) have neither a supremum nor an infimum.

We also wish to stress that in order to show that $u = \sup S$ for some nonempty subset S of $\mathbb{R}$, we need to show that *both* (1) and (2) of Definition 2.3.2(a) hold. It will be instructive to reformulate these statements. First the reader should see that the following two statements about a number u and a set S are equivalent:

(1) u is an upper bound of S,
(1′) $s \leq u$ for all $s \in S$.

Also, the following statements about an upper bound u of a set S are equivalent:

(2) if v is any upper bound of S, then $u \leq v$,
(2′) if $z < u$, then z is not an upper bound of S,
(2″) if $z < u$, then there exists $s_z \in S$ such that $z < s_z$,
(2‴) if $\varepsilon > 0$, then there exists $s_\varepsilon \in S$ such that $u - \varepsilon < s_\varepsilon$.

Therefore, we can state two alternate formulations for the supremum.

2.3.3 Lemma *A number u is the supremum of a nonempty subset S of $\mathbb{R}$ if and only if u satisfies the conditions:*

(1) $s \leq u$ *for all* $s \in S$,
(2) *if* $v < u$, *then there exists* $s' \in S$ *such that* $v < s'$.

We leave it to the reader to write out the details of the proof.

2.3.4 Lemma *An upper bound u of a nonempty set S in $\mathbb{R}$ is the supremum of S if and only if for every $\varepsilon > 0$ there exists an $s_\varepsilon \in S$ such that $u - \varepsilon < s_\varepsilon$.*

Proof. If u is an upper bound of S that satisfies the stated condition and if $v < u$, then we put $\varepsilon := u - v$. Then $\varepsilon > 0$, so there exists $s_\varepsilon \in S$ such that $v = u - \varepsilon < s_\varepsilon$. Therefore, v is not an upper bound of S, and we conclude that $u = \sup S$.

Conversely, suppose that $u = \sup S$ and let $\varepsilon > 0$. Since $u - \varepsilon < u$, then $u - \varepsilon$ is not an upper bound of S. Therefore, some element s_ε of S must be greater than $u - \varepsilon$; that is, $u - \varepsilon < s_\varepsilon$. (See Figure 2.3.2.) Q.E.D.

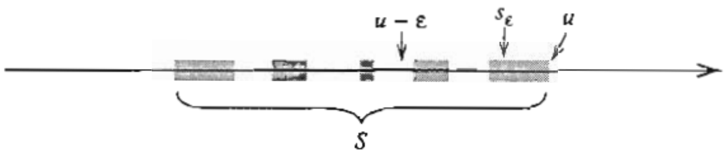


Figure 2.3.2 $u = \sup S$

It is important to realize that the supremum of a set may or may not be an element of the set. Sometimes it is and sometimes it is not, depending on the particular set. We consider a few examples.

2.3.5 Examples **(a)** If a nonempty set S_1 has a finite number of elements, then it can be shown that S_1 has a largest element u and a least element w. Then $u = \sup S_1$ and $w = \inf S_1$, and they are both members of S_1. (This is clear if S_1 has only one element, and it can be proved by induction on the number of elements in S_1; see Exercises 11 and 12.)

(b) The set $S_2 := \{x : 0 \le x \le 1\}$ clearly has 1 for an upper bound. We prove that 1 is its supremum as follows. If $v < 1$, there exists an element $s' \in S_2$ such that $v < s'$. (Name one such element s'.) Therefore v is not an upper bound of S_2 and, since v is an arbitrary number $v < 1$, we conclude that $\sup S_2 = 1$. It is similarly shown that $\inf S_2 = 0$. Note that both the supremum and the infimum of S_2 are contained in S_2.

(c) The set $S_3 := \{x : 0 < x < 1\}$ clearly has 1 for an upper bound. Using the same argument as given in (b), we see that $\sup S_3 = 1$. In this case, the set S_3 does *not* contain its supremum. Similarly, $\inf S_3 = 0$ is not contained in S_3. □

The Completeness Property of $\mathbb{R}$

It is not possible to prove on the basis of the field and order properties of $\mathbb{R}$ that were discussed in Section 2.1 that every nonempty subset of $\mathbb{R}$ that is bounded above has a supremum in $\mathbb{R}$. However, it is a deep and fundamental property of the real number system that this is indeed the case. We will make frequent and essential use of this property, especially in our discussion of limiting processes. The following statement concerning the existence of suprema is our final assumption about $\mathbb{R}$. Thus, we say that $\mathbb{R}$ is a *complete ordered field.*

2.3.6 The Completeness Property of $\mathbb{R}$ *Every nonempty set of real numbers that has an upper bound also has a supremum in $\mathbb{R}$.*

This property is also called the **Supremum Property** of $\mathbb{R}$. The analogous property for infima can be deduced from the Completeness Property as follows. Suppose that S is a nonempty subset of $\mathbb{R}$ that is bounded below. Then the nonempty set $\overline{S} := \{-s : s \in S\}$ is bounded above, and the Supremum Property implies that $u := \sup \overline{S}$ exists in $\mathbb{R}$. The reader should verify in detail that $-u$ is the infimum of S.

Exercises for Section 2.3

1. Let $S_1 := \{x \in \mathbb{R} : x \geq 0\}$. Show in detail that the set S_1 has lower bounds, but no upper bounds. Show that $\inf S_1 = 0$.
2. Let $S_2 = \{x \in \mathbb{R} : x > 0\}$. Does S_2 have lower bounds? Does S_2 have upper bounds? Does $\inf S_2$ exist? Does $\sup S_2$ exist? Prove your statements.
3. Let $S_3 = \{1/n : n \in \mathbb{N}\}$. Show that $\sup S_3 = 1$ and $\inf S_3 \geq 0$. (It will follow from the Archimedean Property in Section 2.4 that $\inf S_3 = 0$.)
4. Let $S_4 := \{1 - (-1)^n/n : n \in \mathbb{N}\}$. Find $\inf S_4$ and $\sup S_4$.
5. Let S be a nonempty subset of $\mathbb{R}$ that is bounded below. Prove that $\inf S = -\sup\{-s : s \in S\}$.
6. If a set $S \subseteq \mathbb{R}$ contains one of its upper bounds, show that this upper bound is the supremum of S.
7. Let $S \subseteq \mathbb{R}$ be nonempty. Show that $u \in \mathbb{R}$ is an upper bound of S if and only if the conditions $t \in \mathbb{R}$ and $t > u$ imply that $t \notin S$.
8. Let $S \subseteq \mathbb{R}$ be nonempty. Show that if $u = \sup S$, then for every number $n \in \mathbb{N}$ the number $u - 1/n$ is not an upper bound of S, but the number $u + 1/n$ is an upper bound of S. (The converse is also true; see Exercise 2.4.3.)
9. Show that if A and B are bounded subsets of $\mathbb{R}$, then $A \cup B$ is a bounded set. Show that $\sup(A \cup B) = \sup\{\sup A, \sup B\}$.
10. Let S be a bounded set in $\mathbb{R}$ and let S_0 be a nonempty subset of S. Show that $\inf S \leq \inf S_0 \leq \sup S_0 \leq \sup S$.
11. Let $S \subseteq \mathbb{R}$ and suppose that $s^* := \sup S$ belongs to S. If $u \notin S$, show that $\sup(S \cup \{u\}) = \sup\{s^*, u\}$.
12. Show that a nonempty finite set $S \subseteq \mathbb{R}$ contains its supremum. [Hint: Use Mathematical Induction and the preceding exercise.]
13. Show that the assertions (1) and (1′) before Lemma 2.3.3 are equivalent.
14. Show that the assertions (2), (2′), (2″), and (2‴) before Lemma 2.3.3 are equivalent.
15. Write out the details of the proof of Lemma 2.3.3.

Section 2.4 Applications of the Supremum Property

We will now discuss how to work with suprema and infima. We will also give some very important applications of these concepts to derive fundamental properties of $\mathbb{R}$. We begin with examples that illustrate useful techniques in applying the ideas of supremum and infimum.

2.4.1 Example (a) It is an important fact that taking suprema and infima of sets is compatible with the algebraic properties of $\mathbb{R}$. As an example, we present here the compatibility of taking suprema and addition.

Let S be a nonempty subset of $\mathbb{R}$ that is bounded above, and let a be any number in $\mathbb{R}$. Define the set $a + S := \{a + s : s \in S\}$. We will prove that

$$\sup(a + S) = a + \sup S.$$

If we let $u := \sup S$, then $x \le u$ for all $x \in S$, so that $a + x \le a + u$. Therefore, $a + u$ is an upper bound for the set $a + S$; consequently, we have $\sup(a + S) \le a + u$.

Now if v is *any* upper bound of the set $a + S$, then $a + x \le v$ for all $x \in S$. Consequently $x \le v - a$ for all $x \in S$, so that $v - a$ is an upper bound of S. Therefore, $u = \sup S \le v - a$, which gives us $a + u \le v$. Since v is any upper bound of $a + S$, we can replace v by $\sup(a + S)$ to get $a + u \le \sup(a + S)$.

Combining these inequalities, we conclude that

$$\sup(a + S) = a + u = a + \sup S.$$

For similar relationships between the suprema and infima of sets and the operations of addition and multiplication, see the exercises.

(b) If the suprema or infima of two sets are involved, it is often necessary to establish results in two stages, working with one set at a time. Here is an example.

Suppose that A and B are nonempty subsets of $\mathbb{R}$ that satisfy the property:

$$a \le b \qquad \text{for all } a \in A \text{ and all } b \in B.$$

We will prove that

$$\sup A \le \inf B.$$

For, given $b \in B$, we have $a \le b$ for all $a \in A$. This means that b is an upper bound of A, so that $\sup A \le b$. Next, since the last inequality holds for all $b \in B$, we see that the number $\sup A$ is a lower bound for the set B. Therefore, we conclude that $\sup A \le \inf B$. $\square$

Functions

The idea of upper bound and lower bound is applied to functions by considering the range of a function. Given a function $f : D \to \mathbb{R}$, we say that f is **bounded above** if the set $f(D) = \{f(x) : x \in D\}$ is bounded above in $\mathbb{R}$; that is, there exists $B \in \mathbb{R}$ such that $f(x) \le B$ for all $x \in D$. Similarly, the function f is **bounded below** if the set $f(D)$ is bounded below. We say that f is **bounded** if it is bounded above and below; this is equivalent to saying that there exists $B \in \mathbb{R}$ such that $|f(x)| \le B$ for all $x \in D$.

The following example illustrates how to work with suprema and infima of functions.

2.4.2 Example Suppose that f and g are real-valued functions with common domain $D \subseteq \mathbb{R}$. We assume that f and g are bounded.

(a) If $f(x) \le g(x)$ for all $x \in D$, then $\sup f(D) \le \sup g(D)$, which is sometimes written:

$$\sup_{x \in D} f(x) \le \sup_{x \in D} g(x).$$

We first note that $f(x) \le g(x) \le \sup g(D)$, which implies that the number $\sup g(D)$ is an upper bound for $f(D)$. Therefore, $\sup f(D) \le \sup g(D)$.

(b) We note that the hypothesis $f(x) \le g(x)$ for all $x \in D$ in part (a) does not imply any relation between $\sup f(D)$ and $\inf g(D)$.

For example, if $f(x) := x^2$ and $g(x) := x$ with $D = \{x : 0 \le x \le 1\}$, then $f(x) \le g(x)$ for all $x \in D$. However, we see that $\sup f(D) = 1$ and $\inf g(D) = 0$. Since $\sup g(D) = 1$, the conclusion of (a) holds.

(c) If $f(x) \le g(y)$ for all $x, y \in D$, then we may conclude that $\sup f(D) \le \inf g(D)$, which we may write as:

$$\sup_{x\in D} f(x) \le \inf_{y\in D} g(y).$$

(Note that the functions in (b) do not satisfy this hypothesis.)

The proof proceeds in two stages as in Example 2.4.1(b). The reader should write out the details of the argument. □

Further relationships between suprema and infima of functions are given in the exercises.

The Archimedean Property

Because of your familiarity with the set $\mathbb{R}$ and the customary picture of the real line, it may seem obvious that the set $\mathbb{N}$ of natural numbers is *not* bounded in $\mathbb{R}$. How can we prove this "obvious" fact? In fact, we cannot do so by using only the Algebraic and Order Properties given in Section 2.1. Indeed, we must use the Completeness Property of $\mathbb{R}$ as well as the Inductive Property of $\mathbb{N}$ (that is, if $n \in \mathbb{N}$, then $n + 1 \in \mathbb{N}$).

The absence of upper bounds for $\mathbb{N}$ means that given any real number x there exists a natural number n (depending on x) such that $x < n$.

2.4.3 Archimedean Property *If $x \in \mathbb{R}$, then there exists $n_x \in \mathbb{N}$ such that $x < n_x$.*

Proof. If the assertion is false, then $n \le x$ for all $n \in \mathbb{N}$; therefore, x is an upper bound of $\mathbb{N}$. Therefore, by the Completeness Property, the nonempty set $\mathbb{N}$ has a supremum $u \in \mathbb{R}$. Subtracting 1 from u gives a number $u - 1$ which is smaller than the supremum u of $\mathbb{N}$. Therefore $u - 1$ is not an upper bound of $\mathbb{N}$, so there exists $m \in \mathbb{N}$ with $u - 1 < m$. Adding 1 gives $u < m + 1$, and since $m + 1 \in \mathbb{N}$, this inequality contradicts the fact that u is an upper bound of $\mathbb{N}$. Q.E.D.

2.4.4 Corollary *If $S := \{1/n : n \in \mathbb{N}\}$, then $\inf S = 0$.*

Proof. Since $S \neq \emptyset$ is bounded below by 0, it has an infimum and we let $w := \inf S$. It is clear that $w \ge 0$. For any $\varepsilon > 0$, the Archimedean Property implies that there exists $n \in \mathbb{N}$ such that $1/\varepsilon < n$, which implies $1/n < \varepsilon$. Therefore we have

$$0 \le w \le 1/n < \varepsilon.$$

But since $\varepsilon > 0$ is arbitrary, it follows from Theorem 2.1.9 that $w = 0$. Q.E.D.

2.4.5 Corollary *If $t > 0$, there exists $n_t \in \mathbb{N}$ such that $0 < 1/n_t < t$.*

Proof. Since $\inf\{1/n : n \in \mathbb{N}\} = 0$ and $t > 0$, then t is not a lower bound for the set $\{1/n : n \in \mathbb{N}\}$. Thus there exists $n_t \in \mathbb{N}$ such that $0 < 1/n_t < t$. Q.E.D.

2.4.6 Corollary *If $y > 0$, there exists $n_y \in \mathbb{N}$ such that $n_y - 1 \le y < n_y$.*

Proof. The Archimedean Property ensures that the subset $E_y := \{m \in \mathbb{N} : y < m\}$ of $\mathbb{N}$ is not empty. By the Well-Ordering Property 1.2.1, E_y has a least element, which we denote by n_y. Then $n_y - 1$ does not belong to E_y, and hence we have $n_y - 1 \le y < n_y$. Q.E.D.

Collectively, the Corollaries 2.4.4–2.4.6 are sometimes referred to as the Archimedean Property of $\mathbb{R}$.

The Existence of $\sqrt{2}$

The importance of the Supremum Property lies in the fact that it guarantees the existence of real numbers under certain hypotheses. We shall make use of it in this way many times. At the moment, we shall illustrate this use by proving the existence of a positive real number x such that $x^2 = 2$; that is, the positive square root of 2. It was shown earlier (see Theorem 2.1.4) that such an x cannot be a rational number; thus, we will be deriving the existence of at least one irrational number.

2.4.7 Theorem *There exists a positive real number x such that $x^2 = 2$.*

Proof. Let $S := \{s \in \mathbb{R}: 0 \le s, s^2 < 2\}$. Since $1 \in S$, the set is not empty. Also, S is bounded above by 2, because if $t > 2$, then $t^2 > 4$ so that $t \notin S$. Therefore the Supremum Property implies that the set S has a supremum in $\mathbb{R}$, and we let $x := \sup S$. Note that $x > 1$.

We will prove that $x^2 = 2$ by ruling out the other two possibilities: $x^2 < 2$ and $x^2 > 2$.

First assume that $x^2 < 2$. We will show that this assumption contradicts the fact that $x = \sup S$ by finding an $n \in \mathbb{N}$ such that $x + 1/n \in S$, thus implying that x is not an upper bound for S. To see how to choose n, note that $1/n^2 \le 1/n$ so that

$$\left(x + \frac{1}{n}\right)^2 = x^2 + \frac{2x}{n} + \frac{1}{n^2} \le x^2 + \frac{1}{n}(2x + 1).$$

Hence if we can choose n so that

$$\frac{1}{n}(2x + 1) < 2 - x^2,$$

then we get $(x + 1/n)^2 < x^2 + (2 - x^2) = 2$. By assumption we have $2 - x^2 > 0$, so that $(2 - x^2)/(2x + 1) > 0$. Hence the Archimedean Property (Corollary 2.4.5) can be used to obtain $n \in \mathbb{N}$ such that

$$\frac{1}{n} < \frac{2 - x^2}{2x + 1}.$$

These steps can be reversed to show that for this choice of n we have $x + 1/n \in S$, which contradicts the fact that x is an upper bound of S. Therefore we cannot have $x^2 < 2$.

Now assume that $x^2 > 2$. We will show that it is then possible to find $m \in \mathbb{N}$ such that $x - 1/m$ is also an upper bound of S, contradicting the fact that $x = \sup S$. To do this, note that

$$\left(x - \frac{1}{m}\right)^2 = x^2 - \frac{2x}{m} + \frac{1}{m^2} > x^2 - \frac{2x}{m}.$$

Hence if we can choose m so that

$$\frac{2x}{m} < x^2 - 2,$$

then $(x - 1/m)^2 > x^2 - (x^2 - 2) = 2$. Now by assumption we have $x^2 - 2 > 0$, so that $(x^2 - 2)/2x > 0$. Hence, by the Archimedean Property, there exists $m \in \mathbb{N}$ such that

$$\frac{1}{m} < \frac{x^2 - 2}{2x}.$$

These steps can be reversed to show that for this choice of m we have $(x - 1/m)^2 > 2$. Now if $s \in S$, then $s^2 < 2 < (x - 1/m)^2$, whence it follows from 2.1.13(a) that $s < x - 1/m$. This implies that $x - 1/m$ is an upper bound for S, which contradicts the fact that $x = \sup S$. Therefore we cannot have $x^2 > 2$.

Since the possibilities $x^2 < 2$ and $x^2 > 2$ have been excluded, we must have $x^2 = 2$. Q.E.D.

By slightly modifying the preceding argument, the reader can show that if $a > 0$, then there is a unique $b > 0$ such that $b^2 = a$. We call b the **positive square root** of a and denote it by $b = \sqrt{a}$ or $b = a^{1/2}$. A slightly more complicated argument involving the binomial theorem can be formulated to establish the existence of a unique **positive nth root** of a, denoted by $\sqrt[n]{a}$ or $a^{1/n}$, for each $n \in \mathbb{N}$.

Remark If in the proof of Theorem 2.4.7 we replace the set S by the set of rational numbers $T := \{r \in \mathbb{Q} : 0 \le r, r^2 < 2\}$, the argument then gives the conclusion that $y := \sup T$ satisfies $y^2 = 2$. Since we have seen in Theorem 2.1.4 that y cannot be a rational number, it follows that the set T that consists of rational numbers does not have a supremum belonging to the set $\mathbb{Q}$. Thus the ordered field $\mathbb{Q}$ of rational numbers does *not* possess the Completeness Property.

Density of Rational Numbers in $\mathbb{R}$

We now know that there exists at least one irrational real number, namely $\sqrt{2}$. Actually there are "more" irrational numbers than rational numbers in the sense that the set of rational numbers is countable (as shown in Section 1.3), while the set of irrational numbers is uncountable (see Section 2.5). However, we next show that in spite of this apparent disparity, the set of rational numbers is "dense" in $\mathbb{R}$ in the sense that given any two real numbers there is a rational number between them (in fact, there are infinitely many such rational numbers).

2.4.8 The Density Theorem *If x and y are any real numbers with $x < y$, then there exists a rational number $r \in \mathbb{Q}$ such that $x < r < y$.*

Proof. It is no loss of generality (why?) to assume that $x > 0$. Since $y - x > 0$, it follows from Corollary 2.4.5 that there exists $n \in \mathbb{N}$ such that $1/n < y - x$. Therefore, we have $nx + 1 < ny$. If we apply Corollary 2.4.6 to $nx > 0$, we obtain $m \in \mathbb{N}$ with $m - 1 \le nx < m$. Therefore, $m \le nx + 1 < ny$, whence $nx < m < ny$. Thus, the rational number $r := m/n$ satisfies $x < r < y$. Q.E.D.

To round out the discussion of the interlacing of rational and irrational numbers, we have the same "betweenness property" for the set of irrational numbers.

2.4.9 Corollary *If x and y are real numbers with $x < y$, then there exists an irrational number z such that $x < z < y$.*

Proof. If we apply the Density Theorem 2.4.8 to the real numbers $x/\sqrt{2}$ and $y/\sqrt{2}$, we obtain a rational number $r \neq 0$ (why?) such that

$$\frac{x}{\sqrt{2}} < r < \frac{y}{\sqrt{2}}.$$

Then $z := r\sqrt{2}$ is irrational (why?) and satisfies $x < z < y$. Q.E.D.

Exercises for Section 2.4

1. Show that $\sup\{1 - 1/n : n \in \mathbb{N}\} = 1$.
2. If $S := \{1/n - 1/m : n, m \in \mathbb{N}\}$, find $\inf S$ and $\sup S$.
3. Let $S \subseteq \mathbb{R}$ be nonempty. Prove that if a number u in $\mathbb{R}$ has the properties: (i) for every $n \in \mathbb{N}$ the number $u - 1/n$ is not an upper bound of S, and (ii) for every number $n \in \mathbb{N}$ the number $u + 1/n$ is an upper bound of S, then $u = \sup S$. (This is the converse of Exercise 2.3.8.)
4. Let S be a nonempty bounded set in $\mathbb{R}$.
   (a) Let $a > 0$, and let $aS := \{as : s \in S\}$. Prove that
   $$\inf(aS) = a \inf S, \qquad \sup(aS) = a \sup S.$$
   (b) Let $b < 0$ and let $bS = \{bs : s \in S\}$. Prove that
   $$\inf(bS) = b \sup S, \qquad \sup(bS) = b \inf S.$$
5. Let X be a nonempty set and let $f : X \to \mathbb{R}$ have bounded range in $\mathbb{R}$. If $a \in \mathbb{R}$, show that Example 2.4.1(a) implies that
   $$\sup\{a + f(x) : x \in X\} = a + \sup\{f(x) : x \in X\}$$
   Show that we also have
   $$\inf\{a + f(x) : x \in X\} = a + \inf\{f(x) : x \in X\}$$
6. Let A and B be bounded nonempty subsets of $\mathbb{R}$, and let $A + B := \{a + b : a \in A, b \in B\}$. Prove that $\sup(A + B) = \sup A + \sup B$ and $\inf(A + B) = \inf A + \inf B$.
7. Let X be a nonempty set, and let f and g be defined on X and have bounded ranges in $\mathbb{R}$. Show that
   $$\sup\{f(x) + g(x) : x \in X\} \leq \sup\{f(x) : x \in X\} + \sup\{g(x) : x \in X\}$$
   and that
   $$\inf\{f(x) : x \in X\} + \inf\{g(x) : x \in X\} \leq \inf\{f(x) + g(x) : x \in X\}.$$
   Give examples to show that each of these inequalities can be either equalities or strict inequalities.
8. Let $X = Y := \{x \in \mathbb{R} : 0 < x < 1\}$. Define $h : X \times Y \to \mathbb{R}$ by $h(x, y) := 2x + y$.
   (a) For each $x \in X$, find $f(x) := \sup\{h(x, y) : y \in Y\}$; then find $\inf\{f(x) : x \in X\}$.
   (b) For each $y \in Y$, find $g(y) := \inf\{h(x, y) : x \in X\}$; then find $\sup\{g(y) : y \in Y\}$. Compare with the result found in part (a).
9. Perform the computations in (a) and (b) of the preceding exercise for the function $h : X \times Y \to \mathbb{R}$ defined by
   $$h(x, y) := \begin{cases} 0 & \text{if } x < y, \\ 1 & \text{if } x \geq y. \end{cases}$$
10. Let X and Y be nonempty sets and let $h : X \times Y \to \mathbb{R}$ have bounded range in $\mathbb{R}$. Let $f : X \to \mathbb{R}$ and $g : Y \to \mathbb{R}$ be defined by
    $$f(x) := \sup\{h(x, y) : y \in Y\}, \qquad g(y) := \inf\{h(x, y) : x \in X\}.$$

Prove that

$$\sup\{g(y) : y \in Y\} \le \inf\{f(x) : x \in X\}$$

We sometimes express this by writing

$$\sup_y \inf_x h(x, y) \le \inf_x \sup_y h(x, y).$$

Note that Exercises 8 and 9 show that the inequality may be either an equality or a strict inequality.

11. Let X and Y be nonempty sets and let $h : X \times Y \to \mathbb{R}$ have bounded range in $\mathbb{R}$. Let $F: X \to \mathbb{R}$ and $G : Y \to \mathbb{R}$ be defined by

$$F(x) := \sup\{h(x, y) : y \in Y\}, \qquad G(y) := \sup\{h(x, y) : x \in X\}.$$

Establish the **Principle of the Iterated Suprema**:

$$\sup\{h(x, y) : x \in X, y \in Y\} = \sup\{F(x) : x \in X\} = \sup\{G(y) : y \in Y\}$$

We sometimes express this in symbols by

$$\sup_{x,y} h(x, y) = \sup_x \sup_y h(x, y) = \sup_y \sup_x h(x, y).$$

12. Given any $x \in \mathbb{R}$, show that there exists a *unique* $n \in \mathbb{Z}$ such that $n - 1 \le x < n$.
13. If $y > 0$, show that there exists $n \in \mathbb{N}$ such that $1/2^n < y$.
14. Modify the argument in Theorem 2.4.7 to show that there exists a positive real number y such that $y^2 = 3$.
15. Modify the argument in Theorem 2.4.7 to show that if $a > 0$, then there exists a positive real number z such that $z^2 = a$.
16. Modify the argument in Theorem 2.4.7 to show that there exists a positive real number u such that $u^3 = 2$.
17. Complete the proof of the Density Theorem 2.4.8 by removing the assumption that $x > 0$.
18. If $u > 0$ is any real number and $x < y$, show that there exists a rational number r such that $x < ru < y$. (Hence the set $\{ru: r \in \mathbb{Q}\}$ is dense in $\mathbb{R}$.)

Section 2.5 Intervals

The Order Relation on $\mathbb{R}$ determines a natural collection of subsets called "intervals". The notations and terminology for these special sets will be familiar from earlier courses. If $a, b \in \mathbb{R}$ satisfy $a < b$, then the **open interval** determined by a and b is the set

$$(a, b) := \{x \in \mathbb{R} : a < x < b\}.$$

The points a and b are called the **endpoints** of the interval; however, the endpoints are not included in an open interval. If both endpoints are adjoined to this open interval, then we obtain the **closed interval** determined by a and b; namely, the set

$$[a, b] := \{x \in \mathbb{R} : a \le x \le b\}.$$

The two **half-open** (or **half-closed**) intervals determined by a and b are $[a, b)$, which includes the endpoint a, and $(a, b]$, which includes the endpoint b.

Each of these four intervals is bounded and has **length** defined by $b - a$. If $a = b$, the corresponding open interval is the empty set $(a, a) = \emptyset$, whereas the corresponding closed interval is the singleton set $[a, a] = \{a\}$.

There are five types of unbounded intervals for which the symbols ∞ (or $+\infty$) and $-\infty$ are used as notational convenience in place of the endpoints. The **infinite open intervals** are the sets of the form

$$(a, \infty) := \{x \in \mathbb{R} : x > a\} \quad \text{and} \quad (-\infty, b) := \{x \in \mathbb{R} : x < b\}.$$

The first set has no upper bounds and the second one has no lower bounds. Adjoining endpoints gives us the **infinite closed intervals**:

$$[a, \infty) := \{x \in \mathbb{R} : a \le x\} \quad \text{and} \quad (-\infty, b] := \{x \in \mathbb{R} : x \le b\}.$$

It is often convenient to think of the entire set $\mathbb{R}$ as an infinite interval; in this case, we write $(-\infty, \infty) := \mathbb{R}$. No point is an endpoint of $(-\infty, \infty)$.

Warning It must be emphasized that ∞ and $-\infty$ are *not* elements of $\mathbb{R}$, but only convenient symbols.

Characterization of Intervals

An obvious property of intervals is that if two points x, y with $x < y$ belong to an interval I, then any point lying between them also belongs to I. That is, if $x < t < y$, then the point t belongs to the same interval as x and y. In other words, if x and y belong to an interval I, then the interval $[x, y]$ is contained in I. We now show that a subset of $\mathbb{R}$ possessing this property must be an interval.

2.5.1 Characterization Theorem *If S is a subset of $\mathbb{R}$ that contains at least two points and has the property*

$$(1) \qquad \text{if} \quad x, y \in S \quad \text{and} \quad x < y, \quad \text{then} \quad [x, y] \subseteq S,$$

then S is an interval.

Proof. There are four cases to consider: (i) S is bounded, (ii) S is bounded above but not below, (iii) S is bounded below but not above, and (iv) S is neither bounded above nor below.

Case (i): Let $a := \inf S$ and $b := \sup S$. Then $S \subseteq [a, b]$ and we will show that $(a, b) \subseteq S$.

If $a < z < b$, then z is not a lower bound of S, so there exists $x \in S$ with $x < z$. Also, z is not an upper bound of S, so there exists $y \in S$ with $z < y$. Therefore $z \in [x, y]$, so property (1) implies that $z \in S$. Since z is an arbitrary element of (a, b), we conclude that $(a, b) \subseteq S$.

Now if $a \in S$ and $b \in S$, then $S = [a, b]$. (Why?) If $a \notin S$ and $b \notin S$, then $S = (a, b)$. The other possibilities lead to either $S = (a, b]$ or $S = [a, b)$.

Case (ii): Let $b := \sup S$. Then $S \subseteq (-\infty, b]$ and we will show that $(-\infty, b) \subseteq S$. For, if $z < b$, then there exist $x, y \in S$ such that $z \in [x, y] \subseteq S$. (Why?) Therefore $(-\infty, b) \subseteq S$. If $b \in S$, then $S = (-\infty, b]$, and if $b \notin S$, then $S = (-\infty, b)$.

Cases (iii) and (iv) are left as exercises. Q.E.D.

Nested Intervals

We say that a sequence of intervals I_n, $n \in \mathbb{N}$, is **nested** if the following chain of inclusions holds (see Figure 2.5.1):

$$I_1 \supseteq I_2 \supseteq \cdots \supseteq I_n \supseteq I_{n+1} \supseteq \cdots$$

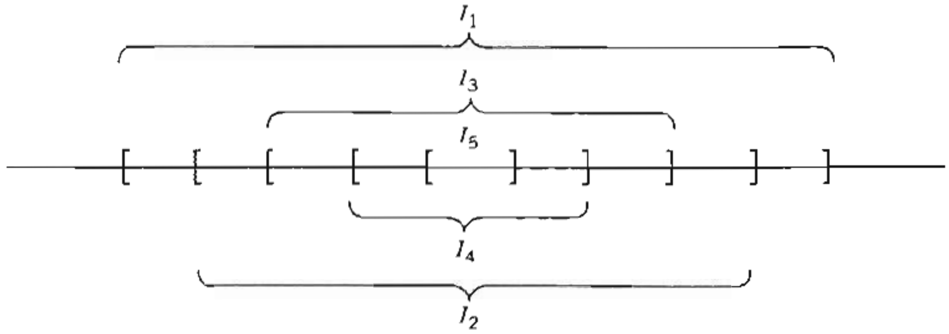


Figure 2.5.1 Nested intervals

For example, if $I_n := [0, 1/n]$ for $n \in \mathbb{N}$, then $I_n \supseteq I_{n+1}$ for each $n \in \mathbb{N}$ so that this sequence of intervals is nested. In this case, the element 0 belongs to all I_n and the Archimedean Property 2.4.5 can be used to show that 0 is the only such common point. (Prove this.) We denote this by writing $\bigcap_{n=1}^{\infty} I_n = \{0\}$.

It is important to realize that, in general, a nested sequence of intervals need *not* have a common point. For example, if $J_n := (0, 1/n)$ for $n \in \mathbb{N}$, then this sequence of intervals is nested, but there is no common point, since for every given $x > 0$, there exists (why?) $m \in \mathbb{N}$ such that $1/m < x$ so that $x \notin J_m$. Similarly, the sequence of intervals $K_n := (n, \infty)$, $n \in \mathbb{N}$, is nested but has no common point. (Why?)

However, it is an important property of $\mathbb{R}$ that every nested sequence of *closed, bounded* intervals does have a common point, as we will now prove. Notice that the completeness of $\mathbb{R}$ plays an essential role in establishing this property.

2.5.2 Nested Intervals Property *If $I_n = [a_n, b_n]$, $n \in \mathbb{N}$, is a nested sequence of closed bounded intervals, then there exists a number $\xi \in \mathbb{R}$ such that $\xi \in I_n$ for all $n \in \mathbb{N}$.*

Proof. Since the intervals are nested, we have $I_n \subseteq I_1$ for all $n \in \mathbb{N}$, so that $a_n \le b_1$ for all $n \in \mathbb{N}$. Hence, the nonempty set $\{a_n : n \in \mathbb{N}\}$ is bounded above, and we let ξ be its supremum. Clearly $a_n \le \xi$ for all $n \in \mathbb{N}$.

We claim also that $\xi \le b_n$ for all n. This is established by showing that for any particular n, the number b_n is an upper bound for the set $\{a_k : k \in \mathbb{N}\}$. We consider two cases. (i) If $n \le k$, then since $I_n \supseteq I_k$, we have $a_k \le b_k \le b_n$. (ii) If $k < n$, then since $I_k \supseteq I_n$, we have $a_k \le a_n \le b_n$. (See Figure 2.5.2.) Thus, we conclude that $a_k \le b_n$ for all k, so that b_n is an upper bound of the set $\{a_k : k \in \mathbb{N}\}$. Hence, $\xi \le b_n$ for each $n \in \mathbb{N}$. Since $a_n \le \xi \le b_n$ for all n, we have $\xi \in I_n$ for all $n \in \mathbb{N}$. Q.E.D.

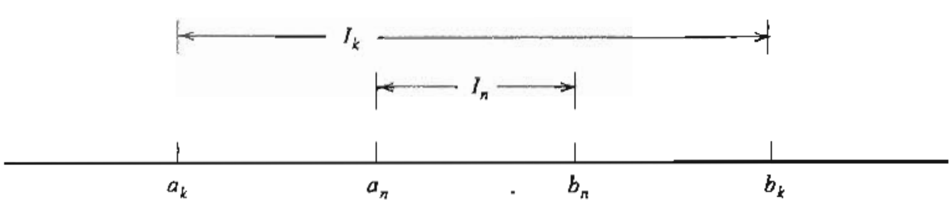


Figure 2.5.2 If $k < n$, then $I_n \subseteq I_k$

2.5.3 Theorem *If $I_n := [a_n, b_n]$, $n \in \mathbb{N}$, is a nested sequence of closed, bounded intervals such that the lengths $b_n - a_n$ of I_n satisfy*

$$\inf\{b_n - a_n : n \in \mathbb{N}\} = 0,$$

then the number ξ contained in I_n for all $n \in \mathbb{N}$ is unique.

Proof. If $\eta := \inf\{b_n : n \in \mathbb{N}\}$, then an argument similar to the proof of 2.5.2 can be used to show that $a_n \le \eta$ for all n, and hence that $\xi \le \eta$. In fact, it is an exercise (see Exercise 10) to show that $x \in I_n$ for all $n \in \mathbb{N}$ if and only if $\xi \le x \le \eta$. If we have $\inf\{b_n - a_n : n \in \mathbb{N}\} = 0$, then for any $\varepsilon > 0$, there exists an $m \in \mathbb{N}$ such that $0 \le \eta - \xi \le b_m - a_m < \varepsilon$. Since this holds for all $\varepsilon > 0$, it follows from Theorem 2.1.9 that $\eta - \xi = 0$. Therefore, we conclude that $\xi = \eta$ is the only point that belongs to I_n for every $n \in \mathbb{N}$. Q.E.D.

The Uncountability of $\mathbb{R}$

The concept of a countable set was discussed in Section 1.3 and the countability of the set $\mathbb{Q}$ of rational numbers was established there. We will now use the Nested Interval Property to prove that the set $\mathbb{R}$ is an *uncountable* set. The proof was given by Georg Cantor in 1874 in the first of his papers on infinite sets. He later published a proof that used decimal representations of real numbers, and that proof will be given later in this section.

2.5.4 Theorem *The set $\mathbb{R}$ of real numbers is not countable.*

Proof. We will prove that the unit interval $I := [0, 1]$ is an uncountable set. This implies that the set $\mathbb{R}$ is an uncountable set, for if $\mathbb{R}$ were countable, then the subset I would also be countable. (See Theorem 1.3.9(a).)

The proof is by contradiction. If we assume that I is countable, then we can enumerate the set as $I = \{x_1, x_2, \cdots, x_n, \cdots\}$. We first select a closed subinterval I_1 of I such that $x_1 \notin I_1$, then select a closed subinterval I_2 of I_1 such that $x_2 \notin I_2$, and so on. In this way, we obtain nonempty closed intervals

$$I_1 \supseteq I_2 \supseteq \cdots \supseteq I_n \supseteq \cdots$$

such that $I_n \subseteq I$ and $x_n \notin I_n$ for all n. The Nested Intervals Property 2.5.2 implies that there exists a point $\xi \in I$ such that $\xi \in I_n$ for all n. Therefore $\xi \ne x_n$ for all $n \in \mathbb{N}$, so the enumeration of I is not a complete listing of the elements of I, as claimed. Hence, I is an uncountable set. Q.E.D.

The fact that the set $\mathbb{R}$ of real numbers is uncountable can be combined with the fact that the set $\mathbb{Q}$ of rational numbers is countable to conclude that the set $\mathbb{R}\backslash\mathbb{Q}$ of irrational numbers is uncountable. Indeed, since the union of two countable sets is countable (see 1.3.7(c)), if $\mathbb{R}\backslash\mathbb{Q}$ is countable, then since $\mathbb{R} = \mathbb{Q} \cup (\mathbb{R}\backslash\mathbb{Q})$, we conclude that $\mathbb{R}$ is also a countable set, which is a contradiction. Therefore, the set of irrational numbers $\mathbb{R}\backslash\mathbb{Q}$ is an uncountable set.

†Binary Representations

We will digress briefly to discuss informally the binary (and decimal) representations of real numbers. It will suffice to consider real numbers between 0 and 1, since the representations for other real numbers can then be obtained by adding a positive or negative number.

†The remainder of this section can be omitted on a first reading.

If $x \in [0, 1]$, we will use a repeated bisection procedure to associate a sequence (a_n) of 0s and 1s as follows. If $x \neq \frac{1}{2}$ belongs to the left subinterval $\left[0, \frac{1}{2}\right]$ we take $a_1 := 0$, while if x belongs to the right subinterval $\left[\frac{1}{2}, 1\right]$ we take $a_1 = 1$. If $x = \frac{1}{2}$, then we may take a_1 to be either 0 or 1. In any case, we have

$$\frac{a_1}{2} \le x \le \frac{a_1 + 1}{2}.$$

We now bisect the interval $\left[\frac{1}{2}a_1, \frac{1}{2}(a_1 + 1)\right]$. If x is not the bisection point and belongs to the left subinterval we take $a_2 := 0$, and if x belongs to the right subinterval we take $a_2 := 1$. If $x = \frac{1}{4}$ or $x = \frac{3}{4}$, we can take a_2 to be either 0 or 1. In any case, we have

$$\frac{a_1}{2} + \frac{a_2}{2^2} \le x \le \frac{a_1}{2} + \frac{a_2 + 1}{2^2}.$$

We continue this bisection procedure, assigning at the nth stage the value $a_n := 0$ if x is not the bisection point and lies in the left subinterval, and assigning the value $a_n := 1$ if x lies in the right subinterval. In this way we obtain a sequence (a_n) of 0s or 1s that correspond to a nested sequence of intervals containing the point x. For each n, we have the inequality

$$\frac{a_1}{2} + \frac{a_2}{2^2} + \cdots + \frac{a_n}{2^n} \le x \le \frac{a_1}{2} + \frac{a_2}{2^2} + \cdots + \frac{a_n + 1}{2^n}. \tag{2}$$

If x is the bisection point at the nth stage, then $x = m/2^n$ with m odd. In this case, we may choose either the left or the right subinterval; however, once this subinterval is chosen, then all subsequent subintervals in the bisection procedure are determined. [For instance, if we choose the left subinterval so that $a_n = 0$, then x is the right endpoint of all subsequent subintervals, and hence $a_k = 1$ for all $k \ge n + 1$. On the other hand, if we choose the right subinterval so that $a_n = 1$, then x is the left endpoint of all subsequent subintervals, and hence $a_k = 0$ for all $k \ge n + 1$. For example, if $x = \frac{3}{4}$, then the two possible sequences for x are $1, 0, 1, 1, 1, \cdots$ and $1, 1, 0, 0, 0, \cdots$.]

To summarize: *If $x \in [0, 1]$, then there exists a sequence (a_n) of* 0s *and* 1s *such that inequality* (2) *holds for all $n \in \mathbb{N}$*. In this case we write

$$x = (.a_1 a_2 \cdots a_n \cdots)_2, \tag{3}$$

and call (3) a **binary representation** of x. This representation is unique except when $x = m/2^n$ for m odd, in which case x has the two representations

$$x = (.a_1 a_2 \cdots a_{n-1} 1000 \cdots)_2 = (.a_1 a_2 \cdots a_{n-1} 0111 \cdots)_2,$$

one ending in 0s and the other ending in 1s.

Conversely, each sequence of 0s *and* 1s *is the binary representation of a unique real number in* [0, 1]. The inequality corresponding to (2) determines a closed interval with length $1/2^n$ and the sequence of these intervals is nested. Therefore, Theorem 2.5.3 implies that there exists a unique real number x satisfying (2) for every $n \in \mathbb{N}$. Consequently, x has the binary representation $(.a_1 a_2 \cdots a_n \cdots)_2$.

Remark The concept of binary representation is extremely important in this era of digital computers. A number is entered in a digital computer on "bits", and each bit can be put in one of two states—either it will pass current or it will not. These two states correspond to the values 1 and 0, respectively. Thus, the binary representation of a number can be stored in a digital computer on a string of bits. Of course, in actual practice, since only finitely many bits can be stored, the binary representations must be truncated. If n binary digits are used for a number $x \in [0, 1]$, then the accuracy is at most $1/2^n$. For example, to assure four-decimal accuracy, it is necessary to use at least 15 binary digits (or 15 bits).

Decimal Representations

Decimal representations of real numbers are similar to binary representations, except that we subdivide intervals into *ten* equal subintervals instead of two.

Thus, given $x \in [0, 1]$, if we subdivide $[0, 1]$ into ten equal subintervals, then x belongs to a subinterval $[b_1/10, (b_1 + 1)/10]$ for some integer b_1 in $\{0, 1, \cdots, 9\}$. Proceeding as in the binary case, we obtain a sequence (b_n) of integers with $0 \le b_n \le 9$ for all $n \in \mathbb{N}$ such that x satisfies

$$\frac{b_1}{10} + \frac{b_2}{10^2} + \cdots + \frac{b_n}{10^n} \le x \le \frac{b_1}{10} + \frac{b_2}{10^2} + \cdots + \frac{b_n + 1}{10^n}. \tag{4}$$

In this case we say that x has a **decimal representation** given by

$$x = .b_1 b_2 \cdots b_n \cdots.$$

If $x \ge 1$ and if $B \in \mathbb{N}$ is such that $B \le x < B + 1$, then $x = B.b_1 b_2 \cdots b_n \cdots$ where the decimal representation of $x - B \in [0, 1]$ is as above. Negative numbers are treated similarly.

The fact that each decimal determines a unique real number follows from Theorem 2.5.3, since each decimal specifies a nested sequence of intervals with lengths $1/10^n$.

The decimal representation of $x \in [0, 1]$ is unique except when x is a subdivision point at some stage, which can be seen to occur when $x = m/10^n$ for some $m, n \in \mathbb{N}$, $1 \le m \le 10^n$. (We may also assume that m is not divisible by 10.) When x is a subdivision point at the nth stage, one choice for b_n corresponds to selecting the left subinterval, which causes all subsequent digits to be 9, and the other choice corresponds to selecting the right subinterval, which causes all subsequent digits to be 0. [For example, if $x = \frac{1}{2}$ then $x = .4999 \cdots = .5000 \cdots$, and if $y = 38/100$ then $y = .37999 \cdots = .38000 \cdots$.]

Periodic Decimals

A decimal $B.b_1 b_2 \cdots b_n \cdots$ is said to be **periodic** (or to be **repeating**), if there exist $k, n \in \mathbb{N}$ such that $b_n = b_{n+m}$ for all $n \ge k$. In this case, the block of digits $b_k b_{k+1} \cdots b_{k+m-1}$ is repeated once the kth digit is reached. The smallest number m with this property is called the **period** of the decimal. For example, $19/88 = .2159090 \cdots 90 \cdots$ has period $m = 2$ with repeating block 90 starting at $k = 4$. A **terminating decimal** is a periodic decimal where the repeated block is simply the digit 0.

We will give an informal proof of the assertion: A *positive real number is rational if and only if its decimal representation is periodic.*

For, suppose that $x = p/q$ where $p, q \in \mathbb{N}$ have no common integer factors. For convenience we will also suppose that $0 < p < q$. We note that the process of "long division" of q into p gives the decimal representation of p/q. Each step in the division process produces a remainder that is an integer from 0 to $q - 1$. Therefore, after at most q steps, some remainder will occur a second time and, at that point, the digits in the quotient will begin to repeat themselves in cycles. Hence, the decimal representation of such a rational number is periodic.

Conversely, if a decimal is periodic, then it represents a rational number. The idea of the proof is best illustrated by an example. Suppose that $x = 7.31414 \cdots 14 \cdots$. We multiply by a power of 10 to move the decimal point to the first repeating block; here obtaining $10x = 73.1414 \cdots$. We now multiply by a power of 10 to move one block to the left of the decimal point; here getting $1000x = 7314.1414 \cdots$. We now subtract to obtain an integer; here getting $1000x - 10x = 7314 - 73 = 7241$, whence $x = 7241/990$, a rational number.

Cantor's Second Proof

We will now give Cantor's second proof of the uncountability of $\mathbb{R}$. This is the elegant "diagonal" argument based on decimal representations of real numbers.

2.5.5 Theorem *The unit interval* $[0, 1] := \{x \in \mathbb{R} : 0 \le x \le 1\}$ *is not countable.*

Proof. The proof is by contradiction. We will use the fact that every real number $x \in [0, 1]$ has a decimal representation $x = 0.b_1 b_2 b_3 \cdots$, where $b_i = 0, 1 \cdots, 9$. Suppose that there is an enumeration $x_1, x_2, x_3 \cdots$ of all numbers in $[0, 1]$, which we display as:

$$\begin{aligned}
x_1 &= 0.b_{11}b_{12}b_{13}\cdots b_{1n}\cdots,\\
x_2 &= 0.b_{21}b_{22}b_{23}\cdots b_{2n}\cdots,\\
x_3 &= 0.b_{31}b_{32}b_{33}\cdots b_{3n}\cdots,\\
&\cdots \qquad \cdots\\
x_n &= 0.b_{n1}b_{n2}b_{n3}\cdots b_{nn}\cdots,\\
&\cdots \qquad \cdots
\end{aligned}$$

We now define a real number $y := 0.y_1 y_2 y_3 \cdots y_n \cdots$ by setting $y_1 := 2$ if $b_{11} \ge 5$ and $y_1 := 7$ if $b_{11} \le 4$; in general, we let

$$y_n := \begin{cases} 2 & \text{if } b_{nn} \ge 5, \\ 7 & \text{if } b_{nn} \le 4. \end{cases}$$

Then $y \in [0, 1]$. Note that the number y is not equal to any of the numbers with two decimal representations, since $y_n \ne 0, 9$ for all $n \in \mathbb{N}$. Further, since y and x_n differ in the nth decimal place, then $y \ne x_n$ for any $n \in \mathbb{N}$. Therefore, y is not included in the enumeration of $[0, 1]$, contradicting the hypothesis. Q.E.D.

Exercises for Section 2.5

1. If $I := [a, b]$ and $I' := [a', b']$ are closed intervals in $\mathbb{R}$, show that $I \subseteq I'$ if and only if $a' \le a$ and $b \le b'$.
2. If $S \subseteq \mathbb{R}$ is nonempty, show that S is bounded if and only if there exists a closed bounded interval I such that $S \subseteq I$.
3. If $S \subseteq \mathbb{R}$ is a nonempty bounded set, and $I_S := [\inf S, \sup S]$, show that $S \subseteq I_S$. Moreover, if J is any closed bounded interval containing S, show that $I_S \subseteq J$.
4. In the proof of Case (ii) of Theorem 2.5.1, explain why x, y exist in S.
5. Write out the details of the proof of case (iv) in Theorem 2.5.1.
6. If $I_1 \supseteq I_2 \supseteq \cdots \supseteq I_n \supseteq \cdots$ is a nested sequence of intervals and if $I_n = [a_n, b_n]$, show that $a_1 \le a_2 \le \cdots \le a_n \le \cdots$ and $b_1 \ge b_2 \ge \cdots \ge b_n \ge \cdots$.
7. Let $I_n := [0, 1/n]$ for $n \in \mathbb{N}$. Prove that $\bigcap_{n=1}^{\infty} I_n = \{0\}$.
8. Let $J_n := (0, 1/n)$ for $n \in \mathbb{N}$. Prove that $\bigcap_{n=1}^{\infty} J_n = \emptyset$.
9. Let $K_n := (n, \infty)$ for $n \in \mathbb{N}$. Prove that $\bigcap_{n=1}^{\infty} K_n = \emptyset$.

10. With the notation in the proofs of Theorems 2.5.2 and 2.5.3, show that we have $\eta \in \bigcap_{n=1}^{\infty} I_n$. Also show that $[\xi, \eta] = \bigcap_{n=1}^{\infty} I_n$.
11. Show that the intervals obtained from the inequalities in (2) form a nested sequence.
12. Give the two binary representations of $\frac{3}{8}$ and $\frac{7}{16}$.
13. (a) Give the first four digits in the binary representation of $\frac{1}{3}$.
    (b) Give the complete binary representation of $\frac{1}{3}$.
14. Show that if $a_k, b_k \in \{0, 1, \cdots, 9\}$ and if
$$\frac{a_1}{10} + \frac{a_2}{10^2} + \cdots + \frac{a_n}{10^n} = \frac{b_1}{10} + \frac{b_2}{10^2} + \cdots + \frac{b_m}{10^m} \neq 0,$$
then $n = m$ and $a_k = b_k$ for $k = 1, \cdots, n$.
15. Find the decimal representation of $-\frac{2}{7}$.
16. Express $\frac{1}{7}$ and $\frac{2}{19}$ as periodic decimals.
17. What rationals are represented by the periodic decimals $1.25137 \cdots 137 \cdots$ and $35.14653 \cdots 653 \cdots$?

CHAPTER 3

SEQUENCES AND SERIES

Now that the foundations of the real number system $\mathbb{R}$ have been laid, we are prepared to pursue questions of a more analytic nature, and we will begin with a study of the convergence of sequences. Some of the early results may be familiar to the reader from calculus, but the presentation here is intended to be rigorous and will lead to certain more profound theorems than are usually discussed in earlier courses.

We will first introduce the meaning of the convergence of a sequence of real numbers and establish some basic, but useful, results about convergent sequences. We then present some deeper results concerning the convergence of sequences. These include the Monotone Convergence Theorem, the Bolzano-Weierstrass Theorem, and the Cauchy Criterion for convergence of sequences. It is important for the reader to learn both the theorems and how the theorems apply to special sequences.

Because of the linear limitations inherent in a book it is necessary to decide where to locate the subject of infinite series. It would be reasonable to follow this chapter with a full discussion of infinite series, but this would delay the important topics of continuity, differentiation, and integration. Consequently, we have decided to compromise. A brief introduction to infinite series is given in Section 3.7 at the end of this chapter, and a more extensive treatment is given later in Chapter 9. Thus readers who want a fuller discussion of series at this point can move to Chapter 9 after completing this chapter.

Augustin-Louis Cauchy

Augustin-Louis Cauchy (1789–1857) was born in Paris just after the start of the French Revolution. His father was a lawyer in the Paris police department, and the family was forced to flee during the Reign of Terror. As a result, Cauchy's early years were difficult and he developed strong anti-revolutionary and pro-royalist feelings. After returning to Paris, Cauchy's father became secretary to the newly-formed Senate, which included the mathematicians Laplace and Lagrange. They were impressed by young Cauchy's mathematical talent and helped him begin his career.

He entered the École Polytechnique in 1805 and soon established a reputation as an exceptional mathematician. In 1815, the year royalty was restored, he was appointed to the faculty of the École Polytechnique, but his strong political views and his uncompromising standards in mathematics often resulted in bad relations with his colleagues. After the July revolution of 1830, Cauchy refused to sign the new loyalty oath and left France for eight years in self-imposed exile. In 1838, he accepted a minor teaching post in Paris, and in 1848 Napoleon III reinstated him to his former position at the École Polytechnique, where he remained until his death.

Cauchy was amazingly versatile and prolific, making substantial contributions to many areas, including real and complex analysis, number theory, differential equations, mathematical physics and probability. He published eight books and 789 papers, and his collected works fill 26 volumes. He was one of the most important mathematicians in the first half of the nineteenth century.

Section 3.1 Sequences and Their Limits

A sequence in a set S is a function whose domain is the set $\mathbb{N}$ of natural numbers, and whose range is contained in the set S. In this chapter, we will be concerned with sequences in $\mathbb{R}$ and will discuss what we mean by the convergence of these sequences.

3.1.1 Definition **A sequence of real numbers** (or a **sequence in $\mathbb{R}$**) is a function defined on the set $\mathbb{N} = \{1, 2, \cdots\}$ of natural numbers whose range is contained in the set $\mathbb{R}$ of real numbers.

In other words, a sequence in $\mathbb{R}$ assigns to each natural number $n = 1, 2, \cdots$ a uniquely determined real number. If $X : \mathbb{N} \to \mathbb{R}$ is a sequence, we will usually denote the value of X at n by the symbol x_n rather than using the function notation $X(n)$. The values x_n are also called the **terms** or the **elements** of the sequence. We will denote this sequence by the notations

$$X, \qquad (x_n), \qquad (x_n : n \in \mathbb{N}).$$

Of course, we will often use other letters, such as $Y = (y_k)$, $Z = (z_i)$, and so on, to denote sequences.

We purposely use parentheses to emphasize that the ordering induced by the natural order of $\mathbb{N}$ is a matter of importance. Thus, we distinguish notationally between the sequence $(x_n : n \in \mathbb{N})$, whose infinitely many terms have an ordering, and the set of values $\{x_n : n \in \mathbb{N}\}$ in the range of the sequence which are not ordered. For example, the sequence $X := ((-1)^n : n \in \mathbb{N})$ has infinitely many terms that alternate between -1 and 1, whereas the set of values $\{(-1)^n : n \in \mathbb{N}\}$ is equal to the set $\{-1, 1\}$, which has only two elements.

Sequences are often defined by giving a formula for the nth term x_n. Frequently, it is convenient to list the terms of a sequence in order, stopping when the rule of formation seems evident. For example, we may define the sequence of reciprocals of the even numbers by writing

$$X := \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{6}, \frac{1}{8}, \cdots\right),$$

though a more satisfactory method is to specify the formula for the general term and write

$$X := \left(\frac{1}{2n} : n \in \mathbb{N}\right)$$

or more simply $X = (1/2n)$.

Another way of defining a sequence is to specify the value of x_1 and give a formula for x_{n+1} $(n \geq 1)$ in terms of x_n. More generally, we may specify x_1 and give a formula for obtaining x_{n+1} from $x_1, x_2, \cdots, x_n$. Sequences defined in this manner are said to be **inductively** (or **recursively**) defined.

3.1.2 Examples (a) If $b \in \mathbb{R}$, the sequence $B := (b, b, b, \cdots)$, all of whose terms equal b, is called the **constant sequence** b. Thus the constant sequence 1 is the sequence $(1, 1, 1, \cdots)$, and the constant sequence 0 is the sequence $(0, 0, 0, \cdots)$.

(b) If $b \in \mathbb{R}$, then $B := (b^n)$ is the sequence $B = (b, b^2, b^3, \cdots, b^n, \cdots)$. In particular, if $b = \frac{1}{2}$, then we obtain the sequence

$$\left(\frac{1}{2^n} : n \in \mathbb{N}\right) = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \cdots, \frac{1}{2^n}, \cdots\right).$$

(c) The sequence of $(2n : n \in \mathbb{N})$ of even natural numbers can be defined inductively by

$$x_1 := 2, \quad x_{n+1} := x_n + 2,$$

or by the definition

$$y_1 := 2, \quad y_{n+1} := y_1 + y_n.$$

(d) The celebrated **Fibonacci sequence** $F := (f_n)$ is given by the inductive definition

$$f_1 := 1, \quad f_2 := 1, \quad f_{n+1} := f_{n-1} + f_n \quad (n \geq 2).$$

Thus each term past the second is the sum of its two immediate predecessors. The first ten terms of F are seen to be $(1, 1, 2, 3, 5, 8, 13, 21, 34, 55, \cdots)$. □

The Limit of a Sequence

There are a number of different limit concepts in real analysis. The notion of limit of a sequence is the most basic, and it will be the focus of this chapter.

3.1.3 Definition A sequence $X = (x_n)$ in $\mathbb{R}$ is said to **converge** to $x \in \mathbb{R}$, or x is said to be a **limit** of (x_n), if for every $\varepsilon > 0$ there exists a natural number $K(\varepsilon)$ such that for all $n \geq K(\varepsilon)$, the terms x_n satisfy $|x_n - x| < \varepsilon$.

If a sequence has a limit, we say that the sequence is **convergent**; if it has no limit, we say that the sequence is **divergent**.

Note The notation $K(\varepsilon)$ is used to emphasize that the choice of K depends on the value of ε. However, it is often convenient to write K instead of $K(\varepsilon)$. In most cases, a "small" value of ε will usually require a "large" value of K to guarantee that the distance $|x_n - x|$ between x_n and x is less than ε for all $n \geq K = K(\varepsilon)$.

When a sequence has limit x, we will use the notation

$$\lim X = x \qquad \text{or} \qquad \lim(x_n) = x.$$

We will sometimes use the symbolism $x_n \to x$, which indicates the intuitive idea that the values x_n "approach" the number x as $n \to \infty$.

3.1.4 Uniqueness of Limits *A sequence in $\mathbb{R}$ can have at most one limit.*

Proof. Suppose that x' and x'' are both limits of (x_n). For each $\varepsilon > 0$ there exist K' such that $|x_n - x'| < \varepsilon/2$ for all $n \geq K'$, and there exists K'' such that $|x_n - x''| < \varepsilon/2$ for all $n \geq K''$. We let K be the larger of K' and K''. Then for $n \geq K$ we apply the Triangle Inequality to get

$$\begin{aligned} |x' - x''| &= |x' - x_n + x_n - x''| \\ &\leq |x' - x_n| + |x_n - x''| < \varepsilon/2 + \varepsilon/2 = \varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ is an arbitrary positive number, we conclude that $x' - x'' = 0$. Q.E.D.

For $x \in \mathbb{R}$ and $\varepsilon > 0$, recall that the ε-neighborhood of x is the set

$$V_\varepsilon(x) := \{u \in \mathbb{R} : |u - x| < \varepsilon\}.$$

(See Section 2.2.) Since $u \in V_\varepsilon(x)$ is equivalent to $|u - x| < \varepsilon$, the definition of convergence of a sequence can be formulated in terms of neighborhoods. We give several different ways of saying that a sequence x_n converges to x in the following theorem.

3.1.5 Theorem *Let $X = (x_n)$ be a sequence of real numbers, and let $x \in \mathbb{R}$. The following statements are equivalent.*

(a) *X converges to x.*

(b) *For every $\varepsilon > 0$, there exists a natural number K such that for all $n \geq K$, the terms x_n satisfy $|x_n - x| < \varepsilon$.*

(c) *For every $\varepsilon > 0$, there exists a natural number K such that for all $n \geq K$, the terms x_n satisfy $x - \varepsilon < x_n < x + \varepsilon$.*

(d) *For every ε-neighborhood $V_\varepsilon(x)$ of x, there exists a natural number K such that for all $n \geq K$, the terms x_n belong to $V_\varepsilon(x)$.*

Proof. The equivalence of (a) and (b) is just the definition. The equivalence of (b), (c), and (d) follows from the following implications:

$$|u - x| < \varepsilon \iff -\varepsilon < u - x < \varepsilon \iff x - \varepsilon < u < x + \varepsilon \iff u \in V_\varepsilon(x).$$

Q.E.D.

With the language of neighborhoods, one can describe the convergence of the sequence $X = (x_n)$ to the number x by saying: *for each ε-neighborhood $V_\varepsilon(x)$ of x, all but a finite number of terms of X belong to $V_\varepsilon(x)$.* The finite number of terms that may not belong to the ε-neighborhood are the terms $x_1, x_2, \cdots, x_{K-1}$.

Remark The definition of the limit of a sequence of real numbers is used to verify that a proposed value x is indeed the limit. It does *not* provide a means for initially determining what that value of x might be. Later results will contribute to this end, but quite often it is necessary in practice to arrive at a conjectured value of the limit by direct calculation of a number of terms of the sequence. Computers can be helpful in this respect, but since they can calculate only a finite number of terms of a sequence, such computations do not in any way constitute a proof of the value of the limit.

The following examples illustrate how the definition is applied to prove that a sequence has a particular limit. In each case, a positive ε is given and we are required to find a K, depending on ε, as required by the definition.

3.1.6 Examples (a) $\lim(1/n) = 0$.

If $\varepsilon > 0$ is given, then $1/\varepsilon > 0$. By the Archimedean Property 2.4.5, there is a natural number $K = K(\varepsilon)$ such that $1/K < \varepsilon$. Then, if $n \geq K$, we have $1/n \leq 1/K < \varepsilon$. Consequently, if $n \geq K$, then

$$\left|\frac{1}{n} - 0\right| = \frac{1}{n} < \varepsilon.$$

Therefore, we can assert that the sequence $(1/n)$ converges to 0.

(b) $\lim(1/(n^2+1)) = 0$.

Let $\varepsilon > 0$ be given. To find K, we first note that if $n \in \mathbb{N}$, then

$$\frac{1}{n^2+1} < \frac{1}{n^2} \leq \frac{1}{n}.$$

Now choose K such that $1/K < \varepsilon$, as in (a) above. Then $n \geq K$ implies that $1/n < \varepsilon$, and therefore

$$\left|\frac{1}{n^2+1} - 0\right| = \frac{1}{n^2+1} < \frac{1}{n} < \varepsilon.$$

Hence, we have shown that the limit of the sequence is zero.

(c) $\lim\left(\dfrac{3n+2}{n+1}\right) = 3$.

Given $\varepsilon > 0$, we want to obtain the inequality

$$\left|\frac{3n+2}{n+1} - 3\right| < \varepsilon \tag{1}$$

when n is sufficiently large. We first simplify the expression on the left:

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

Now if the inequality $1/n < \varepsilon$ is satisfied, then the inequality (1) holds. Thus if $1/K < \varepsilon$, then for any $n \geq K$, we also have $1/n < \varepsilon$ and hence (1) holds. Therefore the limit of the sequence is 3.

(d) If $0 < b < 1$, then $\lim(b^n) = 0$.

We will use elementary properties of the natural logarithm function. If $\varepsilon > 0$ is given, we see that

$$b^n < \varepsilon \iff n \ln b < \ln \varepsilon \iff n > \ln \varepsilon / \ln b.$$

(The last inequality is reversed because $\ln b < 0$.) Thus if we choose K to be a number such that $K > \ln \varepsilon / \ln b$, then we will have $0 < b^n < \varepsilon$ for all $n \geq K$. Thus we have $\lim(b^n) = 0$.

For example, if $b = .8$, and if $\varepsilon = .01$ is given, then we would need $K > \ln .01 / \ln .8 \approx 20.6377$. Thus $K = 21$ would be an appropriate choice for $\varepsilon = .01$. □

Remark **The $K(\varepsilon)$ Game** In the notion of convergence of a sequence, one way to keep in mind the connection between the ε and the K is to think of it as a game called the $K(\varepsilon)$ Game. In this game, Player A asserts that a certain number x is the limit of a sequence (x_n). Player B challenges this assertion by giving Player A a specific value for $\varepsilon > 0$. Player A must respond to the challenge by coming up with a value of K such that $|x_n - x| < \varepsilon$ for all $n > K$. If Player A can always find a value of K that works, then he wins, and the sequence is convergent. However, if Player B can give a specific value of $\varepsilon > 0$ for which Player A cannot respond adequately, then Player B wins, and we conclude that the sequence does not converge to x.

In order to show that a sequence $X = (x_n)$ does *not* converge to the number x, it is enough to produce one number $\varepsilon_0 > 0$ such that no matter what natural number K is chosen, one can find a particular n_K satisfying $n_K \geq K$ such that $|x_{n_K} - x| \geq \varepsilon_0$. (This will be discussed in more detail in Section 3.4.)

3.1.7 Example The sequence $(0, 2, 0, 2, \cdots, 0, 2, \cdots)$ does *not* converge to the number 0.

If Player A asserts that 0 is the limit of the sequence, he will lose the $K(\varepsilon)$ Game when Player B gives him a value of $\varepsilon < 2$. To be definite, let Player B give Player A the value $\varepsilon_0 = 1$. Then no matter what value Player A chooses for K, his response will not be adequate, for Player B will respond by selecting an even number $n > K$. Then the corresponding value is $x_n = 2$ so that $|x_n - 0| = 2 > 1 = \varepsilon_0$. Thus the number 0 is not the limit of the sequence. □

Tails of Sequences

It is important to realize that the convergence (or divergence) of a sequence $X = (x_n)$ depends only on the "ultimate behavior" of the terms. By this we mean that if, for any natural number m, we drop the first m terms of the sequence, then the resulting sequence X_m converges if and only if the original sequence converges, and in this case, the limits are the same. We will state this formally after we introduce the idea of a "tail" of a sequence.

3.1.8 Definition If $X = (x_1, x_2, \cdots, x_n, \cdots)$ is a sequence of real numbers and if m is a given natural number, then the **m-tail** of X is the sequence

$$X_m := (x_{m+n} : n \in \mathbb{N}) = (x_{m+1}, x_{m+2}, \cdots)$$

For example, the 3-tail of the sequence $X = (2, 4, 6, 8, 10, \cdots, 2n, \cdots)$, is the sequence $X_3 = (8, 10, 12, \cdots, 2n + 6, \cdots)$.

3.1.9 Theorem *Let $X = (x_n : n \in \mathbb{N})$ be a sequence of real numbers and let $m \in \mathbb{N}$. Then the m-tail $X_m = (x_{m+n} : n \in \mathbb{N})$ of X converges if and only if X converges. In this case, $\lim X_m = \lim X$.*

Proof. We note that for any $p \in \mathbb{N}$, the pth term of X_m is the $(p + m)$th term of X. Similarly, if $q > m$, then the qth term of X is the $(q - m)$th term of X_m.

Assume X converges to x. Then given any $\varepsilon > 0$, if the terms of X for $n \ge K(\varepsilon)$ satisfy $|x_n - x| < \varepsilon$, then the terms of X_m for $k \ge K(\varepsilon) - m$ satisfy $|x_k - x| < \varepsilon$. Thus we can take $K_m(\varepsilon) = K(\varepsilon) - m$, so that X_m also converges to x.

Conversely, if the terms of X_m for $k \ge K_m(\varepsilon)$ satisfy $|x_k - x| < \varepsilon$, then the terms of X for $n \ge K(\varepsilon) + m$ satisfy $|x_n - x| < \varepsilon$. Thus we can take $K(\varepsilon) = K_m(\varepsilon) + m$.

Therefore, X converges to x if and only if X_m converges to x. Q.E.D.

We shall sometimes say that a sequence X *ultimately* has a certain property if some tail of X has this property. For example, we say that the sequence $(3, 4, 5, 5, 5, \cdots, 5, \cdots)$ is "ultimately constant". On the other hand, the sequence $(3, 5, 3, 5, \cdots, 3, 5, \cdots)$ is not ultimately constant. The notion of convergence can be stated using this terminology: A sequence X converges to x if and only if the terms of X are ultimately in every ε-neighborhood of x. Other instances of this "ultimate terminology" will be noted below.

Further Examples

In establishing that a number x is the limit of a sequence (x_n), we often try to simplify the difference $|x_n - x|$ before considering an $\varepsilon > 0$ and finding a $K(\varepsilon)$ as required by the definition of limit. This was done in some of the earlier examples. The next result is a more formal statement of this idea, and the examples that follow make use of this approach.

3.1.10 Theorem *Let (x_n) be a sequence of real numbers and let $x \in \mathbb{R}$. If (a_n) is a sequence of positive real numbers with $\lim(a_n) = 0$ and if for some constant $C > 0$ and some $m \in \mathbb{N}$ we have*

$$|x_n - x| \le Ca_n \qquad \textit{for all} \quad n \ge m,$$

then it follows that $\lim(x_n) = x$.

Proof. If $\varepsilon > 0$ is given, then since $\lim(a_n) = 0$, we know there exists $K = K(\varepsilon/C)$ such that $n \ge K$ implies

$$a_n = |a_n - 0| < \varepsilon/C.$$

Therefore it follows that if both $n \ge K$ and $n \ge m$, then

$$|x_n - x| \le Ca_n < C(\varepsilon/C) = \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, we conclude that $x = \lim(x_n)$. Q.E.D.

3.1.11 Examples **(a)** If $a > 0$, then $\lim\left(\dfrac{1}{1+na}\right) = 0$.

Since $a > 0$, then $0 < na < 1 + na$, and therefore $0 < 1/(1+na) < 1/(na)$. Thus we have

$$\left|\frac{1}{1+na} - 0\right| \le \left(\frac{1}{a}\right)\frac{1}{n} \qquad \text{for all} \quad n \in \mathbb{N}.$$

Since $\lim(1/n) = 0$, we may invoke Theorem 3.1.10 with $C = 1/a$ and $m = 1$ to infer that $\lim(1/(1+na)) = 0$.

(b) If $0 < b < 1$, then $\lim(b^n) = 0$.

This limit was obtained earlier in Example 3.1.6(d). We will give a second proof that illustrates the use of Bernoulli's Inequality (see Example 2.1.13(c)).

Since $0 < b < 1$, we can write $b = 1/(1+a)$, where $a := (1/b) - 1$ so that $a > 0$. By Bernoulli's Inequality, we have $(1+a)^n \ge 1 + na$. Hence

$$0 < b^n = \frac{1}{(1+a)^n} \le \frac{1}{1+na} < \frac{1}{na}.$$

Thus from Theorem 3.1.10 we conclude that $\lim(b^n) = 0$.

In particular, if $b = .8$, so that $a = .25$, and if we are given $\varepsilon = .01$, then the preceding inequality gives us $K(\varepsilon) = 4/(.01) = 400$. Comparing with Example 3.1.6(d), where we obtained $K = 25$, we see this method of estimation does not give us the "best" value of K. However, for the purpose of establishing the limit, the size of K is immaterial.

(c) If $c > 0$, then $\lim(c^{1/n}) = 1$.

The case $c = 1$ is trivial, since then $(c^{1/n})$ is the constant sequence $(1, 1, \cdots)$, which evidently converges to 1.

If $c > 1$, then $c^{1/n} = 1 + d_n$ for some $d_n > 0$. Hence by Bernoulli's Inequality 2.1.13(c),

$$c = (1 + d_n)^n \ge 1 + nd_n \qquad \text{for} \quad n \in \mathbb{N}.$$

Therefore we have $c - 1 \ge nd_n$, so that $d_n \le (c-1)/n$. Consequently we have

$$\left|c^{1/n} - 1\right| = d_n \le (c-1)\frac{1}{n} \qquad \text{for} \quad n \in \mathbb{N}.$$

We now invoke Theorem 3.1.10 to infer that $\lim(c^{1/n}) = 1$ when $c > 1$.

Now suppose that $0 < c < 1$; then $c^{1/n} = 1/(1 + h_n)$ for some $h_n > 0$. Hence Bernoulli's Inequality implies that

$$c = \frac{1}{(1 + h_n)^n} \le \frac{1}{1 + nh_n} < \frac{1}{nh_n},$$

from which it follows that $0 < h_n < 1/nc$ for $n \in \mathbb{N}$. Therefore we have

$$0 < 1 - c^{1/n} = \frac{h_n}{1 + h_n} < h_n < \frac{1}{nc}$$

so that

$$\left|c^{1/n} - 1\right| < \left(\frac{1}{c}\right)\frac{1}{n} \qquad \text{for} \quad n \in \mathbb{N}.$$

We now apply Theorem 3.1.10 to infer that $\lim(c^{1/n}) = 1$ when $0 < c < 1$.

(d) $\lim(n^{1/n}) = 1$

Since $n^{1/n} > 1$ for $n > 1$, we can write $n^{1/n} = 1 + k_n$ for some $k_n > 0$ when $n > 1$. Hence $n = (1 + k_n)^n$ for $n > 1$. By the Binomial Theorem, if $n > 1$ we have

$$n = 1 + nk_n + \tfrac{1}{2}n(n-1)k_n^2 + \cdots \ge 1 + \tfrac{1}{2}n(n-1)k_n^2,$$

whence it follows that

$$n - 1 \ge \tfrac{1}{2}n(n-1)k_n^2.$$

Hence $k_n^2 \le 2/n$ for $n > 1$. If $\varepsilon > 0$ is given, it follows from the Archimedean Property that there exists a natural number N_ε such that $2/N_\varepsilon < \varepsilon^2$. It follows that if $n \ge \sup\{2, N_\varepsilon\}$ then $2/n < \varepsilon^2$, whence

$$0 < n^{1/n} - 1 = k_n \le (2/n)^{1/2} < \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, we deduce that $\lim(n^{1/n}) = 1$. □

Exercises for Section 3.1

1. The sequence (x_n) is defined by the following formulas for the nth term. Write the first five terms in each case:
   (a) $x_n := 1 + (-1)^n$,
   (b) $x_n := (-1)^n/n$,
   (c) $x_n := \dfrac{1}{n(n+1)}$,
   (d) $x := \dfrac{1}{n^2 + 2}$.
2. The first few terms of a sequence (x_n) are given below. Assuming that the "natural pattern" indicated by these terms persists, give a formula for the nth term x_n.
   (a) $5, 7, 9, 11, \cdots$,
   (b) $1/2, -1/4, 1/8, -1/16, \cdots$,
   (c) $1/2, 2/3, 3/4, 4/5, \cdots$,
   (d) $1, 4, 9, 16, \cdots$.
3. List the first five terms of the following inductively defined sequences.
   (a) $x_1 := 1, \quad x_{n+1} = 3x_n + 1$,
   (b) $y_1 := 2, \quad y_{n+1} = \frac{1}{2}(y_n + 2/y_n)$,
   (c) $z_1 := 1, \quad z_2 := 2, \quad z_{n+2} := (z_{n+1} + z_n)/(z_{n+1} - z_n)$,
   (d) $s_1 = 3, \quad s_2 := 5, \quad s_{n+2} := s_n + s_{n+1}$.
4. For any $b \in \mathbb{R}$, prove that $\lim(b/n) = 0$.

5. Use the definition of the limit of a sequence to establish the following limits.
   (a) $\lim\left(\dfrac{n}{n^2+1}\right) = 0$, (b) $\lim\left(\dfrac{2n}{n+1}\right) = 2$,
   (c) $\lim\left(\dfrac{3n+1}{2n+5}\right) = \dfrac{3}{2}$, (d) $\lim\left(\dfrac{n^2-1}{2n^2+3}\right) = \dfrac{1}{2}$.
6. Show that
   (a) $\lim\left(\dfrac{1}{\sqrt{n+7}}\right) = 0$, (b) $\lim\left(\dfrac{2n}{n+2}\right) = 2$,
   (c) $\lim\left(\dfrac{\sqrt{n}}{n+1}\right) = 0$, (d) $\lim\left(\dfrac{(-1)^n n}{n^2+1}\right) = 0$.
7. Let $x_n := 1/\ln(n+1)$ for $n \in \mathbb{N}$.
   (a) Use the definition of limit to show that $\lim(x_n) = 0$.
   (b) Find a specific value of $K(\varepsilon)$ as required in the definition of limit for each of (i) $\varepsilon = 1/2$, and (ii) $\varepsilon = 1/10$.
8. Prove that $\lim(x_n) = 0$ if and only if $\lim(|x_n|) = 0$. Give an example to show that the convergence of $(|x_n|)$ need not imply the convergence of (x_n).
9. Show that if $x_n \geq 0$ for all $n \in \mathbb{N}$ and $\lim(x_n) = 0$, then $\lim\left(\sqrt{x_n}\right) = 0$.
10. Prove that if $\lim(x_n) = x$ and if $x > 0$, then there exists a natural number M such that $x_n > 0$ for all $n \geq M$.
11. Show that $\lim\left(\dfrac{1}{n} - \dfrac{1}{n+1}\right) = 0$.
12. Show that $\lim(1/3^n) = 0$.
13. Let $b \in \mathbb{R}$ satisfy $0 < b < 1$. Show that $\lim(nb^n) = 0$. [*Hint*: Use the Binomial Theorem as in Example 3.1.11(d).]
14. Show that $\lim\left((2n)^{1/n}\right) = 1$.
15. Show that $\lim(n^2/n!) = 0$.
16. Show that $\lim(2^n/n!) = 0$. [*Hint*: if $n \geq 3$, then $0 < 2^n/n! \leq 2\left(\frac{2}{3}\right)^{n-2}$.]
17. If $\lim(x_n) = x > 0$, show that there exists a natural number K such that if $n \geq K$, then $\frac{1}{2}x < x_n < 2x$.

Section 3.2 Limit Theorems

In this section we will obtain some results that enable us to evaluate the limits of certain sequences of real numbers. These results will expand our collection of convergent sequences rather extensively. We begin by establishing an important property of convergent sequences that will be needed in this and later sections.

3.2.1 Definition A sequence $X = (x_n)$ of real numbers is said to be **bounded** if there exists a real number $M > 0$ such that $|x_n| \leq M$ for all $n \in \mathbb{N}$.

Thus, the sequence (x_n) is bounded if and only if the set $\{x_n : n \in \mathbb{N}\}$ of its values is a bounded subset of $\mathbb{R}$.

3.2.2 Theorem *A convergent sequence of real numbers is bounded.*

Proof. Suppose that $\lim(x_n) = x$ and let $\varepsilon := 1$. Then there exists a natural number $K = K(1)$ such that $|x_n - x| < 1$ for all $n \geq K$. If we apply the Triangle Inequality with $n \geq K$ we obtain

$$|x_n| = |x_n - x + x| \leq |x_n - x| + |x| < 1 + |x|.$$

If we set

$$M := \sup\left\{|x_1|, |x_2|, \cdots, |x_{K-1}|, 1 + |x|\right\},$$

then it follows that $|x_n| \leq M$ for all $n \in \mathbb{N}$. Q.E.D.

We will now examine how the limit process interacts with the operations of addition, subtraction, multiplication, and division of sequences. If $X = (x_n)$ and $Y = (y_n)$ are sequences of real numbers, then we define their **sum** to be the sequence $X + Y := (x_n + y_n)$, their **difference** to be the sequence $X - Y := (x_n - y_n)$, and their **product** to be the sequence $X \cdot Y := (x_n y_n)$. If $c \in \mathbb{R}$, we define the **multiple** of X by c to be the sequence $cX := (cx_n)$. Finally, if $Z = (z_n)$ is a sequence of real numbers with $z_n \neq 0$ for all $n \in \mathbb{N}$, then we define the **quotient** of X and Z to be the sequence $X/Z := (x_n/z_n)$.

For example, if X and Y are the sequences

$$X := (2, 4, 6, \cdots 2n, \cdots), \quad Y := \left(\frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \cdots, \frac{1}{n}, \cdots\right),$$

then we have

$$\begin{aligned} X + Y &= \left(\frac{3}{1}, \frac{9}{2}, \frac{19}{3}, \cdots, \frac{2n^2+1}{n}, \cdots\right), \\ X - Y &= \left(\frac{1}{1}, \frac{7}{2}, \frac{17}{3}, \cdots, \frac{2n^2-1}{n}, \cdots\right), \\ X \cdot Y &= (2, 2, 2, \cdots, 2, \cdots), \\ 3X &= (6, 12, 18, \cdots, 6n, \cdots), \\ X/Y &= (2, 8, 18, \cdots, 2n^2, \cdots). \end{aligned}$$

We note that if Z is the sequence

$$Z := (0, 2, 0, \cdots, 1 + (-1)^n, \cdots),$$

then we can define $X + Z$, $X - Z$ and $X \cdot Z$, but X/Z is not defined since some of the terms of Z are zero.

We now show that sequences obtained by applying these operations to convergent sequences give rise to new sequences whose limits can be predicted.

3.2.3 Theorem (a) *Let $X = (x_n)$ and $Y = (y_n)$ be sequences of real numbers that converge to x and y, respectively, and let $c \in \mathbb{R}$. Then the sequences $X + Y$, $X - Y$, $X \cdot Y$, and cX converge to $x + y$, $x - y$, xy, and cx, respectively.*

(b) *If $X = (x_n)$ converges to x and $Z = (z_n)$ is a sequence of nonzero real numbers that converges to z and if $z \neq 0$, then the quotient sequence X/Z converges to x/z.*

Proof. (a) To show that $\lim(x_n + y_n) = x + y$, we need to estimate the magnitude of $|(x_n + y_n) - (x + y)|$. To do this we use the Triangle Inequality 2.2.3 to obtain

$$\begin{aligned} |(x_n + y_n) - (x + y)| &= |(x_n - x) + (y_n - y)| \\ &\leq |x_n - x| + |y_n - y|. \end{aligned}$$

By hypothesis, if $\varepsilon > 0$ there exists a natural number K_1 such that if $n \geq K_1$, then $|x_n - x| < \varepsilon/2$; also there exists a natural number K_2 such that if $n \geq K_2$, then $|y_n - y| < \varepsilon/2$. Hence if $K(\varepsilon) := \sup\{K_1, K_2\}$, it follows that if $n \geq K(\varepsilon)$ then

$$\begin{aligned} |(x_n + y_n) - (x + y)| &\leq |x_n - x| + |y_n - y| \\ &< \tfrac{1}{2}\varepsilon + \tfrac{1}{2}\varepsilon = \varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, we infer that $X + Y = (x_n + y_n)$ converges to $x + y$.

Precisely the same argument can be used to show that $X - Y = (x_n - y_n)$ converges to $x - y$.

To show that $X \cdot Y = (x_n y_n)$ converges to xy, we make the estimate

$$\begin{aligned} |x_n y_n - xy| &= |(x_n y_n - x_n y) + (x_n y - xy)| \\ &\leq |x_n(y_n - y)| + |(x_n - x)y| \\ &= |x_n||y_n - y| + |x_n - x||y|. \end{aligned}$$

According to Theorem 3.2.2 there exists a real number $M_1 > 0$ such that $|x_n| \leq M_1$ for all $n \in \mathbb{N}$ and we set $M := \sup\{M_1, |y|\}$. Hence we have the estimate

$$|x_n y_n - xy| \leq M|y_n - y| + M|x_n - x|.$$

From the convergence of X and Y we conclude that if $\varepsilon > 0$ is given, then there exist natural numbers K_1 and K_2 such that if $n \geq K_1$ then $|x_n - x| < \varepsilon/2M$, and if $n \geq K_2$ then $|y_n - y| < \varepsilon/2M$. Now let $K(\varepsilon) = \sup\{K_1, K_2\}$; then, if $n \geq K(\varepsilon)$ we infer that

$$\begin{aligned} |x_n y_n - xy| &\leq M|y_n - y| + M|x_n - x| \\ &< M(\varepsilon/2M) + M(\varepsilon/2M) = \varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, this proves that the sequence $X \cdot Y = (x_n y_n)$ converges to xy.

The fact that $cX = (cx_n)$ converges to cx can be proved in the same way; it can also be deduced by taking Y to be the constant sequence $(c, c, c, \cdots)$. We leave the details to the reader.

(b) We next show that if $Z = (z_n)$ is a sequence of nonzero numbers that converges to a nonzero limit z, then the sequence $(1/z_n)$ of reciprocals converges to $1/z$. First let $\alpha := \frac{1}{2}|z|$ so that $\alpha > 0$. Since $\lim(z_n) = z$, there exists a natural number K_1 such that if $n \geq K_1$ then $|z_n - z| < \alpha$. It follows from Corollary 2.2.4(a) of the Triangle Inequality that $-\alpha \leq -|z_n - z| \leq |z_n| - |z|$ for $n \geq K_1$, whence it follows that $\frac{1}{2}|z| = |z| - \alpha \leq |z_n|$ for $n \geq K_1$. Therefore $1/|z_n| \leq 2/|z|$ for $n \geq K_1$ so we have the estimate

$$\begin{aligned} \left|\frac{1}{z_n} - \frac{1}{z}\right| &= \left|\frac{z - z_n}{z_n z}\right| = \frac{1}{|z_n z|}|z - z_n| \\ &\leq \frac{2}{|z|^2}|z - z_n| \qquad \text{for all} \quad n \geq K_1. \end{aligned}$$

Now, if $\varepsilon > 0$ is given, there exists a natural number K_2 such that if $n \geq K_2$ then $|z_n - z| < \frac{1}{2}\varepsilon|z|^2$. Therefore, it follows that if $K(\varepsilon) = \sup\{K_1, K_2\}$, then

$$\left|\frac{1}{z_n} - \frac{1}{z}\right| < \varepsilon \qquad \text{for all} \quad n > K(\varepsilon).$$

Since $\varepsilon > 0$ is arbitrary, it follows that

$$\lim\left(\frac{1}{z_n}\right) = \frac{1}{z}.$$

The proof of (b) is now completed by taking Y to be the sequence $(1/z_n)$ and using the fact that $X \cdot Y = (x_n/z_n)$ converges to $x(1/z) = x/z$. Q.E.D.

Some of the results of Theorem 3.2.3 can be extended, by Mathematical Induction, to a finite number of convergent sequences. For example, if $A = (a_n)$, $B = (b_n)$, $\cdots$, $Z = (z_n)$ are convergent sequences of real numbers, then their sum $A + B + \cdots + Z = (a_n + b_n + \cdots + z_n)$ is a convergent sequence and

(1) $$\lim(a_n + b_n + \cdots + z_n) = \lim(a_n) + \lim(b_n) + \cdots + \lim(z_n).$$

Also their product $A \cdot B \cdots Z := (a_n b_n \cdots z_n)$ is a convergent sequence and

(2) $$\lim(a_n b_n \cdots z_n) = \left(\lim(a_n)\right)\left(\lim(b_n)\right)\cdots\left(\lim(z_n)\right).$$

Hence, if $k \in \mathbb{N}$ and if $A = (a_n)$ is a convergent sequence, then

(3) $$\lim(a_n^k) = \left(\lim(a_n)\right)^k.$$

We leave the proofs of these assertions to the reader.

3.2.4 Theorem *If $X = (x_n)$ is a convergent sequence of real numbers and if $x_n \geq 0$ for all $n \in \mathbb{N}$, then $x = \lim(x_n) \geq 0$.*

Proof. Suppose the conclusion is not true and that $x < 0$; then $\varepsilon := -x$ is positive. Since X converges to x, there is a natural number K such that

$$x - \varepsilon < x_n < x + \varepsilon \qquad \text{for all} \quad n \geq K.$$

In particular, we have $x_K < x + \varepsilon = x + (-x) = 0$. But this contradicts the hypothesis that $x_n \geq 0$ for all $n \in \mathbb{N}$. Therefore, this contradiction implies that $x \geq 0$. Q.E.D.

We now give a useful result that is formally stronger than Theorem 3.2.4.

3.2.5 Theorem *If $X = (x_n)$ and $Y = (y_n)$ are convergent sequences of real numbers and if $x_n \leq y_n$ for all $n \in \mathbb{N}$, then $\lim(x_n) \leq \lim(y_n)$.*

Proof. Let $z_n := y_n - x_n$ so that $Z := (z_n) = Y - X$ and $z_n \geq 0$ for all $n \in \mathbb{N}$. It follows from Theorems 3.2.4 and 3.2.3 that

$$0 \leq \lim Z = \lim(y_n) - \lim(x_n),$$

so that $\lim(x_n) \leq \lim(y_n)$. Q.E.D.

The next result asserts that if all the terms of a convergent sequence satisfy an inequality of the form $a \leq x_n \leq b$, then the limit of the sequence satisfies the same inequality. Thus if the sequence is convergent, one may "pass to the limit" in an inequality of this type.

3.2.6 Theorem *If $X = (x_n)$ is a convergent sequence and if $a \leq x_n \leq b$ for all $n \in \mathbb{N}$, then $a \leq \lim(x_n) \leq b$.*

Proof. Let Y be the constant sequence $(b, b, b, \cdots)$. Theorem 3.2.5 implies that $\lim X \leq \lim Y = b$. Similarly one shows that $a \leq \lim X$. Q.E.D.

The next result asserts that if a sequence Y is squeezed between two sequences that converge to the *same limit*, then it must also converge to this limit.

3.2.7 Squeeze Theorem *Suppose that $X = (x_n)$, $Y = (y_n)$, and $Z = (z_n)$ are sequences of real numbers such that*

$$x_n \le y_n \le z_n \qquad \textit{for all} \quad n \in \mathbb{N},$$

and that $\lim(x_n) = \lim(z_n)$. Then $Y = (y_n)$ is convergent and

$$\lim(x_n) = \lim(y_n) = \lim(z_n).$$

Proof. Let $w := \lim(x_n) = \lim(z_n)$. If $\varepsilon > 0$ is given, then it follows from the convergence of X and Z to w that there exists a natural number K such that if $n \ge K$ then

$$|x_n - w| < \varepsilon \qquad \text{and} \qquad |z_n - w| < \varepsilon.$$

Since the hypothesis implies that

$$x_n - w \le y_n - w \le z_n - w \qquad \text{for all} \quad n \in \mathbb{N},$$

it follows (why?) that

$$-\varepsilon < y_n - w < \varepsilon$$

for all $n \ge K$. Since $\varepsilon > 0$ is arbitrary, this implies that $\lim(y_n) = w$. Q.E.D.

Remark Since any tail of a convergent sequence has the same limit, the hypotheses of Theorems 3.2.4, 3.2.5, 3.2.6, and 3.2.7 can be weakened to apply to the tail of a sequence. For example, in Theorem 3.2.4, if $X = (x_n)$ is "ultimately positive" in the sense that there exists $m \in \mathbb{N}$ such that $x_n \ge 0$ for all $n \ge m$, then the same conclusion that $x \ge 0$ will hold. Similar modifications are valid for the other theorems, as the reader should verify.

3.2.8 Examples (a) The sequence (n) is divergent.

It follows from Theorem 3.2.2 that if the sequence $X := (n)$ is convergent, then there exists a real number $M > 0$ such that $n = |n| < M$ for all $n \in \mathbb{N}$. But this violates the Archimedean Property 2.4.3.

(b) The sequence $((-1)^n)$ is divergent.

This sequence $X = ((-1)^n)$ is bounded (take $M := 1$), so we cannot invoke Theorem 3.2.2. However, assume that $a := \lim X$ exists. Let $\varepsilon := 1$ so that there exists a natural number K_1 such that

$$|(-1) - a| < 1 \qquad \text{for all} \quad n \ge K_1.$$

If n is an odd natural number with $n \ge K_1$, this gives $|-1 - a| < 1$, so that $-2 < a < 0$. (Why?) On the other hand, if n is an even natural number with $n \ge K_1$, this inequality gives $|1 - a| < 1$ so that $0 < a < 2$. Since a cannot satisfy both of these inequalities, the hypothesis that X is convergent leads to a contradiction. Therefore the sequence X is divergent.

(c) $\lim \left(\dfrac{2n+1}{n} \right) = 2.$

If we let $X := (2)$ and $Y := (1/n)$, then $((2n+1)/n) = X + Y$. Hence it follows from Theorem 3.2.3(a) that $\lim(X + Y) = \lim X + \lim Y = 2 + 0 = 2$.

(d) $\lim \left(\dfrac{2n+1}{n+5} \right) = 2.$

Since the sequences $(2n+1)$ and $(n+5)$ are not convergent (why?), it is not possible to use Theorem 3.2.3(b) directly. However, if we write

$$\frac{2n+1}{n+5} = \frac{2+1/n}{1+5/n},$$

we can obtain the given sequence as one to which Theorem 3.2.3(b) applies when we take $X := (2+1/n)$ and $Z := (1+5/n)$. (Check that all hypotheses are satisfied.) Since $\lim X = 2$ and $\lim Z = 1 \neq 0$, we deduce that $\lim\big((2n+1)/(n+5)\big) = 2/1 = 2$.

(e) $\lim\left(\dfrac{2n}{n^2+1}\right) = 0.$

Theorem 3.2.3(b) does not apply directly. (Why?) We note that

$$\frac{2n}{n^2+1} = \frac{2}{n+1/n},$$

but Theorem 3.2.3(b) does not apply here either, because $(n+1/n)$ is not a convergent sequence. (Why not?) However, if we write

$$\frac{2n}{n^2+1} = \frac{2/n}{1+1/n^2},$$

then we can apply Theorem 3.2.3(b), since $\lim(2/n) = 0$ and $\lim(1+1/n^2) = 1 \neq 0$. Therefore $\lim(2n/(n^2+1)) = 0/1 = 0$.

(f) $\lim\left(\dfrac{\sin n}{n}\right) = 0.$

We cannot apply Theorem 3.2.3(b) directly, since the sequence (n) is not convergent [neither is the sequence $(\sin n)$]. It does not appear that a simple algebraic manipulation will enable us to reduce the sequence into one to which Theorem 3.2.3 will apply. However, if we note that $-1 \le \sin n \le 1$, then it follows that

$$-\frac{1}{n} \le \frac{\sin n}{n} \le \frac{1}{n} \qquad \text{for all} \quad n \in \mathbb{N}.$$

Hence we can apply the Squeeze Theorem 3.2.7 to infer that $\lim(n^{-1}\sin n) = 0$. (We note that Theorem 3.1.10 could also be applied to this sequence.)

(g) Let $X = (x_n)$ be a sequence of real numbers that converges to $x \in \mathbb{R}$. Let p be a polynomial; for example, let

$$p(t) := a_k t^k + a_{k-1}t^{k-1} + \cdots + a_1 t + a_0,$$

where $k \in \mathbb{N}$ and $a_j \in \mathbb{R}$ for $j = 0, 1, \cdots, k$. It follows from Theorem 3.2.3 that the sequence $(p(x_n))$ converges to $p(x)$. We leave the details to the reader as an exercise.

(h) Let $X = (x_n)$ be a sequence of real numbers that converges to $x \in \mathbb{R}$. Let r be a rational function (that is, $r(t) := p(t)/q(t)$, where p and q are polynomials). Suppose that $q(x_n) \neq 0$ for all $n \in \mathbb{N}$ and that $q(x) \neq 0$. Then the sequence $(r(x_n))$ converges to $r(x) = p(x)/q(x)$. We leave the details to the reader as an exercise. □

We conclude this section with several results that will be useful in the work that follows.

3.2.9 Theorem *Let the sequence $X = (x_n)$ converge to x. Then the sequence $(|x_n|)$ of absolute values converges to $|x|$. That is, if $x = \lim(x_n)$, then $|x| = \lim(|x_n|)$.*

Proof. It follows from the Triangle Inequality (see Corollary 2.2.4(a)) that

$$\big||x_n| - |x|\big| \le |x_n - x| \qquad \text{for all} \quad n \in \mathbb{N}.$$

The convergence of $(|x_n|)$ to $|x|$ is then an immediate consequence of the convergence of (x_n) to x. Q.E.D.

3.2.10 Theorem *Let $X = (x_n)$ be a sequence of real numbers that converges to x and suppose that $x_n \ge 0$. Then the sequence $(\sqrt{x_n})$ of positive square roots converges and* $\lim(\sqrt{x_n}) = \sqrt{x}$.

Proof. It follows from Theorem 3.2.4 that $x = \lim(x_n) \ge 0$ so the assertion makes sense. We now consider the two cases: (i) $x = 0$ and (ii) $x > 0$.

Case (i) If $x = 0$, let $\varepsilon > 0$ be given. Since $x_n \to 0$ there exists a natural number K such that if $n \ge K$ then

$$0 \le x_n = x_n - 0 < \varepsilon^2.$$

Therefore [see Example 2.1.13(a)], $0 \le \sqrt{x_n} < \varepsilon$ for $n \ge K$. Since $\varepsilon > 0$ is arbitrary, this implies that $\sqrt{x_n} \to 0$.

Case (ii) If $x > 0$, then $\sqrt{x} > 0$ and we note that

$$\sqrt{x_n} - \sqrt{x} = \frac{(\sqrt{x_n} - \sqrt{x})(\sqrt{x_n} + \sqrt{x})}{\sqrt{x_n} + \sqrt{x}} = \frac{x_n - x}{\sqrt{x_n} + \sqrt{x}}$$

Since $\sqrt{x_n} + \sqrt{x} \ge \sqrt{x} > 0$, it follows that

$$\left|\sqrt{x_n} - \sqrt{x}\right| \le \left(\frac{1}{\sqrt{x}}\right)|x_n - x|.$$

The convergence of $\sqrt{x_n} \to \sqrt{x}$ follows from the fact that $x_n \to x$. Q.E.D.

For certain types of sequences, the following result provides a quick and easy "ratio test" for convergence. Related results can be found in the exercises.

3.2.11 Theorem *Let (x_n) be a sequence of positive real numbers such that $L := \lim(x_{n+1}/x_n)$ exists. If $L < 1$, then (x_n) converges and $\lim(x_n) = 0$.*

Proof. By 3.2.4 it follows that $L \ge 0$. Let r be a number such that $L < r < 1$, and let $\varepsilon := r - L > 0$. There exists a number $K \in \mathbb{N}$ such that if $n \ge K$ then

$$\left|\frac{x_{n+1}}{x_n} - L\right| < \varepsilon.$$

It follows from this (why?) that if $n \ge K$, then

$$\frac{x_{n+1}}{x_n} < L + \varepsilon = L + (r - L) = r.$$

Therefore, if $n \ge K$, we obtain

$$0 < x_{n+1} < x_n r < x_{n-1} r^2 < \cdots < x_K r^{n-K+1}.$$

If we set $C := x_K / r^K$, we see that $0 < x_{n+1} < C r^{n+1}$ for all $n \ge K$. Since $0 < r < 1$, it follows from 3.1.11(b) that $\lim(r^n) = 0$ and therefore from Theorem 3.1.10 that $\lim(x_n) = 0$. Q.E.D.

As an illustration of the utility of the preceding theorem, consider the sequence (x_n) given by $x_n := n/2^n$. We have

$$\frac{x_{n+1}}{x_n} = \frac{n+1}{2^{n+1}} \cdot \frac{2^n}{n} = \frac{1}{2}\left(1 + \frac{1}{n}\right),$$

so that $\lim(x_{n+1}/x_n) = \frac{1}{2}$. Since $\frac{1}{2} < 1$, it follows from Theorem 3.2.11 that $\lim(n/2^n) = 0$.

Exercises for Section 3.2

1. For x_n given by the following formulas, establish either the convergence or the divergence of the sequence $X = (x_n)$.
   (a) $x_n := \dfrac{n}{n+1}$, (b) $x_n := \dfrac{(-1)^n n}{n+1}$,
   (c) $x_n := \dfrac{n^2}{n+1}$, (d) $x_n := \dfrac{2n^2+3}{n^2+1}$.
2. Give an example of two divergent sequences X and Y such that:
   (a) their sum $X + Y$ converges, (b) their product XY converges.
3. Show that if X and Y are sequences such that X and $X + Y$ are convergent, then Y is convergent.
4. Show that if X and Y are sequences such that X converges to $x \neq 0$ and XY converges, then Y converges.
5. Show that the following sequences are not convergent.
   (a) (2^n), (b) $((-1)^n n^2)$.
6. Find the limits of the following sequences:
   (a) $\lim\left((2 + 1/n)^2\right)$, (b) $\lim\left(\dfrac{(-1)^n}{n+2}\right)$,
   (c) $\lim\left(\dfrac{\sqrt{n}-1}{\sqrt{n}+1}\right)$, (d) $\lim\left(\dfrac{n+1}{n\sqrt{n}}\right)$.
7. If (b_n) is a bounded sequence and $\lim(a_n) = 0$, show that $\lim(a_n b_n) = 0$. Explain why Theorem 3.2.3 *cannot* be used.
8. Explain why the result in equation (3) before Theorem 3.2.4 *cannot* be used to evaluate the limit of the sequence $\left((1 + 1/n)^n\right)$.
9. Let $y_n := \sqrt{n+1} - \sqrt{n}$ for $n \in \mathbb{N}$. Show that (y_n) and $(\sqrt{n}\, y_n)$ converge. Find their limits.
10. Determine the following limits.
    (a) $\lim\left((3\sqrt{n})^{1/2n}\right)$, (b) $\lim\left((n+1)^{1/\ln(n+1)}\right)$.
11. If $0 < a < b$, determine $\lim\left(\dfrac{a^{n+1} + b^{n+1}}{a^n + b^n}\right)$.
12. If $a > 0$, $b > 0$, show that $\lim\left(\sqrt{(n+a)(n+b)} - n\right) = (a+b)/2$.
13. Use the Squeeze Theorem 3.2.7 to determine the limits of the following.
    (a) $\left(n^{1/n^2}\right)$, (b) $\left((n!)^{1/n^2}\right)$.
14. Show that if $z_n := (a^n + b^n)^{1/n}$ where $0 < a < b$, then $\lim(z_n) = b$.
15. Apply Theorem 3.2.11 to the following sequences, where a, b satisfy $0 < a < 1$, $b > 1$.
    (a) (a^n), (b) $(b^n/2^n)$,
    (c) (n/b^n), (d) $(2^{3n}/3^{2n})$.

16. (a) Give an example of a convergent sequence (x_n) of positive numbers with $\lim(x_{n+1}/x_n) = 1$.
(b) Give an example of a divergent sequence with this property. (Thus, this property cannot be used as a test for convergence.)

17. Let $X = (x_n)$ be a sequence of positive real numbers such that $\lim(x_{n+1}/x_n) = L > 1$. Show that X is not a bounded sequence and hence is not convergent.

18. Discuss the convergence of the following sequences, where a, b satisfy $0 < a < 1, b > 1$.
(a) (n^2a^n), (b) (b^n/n^2),
(c) $(b^n/n!)$, (d) $(n!/n^n)$.

19. Let (x_n) be a sequence of positive real numbers such that $\lim(x_n^{1/n}) = L < 1$. Show that there exists a number r with $0 < r < 1$ such that $0 < x_n < r^n$ for all sufficiently large $n \in \mathbb{N}$. Use this to show that $\lim(x_n) = 0$.

20. (a) Give an example of a convergent sequence (x_n) of positive numbers with $\lim(x_n^{1/n}) = 1$.
(b) Give an example of a divergent sequence (x_n) of positive numbers with $\lim(x_n^{1/n}) = 1$. (Thus, this property cannot be used as a test for convergence.)

21. Suppose that (x_n) is a convergent sequence and (y_n) is such that for any $\varepsilon > 0$ there exists M such that $|x_n - y_n| < \varepsilon$ for all $n \geq M$. Does it follow that (y_n) is convergent?

22. Show that if (x_n) and (y_n) are convergent sequences, then the sequences (u_n) and (v_n) defined by $u_n := \max\{x_n, y_n\}$ and $v_n := \min\{x_n, y_n\}$ are also convergent. (See Exercise 2.2.16.)

23. Show that if (x_n), (y_n), (z_n) are convergent sequences, then the sequence (w_n) defined by $w_n := \text{mid}\{x_n, y_n, z_n\}$ is also convergent. (See Exercise 2.2.17.)

Section 3.3 Monotone Sequences

Until now, we have obtained several methods of showing that a sequence $X = (x_n)$ of real numbers is convergent:

(i) We can use Definition 3.1.3 or Theorem 3.1.5 directly. This is often (but not always) difficult to do.

(ii) We can dominate $|x_n - x|$ by a multiple of the terms in a sequence (a_n) known to converge to 0, and employ Theorem 3.1.10.

(iii) We can identify X as a sequence obtained from other sequences that are known to be convergent by taking tails, algebraic combinations, absolute values, or square roots, and employ Theorems 3.1.9, 3.2.3, 3.2.9, or 3.2.10.

(iv) We can "squeeze" X between two sequences that converge to the same limit and use Theorem 3.2.7.

(v) We can use the "ratio test" of Theorem 3.2.11.

Except for (iii), all of these methods require that we already know (or at least suspect) the value of the limit, and we then verify that our suspicion is correct.

There are many instances, however, in which there is no obvious candidate for the limit of a sequence, even though a preliminary analysis may suggest that convergence is likely. In this and the next two sections, we shall establish results that can be used to show a sequence is convergent even though the value of the limit is not known. The method we introduce in this section is more restricted in scope than the methods we give in the next two, but it is much easier to employ. It applies to sequences that are monotone in the following sense.

3.3.1 Definition Let $X = (x_n)$ be a sequence of real numbers. We say that X is **increasing** if it satisfies the inequalities

$$x_1 \leq x_2 \leq \cdots \leq x_n \leq x_{n+1} \leq \cdots.$$

We say that X is **decreasing** if it satisfies the inequalities

$$x_1 \geq x_2 \geq \cdots \geq x_n \geq x_{n+1} \geq \cdots.$$

We say that X is **monotone** if it is either increasing or decreasing.

The following sequences are increasing:

$$(1, 2, 3, 4, \cdots, n, \cdots), \qquad (1, 2, 2, 3, 3, 3, \cdots),$$
$$(a, a^2, a^3, \cdots, a^n, \cdots) \quad \text{if} \quad a > 1.$$

The following sequences are decreasing:

$$(1, 1/2, 1/3, \cdots, 1/n, \cdots), \qquad (1, 1/2, 1/2^2, \cdots, 1/2^{n-1}, \cdots),$$
$$(b, b^2, b^3, \cdots, b^n, \cdots) \quad \text{if} \quad 0 < b < 1.$$

The following sequences are not monotone:

$$\left(+1, -1, +1, \cdots, (-1)^{n+1}, \cdots\right), \qquad \left(-1, +2, -3, \cdots, (-1)^n n \cdots\right)$$

The following sequences are not monotone, but they are "ultimately" monotone:

$$(7, 6, 2, 1, 2, 3, 4, \cdots), \qquad (-2, 0, 1, 1/2, 1/3, 1/4, \cdots).$$

3.3.2 Monotone Convergence Theorem *A monotone sequence of real numbers is convergent if and only if it is bounded. Further:*

(a) *If $X = (x_n)$ is a bounded increasing sequence, then*

$$\lim(x_n) = \sup\{x_n : n \in \mathbb{N}\}.$$

(b) *If $Y = (y_n)$ is a bounded decreasing sequence, then*

$$\lim(y_n) = \inf\{y_n : n \in \mathbb{N}\}.$$

Proof. It was seen in Theorem 3.2.2 that a convergent sequence must be bounded.

Conversely, let X be a bounded monotone sequence. Then X is either increasing or decreasing.

(a) We first treat the case where $X = (x_n)$ is a bounded, increasing sequence. Since X is bounded, there exists a real number M such that $x_n \leq M$ for all $n \in \mathbb{N}$. According to the Completeness Property 2.3.6, the supremum $x^* = \sup\{x_n : n \in \mathbb{N}\}$ exists in $\mathbb{R}$; we will show that $x^* = \lim(x_n)$.

If $\varepsilon > 0$ is given, then $x^* - \varepsilon$ is not an upper bound of the set $\{x_n : n \in \mathbb{N}\}$, and hence there exists a member of set x_K such that $x^* - \varepsilon < x_K$. The fact that X is an increasing sequence implies that $x_K \leq x_n$ whenever $n \geq K$, so that

$$x^* - \varepsilon < x_K \leq x_n \leq x^* < x^* + \varepsilon \qquad \text{for all} \quad n \geq K.$$

Therefore we have

$$|x_n - x^*| < \varepsilon \qquad \text{for all} \quad n \geq K.$$

Since $\varepsilon > 0$ is arbitrary, we conclude that (x_n) converges to x^*.

(b) If $Y = (y_n)$ is a bounded decreasing sequence, then it is clear that $X := -Y = (-y_n)$ is a bounded increasing sequence. It was shown in part (a) that $\lim X = \sup\{-y_n : n \in \mathbb{N}\}$. Now $\lim X = -\lim Y$ and also, by Exercise 2.4.4(b), we have

$$\sup\{-y_n : n \in \mathbb{N}\} = -\inf\{y_n : n \in \mathbb{N}\}.$$

Therefore $\lim Y = -\lim X = \inf\{y_n : n \in \mathbb{N}\}$. Q.E.D.

The Monotone Convergence Theorem establishes the existence of the limit of a bounded monotone sequence. It also gives us a way of calculating the limit of the sequence *provided* we can evaluate the supremum in case (a), or the infimum in case (b). Sometimes it is difficult to evaluate this supremum (or infimum), but once we know that it exists, it is often possible to evaluate the limit by other methods.

3.3.3 Examples (a) $\lim(1/\sqrt{n}) = 0$.

It is possible to handle this sequence by using Theorem 3.2.10; however, we shall use the Monotone Convergence Theorem. Clearly 0 is a lower bound for the set $\{1/\sqrt{n}: n \in \mathbb{N}\}$, and it is not difficult to show that 0 is the infimum of the set $\{1/\sqrt{n}: n \in \mathbb{N}\}$; hence $0 = \lim(1/\sqrt{n})$.

On the other hand, once we know that $X := (1/\sqrt{n})$ is bounded and decreasing, we know that it converges to some real number x. Since $X = (1/\sqrt{n})$ converges to x, it follows from Theorem 3.2.3 that $X \cdot X = (1/n)$ converges to x^2. Therefore $x^2 = 0$, whence $x = 0$.

(b) Let $x_n := 1 + 1/2 + 1/3 + \cdots + 1/n$ for $n \in \mathbb{N}$.

Since $x_{n+1} = x_n + 1/(n+1) > x_n$, we see that (x_n) is an increasing sequence. By the Monotone Convergence Theorem 3.3.2, the question of whether the sequence is convergent or not is reduced to the question of whether the sequence is bounded or not. Attempts to use direct numerical calculations to arrive at a conjecture concerning the possible boundedness of the sequence (x_n) lead to inconclusive frustration. A computer run will reveal the approximate values $x_n \approx 11.4$ for $n = 50{,}000$, and $x_n \approx 12.1$ for $n = 100{,}000$. Such numerical facts may lead the casual observer to conclude that the sequence is bounded. However, the sequence is in fact divergent, which is established by noting that

$$\begin{aligned} x_{2^n} &= 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) + \cdots + \left(\frac{1}{2^{n-1}+1} + \cdots + \frac{1}{2^n}\right) \\ &> 1 + \frac{1}{2} + \left(\frac{1}{4} + \frac{1}{4}\right) + \cdots + \left(\frac{1}{2^n} + \cdots + \frac{1}{2^n}\right) \\ &= 1 + \frac{1}{2} + \frac{1}{2} + \cdots + \frac{1}{2} \\ &= 1 + \frac{n}{2}. \end{aligned}$$

Since (x_n) is unbounded, Theorem 3.2.2 implies that it is divergent.

The terms x_n increase extremely slowly. For example, it can be shown that to achieve $x_n > 50$ would entail approximately 5.2×10^{21} additions, and a normal computer performing 400 million additions a second would require more than 400,000 years to perform the calculation (there are 31,536,000 seconds in a year). Even a supercomputer that can perform more than a trillion additions a second, would take more than 164 years to reach that modest goal. □

Sequences that are defined inductively must be treated differently. If such a sequence is known to converge, then the value of the limit can sometimes be determined by using the inductive relation.

For example, suppose that convergence has been established for the sequence (x_n) defined by

$$x_1 = 2, \qquad x_{n+1} = 2 + \frac{1}{x_n}, \quad n \in \mathbb{N}.$$

If we let $x = \lim(x_n)$, then we also have $x = \lim(x_{n+1})$ since the 1-tail (x_{n+1}) converges to the same limit. Further, we see that $x_n \geq 2$, so that $x \neq 0$ and $x_n \neq 0$ for all $n \in \mathbb{N}$. Therefore, we may apply the limit theorems for sequences to obtain

$$x = \lim(x_{n+1}) = 2 + \frac{1}{\lim(x_n)} = 2 + \frac{1}{x}.$$

Thus, the limit x is a solution of the quadratic equation $x^2 - 2x - 1 = 0$, and since x must be positive, we find that the limit of the sequence is $x = 1 + \sqrt{2}$.

Of course, the issue of convergence must not be ignored or casually assumed. For example, if we assumed the sequence (y_n) defined by $y_1 := 1$, $y_{n+1} := 2y_n + 1$ is convergent with limit y, then we would obtain $y = 2y + 1$, so that $y = -1$. Of course, this is absurd.

In the following examples, we employ this method of evaluating limits, but only after carefully establishing convergence using the Monotone Convergence Theorem. Additional examples of this type will be given in Section 3.5.

3.3.4 Examples (a) Let $Y = (y_n)$ be defined inductively by $y_1 := 1$, $y_{n+1} := \frac{1}{4}(2y_n + 3)$ for $n \geq 1$. We shall show that $\lim Y = 3/2$.

Direct calculation shows that $y_2 = 5/4$. Hence we have $y_1 < y_2 < 2$. We show, by Induction, that $y_n < 2$ for all $n \in \mathbb{N}$. Indeed, this is true for $n = 1, 2$. If $y_k < 2$ holds for some $k \in \mathbb{N}$, then

$$y_{k+1} = \tfrac{1}{4}(2y_k + 3) < \tfrac{1}{4}(4 + 3) = \tfrac{7}{4} < 2,$$

so that $y_{k+1} < 2$. Therefore $y_n < 2$ for all $n \in \mathbb{N}$.

We now show, by Induction, that $y_n < y_{n+1}$ for all $n \in \mathbb{N}$. The truth of this assertion has been verified for $n = 1$. Now suppose that $y_k < y_{k+1}$ for some k; then $2y_k + 3 < 2y_{k+1} + 3$, whence it follows that

$$y_{k+1} = \tfrac{1}{4}(2y_k + 3) < \tfrac{1}{4}(2y_{k+1} + 3) = y_{k+2}.$$

Thus $y_k < y_{k+1}$ implies that $y_{k+1} < y_{k+2}$. Therefore $y_n < y_{n+1}$ for all $n \in \mathbb{N}$.

We have shown that the sequence $Y = (y_n)$ is increasing and bounded above by 2. It follows from the Monotone Convergence Theorem that Y converges to a limit that is at most 2. In this case it is not so easy to evaluate $\lim(y_n)$ by calculating $\sup\{y_n : n \in \mathbb{N}\}$. However, there is another way to evaluate its limit. Since $y_{n+1} = \frac{1}{4}(2y_n + 3)$ for all $n \in \mathbb{N}$, the nth term in the 1-tail Y_1 of Y has a simple algebraic relation to the nth term of Y. Since, by Theorem 3.1.9, we have $y := \lim Y_1 = \lim Y$, it therefore follows from Theorem 3.2.3 (why?) that

$$y = \tfrac{1}{4}(2y + 3),$$

from which it follows that $y = 3/2$.

(b) Let $Z = (z_n)$ be the sequence of real numbers defined by $z_1 := 1$, $z_{n+1} := \sqrt{2z_n}$ for $n \in \mathbb{N}$. We will show that $\lim(z_n) = 2$.

Note that $z_1 = 1$ and $z_2 = \sqrt{2}$; hence $1 \leq z_1 < z_2 < 2$. We claim that the sequence Z is increasing and bounded above by 2. To show this we will show, by Induction, that

$1 \le z_n < z_{n+1} < 2$ for all $n \in \mathbb{N}$. This fact has been verified for $n = 1$. Suppose that it is true for $n = k$; then $2 \le 2z_k < 2z_{k+1} < 4$, whence it follows (why?) that

$$1 < \sqrt{2} \le z_{k+1} = \sqrt{2z_k} < z_{k+2} = \sqrt{2z_{k+1}} < \sqrt{4} = 2.$$

[In this last step we have used Example 2.1.13(a).] Hence the validity of the inequality $1 \le z_k < z_{k+1} < 2$ implies the validity of $1 \le z_{k+1} < z_{k+2} < 2$. Therefore $1 \le z_n < z_{n+1} < 2$ for all $n \in \mathbb{N}$.

Since $Z = (z_n)$ is a bounded increasing sequence, it follows from the Monotone Convergence Theorem that it converges to a number $z := \sup\{z_n\}$. It may be shown directly that $\sup\{z_n\} = 2$, so that $z = 2$. Alternatively we may use the method employed in part (a). The relation $z_{n+1} = \sqrt{2z_n}$ gives a relation between the nth term of the 1-tail Z_1 of Z and the nth term of Z. By Theorem 3.1.9, we have $\lim Z_1 = z = \lim Z$. Moreover, by Theorems 3.2.3 and 3.2.10, it follows that the limit z must satisfy the relation

$$z = \sqrt{2z}.$$

Hence z must satisfy the equation $z^2 = 2z$ which has the roots $z = 0, 2$. Since the terms of $z = (z_n)$ all satisfy $1 \le z_n \le 2$, it follows from Theorem 3.2.6 that we must have $1 \le z \le 2$. Therefore $z = 2$. □

The Calculation of Square Roots

We now give an application of the Monotone Convergence Theorem to the calculation of square roots of positive numbers.

3.3.5 Example Let $a > 0$; we will construct a sequence (s_n) of real numbers that converges to $\sqrt{a}$.

Let $s_1 > 0$ be arbitrary and define $s_{n+1} := \frac{1}{2}(s_n + a/s_n)$ for $n \in \mathbb{N}$. We now show that the sequence (s_n) converges to $\sqrt{a}$. (This process for calculating square roots was known in Mesopotamia before 1500 B.C.)

We first show that $s_n^2 \ge a$ for $n \ge 2$. Since s_n satisfies the quadratic equation $s_n^2 - 2s_{n+1}s_n + a = 0$, this equation has a real root. Hence the discriminant $4s_{n+1}^2 - 4a$ must be nonnegative; that is, $s_{n+1}^2 \ge a$ for $n \ge 1$.

To see that (s_n) is ultimately decreasing, we note that for $n \ge 2$ we have

$$s_n - s_{n+1} = s_n - \frac{1}{2}\left(s_n + \frac{a}{s_n}\right) = \frac{1}{2} \cdot \frac{(s_n^2 - a)}{s_n} \ge 0.$$

Hence, $s_{n+1} \le s_n$ for all $n \ge 2$. The Monotone Convergence Theorem implies that $s := \lim(s_n)$ exists. Moreover, from Theorem 3.2.3, the limit s must satisfy the relation

$$s = \frac{1}{2}\left(s + \frac{a}{s}\right),$$

whence it follows (why?) that $s = a/s$ or $s^2 = a$. Thus $s = \sqrt{a}$.

For the purposes of calculation, it is often important to have an estimate of *how rapidly* the sequence (s_n) converges to $\sqrt{a}$. As above, we have $\sqrt{a} \le s_n$ for all $n \ge 2$, whence it follows that $a/s_n \le \sqrt{a} \le s_n$. Thus we have

$$0 \le s_n - \sqrt{a} \le s_n - a/s_n = (s_n^2 - a)/s_n \qquad \text{for} \quad n \ge 2.$$

Using this inequality we can calculate $\sqrt{a}$ to any desired degree of accuracy. □

Euler's Number

We conclude this section by introducing a sequence that converges to one of the most important "transcendental" numbers in mathematics, second in importance only to π.

3.3.6 Example Let $e_n := (1 + 1/n)^n$ for $n \in \mathbb{N}$. We will now show that the sequence $E = (e_n)$ is bounded and increasing; hence it is convergent. The limit of this sequence is the famous *Euler number* e, whose approximate value is 2.718 281 828 459 045 $\cdots$, which is taken as the base of the "natural" logarithm.

If we apply the Binomial Theorem, we have

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

If we divide the powers of n into the terms in the numerators of the binomial coefficients, we get

$$e_n = 1 + 1 + \frac{1}{2!}\left(1 - \frac{1}{n}\right) + \frac{1}{3!}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) + \cdots + \frac{1}{n!}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{n-1}{n}\right).$$

Similarly we have

$$\begin{aligned} e_{n+1} = 1 + 1 &+ \frac{1}{2!}\left(1 - \frac{1}{n+1}\right) + \frac{1}{3!}\left(1 - \frac{1}{n+1}\right)\left(1 - \frac{2}{n+1}\right) \\ &+ \cdots + \frac{1}{n!}\left(1 - \frac{1}{n+1}\right)\left(1 - \frac{2}{n+1}\right) \cdots \left(1 - \frac{n-1}{n+1}\right) \\ &+ \frac{1}{(n+1)!}\left(1 - \frac{1}{n+1}\right)\left(1 - \frac{2}{n+1}\right) \cdots \left(1 - \frac{n}{n+1}\right). \end{aligned}$$

Note that the expression for e_n contains $n + 1$ terms, while that for e_{n+1} contains $n + 2$ terms. Moreover, each term appearing in e_n is less than or equal to the corresponding term in e_{n+1}, and e_{n+1} has one more positive term. Therefore we have $2 \le e_1 < e_2 < \cdots < e_n < e_{n+1} < \cdots$, so that the terms of E are increasing.

To show that the terms of E are bounded above, we note that if $p = 1, 2, \cdots, n$, then $(1 - p/n) < 1$. Moreover $2^{p-1} \le p!$ [see 1.2.4(e)] so that $1/p! \le 1/2^{p-1}$. Therefore, if $n > 1$, then we have

$$2 < e_n < 1 + 1 + \frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^{n-1}}.$$

Since it can be verified that [see 1.2.4(f)]

$$\frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^{n-1}} = 1 - \frac{1}{2^{n-1}} < 1,$$

we deduce that $2 < e_n < 3$ for all $n \in \mathbb{N}$. The Monotone Convergence Theorem implies that the sequence E converges to a real number that is between 2 and 3. We define the number e to be the limit of this sequence.

By refining our estimates we can find closer rational approximations to e, but we cannot evaluate it *exactly*, since e is an irrational number. However, it is possible to calculate e to as many decimal places as desired. The reader should use a calculator (or a computer) to evaluate e_n for "large" values of n. □

Leonhard Euler

Leonhard Euler (1707–1783) was born near Basel, Switzerland. His clergyman father hoped that his son would follow him into the ministry, but when Euler entered the University of Basel at age 14, his mathematical talent was noted by Johann Bernoulli, who became his mentor. In 1727, Euler went to Russia to join Johann's son, Daniel, at the new St. Petersburg Academy. There he met and married Katharina Gsell, the daughter of a Swiss artist. During their long marriage they had 13 children, but only five survived childhood.

In 1741, Euler accepted an offer from Frederick the Great to join the Berlin Academy, where he stayed for 25 years. During this period he wrote landmark books on calculus and a steady stream of papers. In response to a request for instruction in science from the Princess of Anhalt-Dessau, he wrote a multi-volume work on science that became famous under the title *Letters to a German Princess.*

In 1766, he returned to Russia at the invitation of Catherine the Great. His eyesight had deteriorated over the years, and soon after his return to Russia he became totally blind. Incredibly, his blindness made little impact on his mathematical output, for he wrote several books and over 400 papers while blind. He remained busy and active until the day of his death.

Euler's productivity was remarkable: he wrote textbooks on physics, algebra, calculus, real and complex analysis, analytic and differential geometry, and the calculus of variations. He also wrote hundreds of original papers, many of which won prizes. A current edition of his collected works consists of 74 volumes.

Exercises for Section 3.3

1. Let $x_1 := 8$ and $x_{n+1} := \frac{1}{2}x_n + 2$ for $n \in \mathbb{N}$. Show that (x_n) is bounded and monotone. Find the limit.

2. Let $x_1 > 1$ and $x_{n+1} := 2 - 1/x_n$ for $n \in \mathbb{N}$. Show that (x_n) is bounded and monotone. Find the limit.

3. Let $x_1 \geq 2$ and $x_{n+1} := 1 + \sqrt{x_n - 1}$ for $n \in \mathbb{N}$. Show that (x_n) is decreasing and bounded below by 2. Find the limit.

4. Let $x_1 := 1$ and $x_{n+1} := \sqrt{2 + x_n}$ for $n \in \mathbb{N}$. Show that (x_n) converges and find the limit.

5. Let $y_1 := \sqrt{p}$, where $p > 0$, and $y_{n+1} := \sqrt{p + y_n}$ for $n \in \mathbb{N}$. Show that (y_n) converges and find the limit. [*Hint:* One upper bound is $1 + 2\sqrt{p}$.]

6. Let $a > 0$ and let $z_1 > 0$. Define $z_{n+1} := \sqrt{a + z_n}$ for $n \in \mathbb{N}$. Show that (z_n) converges and find the limit.

7. Let $x_1 := a > 0$ and $x_{n+1} := x_n + 1/x_n$ for $n \in \mathbb{N}$. Determine if (x_n) converges or diverges.

8. Let (a_n) be an increasing sequence, (b_n) a decreasing sequence, and assume that $a_n \leq b_n$ for all $n \in \mathbb{N}$. Show that $\lim(a_n) \leq \lim(b_n)$, and thereby deduce the Nested Intervals Property 2.5.2 from the Monotone Convergence Theorem 3.3.2.

9. Let A be an infinite subset of $\mathbb{R}$ that is bounded above and let $u := \sup A$. Show there exists an increasing sequence (x_n) with $x_n \in A$ for all $n \in \mathbb{N}$ such that $u = \lim(x_n)$.

10. Let (x_n) be a bounded sequence, and for each $n \in \mathbb{N}$ let $s_n := \sup\{x_k : k \geq n\}$ and $t_n := \inf\{x_k : k \geq n\}$. Prove that (s_n) and (t_n) are monotone and convergent. Also prove that if $\lim(s_n) = \lim(t_n)$, then (x_n) is convergent. [One calls $\lim(s_n)$ the **limit superior** of (x_n), and $\lim(t_n)$ the **limit inferior** of (x_n).]

11. Establish the convergence or the divergence of the sequence (y_n), where

$$y_n := \frac{1}{n+1} + \frac{1}{n+2} + \cdots + \frac{1}{2n} \qquad \text{for} \quad n \in \mathbb{N}.$$

12. Let $x_n := 1/1^2 + 1/2^2 + \cdots + 1/n^2$ for each $n \in \mathbb{N}$. Prove that (x_n) is increasing and bounded, and hence converges. [*Hint*: Note that if $k \geq 2$, then $1/k^2 \leq 1/k(k-1) = 1/(k-1) - 1/k$.]

13. Establish the convergence and find the limits of the following sequences.
(a) $\big((1+1/n)^{n+1}\big)$, (b) $\big((1+1/n)^{2n}\big)$,
(c) $\left(\left(1+\dfrac{1}{n+1}\right)^n\right)$, (d) $\big((1-1/n)^n\big)$.

14. Use the method in Example 3.3.5 to calculate $\sqrt{2}$, correct to within 4 decimals.

15. Use the method in Example 3.3.5 to calculate $\sqrt{5}$, correct to within 5 decimals.

16. Calculate the number e_n in Example 3.3.6 for $n = 2, 4, 8, 16$.

17. Use a calculator to compute e_n for $n = 50$, $n = 100$, and $n = 1{,}000$.

Section 3.4 Subsequences and the Bolzano-Weierstrass Theorem

In this section we will introduce the notion of a subsequence of a sequence of real numbers. Informally, a subsequence of a sequence is a selection of terms from the given sequence such that the selected terms form a new sequence. Usually the selection is made for a definite purpose. For example, subsequences are often useful in establishing the convergence or the divergence of the sequence. We will also prove the important existence theorem known as the Bolzano-Weierstrass Theorem, which will be used to establish a number of significant results.

3.4.1 Definition Let $X = (x_n)$ be a sequence of real numbers and let $n_1 < n_2 < \cdots < n_k < \cdots$ be a strictly increasing sequence of natural numbers. Then the sequence $X' = (x_{n_k})$ given by

$$\left(x_{n_1}, x_{n_2}, \cdots, x_{n_k}, \cdots\right)$$

is called a **subsequence** of X.

For example, if $X := \left(\frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \cdots\right)$, then the selection of even indexed terms produces the subsequence

$$X' = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{6}, \cdots, \frac{1}{2k}, \cdots\right),$$

where $n_1 = 2, n_2 = 4, \cdots, n_k = 2k, \cdots$. Other subsequences of $X = (1/n)$ are the following:

$$\left(\frac{1}{1}, \frac{1}{3}, \frac{1}{5}, \cdots, \frac{1}{2k-1}, \cdots\right), \quad \left(\frac{1}{2!}, \frac{1}{4!}, \frac{1}{6!}, \cdots, \frac{1}{(2k)!}, \cdots\right).$$

The following sequences are *not* subsequences of $X = (1/n)$:

$$\left(\frac{1}{2}, \frac{1}{1}, \frac{1}{4}, \frac{1}{3}, \frac{1}{6}, \frac{1}{5}, \cdots\right), \quad \left(\frac{1}{1}, 0, \frac{1}{3}, 0, \frac{1}{5}, 0, \cdots\right).$$

A tail of a sequence (see 3.1.8) is a special type of subsequence. In fact, the m-tail corresponds to the sequence of indices

$$n_1 = m + 1, n_2 = m + 2, \cdots, n_k = m + k, \cdots.$$

But, clearly, not every subsequence of a given sequence need be a tail of the sequence.

Subsequences of convergent sequences also converge to the same limit, as we now show.

3.4.2 Theorem *If a sequence $X = (x_n)$ of real numbers converges to a real number x, then any subsequence $X' = (x_{n_k})$ of X also converges to x.*

Proof. Let $\varepsilon > 0$ be given and let $K(\varepsilon)$ be such that if $n \geq K(\varepsilon)$, then $|x_n - x| < \varepsilon$. Since $n_1 < n_2 < \cdots < n_k < \cdots$ is an increasing sequence of natural numbers, it is easily proved (by Induction) that $n_k \geq k$. Hence, if $k \geq K(\varepsilon)$, we also have $n_k \geq k \geq K(\varepsilon)$ so that $|x_{n_k} - x| < \varepsilon$. Therefore the subsequence (x_{n_k}) also converges to x. Q.E.D.

3.4.3 Example **(a)** $\lim(b^n) = 0$ if $0 < b < 1$.

We have already seen, in Example 3.1.11(b), that if $0 < b < 1$ and if $x_n := b^n$, then it follows from Bernoulli's Inequality that $\lim(x_n) = 0$. Alternatively, we see that since $0 < b < 1$, then $x_{n+1} = b^{n+1} < b^n = x_n$ so that the sequence (x_n) is decreasing. It is also clear that $0 \leq x_n \leq 1$, so it follows from the Monotone Convergence Theorem 3.3.2 that the sequence is convergent. Let $x := \lim x_n$. Since (x_{2n}) is a subsequence of (x_n) it follows from Theorem 3.4.2 that $x = \lim(x_{2n})$. Moreover, it follows from the relation $x_{2n} = b^{2n} = (b^n)^2 = x_n^2$ and Theorem 3.2.3 that

$$x = \lim(x_{2n}) = \left(\lim(x_n)\right)^2 = x^2.$$

Therefore we must either have $x = 0$ or $x = 1$. Since the sequence (x_n) is decreasing and bounded above by $b < 1$, we deduce that $x = 0$.

(b) $\lim(c^{1/n}) = 1$ for $c > 1$.

This limit has been obtained in Example 3.1.11(c) for $c > 0$, using a rather ingenious argument. We give here an alternative approach for the case $c > 1$. Note that if $z_n := c^{1/n}$, then $z_n > 1$ and $z_{n+1} < z_n$ for all $n \in \mathbb{N}$. (Why?) Thus by the Monotone Convergence Theorem, the limit $z := \lim(z_n)$ exists. By Theorem 3.4.2, it follows that $z = \lim(z_{2n})$. In addition, it follows from the relation

$$z_{2n} = c^{1/2n} = (c^{1/n})^{1/2} = z_n^{1/2}$$

and Theorem 3.2.10 that

$$z = \lim(z_{2n}) = \left(\lim(z_n)\right)^{1/2} = z^{1/2}.$$

Therefore we have $z^2 = z$ whence it follows that either $z = 0$ or $z = 1$. Since $z_n > 1$ for all $n \in \mathbb{N}$, we deduce that $z = 1$.

We leave it as an exercise to the reader to consider the case $0 < c < 1$. □

The following result is based on a careful negation of the definition of $\lim(x_n) = x$. It leads to a convenient way to establish the divergence of a sequence.

3.4.4 Theorem *Let $X = (x_n)$ be a sequence of real numbers. Then the following are equivalent:*

(i) *The sequence $X = (x_n)$ does not converge to $x \in \mathbb{R}$.*

(ii) *There exists an* $\varepsilon_0 > 0$ *such that for any* $k \in \mathbb{N}$, *there exists* $n_k \in \mathbb{N}$ *such that* $n_k \geq k$ *and* $|x_{n_k} - x| \geq \varepsilon_0$.
(iii) *There exists an* $\varepsilon_0 > 0$ *and a subsequence* $X' = (x_{n_k})$ *of* X *such that* $|x_{n_k} - x| \geq \varepsilon_0$ *for all* $k \in \mathbb{N}$.

Proof. (i) $\Rightarrow$ (ii) If (x_n) does not converge to x, then for some $\varepsilon_0 > 0$ it is impossible to find a natural number k such that for all $n \geq k$ the terms x_n satisfy $|x_n - x| < \varepsilon_0$. That is, for each $k \in \mathbb{N}$ it is *not true* that for *all* $n \geq k$ the inequality $|x_n - x| < \varepsilon_0$ holds. In other words, for each $k \in \mathbb{N}$ there exists a natural number $n_k \geq k$ such that $|x_{n_k} - x| \geq \varepsilon_0$.

(ii) $\Rightarrow$ (iii) Let ε_0 be as in (ii) and let $n_1 \in \mathbb{N}$ be such that $n_1 \geq 1$ and $|x_{n_1} - x| \geq \varepsilon_0$. Now let $n_2 \in \mathbb{N}$ be such that $n_2 > n_1$ and $|x_{n_2} - x| \geq \varepsilon_0$; let $n_3 \in \mathbb{N}$ be such that $n_3 > n_2$ and $|x_{n_3} - x| \geq \varepsilon_0$. Continue in this way to obtain a subsequence $X' = (x_{n_k})$ of X such that $|x_{n_k} - x| \geq \varepsilon_0$ for all $k \in \mathbb{N}$.

(iii) $\Rightarrow$ (i) Suppose $X = (x_n)$ has a subsequence $X' = (x_{n_k})$ satisfying the condition in (iii). Then X cannot converge to x; for if it did, then, by Theorem 3.4.2, the subsequence X' would also converge to x. But this is impossible, since none of the terms of X' belongs to the ε_0-neighborhood of x. Q.E.D.

Since all subsequences of a convergent sequence must converge to the same limit, we have part (i) in the following result. Part (ii) follows from the fact that a convergent sequence is bounded.

3.4.5 Divergence Criteria *If a sequence* $X = (x_n)$ *of real numbers has either of the following properties, then* X *is divergent.*
(i) X *has two convergent subsequences* $X' = (x_{n_k})$ *and* $X'' = (x_{r_k})$ *whose limits are not equal.*
(ii) X *is unbounded.*

3.4.6 Examples (a) The sequence $X := ((-1^n))$ is divergent.

The subsequence $X' := ((-1)^{2n}) = (1, 1, \cdots)$ converges to 1, and the subsequence $X'' := ((-1)^{2n-1}) = (-1, -1, \cdots)$ converges to -1. Therefore, we conclude from Theorem 3.4.5(i) that X is divergent.

(b) The sequence $\left(1, \frac{1}{2}, 3, \frac{1}{4}, \cdots\right)$ is divergent.

This is the sequence $Y = (y_n)$, where $y_n = n$ if n is odd, and $y_n = 1/n$ if n is even. It can easily be seen that Y is not bounded. Hence, by Theorem 3.4.5(ii), the sequence is divergent.

(c) The sequence $S := (\sin n)$ is divergent.

This sequence is not so easy to handle. In discussing it we must, of course, make use of elementary properties of the sine function. We recall that $\sin(\pi/6) = \frac{1}{2} = \sin(5\pi/6)$ and that $\sin x > \frac{1}{2}$ for x in the interval $I_1 := (\pi/6, 5\pi/6)$. Since the length of I_1 is $5\pi/6 - \pi/6 = 2\pi/3 > 2$, there are at least two natural numbers lying inside I_1; we let n_1 be the first such number. Similarly, for each $k \in \mathbb{N}$, $\sin x > \frac{1}{2}$ for x in the interval

$$I_k := \Big(\pi/6 + 2\pi(k-1), 5\pi/6 + 2\pi(k-1)\Big).$$

Since the length of I_k is greater than 2, there are at least two natural numbers lying inside I_k; we let n_k be the first one. The subsequence $S' := (\sin n_k)$ of S obtained in this way has the property that all of its values lie in the interval $\left[\frac{1}{2}, 1\right]$.

Similarly, if $k \in \mathbb{N}$ and J_k is the interval

$$J_k := \left(7\pi/6 + 2\pi(k-1),\ 11\pi/6 + 2\pi(k-1)\right),$$

then it is seen that $\sin x < -\frac{1}{2}$ for all $x \in J_k$ and the length of J_k is greater than 2. Let m_k be the first natural number lying in J_k. Then the subsequence $S'' := (\sin m_k)$ of S has the property that all of its values lie in the interval $\left[-1, -\frac{1}{2}\right]$.

Given any real number c, it is readily seen that at least one of the subsequences S' and S'' lies entirely outside of the $\frac{1}{2}$-neighborhood of c. Therefore c cannot be a limit of S. Since $c \in \mathbb{R}$ is arbitrary, we deduce that S is divergent. □

The Existence of Monotone Subsequences

While not every sequence is a monotone sequence, we will now show that every sequence has a monotone subsequence.

3.4.7 Monotone Subsequence Theorem *If $X = (x_n)$ is a sequence of real numbers, then there is a subsequence of X that is monotone.*

Proof. For the purpose of this proof, we will say that the mth term x_m is a "peak" if $x_m \geq x_n$ for all n such that $n \geq m$. (That is, x_m is never exceeded by any term that follows it in the sequence.) Note that, in a decreasing sequence, every term is a peak, while in an increasing sequence, no term is a peak.

We will consider two cases, depending on whether X has infinitely many, or finitely many, peaks.

Case 1: X has infinitely many peaks. In this case, we list the peaks by increasing subscripts: $x_{m_1}, x_{m_2}, \cdots, x_{m_k}, \cdots$. Since each term is a peak, we have

$$x_{m_1} \geq x_{m_2} \geq \cdots \geq x_{m_k} \geq \cdots.$$

Therefore, the subsequence (x_{m_k}) of peaks is a decreasing subsequence of X.

Case 2: X has a finite number (possibly zero) of peaks. Let these peaks be listed by increasing subscripts: $x_{m_1}, x_{m_2}, \cdots, x_{m_r}$. Let $s_1 := m_r + 1$ be the first index beyond the last peak. Since x_{s_1} is not a peak, there exists $s_2 > s_1$ such that $x_{s_1} < x_{s_2}$. Since x_{s_2} is not a peak, there exists $s_3 > s_2$ such that $x_{s_2} < x_{s_3}$. Continuing in this way, we obtain an increasing subsequence (x_{s_k}) of X. Q.E.D.

It is not difficult to see that a given sequence may have one subsequence that is increasing, and another subsequence that is decreasing.

The Bolzano-Weierstrass Theorem

We will now use the Monotone Subsequence Theorem to prove the Bolzano-Weierstrass Theorem, which states that every bounded sequence has a convergent subsequence. Because of the importance of this theorem we will also give a second proof of it based on the Nested Interval Property.

3.4.8 The Bolzano-Weierstrass Theorem *A bounded sequence of real numbers has a convergent subsequence.*

First Proof. It follows from the Monotone Subsequence Theorem that if $X = (x_n)$ is a bounded sequence, then it has a subsequence $X' = (x_{n_k})$ that is monotone. Since this subsequence is also bounded, it follows from the Monotone Convergence Theorem 3.3.2 that the subsequence is convergent. Q.E.D.

Second Proof. Since the set of values $\{x_n : n \in \mathbb{N}\}$ is bounded, this set is contained in an interval $I_1 := [a, b]$. We take $n_1 := 1$.

We now bisect I_1 into two equal subintervals I_1' and I_1'', and divide the set of indices $\{n \in \mathbb{N} : n > 1\}$ into two parts:

$$A_1 := \{n \in \mathbb{N} : n > n_1, x_n \in I_1'\}, \qquad B_1 = \{n \in \mathbb{N} : n > n_1, x_n \in I_1''\}.$$

If A_1 is infinite, we take $I_2 := I_1'$ and let n_2 be the smallest natural number in A_1. (See 1.2.1.) If A_1 is a finite set, then B_1 must be infinite, and we take $I_2 := I_1''$ and let n_2 be the smallest natural number in B_1.

We now bisect I_2 into two equal subintervals I_2' and I_2'', and divide the set $\{n \in \mathbb{N} : n > n_2\}$ into two parts:

$$A_2 = \{n \in \mathbb{N} : n > n_2, x_n \in I_2'\}, \qquad B_2 := \{n \in \mathbb{N} : n > n_2, x_n \in I_2''\}$$

If A_2 is infinite, we take $I_3 := I_2'$ and let n_3 be the smallest natural number in A_2. If A_2 is a finite set, then B_2 must be infinite, and we take $I_3 := I_2''$ and let n_3 be the smallest natural number in B_2.

We continue in this way to obtain a sequence of nested intervals $I_1 \supseteq I_2 \supseteq \cdots \supseteq I_k \supseteq \cdots$ and a subsequence (x_{n_k}) of X such that $x_{n_k} \in I_k$ for $k \in \mathbb{N}$. Since the length of I_k is equal to $(b-a)/2^{k-1}$, it follows from Theorem 2.5.3 that there is a (unique) common point $\xi \in I_k$ for all $k \in \mathbb{N}$. Moreover, since x_{n_k} and ξ both belong to I_k, we have

$$|x_{n_k} - \xi| \le (b-a)/2^{k-1},$$

whence it follows that the subsequence (x_{n_k}) of X converges to ξ. Q.E.D.

Theorem 3.4.8 is sometimes called the Bolzano-Weierstrass Theorem *for sequences*, because there is another version of it that deals with bounded sets in $\mathbb{R}$ (see Exercise 11.2.6).

It is readily seen that a bounded sequence can have various subsequences that converge to different limits or even diverge. For example, the sequence $((-1)^n)$ has subsequences that converge to -1, other subsequences that converge to $+1$, and it has subsequences that diverge.

Let X be a sequence of real numbers and let X' be a subsequence of X. Then X' is a sequence in its own right, and so it has subsequences. We note that if X'' is a subsequence of X', then it is also a subsequence of X.

3.4.9 Theorem *Let $X = (x_n)$ be a bounded sequence of real numbers and let $x \in \mathbb{R}$ have the property that every convergent subsequence of X converges to x. Then the sequence X converges to x.*

Proof. Suppose $M > 0$ is a bound for the sequence X so that $|x_n| \le M$ for all $n \in \mathbb{N}$. If X does not converge to x, then Theorem 3.4.4 implies that there exist $\varepsilon_0 > 0$ and a subsequence $X' = (x_{n_k})$ of X such that

$$(1) \qquad |x_{n_k} - x| \ge \varepsilon_0 \qquad \text{for all} \quad k \in \mathbb{N}.$$

Since X' is a subsequence of X, the number M is also a bound for X'. Hence the Bolzano-Weierstrass Theorem implies that X' has a convergent subsequence X''. Since X'' is also a subsequence of X, it converges to x by hypothesis. Thus, its terms ultimately belong to the ε_0-neighborhood of x, contradicting (1). Q.E.D.

Exercises for Section 3.4

1. Give an example of an unbounded sequence that has a convergent subsequence.
2. Use the method of Example 3.4.3(b) to show that if $0 < c < 1$, then $\lim(c^{1/n}) = 1$.
3. Let (f_n) be the Fibonacci sequence of Example 3.1.2(d), and let $x_n := f_{n+1}/f_n$. Given that $\lim(x_n) = L$ exists, determine the value of L.
4. Show that the following sequences are divergent.
   (a) $\left(1 - (-1)^n + 1/n\right)$, (b) $(\sin n\pi/4)$.
5. Let $X = (x_n)$ and $Y = (y_n)$ be given sequences, and let the "shuffled" sequence $Z = (z_n)$ be defined by $z_1 := x_1, z_2 := y_1, \cdots, z_{2n-1} := x_n, z_{2n} := y_n, \cdots$. Show that Z is convergent if and only if both X and Y are convergent and $\lim X = \lim Y$.
6. Let $x_n := n^{1/n}$ for $n \in \mathbb{N}$.
   (a) Show that $x_{n+1} < x_n$ if and only if $(1 + 1/n)^n < n$, and infer that the inequality is valid for $n \geq 3$. (See Example 3.3.6.) Conclude that (x_n) is ultimately decreasing and that $x := \lim(x_n)$ exists.
   (b) Use the fact that the subsequence (x_{2n}) also converges to x to conclude that $x = 1$.
7. Establish the convergence and find the limits of the following sequences:
   (a) $\left((1 + 1/n^2)^{n^2}\right)$, (b) $\left((1 + 1/2n)^n\right)$,
   (c) $\left((1 + 1/n^2)^{2n^2}\right)$, (d) $\left((1 + 2/n)^n\right)$.
8. Determine the limits of the following.
   (a) $\left((3n)^{1/2n}\right)$, (b) $\left((1 + 1/2n)^{3n}\right)$.
9. Suppose that every subsequence of $X = (x_n)$ has a subsequence that converges to 0. Show that $\lim X = 0$.
10. Let (x_n) be a bounded sequence and for each $n \in \mathbb{N}$ let $s_n := \sup\{x_k : k \geq n\}$ and $S := \inf\{s_n\}$. Show that there exists a subsequence of (x_n) that converges to S.
11. Suppose that $x_n \geq 0$ for all $n \in \mathbb{N}$ and that $\lim\left((-1)^n x_n\right)$ exists. Show that (x_n) converges.
12. Show that if (x_n) is unbounded, then there exists a subsequence (x_{n_k}) such that $\lim(1/x_{n_k}) = 0$.
13. If $x_n := (-1)^n/n$, find the subsequence of (x_n) that is constructed in the second proof of the Bolzano-Weierstrass Theorem 3.4.8, when we take $I_1 := [-1, 1]$.
14. Let (x_n) be a bounded sequence and let $s := \sup\{x_n : n \in \mathbb{N}\}$. Show that if $s \notin \{x_n : n \in \mathbb{N}\}$, then there is a subsequence of (x_n) that converges to s.
15. Let (I_n) be a nested sequence of closed bounded intervals. For each $n \in \mathbb{N}$, let $x_n \in I_n$. Use the Bolzano-Weierstrass Theorem to give a proof of the Nested Intervals Property 2.5.2.
16. Give an example to show that Theorem 3.4.9 fails if the hypothesis that X is a bounded sequence is dropped.

Section 3.5 The Cauchy Criterion

The Monotone Convergence Theorem is extraordinarily useful and important, but it has the significant drawback that it applies only to sequences that are monotone. It is important for us to have a condition implying the convergence of a sequence that does not require us to know the value of the limit in advance, and is not restricted to monotone sequences. The Cauchy Criterion, which will be established in this section, is such a condition.

3.5.1 Definition A sequence $X = (x_n)$ of real numbers is said to be a **Cauchy sequence** if for every $\varepsilon > 0$ there exists a natural number $H(\varepsilon)$ such that for all natural numbers $n, m \geq H(\varepsilon)$, the terms x_n, x_m satisfy $|x_n - x_m| < \varepsilon$.

The significance of the concept of Cauchy sequence lies in the main theorem of this section, which asserts that a sequence of real numbers is convergent if and only if it is a Cauchy sequence. This will give us a method of proving a sequence converges without knowing the limit of the sequence.

However, we will first highlight the definition of Cauchy sequence in the following examples.

3.5.2 Examples (a) The sequence $(1/n)$ is a Cauchy sequence.

If $\varepsilon > 0$ is given, we choose a natural number $H = H(\varepsilon)$ such that $H > 2/\varepsilon$. Then if $m, n \geq H$, we have $1/n \leq 1/H < \varepsilon/2$ and similarly $1/m < \varepsilon/2$. Therefore, it follows that if $m, n \geq H$, then

$$\left|\frac{1}{n} - \frac{1}{m}\right| \leq \frac{1}{n} + \frac{1}{m} < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, we conclude that $(1/n)$ is a Cauchy sequence.

(b) The sequence $(1 + (-1)^n)$ is *not* a Cauchy sequence.

The negation of the definition of Cauchy sequence is: There exists $\varepsilon_0 > 0$ such that for every H there exist at least one $n > H$ and at least one $m > H$ such that $|x_n - x_m| \geq \varepsilon_0$. For the terms $x_n := 1 + (-1)^n$, we observe that if n is even, then $x_n = 2$ and $x_{n+1} = 0$. If we take $\varepsilon_0 = 2$, then for any H we can choose an even number $n > H$ and let $m := n + 1$ to get

$$|x_n - x_{n+1}| = 2 = \varepsilon_0.$$

We conclude that (x_n) is not a Cauchy sequence. □

Remark We emphasize that to prove a sequence (x_n) is a Cauchy sequence, we may not assume a relationship between m and n, since the required inequality $|x_n - x_m| < \varepsilon$ must hold for *all* $n, m \geq H(\varepsilon)$. But to prove a sequence is *not* a Cauchy sequence, we may specify a relation between n and m as long as arbitrarily large values of n and m can be chosen so that $|x_n - x_m| \geq \varepsilon_0$.

Our goal is to show that the Cauchy sequences are precisely the convergent sequences. We first prove that a convergent sequence is a Cauchy sequence.

3.5.3 Lemma *If $X = (x_n)$ is a convergent sequence of real numbers, then X is a Cauchy sequence.*

Proof. If $x := \lim X$, then given $\varepsilon > 0$ there is a natural number $K(\varepsilon/2)$ such that if $n \geq K(\varepsilon/2)$ then $|x_n - x| < \varepsilon/2$. Thus, if $H(\varepsilon) := K(\varepsilon/2)$ and if $n, m \geq H(\varepsilon)$, then we have

$$\begin{aligned} |x_n - x_m| &= |(x_n - x) + (x - x_m)| \\ &\leq |x_n - x| + |x_m - x| < \varepsilon/2 + \varepsilon/2 = \varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, it follows that (x_n) is a Cauchy sequence. Q.E.D.

In order to establish that a Cauchy sequence is convergent, we will need the following result. (See Theorem 3.2.2.)

3.5.4 Lemma *A Cauchy sequence of real numbers is bounded.*

Proof. Let $X := (x_n)$ be a Cauchy sequence and let $\varepsilon := 1$. If $H := H(1)$ and $n \geq H$, then $|x_n - x_H| < 1$. Hence, by the Triangle Inequality, we have $|x_n| \leq |x_H| + 1$ for all $n \geq H$. If we set

$$M := \sup\left\{|x_1|, |x_2|, \cdots, |x_{H-1}|, |x_H| + 1\right\},$$

then it follows that $|x_n| \leq M$ for all $n \in \mathbb{N}$. Q.E.D.

We now present the important Cauchy Convergence Criterion.

3.5.5 Cauchy Convergence Criterion *A sequence of real numbers is convergent if and only if it is a Cauchy sequence.*

Proof. We have seen, in Lemma 3.5.3, that a convergent sequence is a Cauchy sequence.

Conversely, let $X = (x_n)$ be a Cauchy sequence; we will show that X is convergent to some real number. First we observe from Lemma 3.5.4 that the sequence X is bounded. Therefore, by the Bolzano-Weierstrass Theorem 3.4.8, there is a subsequence $X' = (x_{n_k})$ of X that converges to some real number x^*. We shall complete the proof by showing that X converges to x^*.

Since $X = (x_n)$ is a Cauchy sequence, given $\varepsilon > 0$ there is a natural number $H(\varepsilon/2)$ such that if $n, m \geq H(\varepsilon/2)$ then

$$|x_n - x_m| < \varepsilon/2. \tag{1}$$

Since the subsequence $X' = (x_{n_k})$ converges to x^*, there is a natural number $K \geq H(\varepsilon/2)$ belonging to the set $\{n_1, n_2, \cdots\}$ such that

$$|x_K - x^*| < \varepsilon/2.$$

Since $K \geq H(\varepsilon/2)$, it follows from (1) with $m = K$ that

$$|x_n - x_K| < \varepsilon/2 \qquad \text{for} \quad n \geq H(\varepsilon/2).$$

Therefore, if $n \geq H(\varepsilon/2)$, we have

$$\begin{aligned} |x_n - x^*| &= |(x_n - x_K) + (x_K - x^*)| \\ &\leq |x_n - x_K| + |x_K - x^*| \\ &< \varepsilon/2 + \varepsilon/2 = \varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, we infer that $\lim(x_n) = x^*$. Therefore the sequence X is convergent. Q.E.D.

We will now give some examples of applications of the Cauchy Criterion.

3.5.6 Examples **(a)** Let $X = (x_n)$ be defined by

$$x_1 := 1, \qquad x_2 := 2, \qquad \text{and} \qquad x_n := \frac{1}{2}\left(x_{n-2} + x_{n-1}\right) \qquad \text{for} \quad n > 2.$$

n be shown by Induction that $1 \leq x_n \leq 2$ for all $n \in \mathbb{N}$. (Do so.) Some calculation shows at the sequence X is not monotone. However, since the terms are formed by averaging, it readily seen that

$$|x_n - x_{n+1}| = \frac{1}{2^{n-1}} \quad \text{for} \quad n \in \mathbb{N}.$$

Prove this by Induction.) Thus, if $m > n$, we may employ the Triangle Inequality to obtain

$$\begin{aligned} |x_n - x_m| &\leq |x_n - x_{n+1}| + |x_{n+1} - x_{n+2}| + \cdots + |x_{m-1} - x_m| \\ &= \frac{1}{2^{n-1}} + \frac{1}{2^n} + \cdots + \frac{1}{2^{m-2}} \\ &= \frac{1}{2^{n-1}}\left(1 + \frac{1}{2} + \cdots + \frac{1}{2^{m-n-1}}\right) < \frac{1}{2^{n-2}}. \end{aligned}$$

Therefore, given $\varepsilon > 0$, if n is chosen so large that $1/2^n < \varepsilon/4$ and if $m \geq n$, then it follows that $|x_n - x_m| < \varepsilon$. Therefore, X is a Cauchy sequence in $\mathbb{R}$. By the Cauchy Criterion 3.5.5 we infer that the sequence X converges to a number x.

To evaluate the limit x, we might first "pass to the limit" in the rule of definition $= \frac{1}{2}(x_{n-1} + x_{n-2})$ to conclude that x must satisfy the relation $x = \frac{1}{2}(x + x)$, which is true, but not informative. Hence we must try something else.

Since X converges to x, so does the subsequence X' with odd indices. By Induction, the reader can establish that [see 1.2.4(f)]

$$\begin{aligned} x_{2n+1} &= 1 + \frac{1}{2} + \frac{1}{2^3} + \cdots + \frac{1}{2^{2n-1}} \\ &= 1 + \frac{2}{3}\left(1 - \frac{1}{4^n}\right). \end{aligned}$$

It follows from this (how?) that $x = \lim X = \lim X' = 1 + \frac{2}{3} = \frac{5}{3}$.

(b) Let $Y = (y_n)$ be the sequence of real numbers given by

$$y_1 := \frac{1}{1!}, \quad y_2 := \frac{1}{1!} - \frac{1}{2!}, \ldots, \quad y_n := \frac{1}{1!} - \frac{1}{2!} + \cdots + \frac{(-1)^{n+1}}{n!}, \cdots$$

Clearly, Y is not a monotone sequence. However, if $m > n$, then

$$y_m - y_n = \frac{(-1)^{n+2}}{(n+1)!} + \frac{(-1)^{n+3}}{(n+2)!} + \cdots + \frac{(-1)^{m+1}}{m!}.$$

Since $2^{r-1} \leq r!$ [see 1.2.4(e)], it follows that if $m > n$, then (why?)

$$\begin{aligned} |y_m - y_n| &\leq \frac{1}{(n+1)!} + \frac{1}{(n+2)!} + \cdots + \frac{1}{m!} \\ &\leq \frac{1}{2^n} + \frac{1}{2^{n+1}} + \cdots + \frac{1}{2^{m-1}} < \frac{1}{2^{n-1}}. \end{aligned}$$

Therefore, it follows that (y_n) is a Cauchy sequence. Hence it converges to a limit y. At the present moment we cannot evaluate y directly; however, passing to the limit (with respect to m) in the above inequality, we obtain

$$|y_n - y| \leq 1/2^{n-1}.$$

Hence we can calculate y to any desired accuracy by calculating the terms y_n for sufficiently large n. The reader should do this and show that y is approximately equal to 0.632 120 559. (The exact value of y is $1 - 1/e$.)

(c) The sequence $\left(\frac{1}{1}+\frac{1}{2}+\cdots+\frac{1}{n}\right)$ diverges.

Let $H := (h_n)$ be the sequence defined by

$$h_n := \frac{1}{1}+\frac{1}{2}+\cdots+\frac{1}{n} \qquad \text{for} \quad n \in \mathbb{N},$$

which was considered in 3.3.3(b). If $m > n$, then

$$h_m - h_n = \frac{1}{n+1}+\cdots+\frac{1}{m}.$$

Since each of these $m-n$ terms exceeds $1/m$, then $h_m - h_n > (m-n)/m = 1 - n/m$. In particular, if $m = 2n$ we have $h_{2n} - h_n > \frac{1}{2}$. This shows that H is not a Cauchy sequence (why?); therefore H is *not* a convergent sequence. (In terms that will be introduced in Section 3.7, we have just proved that the "harmonic series" $\sum_{n=1}^{\infty} 1/n$ is divergent.) □

3.5.7 Definition We say that a sequence $X = (x_n)$ of real numbers is **contractive** if there exists a constant C, $0 < C < 1$, such that

$$|x_{n+2} - x_{n+1}| \le C|x_{n+1} - x_n|$$

for all $n \in \mathbb{N}$. The number C is called the **constant** of the contractive sequence.

3.5.8 Theorem *Every contractive sequence is a Cauchy sequence, and therefore is convergent.*

Proof. If we successively apply the defining condition for a contractive sequence, we can work our way back to the beginning of the sequence as follows:

$$\begin{aligned}|x_{n+2} - x_{n+1}| &\le C|x_{n+1} - x_n| \le C^2|x_n - x_{n-1}| \\ &\le C^3|x_{n-1} - x_{n-2}| \le \cdots \le C^n|x_2 - x_1|.\end{aligned}$$

For $m > n$, we estimate $|x_m - x_n|$ by first applying the Triangle Inequality and then using the formula for the sum of a geometric progression (see 1.2.4(f)). This gives

$$\begin{aligned}|x_m - x_n| &\le |x_m - x_{m-1}| + |x_{m-1} - x_{m-2}| + \cdots + |x_{n+1} - x_n| \\ &\le \left(C^{m-2} + C^{m-3} + \cdots + C^{n-1}\right)|x_2 - x_1| \\ &= C^{n-1}\left(\frac{1 - C^{m-n}}{1 - C}\right)|x_2 - x_1| \\ &\le C^{n-1}\left(\frac{1}{1 - C}\right)|x_2 - x_1|.\end{aligned}$$

Since $0 < C < 1$, we know $\lim(C^n) = 0$ [see 3.1.11(b)]. Therefore, we infer that (x_n) is a Cauchy sequence. It now follows from the Cauchy Convergence Criterion 3.5.5. that (x_n) is a convergent sequence. Q.E.D.

In the process of calculating the limit of a contractive sequence, it is often very important to have an estimate of the error at the nth stage. In the next result we give two such estimates: the first one involves the first two terms in the sequence and n; the second one involves the difference $x_n - x_{n-1}$.

3.5.9 Corollary *If $X := (x_n)$ is a contractive sequence with constant C, $0 < C < 1$, and if $x^* := \lim X$, then*

(i) $|x^* - x_n| \le \dfrac{C^{n-1}}{1-C}|x_2 - x_1|$,

(ii) $|x^* - x_n| \le \dfrac{C}{1-C}|x_n - x_{n-1}|$.

Proof. From the preceding proof, if $m > n$, then $|x_m - x_n| \le (C^{n-1}/(1-C))|x_2 - x_1|$. If we let $m \to \infty$ in this inequality, we obtain (i).

To prove (ii), recall that if $m > n$, then

$$|x_m - x_n| \le |x_m - x_{m-1}| + \cdots + |x_{n+1} - x_n|.$$

Since it is readily established, using Induction, that

$$|x_{n+k} - x_{n+k-1}| \le C^k |x_n - x_{n-1}|,$$

we infer that

$$\begin{aligned}|x_m - x_n| &\le (C^{m-n} + \cdots + C^2 + C)|x_n - x_{n-1}| \\ &\le \frac{C}{1-C}|x_n - x_{n-1}|\end{aligned}$$

We now let $m \to \infty$ in this inequality to obtain assertion (ii). Q.E.D.

3.5.10 Example We are told that the cubic equation $x^3 - 7x + 2 = 0$ has a solution between 0 and 1 and we wish to approximate this solution. This can be accomplished by means of an iteration procedure as follows. We first rewrite the equation as $x = (x^3 + 2)/7$ and use this to define a sequence. We assign to x_1 an arbitrary value between 0 and 1, and then define

$$x_{n+1} := \tfrac{1}{7}(x_n^3 + 2) \qquad \text{for} \quad n \in \mathbb{N}.$$

Because $0 < x_1 < 1$, it follows that $0 < x_n < 1$ for all $n \in \mathbb{N}$. (Why?) Moreover, we have

$$\begin{aligned}|x_{n+2} - x_{n+1}| &= \left|\tfrac{1}{7}(x_{n+1}^3 + 2) - \tfrac{1}{7}(x_n^3 + 2)\right| = \tfrac{1}{7}|x_{n+1}^3 - x_n^3| \\ &= \tfrac{1}{7}|x_{n+1}^2 + x_{n+1}x_n + x_n^2||x_{n+1} - x_n| \le \tfrac{3}{7}|x_{n+1} - x_n|.\end{aligned}$$

Therefore, (x_n) is a contractive sequence and hence there exists r such that $\lim(x_n) = r$. If we pass to the limit on both sides of the equality $x_{n+1} = (x_n^3 + 2)/7$, we obtain $r = (r^3 + 2)/7$ and hence $r^3 - 7r + 2 = 0$. Thus r is a solution of the equation.

We can approximate r by choosing x_1 and calculating $x_2, x_3, \cdots$ successively. For example, if we take $x_1 = 0.5$, we obtain (to nine decimal places):

$$\begin{aligned} &x_2 = 0.303\,571\,429, && x_3 = 0.289\,710\,830, \\ &x_4 = 0.289\,188\,016, && x_5 = 0.289\,169\,244, \\ &x_6 = 0.289\,168\,571, && \text{etc.}\end{aligned}$$

To estimate the accuracy, we note that $|x_2 - x_1| < 0.2$. Thus, after n steps it follows from Corollary 3.5.9(i) that we are sure that $|x^* - x_n| \le 3^{n-1}/(7^{n-2} \cdot 20)$. Thus, when $n = 6$, we are sure that

$$|x^* - x_6| \le 3^5/(7^4 \cdot 20) = 243/48\,020 < 0.0051.$$

Actually the approximation is substantially better than this. In fact, since $|x_6 - x_5| < 0.000\,0005$, it follows from 3.5.9(ii) that $|x^* - x_6| \le \frac{3}{4}|x_6 - x_5| < 0.000\,0004$. Hence the first five decimal places of x_6 are correct. □

Exercises for Section 3.5

1. Give an example of a bounded sequence that is not a Cauchy sequence.
2. Show directly from the definition that the following are Cauchy sequences.
   (a) $\left(\dfrac{n+1}{n}\right)$, (b) $\left(1+\dfrac{1}{2!}+\cdots+\dfrac{1}{n!}\right)$.
3. Show directly from the definition that the following are not Cauchy sequences.
   (a) $\left((-1)^n\right)$, (b) $\left(n+\dfrac{(-1)^n}{n}\right)$, (c) $(\ln n)$.
4. Show directly from the definition that if (x_n) and (y_n) are Cauchy sequences, then $(x_n + y_n)$ and $(x_n y_n)$ are Cauchy sequences.
5. If $x_n := \sqrt{n}$, show that (x_n) satisfies $\lim |x_{n+1} - x_n| = 0$, but that it is not a Cauchy sequence.
6. Let p be a given natural number. Give an example of a sequence (x_n) that is not a Cauchy sequence, but that satisfies $\lim |x_{n+p} - x_n| = 0$.
7. Let (x_n) be a Cauchy sequence such that x_n is an integer for every $n \in \mathbb{N}$. Show that (x_n) is ultimately constant.
8. Show directly that a bounded, monotone increasing sequence is a Cauchy sequence.
9. If $0 < r < 1$ and $|x_{n+1} - x_n| < r^n$ for all $n \in \mathbb{N}$, show that (x_n) is a Cauchy sequence.
10. If $x_1 < x_2$ are arbitrary real numbers and $x_n := \frac{1}{2}(x_{n-2} + x_{n-1})$ for $n > 2$, show that (x_n) is convergent. What is its limit?
11. If $y_1 < y_2$ are arbitrary real numbers and $y_n := \frac{1}{3}y_{n-1} + \frac{2}{3}y_{n-2}$ for $n > 2$, show that (y_n) is convergent. What is its limit?
12. If $x_1 > 0$ and $x_{n+1} := (2 + x_n)^{-1}$ for $n \geq 1$, show that (x_n) is a contractive sequence. Find the limit.
13. If $x_1 := 2$ and $x_{n+1} := 2 + 1/x_n$ for $n \geq 1$, show that (x_n) is a contractive sequence. What is its limit?
14. The polynomial equation $x^3 - 5x + 1 = 0$ has a root r with $0 < r < 1$. Use an appropriate contractive sequence to calculate r within 10^{-4}.

Section 3.6 Properly Divergent Sequences

For certain purposes it is convenient to define what is meant for a sequence (x_n) of real numbers to "tend to $\pm\infty$".

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

(i) We say that (x_n) **tends to** $+\infty$, and write $\lim(x_n) = +\infty$, if for every $\alpha \in \mathbb{R}$ there exists a natural number $K(\alpha)$ such that if $n \geq K(\alpha)$, then $x_n > \alpha$.

(ii) We say that (x_n) **tends to** $-\infty$, and write $\lim(x_n) = -\infty$, if for every $\beta \in \mathbb{R}$ there exists a natural number $K(\beta)$ such that if $n \geq K(\beta)$, then $x_n < \beta$.

We say that (x_n) is **properly divergent** in case we have either $\lim(x_n) = +\infty$ or $\lim(x_n) = -\infty$.

The reader should realize that we are using the symbols $+\infty$ and $-\infty$ purely as a convenient *notation* in the above expressions. Results that have been proved in earlier sections for conventional limits $\lim(x_n) = L$ (for $L \in \mathbb{R}$) *may not* remain true when $\lim(x_n) = \pm\infty$.

3.6.2 Examples (a) $\lim(n) = +\infty$.

In fact, if $\alpha \in \mathbb{R}$ is given, let $K(\alpha)$ be any natural number such that $K(\alpha) > \alpha$.

(b) $\lim(n^2) = +\infty$.

If $K(\alpha)$ is a natural number such that $K(\alpha) > \alpha$, and if $n \geq K(\alpha)$ then we have $n^2 \geq n > \alpha$.

(c) If $c > 1$, then $\lim(c^n) = +\infty$.

Let $c = 1 + b$, where $b > 0$. If $\alpha \in \mathbb{R}$ is given, let $K(\alpha)$ be a natural number such that $K(\alpha) > \alpha/b$. If $n \geq K(\alpha)$ it follows from Bernoulli's Inequality that

$$c^n = (1+b)^n \geq 1 + nb > 1 + \alpha > \alpha.$$

Therefore $\lim(c^n) = +\infty$. □

Monotone sequences are particularly simple in regard to their convergence. We have seen in the Monotone Convergence Theorem 3.3.2 that a monotone sequence is convergent if and only if it is bounded. The next result is a reformulation of that result.

3.6.3 Theorem *A monotone sequence of real numbers is properly divergent if and only if it is unbounded.*

(a) *If (x_n) is an unbounded increasing sequence, then* $\lim(x_n) = +\infty$.

(b) *If (x_n) is an unbounded decreasing sequence, then* $\lim(x_n) = -\infty$.

Proof. (a) Suppose that (x_n) is an increasing sequence. We know that if (x_n) is bounded, then it is convergent. If (x_n) is unbounded, then for any $\alpha \in \mathbb{R}$ there exists $n(\alpha) \in \mathbb{N}$ such that $\alpha < x_{n(\alpha)}$. But since (x_n) is increasing, we have $\alpha < x_n$ for all $n \geq n(\alpha)$. Since α is arbitrary, it follows that $\lim(x_n) = +\infty$.

Part (b) is proved in a similar fashion. Q.E.D.

The following "comparison theorem" is frequently used in showing that a sequence is properly divergent. [In fact, we implicitly used it in Example 3.6.2(c).]

3.6.4 Theorem *Let (x_n) and (y_n) be two sequences of real numbers and suppose that*

$$(1) \qquad x_n \leq y_n \qquad \textit{for all} \quad n \in \mathbb{N}.$$

(a) *If* $\lim(x_n) = +\infty$, *then* $\lim(y_n) = +\infty$.

(b) *If* $\lim(y_n) = -\infty$, *then* $\lim(x_n) = -\infty$.

Proof. (a) If $\lim(x_n) = +\infty$, and if $\alpha \in \mathbb{R}$ is given, then there exists a natural number $K(\alpha)$ such that if $n \geq K(\alpha)$, then $\alpha < x_n$. In view of (1), it follows that $\alpha < y_n$ for all $n \geq K(\alpha)$. Since α is arbitrary, it follows that $\lim(y_n) = +\infty$.

The proof of (b) is similar. Q.E.D.

Remarks (a) Theorem 3.6.4 remains true if condition (1) is ultimately true; that is, if there exists $m \in \mathbb{N}$ such that $x_n \leq y_n$ for all $n \geq m$.

(b) If condition (1) of Theorem 3.6.4 holds and if $\lim(y_n) = +\infty$, it does *not* follow that $\lim(x_n) = +\infty$. Similarly, if (1) holds and if $\lim(x_n) = -\infty$, it does *not* follow that $\lim(y_n) = -\infty$. In using Theorem 3.6.4 to show that a sequence tends to $+\infty$ [respectively, $-\infty$] we need to show that the terms of the sequence are ultimately greater [respectively, less] than or equal to the corresponding terms of a sequence that is known to tend to $+\infty$ [respectively, $-\infty$].

Since it is sometimes difficult to establish an inequality such as (1), the following "limit comparison theorem" is often more convenient to use than Theorem 3.6.4.

3.6.5 Theorem *Let (x_n) and (y_n) be two sequences of positive real numbers and suppose that for some $L \in \mathbb{R}$, $L > 0$, we have*

$$\lim(x_n/y_n) = L. \tag{2}$$

Then $\lim(x_n) = +\infty$ if and only if $\lim(y_n) = +\infty$.

Proof. If (2) holds, there exists $K \in \mathbb{N}$ such that

$$\tfrac{1}{2}L < x_n/y_n < \tfrac{3}{2}L \qquad \text{for all} \quad n \geq K.$$

Hence we have $\left(\tfrac{1}{2}L\right) y_n < x_n < \left(\tfrac{3}{2}L\right) y_n$ for all $n \geq K$. The conclusion now follows from a slight modification of Theorem 3.6.4. We leave the details to the reader. Q.E.D.

The reader can show that the conclusion need not hold if either $L = 0$ or $L = +\infty$. However, there are some partial results that can be established in these cases, as will be seen in the exercises.

Exercises for Section 3.6

1. Show that if (x_n) is an unbounded sequence, then there exists a properly divergent subsequence.
2. Give examples of properly divergent sequences (x_n) and (y_n) with $y_n \neq 0$ for all $n \in \mathbb{N}$ such that:
   (a) (x_n/y_n) is convergent, (b) (x_n/y_n) is properly divergent.
3. Show that if $x_n > 0$ for all $n \in \mathbb{N}$, then $\lim(x_n) = 0$ if and only if $\lim(1/x_n) = +\infty$.
4. Establish the proper divergence of the following sequences.
   (a) $(\sqrt{n})$, (b) $(\sqrt{n+1})$,
   (c) $(\sqrt{n-1})$, (d) $(n/\sqrt{n+1})$.
5. Is the sequence $(n \sin n)$ properly divergent?
6. Let (x_n) be properly divergent and let (y_n) be such that $\lim(x_n y_n)$ belongs to $\mathbb{R}$. Show that (y_n) converges to 0.
7. Let (x_n) and (y_n) be sequences of positive numbers such that $\lim(x_n/y_n) = 0$.
   (a) Show that if $\lim(x_n) = +\infty$, then $\lim(y_n) = +\infty$.
   (b) Show that if (y_n) is bounded, then $\lim(x_n) = 0$.
8. Investigate the convergence or the divergence of the following sequences:
   (a) $\left(\sqrt{n^2+2}\right)$, (b) $\left(\sqrt{n}/\left(n^2+1\right)\right)$,
   (c) $\left(\sqrt{n^2+1}/\sqrt{n}\right)$, (d) $\left(\sin\sqrt{n}\right)$.
9. Let (x_n) and (y_n) be sequences of positive numbers such that $\lim(x_n/y_n) = +\infty$,
   (a) Show that if $\lim(y_n) = +\infty$, then $\lim(x_n) = +\infty$.
   (b) Show that if (x_n) is bounded, then $\lim(y_n) = 0$.
10. Show that if $\lim(a_n/n) = L$, where $L > 0$, then $\lim(a_n) = +\infty$.

Section 3.7 Introduction to Infinite Series

We will now give a brief introduction to infinite series of real numbers. This is a topic that will be discussed in more detail in Chapter 9, but because of its importance, we will establish a few results here. These results will be seen to be immediate consequences of theorems we have met in this chapter.

In elementary texts, an infinite series is sometimes "defined" to be "an expression of the form"

$$x_1 + x_2 + \cdots + x_n + \cdots. \tag{1}$$

However, this "definition" lacks clarity, since there is *a priori* no particular value that we can attach to this array of symbols, which calls for an *infinite* number of additions to be performed.

3.7.1 Definition If $X := (x_n)$ is a sequence in $\mathbb{R}$, then the **infinite series** (or simply the **series**) **generated by** X is the sequence $S := (s_k)$ defined by

$$\begin{aligned}
s_1 &:= x_1 \\
s_2 &:= s_1 + x_2 \quad (= x_1 + x_2) \\
&\cdots \\
s_k &:= s_{k-1} + x_k \quad (= x_1 + x_2 + \cdots + x_k) \\
&\cdots
\end{aligned}$$

The numbers x_n are called the **terms** of the series and the numbers s_k are called the **partial sums** of this series. If $\lim S$ exists, we say that this series is **convergent** and call this limit the **sum** or the **value** of this series. If this limit does not exist, we say that the series S is **divergent**.

It is convenient to use symbols such as

$$\sum (x_n) \quad \text{or} \quad \sum x_n \quad \text{or} \quad \sum_{n=1}^{\infty} x_n \tag{2}$$

to denote both the infinite series S generated by the sequence $X = (x_n)$ and also to denote the value $\lim S$, in case this limit exists. Thus the symbols in (2) may be regarded merely as a way of exhibiting an infinite series whose convergence or divergence is to be investigated. In practice, this double use of these notations does not lead to any confusion, provided it is understood that the convergence (or divergence) of the series must be established.

Just as a sequence may be indexed such that its first element is not x_1, but is x_0, or x_5 or x_{99}, we will denote the series having these numbers as their first element by the symbols

$$\sum_{n=0}^{\infty} x_n \quad \text{or} \quad \sum_{n=5}^{\infty} x_n \quad \text{or} \quad \sum_{n=99}^{\infty} x_n.$$

It should be noted that when the first term in the series is x_N, then the first partial sum is denoted by s_N.

Warning The reader should guard against confusing the words "sequence" and "series". In nonmathematical language, these words are interchangeable; however, in mathematics, these words are not synonyms. Indeed, a series is a sequence $S = (s_k)$ obtained from a given sequence $X = (x_n)$ according to the special procedure given in Definition 3.7.1.

3.7.2 Examples **(a)** Consider the sequence $X := (r^n)_{n=0}^{\infty}$ where $r \in \mathbb{R}$, which generates the **geometric series**:

$$\sum_{n=0}^{\infty} r^n = 1 + r + r^2 + \cdots + r^n + \cdots. \tag{3}$$

We will show that if $|r| < 1$, then this series converges to $1/(1-r)$. (See also Example 1.2.4(f).) Indeed, if $s_n := 1 + r + r^2 + \cdots + r^n$ for $n \geq 0$, and if we multiply s_n by r and subtract the result from s_n, we obtain (after some simplification):

$$s_n(1-r) = 1 - r^{n+1}.$$

Therefore, we have

$$s_n - \frac{1}{1-r} = -\frac{r^{n+1}}{1-r},$$

from which it follows that

$$\left| s_n - \frac{1}{1-r} \right| \leq \frac{|r|^{n+1}}{|1-r|}.$$

Since $|r|^{n+1} \to 0$ when $|r| < 1$, it follows that the geometric series (3) converges to $1/(1-r)$ when $|r| < 1$.

(b) Consider the series generated by $((-1)^n)_{n=0}^{\infty}$; that is, the series:

$$\sum_{n=0}^{\infty} (-1)^n = (+1) + (-1) + (+1) + (-1) + \cdots. \tag{4}$$

It is easily seen (by Mathematical Induction) that $s_n = 1$ if $n \geq 0$ is even and $s_n = 0$ if n is odd; therefore, the sequence of partial sums is $(1, 0, 1, 0, \cdots)$. Since this sequence is not convergent, the series (4) is divergent.

(c) Consider the series

$$\sum_{n=1}^{\infty} \frac{1}{n(n+1)} = \frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \frac{1}{3 \cdot 4} + \cdots. \tag{5}$$

By a stroke of insight, we note that

$$\frac{1}{k(k+1)} = \frac{1}{k} - \frac{1}{k+1}.$$

Hence, on adding these terms from $k = 1$ to $k = n$ and noting the telescoping that takes place, we obtain

$$s_n = \frac{1}{1} - \frac{1}{n+1},$$

whence it follows that $s_n \to 1$. Therefore the series (5) converges to 1. □

We now present a very useful and simple *necessary* condition for the convergence of a series. It is far from being sufficient, however.

3.7.3 The nth Term Test *If the series $\sum x_n$ converges, then $\lim(x_n) = 0$.*

Proof. By Definition 3.7.1, the convergence of $\sum x_n$ requires that $\lim(s_k)$ exists. Since $x_n = s_n - s_{n-1}$, then $\lim(x_n) = \lim(s_n) - \lim(s_{n-1}) = 0$. Q.E.D.

Since the following Cauchy Criterion is precisely a reformulation of Theorem 3.5.5, we will omit its proof.

3.7.4 Cauchy Criterion for Series *The series $\sum x_n$ converges if and only if for every $\varepsilon > 0$ there exists $M(\varepsilon) \in \mathbb{N}$ such that if $m > n \geq M(\varepsilon)$, then*

(6)
$$|s_m - s_n| = |x_{n+1} + x_{n+2} + \cdots + x_m| < \varepsilon.$$

The next result, although limited in scope, is of great importance and utility.

3.7.5 Theorem *Let (x_n) be a sequence of nonnegative real numbers. Then the series $\sum x_n$ converges if and only if the sequence $S = (s_k)$ of partial sums is bounded. In this case,*

$$\sum_{n=1}^{\infty} x_n = \lim(s_k) = \sup\{s_k : k \in \mathbb{N}\}.$$

Proof. Since $x_n > 0$, the sequence S of partial sums is monotone increasing:

$$s_1 \leq s_2 \leq \cdots \leq s_k \leq \cdots.$$

By the Monotone Convergence Theorem 3.3.2, the sequence $S = (s_k)$ converges if and only if it is bounded, in which case its limit equals $\sup\{s_k\}$. Q.E.D.

3.7.6 Examples **(a)** The geometric series (3) diverges if $|r| \geq 1$.

This follows from the fact that the terms r^n do not approach 0 when $|r| \geq 1$.

(b) The **harmonic series** $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n}$ diverges.

Since the terms $1/n \to 0$, we cannot use the nth Term Test 3.7.3 to establish this divergence. However, it was seen in Examples 3.3.3(b) and 3.5.6(c) that the sequence (s_n) of partial sums is not bounded. Therefore, it follows from Theorem 3.7.5 that the harmonic series is divergent.

(c) The 2-**series** $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n^2}$ is convergent.

Since the partial sums are monotone, it suffices (why?) to show that some subsequence of (s_k) is bounded. If $k_1 := 2^1 - 1 = 1$, then $s_{k_1} = 1$. If $k_2 := 2^2 - 1 = 3$, then

$$s_{k_2} = \frac{1}{1} + \left(\frac{1}{2^2} + \frac{1}{3^2}\right) < 1 + \frac{2}{2^2} = 1 + \frac{1}{2},$$

and if $k_3 := 2^3 - 1 = 7$, then we have

$$s_{k_3} = s_{k_2} + \left(\frac{1}{4^2} + \frac{1}{5^2} + \frac{1}{6^2} + \frac{1}{7^2}\right) < s_{k_2} + \frac{4}{4^2} < 1 + \frac{1}{2} + \frac{1}{2^2}.$$

By Mathematical Induction, we find that if $k_j := 2^j - 1$, then

$$0 < s_{k_j} < 1 + \tfrac{1}{2} + \left(\tfrac{1}{2}\right)^2 + \cdots + \left(\tfrac{1}{2}\right)^{j-1}.$$

Since the term on the right is a partial sum of a geometric series with $r = \frac{1}{2}$, it is dominated by $1/(1-\frac{1}{2}) = 2$, and Theorem 3.7.5 implies that the 2-series converges.

(d) The p**-series** $\sum_{n=1}^{\infty} \frac{1}{n^p}$ converges when $p > 1$.

Since the argument is very similar to the special case considered in part (c), we will leave some of the details to the reader. As before, if $k_1 := 2^1 - 1 = 1$, then $s_{k_1} = 1$. If $k_2 := 2^2 - 1 = 3$, then since $2^p < 3^p$, we have

$$s_{k_2} = \frac{1}{1^p} + \left(\frac{1}{2^p} + \frac{1}{3^p}\right) < 1 + \frac{2}{2^p} = 1 + \frac{1}{2^{p-1}}.$$

Further, if $k_3 := 2^3 - 1$, then (how?) it is seen that

$$s_{k_3} < s_{k_2} + \frac{4}{4^p} < 1 + \frac{1}{2^{p-1}} + \frac{1}{4^{p-1}}.$$

Finally, we let $r := 1/2^{p-1}$; since $p > 1$, we have $0 < r < 1$. Using Mathematical Induction, we show that if $k_j := 2^j - 1$, then

$$0 < s_{k_j} < 1 + r + r^2 + \cdots + r^{j-1} < \frac{1}{1-r}.$$

Therefore, Theorem 3.7.5 implies that the p-series converges when $p > 1$.

(e) The p**-series** $\sum_{n=1}^{\infty} \frac{1}{n^p}$ diverges when $0 < p \le 1$.

We will use the elementary inequality $n^p \le n$ when $n \in \mathbb{N}$ and $0 < p \le 1$. It follows that

$$\frac{1}{n} \le \frac{1}{n^p} \qquad \text{for} \quad n \in \mathbb{N}.$$

Since the partial sums of the harmonic series are not bounded, this inequality shows that the partial sums of the p-series are not bounded when $0 < p \le 1$. Hence the p-series diverges for these values of p.

(f) The **alternating harmonic series,** given by

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} = \frac{1}{1} - \frac{1}{2} + \frac{1}{3} - \cdots + \frac{(-1)^{n+1}}{n} + \cdots \tag{7}$$

is convergent.

The reader should compare this series with the harmonic series in (b), which is divergent. Thus, the subtraction of some of the terms in (7) is essential if this series is to converge. Since we have

$$s_{2n} = \left(\frac{1}{1} - \frac{1}{2}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \cdots + \left(\frac{1}{2n-1} - \frac{1}{2n}\right),$$

it is clear that the "even" subsequence (s_{2n}) is increasing. Similarly, the "odd" subsequence (s_{2n+1}) is decreasing since

$$s_{2n+1} = \frac{1}{1} - \left(\frac{1}{2} - \frac{1}{3}\right) - \left(\frac{1}{4} - \frac{1}{5}\right) - \cdots - \left(\frac{1}{2n} - \frac{1}{2n+1}\right).$$

Since $0 < s_{2n} < s_{2n} + 1/(2n+1) = s_{2n+1} \le 1$, both of these subsequences are bounded below by 0 and above by 1. Therefore they are both convergent and to the same value. Thus the sequence (s_n) of partial sums converges, proving that the alternating harmonic series (7) converges. (It is far from obvious that the limit of this series is equal to $\ln 2$.) □

Comparison Tests

Our first test shows that if the terms of a nonnegative series are dominated by the corresponding terms of a *convergent series,* then the first series is convergent.

3.7.7 Comparison Test *Let $X := (x_n)$ and $Y := (y_n)$ be real sequences and suppose that for some $K \in \mathbb{N}$ we have*

(8) $$0 \le x_n \le y_n \qquad \text{for} \quad n \ge K.$$

(a) *Then the convergence of $\sum y_n$ implies the convergence of $\sum x_n$.*

(b) *The divergence of $\sum x_n$ imples the divergence of $\sum y_n$.*

Proof. (a) Suppose that $\sum y_n$ converges and, given $\varepsilon > 0$, let $M(\varepsilon) \in \mathbb{N}$ be such that if $m > n \ge M(\varepsilon)$, then

$$y_{n+1} + \cdots + y_m < \varepsilon.$$

If $m > \sup\{K, M(\varepsilon)\}$, then it follows that

$$0 \le x_{n+1} + \cdots + x_m \le y_{n+1} + \cdots + y_m < \varepsilon,$$

from which the convergence of $\sum x_n$ follows.

(b) This statement is the contrapositive of (a). Q.E.D.

Since it is sometimes difficult to establish the inequalities (8), the next result is frequently very useful.

3.7.8 Limit Comparison Test *Suppose that $X := (x_n)$ and $Y := (y_n)$ are strictly positive sequences and suppose that the following limit exists in $\mathbb{R}$:*

(9) $$r := \lim \left(\frac{x_n}{y_n}\right).$$

(a) *If $r \neq 0$ then $\sum x_n$ is convergent if and only if $\sum y_n$ is convergent.*

(b) *If $r = 0$ and if $\sum y_n$ is convergent, then $\sum x_n$ is convergent.*

Proof. (a) It follows from (9) and Exercise 3.1.17 that there exists $K \in \mathbb{N}$ such that $\frac{1}{2}r \le x_n/y_n \le 2r$ for $n \ge K$, whence

$$\left(\tfrac{1}{2}r\right) y_n \le x_n \le (2r) y_n \qquad \text{for} \quad n \ge K.$$

If we apply the Comparison Test 3.7.7 twice, we obtain the assertion in (a).

(b) If $r = 0$, then there exists $K \in \mathbb{N}$ such that

$$0 < x_n \le y_n \qquad \text{for} \quad n \ge K,$$

so that Theorem 3.7.7(a) applies. Q.E.D.

Remark The Comparison Tests 3.7.7 and 3.7.8 depend on having a stock of series that one knows to be convergent (or divergent). The reader will find that the p-series is often useful for this purpose.

3.7.9 Examples (a) The series $\sum_{n=1}^{\infty} \frac{1}{n^2+n}$ converges.

It is clear that the inequality

$$0 < \frac{1}{n^2+n} < \frac{1}{n^2} \qquad \text{for} \quad n \in \mathbb{N}$$

is valid. Since the series $\sum 1/n^2$ is convergent (by Example 3.7.6(c)), we can apply the Comparison Test 3.7.7 to obtain the convergence of the given series.

(b) The series $\sum_{n=1}^{\infty} \frac{1}{n^2-n+1}$ is convergent.

If the inequality

$$\frac{1}{n^2-n+1} \leq \frac{1}{n^2} \tag{10}$$

were true, we could argue as in (a). However, (10) is *false* for all $n \in \mathbb{N}$. The reader can probably show that the inequality

$$0 < \frac{1}{n^2-n+1} \leq \frac{2}{n^2}$$

is valid for all $n \in \mathbb{N}$, and this inequality will work just as well. However, it might take some experimentation to think of such an inequality and then establish it.

Instead, if we take $x_n := 1/(n^2-n+1)$ and $y_n := 1/n^2$, then we have

$$\frac{x_n}{y_n} = \frac{n^2}{n^2-n+1} = \frac{1}{1-(1/n)+(1/n^2)} \to 1.$$

Therefore, the convergence of the given series follows from the Limit Comparison Test 3.7.8(a).

(c) The series $\sum_{n=1}^{\infty} \frac{1}{\sqrt{n+1}}$ is divergent.

This series closely resembles the series $\sum 1/\sqrt{n}$ which is a p-series with $p = \frac{1}{2}$; by Example 3.7.6(e), it is divergent. If we let $x_n := 1/\sqrt{n+1}$ and $y_n := 1/\sqrt{n}$, then we have

$$\frac{x_n}{y_n} = \frac{\sqrt{n}}{\sqrt{n+1}} = \frac{1}{\sqrt{1+1/n}} \to 1.$$

Therefore the Limit Comparison Test 3.7.8(a) applies.

(d) The series $\sum_{n=1}^{\infty} \frac{1}{n!}$ is convergent.

It would be possible to establish this convergence by showing (by Induction) that $n^2 < n!$ for $n \geq 4$, whence it follows that

$$0 < \frac{1}{n!} < \frac{1}{n^2} \qquad \text{for} \quad n \geq 4.$$

Alternatively, if we let $x := 1/n!$ and $y_n := 1/n^2$, then (when $n \geq 4$) we have

$$0 \leq \frac{x_n}{y_n} = \frac{n^2}{n!} = \frac{n}{1 \cdot 2 \cdots (n-1)} < \frac{1}{n-2} \to 0.$$

Therefore the Limit Comparison Test 3.7.8(b) applies. (Note that this test was a bit troublesome to apply since we do not presently know the convergence of any series for which the limit of x_n/y_n is really easy to determine.) □

Exercises for Section 3.7

1. Let $\sum a_n$ be a given series and let $\sum b_n$ be the series in which the terms are the same and in the same order as in $\sum a_n$ except that the terms for which $a_n = 0$ have been omitted. Show that $\sum a_n$ converges to A if and only if $\sum b_n$ converges to A.

2. Show that the convergence of a series is not affected by changing a *finite* number of its terms. (Of course, the value of the sum may be changed.)

3. By using partial fractions, show that

   (a) $\displaystyle\sum_{n=0}^{\infty} \frac{1}{(n+1)(n+2)} = 1$, (b) $\displaystyle\sum_{n=0}^{\infty} \frac{1}{(\alpha+n)(\alpha+n+1)} = \frac{1}{\alpha} > 0$, if $\alpha > 0$.

   (c) $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n(n+1)(n+2)} = \frac{1}{4}$.

4. If $\sum x_n$ and $\sum y_n$ are convergent, show that $\sum(x_n + y_n)$ is convergent.

5. Can you give an example of a convergent series $\sum x_n$ and a divergent series $\sum y_n$ such that $\sum(x_n + y_n)$ is convergent? Explain.

6. (a) Show that the series $\displaystyle\sum_{n=1}^{\infty} \cos n$ is divergent.

   (b) Show that the series $\displaystyle\sum_{n=1}^{\infty} (\cos n)/n^2$ is convergent.

7. Use an argument similar to that in Example 3.7.6(f) to show that the series $\displaystyle\sum_{n=1}^{\infty} \frac{(-1)^n}{\sqrt{n}}$ is convergent.

8. If $\sum a_n$ with $a_n > 0$ is convergent, then is $\sum a_n^2$ always convergent? Either prove it or give a counterexample.

9. If $\sum a_n$ with $a_n > 0$ is convergent, then is $\sum \sqrt{a_n}$ always convergent? Either prove it or give a counterexample.

10. If $\sum a_n$ with $a_n > 0$ is convergent, then is $\sum \sqrt{a_n a_{n+1}}$ always convergent? Either prove it or give a counterexample.

11. If $\sum a_n$ with $a_n > 0$ is convergent, and if $b_n := (a_1 + \cdots + a_n)/n$ for $n \in \mathbb{N}$, then show that $\sum b_n$ is always divergent.

12. Let $\displaystyle\sum_{n=1}^{\infty} a(n)$ be such that $(a(n))$ is a decreasing sequence of strictly positive numbers. If $s(n)$ denotes the nth partial sum, show (by grouping the terms in $s(2^n)$ in two different ways) that

    $$\tfrac{1}{2}\left(a(1) + 2a(2) + \cdots + 2^n a(2^n)\right) \le s(2^n) \le \left(a(1) + 2a(2) + \cdots + 2^{n-1}a(2^{n-1})\right) + a(2^n).$$

    Use these inequalities to show that $\displaystyle\sum_{n=1}^{\infty} a(n)$ converges if and only if $\displaystyle\sum_{n=1}^{\infty} 2^n a(2^n)$ converges. This result is often called the **Cauchy Condensation Test**; it is very powerful.

13. Use the Cauchy Condensation Test to discuss the p-series $\displaystyle\sum_{n=1}^{\infty} (1/n^p)$ for $p > 0$.

14. Use the Cauchy Condensation Test to establish the divergence of the series:

    (a) $\displaystyle\sum \frac{1}{n \ln n}$, (b) $\displaystyle\sum \frac{1}{n(\ln n)(\ln \ln n)}$,

    (c) $\displaystyle\sum \frac{1}{n(\ln n)(\ln \ln n)(\ln \ln \ln n)}$.

15. Show that if $c > 1$, then the following series are convergent:

    (a) $\displaystyle\sum \frac{1}{n(\ln n)^c}$, (b) $\displaystyle\sum \frac{1}{n(\ln n)(\ln \ln n)^c}$.

CHAPTER 4

LIMITS

"Mathematical analysis" is generally understood to refer to that area of mathematics in which systematic use is made of various limiting concepts. In the preceding chapter we studied one of these basic limiting concepts: the limit of a sequence of real numbers. In this chapter we will encounter the notion of the limit of a function.

The rudimentary notion of a limiting process emerged in the 1680s as Isaac Newton (1642–1727) and Gottfried Leibniz (1646–1716) struggled with the creation of the Calculus. Though each person's work was initially unknown to the other and their creative insights were quite different, both realized the need to formulate a notion of function and the idea of quantities being "close to" one another. Newton used the word "fluent" to denote a relationship between variables, and in his major work *Principia* in 1687 he discussed limits "to which they approach nearer than by any given difference, but never go beyond, nor in effect attain to, till the quantities are diminished *in infinitum*". Leibniz introduced the term "function" to indicate a quantity that depended on a variable, and he invented "infinitesimally small" numbers as a way of handling the concept of a limit. The term "function" soon became standard terminology, and Leibniz also introduced the term "calculus" for this new method of calculation.

In 1748, Leonhard Euler (1707–1783) published his two-volume treatise *Introductio in Analysin Infinitorum,* in which he discussed power series, the exponential and logarithmic functions, the trigonometric functions, and many related topics. This was followed by *Institutiones Calculi Differentialis* in 1755 and the three-volume *Institutiones Calculi Integralis* in 1768–70. These works remained the standard textbooks on calculus for many years. But the concept of limit was very intuitive and its looseness led to a number of problems. Verbal descriptions of the limit concept were proposed by other mathematicians of the era, but none was adequate to provide the basis for rigorous proofs.

In 1821, Augustin-Louis Cauchy (1789–1857) published his lectures on analysis in his *Cours d'Analyse,* which set the standard for mathematical exposition for many years. He was concerned with rigor and in many ways raised the level of precision in mathematical discourse. He formulated definitions and presented arguments with greater care than his predecessors, but the concept of limit still remained elusive. In an early chapter he gave the following definition:

> If the successive values attributed to the same variable approach indefinitely a fixed value, such that they finally differ from it by as little as one wishes, this latter is called the limit of all the others.

The final steps in formulating a precise definition of limit were taken by Karl Weierstrass (1815–1897). He insisted on precise language and rigorous proofs, and his definition of limit is the one we use today.

Gottfried Leibniz

Gottfried Wilhelm Leibniz (1646–1716) was born in Leipzig, Germany. He was six years old when his father, a professor of philosophy, died and left his son the key to his library and a life of books and learning. Leibniz entered the University of Leipzig at age 15, graduated at age 17, and received a Doctor of Law degree from the University of Altdorf four years later. He wrote on legal matters, but was more interested in philosophy. He also developed original theories about language and the nature of the universe. In 1672, he went to Paris as a diplomat for four years. While there he began to study mathematics with the Dutch mathematician Christiaan Huygens. His travels to London to visit the Royal Academy further stimulated his interest in mathematics. His background in philosophy led him to very original, though not always rigorous, results.

Unaware of Newtons's unpublished work, Leibniz published papers in the 1680s that presented a method of finding areas that is known today as the Fundamental Theorem of Calculus. He coined the term "calculus" and invented the dy/dx and elongated S notations that are used today. Unfortunately, some followers of Newton accused Leibniz of plagiarism, resulting in a dispute that lasted until Leibniz's death. Their approaches to calculus were quite different and it is now evident that their discoveries were made independently. Leibniz is now renowned for his work in philosophy, but his mathematical fame rests on his creation of the calculus.

Section 4.1 Limits of Functions

In this section we will introduce the important notion of the limit of a function. The intuitive idea of the function f having a limit L at the point c is that the values $f(x)$ are close to L when x is close to (but different from) c. But it is necessary to have a technical way of working with the idea of "close to" and this is accomplished in the ε-δ definition given below.

In order for the idea of the limit of a function f at a point c to be meaningful, it is necessary that f be defined at points near c. It need not be defined at the point c, but it should be defined at enough points close to c to make the study interesting. This is the reason for the following definition.

4.1.1 Definition Let $A \subseteq \mathbb{R}$. A point $c \in \mathbb{R}$ is a **cluster point** of A if for every $\delta > 0$ there exists at least one point $x \in A$, $x \neq c$ such that $|x - c| < \delta$.

This definition is rephrased in the language of neighborhoods as follows: A point c is a cluster point of the set A if every δ-neighborhood $V_\delta(c) = (c - \delta, c + \delta)$ of c contains at least one point of A distinct from c.

Note The point c may or may not be a member of A, but even if it is in A, it is ignored when deciding whether it is a cluster point of A or not, since we explicitly require that there be points in $V_\delta(c) \cap A$ distinct from c in order for c to be a cluster point of A.

For example, if $A := \{1, 2\}$, then the point 1 is not a cluster point of A, since choosing $\delta := \frac{1}{2}$ gives a neighborhood of 1 that contains no points of A distinct from 1. The same is true for the point 2, so we see that A has no cluster points.

4.1.2 Theorem *A number $c \in \mathbb{R}$ is a cluster point of a subset A of $\mathbb{R}$ if and only if there exists a sequence (a_n) in A such that $\lim(a_n) = c$ and $a_n \neq c$ for all $n \in \mathbb{N}$.*

Proof. If c is a cluster point of A, then for any $n \in \mathbb{N}$ the $(1/n)$-neighborhood $V_{1/n}(c)$ contains at least one point a_n in A distinct from c. Then $a_n \in A$, $a_n \neq c$, and $|a_n - c| < 1/n$ implies $\lim(a_n) = c$.

Conversely, if there exists a sequence (a_n) in $A \backslash \{c\}$ with $\lim(a_n) = c$, then for any $\delta > 0$ there exists K such that if $n \geq K$, then $a_n \in V_\delta(c)$. Therefore the δ-neighborhood $V_\delta(c)$ of c contains the points a_n, for $n \geq K$, which belong to A and are distinct from c.

Q.E.D.

The next examples emphasize that a cluster point of a set may or may not belong to the set.

4.1.3 Examples (a) For the open interval $A_1 := (0, 1)$, every point of the closed interval $[0,1]$ is a cluster point of A_1. Note that the points 0,1 are cluster points of A_1, but do not belong to A_1. All the points of A_1 are cluster points of A_1.

(b) A finite set has no cluster points.

(c) The infinite set $\mathbb{N}$ has no cluster points.

(d) The set $A_4 := \{1/n : n \in \mathbb{N}\}$ has only the point 0 as a cluster point. None of the points in A_4 is a cluster point of A_4.

(e) If $I := [0, 1]$, then the set $A_5 := I \cap \mathbb{Q}$ consists of all the rational numbers in I. It follows from the Density Theorem 2.4.8 that every point in I is a cluster point of A_5. □

Having made this brief detour, we now return to the concept of the limit of a function at a cluster point of its domain.

The Definition of the Limit

We now state the precise definition of the limit of a function f at a point c. It is important to note that in this definition, it is immaterial whether f is defined at c or not. In any case, we exclude c from consideration in the determination of the limit.

4.1.4 Definition Let $A \subseteq \mathbb{R}$, and let c be a cluster point of A. For a function $f : A \to \mathbb{R}$, a real number L is said to be a **limit of f at c** if, given any $\varepsilon > 0$ there exists a $\delta > 0$ such that if $x \in A$ and $0 < |x - c| < \delta$, then $|f(x) - L| < \varepsilon$.

Remarks (a) Since the value of δ usually depends on ε, we will sometimes write $\delta(\varepsilon)$ instead of δ to emphasize this dependence.

(b) The inequality $0 < |x - c|$ is equivalent to saying $x \neq c$.

If L is a limit of f at c, then we also say that f **converges to L at c**. We often write

$$L = \lim_{x \to c} f(x) \quad \text{or} \quad L = \lim_{x \to c} f.$$

We also say that "$f(x)$ approaches L as x approaches c". (But it should be noted that the points do not actually move anywhere.) The symbolism

$$f(x) \to L \qquad \text{as} \quad x \to c$$

is also used sometimes to express the fact that f has limit L at c.

If the limit of f at c does not exist, we say that f **diverges** at c.

Our first result is that the value L of the limit is uniquely determined. This uniqueness is not part of the definition of limit, but must be deduced.

4.1.5 Theorem *If $f : A \to \mathbb{R}$ and if c is a cluster point of A, then f can have only one limit at c.*

Proof. Suppose that numbers L and L' satisfy Definition 4.1.4. For any $\varepsilon > 0$, there exists $\delta(\varepsilon/2) > 0$ such that if $x \in A$ and $0 < |x - c| < \delta(\varepsilon/2)$, then $|f(x) - L| < \varepsilon/2$. Also there exists $\delta'(\varepsilon/2)$ such that if $x \in A$ and $0 < |x - c| < \delta'(\varepsilon/2)$, then $|f(x) - L'| < \varepsilon/2$. Now let $\delta := \inf\{\delta(\varepsilon/2), \delta'(\varepsilon/2)\}$. Then if $x \in A$ and $0 < |x - c| < \delta$, the Triangle Inequality implies that

$$|L - L'| \leq |L - f(x)| + |f(x) - L'| < \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, we conclude that $L - L' = 0$, so that $L = L'$. Q.E.D.

The definition of limit can be very nicely described in terms of neighborhoods. (See Figure 4.1.1.) We observe that because

$$V_\delta(c) = (c - \delta, c + \delta) = \{x : |x - c| < \delta\},$$

the inequality $0 < |x - c| < \delta$ is equivalent to saying that $x \neq c$ and x belongs to the δ-neighborhood $V_\delta(c)$ of c. Similarly, the inequality $|f(x) - L| < \varepsilon$ is equivalent to saying that $f(x)$ belongs to the ε-neighborhood $V_\varepsilon(L)$ of L. In this way, we obtain the following result. The reader should write out a detailed argument to establish the theorem.

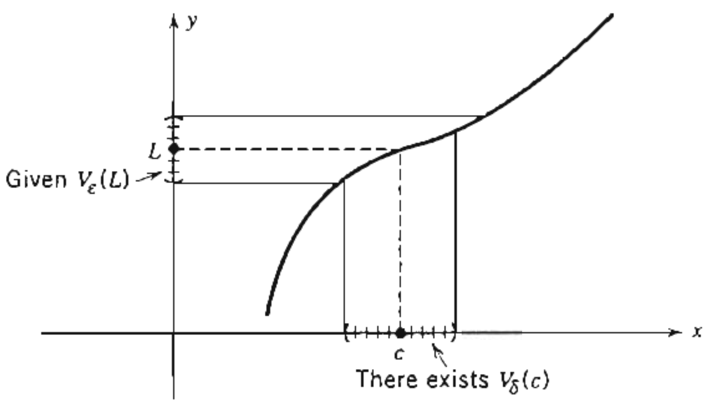


Figure 4.1.1 The limit of f at c is L.

4.1.6 Theorem *Let $f: A \to \mathbb{R}$ and let c be a cluster point of A. Then the following statements are equivalent.*

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

(ii) *Given any ε-neighborhood $V_\varepsilon(L)$ of L, there exists a δ-neighborhood $V_\delta(c)$ of c such that if $x \neq c$ is any point in $V_\delta(c) \cap A$, then $f(x)$ belongs $V_\varepsilon(L)$.*

We now give some examples that illustrate how the definition of limit is applied.

4.1.7 Examples **(a)** $\lim\limits_{x \to c} b = b$.

To be more explicit, let $f(x) := b$ for all $x \in \mathbb{R}$. We want to show that $\lim\limits_{x \to c} f(x) = b$. If $\varepsilon > 0$ is given, we let $\delta := 1$. (In fact, any strictly positive δ will serve the purpose.) Then if $0 < |x - c| < 1$, we have $|f(x) - b| = |b - b| = 0 < \varepsilon$. Since $\varepsilon > 0$ is arbitrary, we conclude from Definition 4.1.4 that $\lim\limits_{x \to c} f(x) = b$.

(b) $\lim_{x \to c} x = c$.

Let $g(x) := x$ for all $x \in \mathbb{R}$. If $\varepsilon > 0$, we choose $\delta(\varepsilon) := \varepsilon$. Then if $0 < |x - c| < \delta(\varepsilon)$, we have $|g(x) - c| = |x - c| < \varepsilon$. Since $\varepsilon > 0$ is arbitrary, we deduce that $\lim_{x \to c} g = c$.

(c) $\lim_{x \to c} x^2 = c^2$.

Let $h(x) := x^2$ for all $x \in \mathbb{R}$. We want to make the difference

$$\left|h(x) - c^2\right| = \left|x^2 - c^2\right|$$

less than a preassigned $\varepsilon > 0$ by taking x sufficiently close to c. To do so, we note that $x^2 - c^2 = (x + c)(x - c)$. Moreover, if $|x - c| < 1$, then

$$|x| \le |c| + 1 \qquad \text{so that} \qquad |x + c| \le |x| + |c| \le 2|c| + 1.$$

Therefore, if $|x - c| < 1$, we have

$$(1) \qquad \left|x^2 - c^2\right| = |x + c|\,|x - c| \le (2|c| + 1)\,|x - c|.$$

Moreover this last term will be less than ε provided we take $|x - c| < \varepsilon/(2|c| + 1)$. Consequently, if we choose

$$\delta(\varepsilon) := \inf\left\{1, \frac{\varepsilon}{2|c| + 1}\right\},$$

then if $0 < |x - c| < \delta(\varepsilon)$, it will follow first that $|x - c| < 1$ so that (1) is valid, and therefore, since $|x - c| < \varepsilon/(2|c| + 1)$ that

$$\left|x^2 - c^2\right| \le (2|c| + 1)\,|x - c| < \varepsilon.$$

Since we have a way of choosing $\delta(\varepsilon) > 0$ for an arbitrary choice of $\varepsilon > 0$, we infer that $\lim_{x \to c} h(x) = \lim_{x \to c} x^2 = c^2$.

(d) $\lim_{x \to c} \frac{1}{x} = \frac{1}{c}$ if $c > 0$.

Let $\varphi(x) := 1/x$ for $x > 0$ and let $c > 0$. To show that $\lim_{x \to c} \varphi = 1/c$ we wish to make the difference

$$\left|\varphi(x) - \frac{1}{c}\right| = \left|\frac{1}{x} - \frac{1}{c}\right|$$

less than a preassigned $\varepsilon > 0$ by taking x sufficiently close to $c > 0$. We first note that

$$\left|\frac{1}{x} - \frac{1}{c}\right| = \left|\frac{1}{cx}(c - x)\right| = \frac{1}{cx}|x - c|$$

for $x > 0$. It is useful to get an upper bound for the term $1/(cx)$ that holds in some neighborhood of c. In particular, if $|x - c| < \frac{1}{2}c$, then $\frac{1}{2}c < x < \frac{3}{2}c$ (why?), so that

$$0 < \frac{1}{cx} < \frac{2}{c^2} \qquad \text{for} \quad |x - c| < \tfrac{1}{2}c.$$

Therefore, for these values of x we have

$$(2) \qquad \left|\varphi(x) - \frac{1}{c}\right| \le \frac{2}{c^2}|x - c| < \varepsilon$$

In order to make this last term less than ε it suffices to take $|x - c| < \frac{1}{2}c^2\varepsilon$. Consequently, if we choose

$$\delta(\varepsilon) := \inf\left\{\tfrac{1}{2}c, \tfrac{1}{2}c^2\varepsilon\right\},$$

then if $0 < |x - c| < \delta(\varepsilon)$, it will follow first that $|x - c| < \frac{1}{2}c$ so that (2) is valid, and therefore, since $|x - c| < \left(\frac{1}{2}c^2\right)\varepsilon$, that

$$\left|\varphi(x) - \frac{1}{c}\right| = \left|\frac{1}{x} - \frac{1}{c}\right| < \varepsilon.$$

Since we have a way of choosing $\delta(\varepsilon) > 0$ for an arbitrary choice of $\varepsilon > 0$, we infer that $\lim\limits_{x \to c} \varphi = 1/c$.

(e) $\lim\limits_{x \to 2} \dfrac{x^3 - 4}{x^2 + 1} = \dfrac{4}{5}$.

Let $\psi(x) := (x^3 - 4)/(x^2 + 1)$ for $x \in \mathbb{R}$. Then a little algebraic manipulation gives us

$$\begin{aligned} \left|\psi(x) - \frac{4}{5}\right| &= \frac{|5x^3 - 4x^2 - 24|}{5(x^2 + 1)} \\ &= \frac{|5x^2 + 6x + 12|}{5(x^2 + 1)} \cdot |x - 2|. \end{aligned}$$

To get a bound on the coefficient of $|x - 2|$, we restrict x by the condition $1 < x < 3$. For x in this interval, we have $5x^2 + 6x + 12 \le 5 \cdot 3^2 + 6 \cdot 3 + 12 = 75$ and $5(x^2 + 1) \ge 5(1 + 1) = 10$, so that

$$\left|\psi(x) - \frac{4}{5}\right| \le \frac{75}{10}|x - 2| = \frac{15}{2}|x - 2|.$$

Now for given $\varepsilon > 0$, we choose

$$\delta(\varepsilon) := \inf\left\{1, \tfrac{2}{15}\varepsilon\right\}.$$

Then if $0 < |x - 2| < \delta(\varepsilon)$, we have $|\psi(x) - (4/5)| \le (15/2)|x - 2| < \varepsilon$. Since $\varepsilon > 0$ is arbitrary, the assertion is proved. ☐

Sequential Criterion for Limits

The following important formulation of limit of a function is in terms of limits of sequences. This characterization permits the theory of Chapter 3 to be applied to the study of limits of functions.

4.1.8 Theorem *(Sequential Criterion) Let $f: A \to \mathbb{R}$ and let c be a cluster point of A. Then the following are equivalent.*

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

(ii) *For every sequence (x_n) in A that converges to c such that $x_n \ne c$ for all $n \in \mathbb{N}$, the sequence $\left(f(x_n)\right)$ converges to L.*

Proof. (i) $\Rightarrow$ (ii). Assume f has limit L at c, and suppose (x_n) is a sequence in A with $\lim(x_n) = c$ and $x_n \ne c$ for all n. We must prove that the sequence $\left(f(x_n)\right)$ converges to L. Let $\varepsilon > 0$ be given. Then by Definition 4.1.4, there exists $\delta > 0$ such that if $x \in A$ satisfies $0 < |x - c| < \delta$, then $f(x)$ satisfies $|f(x) - L| < \varepsilon$. We now apply the definition of convergent sequence for the given δ to obtain a natural number $K(\delta)$ such that if $n > K(\delta)$ then $|x_n - c| < \delta$. But for each such x_n we have $|f(x_n) - L| < \varepsilon$. Thus if $n > K(\delta)$, then $|f(x_n) - L| < \varepsilon$. Therefore, the sequence $\left(f(x_n)\right)$ converges to L.

(ii) $\Rightarrow$ (i). [The proof is a contrapositive argument.] If (i) is not true, then there exists an ε_0-neighborhood $V_{\varepsilon_0}(L)$ such that no matter what δ-neighborhood of c we pick, there will be at least one number x_δ in $A \cap V_\delta(c)$ with $x_\delta \neq c$ such that $f(x_\delta) \notin V_{\varepsilon_0}(L)$. Hence for every $n \in \mathbb{N}$, the $(1/n)$-neighborhood of c contains a number x_n such that

$$0 < |x_n - c| < 1/n \quad \text{and} \quad x_n \in A,$$

but such that

$$|f(x_n) - L| \geq \varepsilon_0 \quad \text{for all} \quad n \in \mathbb{N}.$$

We conclude that the sequence (x_n) in $A \backslash \{c\}$ converges to c, but the sequence $(f(x_n))$ does not converge to L. Therefore we have shown that if (i) is not true, then (ii) is not true. We conclude that (ii) implies (i). Q.E.D.

We shall see in the next section that many of the basic limit properties of functions can be established by using corresponding properties for convergent sequences. For example, we know from our work with sequences that if (x_n) is any sequence that converges to a number c, then (x_n^2) converges to c^2. Therefore, by the sequential criterion, we can conclude that the function $h(x) := x^2$ has limit $\lim\limits_{x \to c} h(x) = c^2$.

Divergence Criteria

It is often important to be able to show (i) that a certain number is *not* the limit of a function at a point, or (ii) that the function *does not have* a limit at a point. The following result is a consequence of (the proof of) Theorem 4.1.8. We leave the details of its proof as an important exercise.

4.1.9 Divergence Criteria *Let $A \subseteq \mathbb{R}$, let $f: A \to \mathbb{R}$ and let $c \in \mathbb{R}$ be a cluster point of A.*

(a) *If $L \in \mathbb{R}$, then f does* **not** *have limit L at c if and only if there exists a sequence (x_n) in A with $x_n \neq c$ for all $n \in \mathbb{N}$ such that the sequence (x_n) converges to c but the sequence $(f(x_n))$ does* **not** *converge to L.*

(b) *The function f does* **not** *have a limit at c if and only if there exists a sequence (x_n) in A with $x_n \neq c$ for all $n \in \mathbb{N}$ such that the sequence (x_n) converges to c but the sequence $(f(x_n))$ does* **not** *converge in $\mathbb{R}$.*

We now give some applications of this result to show how it can be used.

4.1.10 Example **(a)** $\lim\limits_{x \to 0}(1/x)$ does not exist in $\mathbb{R}$.

As in Example 4.1.7(d), let $\varphi(x) := 1/x$ for $x > 0$. However, here we consider $c = 0$. The argument given in Example 4.1.7(d) breaks down if $c = 0$ since we cannot obtain a bound such as that in (2) of that example. Indeed, if we take the sequence (x_n) with $x_n := 1/n$ for $n \in \mathbb{N}$, then $\lim(x_n) = 0$, but $\varphi(x_n) = 1/(1/n) = n$. As we know, the sequence $(\varphi(x_n)) = (n)$ is not convergent in $\mathbb{R}$, since it is not bounded. Hence, by Theorem 4.1.9(b), $\lim\limits_{x \to 0}(1/x)$ does not exist in $\mathbb{R}$.

(b) $\lim\limits_{x \to 0} \operatorname{sgn}(x)$ does not exist.

Let the **signum function** sgn be defined by

$$\operatorname{sgn}(x) := \begin{cases} +1 & \text{for} \quad x > 0, \\ 0 & \text{for} \quad x = 0, \\ -1 & \text{for} \quad x < 0. \end{cases}$$

Note that $\operatorname{sgn}(x) = x/|x|$ for $x \neq 0$. (See Figure 4.1.2.) We shall show that sgn does not have a limit at $x = 0$. We shall do this by showing that there is a sequence (x_n) such that $\lim(x_n) = 0$, but such that $(\operatorname{sgn}(x_n))$ does not converge.

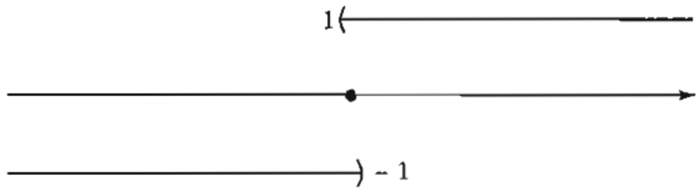


Figure 4.1.2 The signum function.

Indeed, let $x_n := (-1)^n/n$ for $n \in \mathbb{N}$ so that $\lim(x_n) = 0$. However, since

$$\operatorname{sgn}(x_n) = (-1)^n \qquad \text{for} \quad n \in \mathbb{N},$$

it follows from Example 3.4.6(a) that $(\operatorname{sgn}(x_n))$ does not converge. Therefore $\lim\limits_{x\to 0} \operatorname{sgn}(x)$ does not exist.

(c)† $\lim\limits_{x\to 0} \sin(1/x)$ does not exist in $\mathbb{R}$.

Let $g(x) := \sin(1/x)$ for $x \neq 0$. (See Figure 4.1.3.) We shall show that g does not have a limit at $c = 0$, by exhibiting two sequences (x_n) and (y_n) with $x_n \neq 0$ and $y_n \neq 0$ for all $n \in \mathbb{N}$ and such that $\lim(x_n) = 0$ and $\lim(y_n) = 0$, but such that $\lim\big(g(x_n)\big) \neq \lim\big(g(y_n)\big)$. In view of Theorem 4.1.9 this implies that $\lim\limits_{x\to 0} g$ cannot exist. (Explain why.)

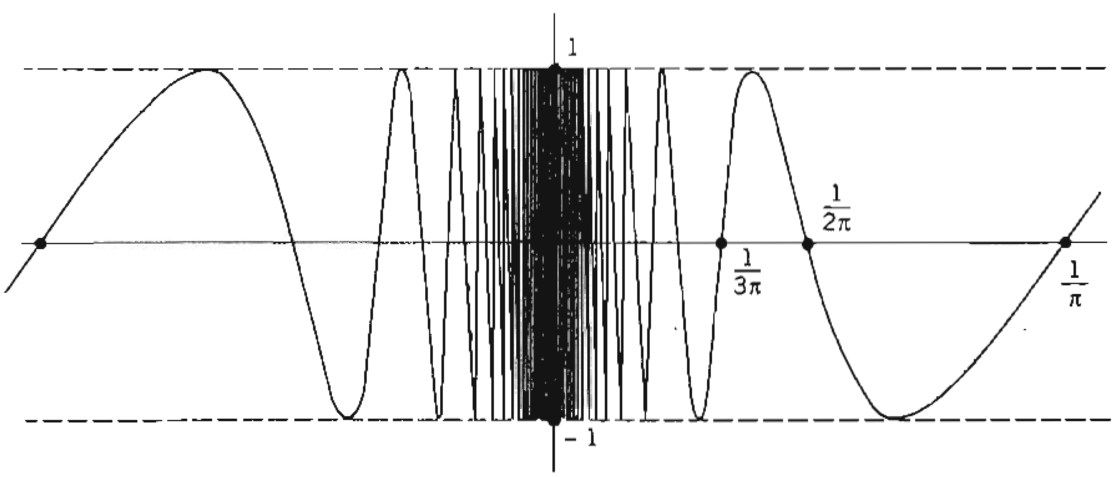


Figure 4.1.3 The function $g(x) = \sin(1/x)$ $(x \neq 0)$.

Indeed, we recall from calculus that $\sin t = 0$ if $t = n\pi$ for $n \in \mathbb{Z}$, and that $\sin t = +1$ if $t = \frac{1}{2}\pi + 2\pi n$ for $n \in \mathbb{Z}$. Now let $x_n := 1/n\pi$ for $n \in \mathbb{N}$; then $\lim(x_n) = 0$ and $g(x_n) = \sin n\pi = 0$ for all $n \in \mathbb{N}$, so that $\lim\big(g(x_n)\big) = 0$. On the other hand, let $y_n := \left(\frac{1}{2}\pi + 2\pi n\right)^{-1}$ for $n \in \mathbb{N}$; then $\lim(y_n) = 0$ and $g(y_n) = \sin\left(\frac{1}{2}\pi + 2\pi n\right) = 1$ for all $n \in \mathbb{N}$, so that $\lim\big(g(y_n)\big) = 1$. We conclude that $\lim\limits_{x\to 0} \sin(1/x)$ does not exist. $\square$

†In order to have some interesting applications in this and later examples, we shall make use of well-known properties of trigonometric and exponential functions that will be established in Chapter 8.

Exercises for Section 4.1

1. Determine a condition on $|x - 1|$ that will assure that:
   (a) $|x^2 - 1| < \frac{1}{2}$, (b) $|x^2 - 1| < 1/10^{-3}$,
   (c) $|x^2 - 1| < 1/n$ for a given $n \in \mathbb{N}$, (d) $|x^3 - 1| < 1/n$ for a given $n \in \mathbb{N}$.

2. Determine a condition on $|x - 4|$ that will assure that:
   (a) $|\sqrt{x} - 2| < \frac{1}{2}$, (b) $|\sqrt{x} - 2| < 10^{-2}$.

3. Let c be a cluster point of $A \subseteq \mathbb{R}$ and let $f\colon A \to \mathbb{R}$. Prove that $\lim_{x\to c} f(x) = L$ if and only if $\lim_{x\to c} |f(x) - L| = 0$.

4. Let $f : \mathbb{R} \to \mathbb{R}$ and let $c \in \mathbb{R}$. Show that $\lim_{x\to c} f(x) = L$ if and only if $\lim_{x\to 0} f(x + c) = L$.

5. Let $I := (0, a)$ where $a > 0$, and let $g(x) := x^2$ for $x \in I$. For any points $x, c \in I$, show that $|g(x) - c^2| \le 2a|x - c|$. Use this inequality to prove that $\lim_{x\to c} x^2 = c^2$ for any $c \in I$.

6. Let I be an interval in $\mathbb{R}$, let $f : I \to \mathbb{R}$, and let $c \in I$. Suppose there exist constants K and L such that $|f(x) - L| \le K|x - c|$ for $x \in I$. Show that $\lim_{x\to c} f(x) = L$.

7. Show that $\lim_{x\to c} x^3 = c^3$ for any $c \in \mathbb{R}$.

8. Show that $\lim_{x\to c} \sqrt{x} = \sqrt{c}$ for any $c > 0$.

9. Use either the ε-δ definition of limit or the Sequential Criterion for limits, to establish the following limits.
   (a) $\lim_{x\to 2} \dfrac{1}{1 - x} = -1$, (b) $\lim_{x\to 1} \dfrac{x}{1 + x} = \dfrac{1}{2}$,
   (c) $\lim_{x\to 0} \dfrac{x^2}{|x|} = 0$, (d) $\lim_{x\to 1} \dfrac{x^2 - x + 1}{x + 1} = \dfrac{1}{2}$.

10. Use the definition of limit to show that
    (a) $\lim_{x\to 2} (x^2 + 4x) = 12$, (b) $\lim_{x\to -1} \dfrac{x + 5}{2x + 3} = 4$.

11. Show that the following limits do *not* exist.
    (a) $\lim_{x\to 0} \dfrac{1}{x^2}$ $(x > 0)$, (b) $\lim_{x\to 0} \dfrac{1}{\sqrt{x}}$ $(x > 0)$,
    (c) $\lim_{x\to 0} (x + \operatorname{sgn}(x))$, (d) $\lim_{x\to 0} \sin(1/x^2)$.

12. Suppose the function $f : \mathbb{R} \to \mathbb{R}$ has limit L at 0, and let $a > 0$. If $g : \mathbb{R} \to \mathbb{R}$ is defined by $g(x) := f(ax)$ for $x \in \mathbb{R}$, show that $\lim_{x\to 0} g(x) = L$.

13. Let $c \in \mathbb{R}$ and let $f : \mathbb{R} \to \mathbb{R}$ be such that $\lim_{x\to c} \big(f(x)\big)^2 = L$.
    (a) Show that if $L = 0$, then $\lim_{x\to c} f(x) = 0$.
    (b) Show by example that if $L \ne 0$, then f may not have a limit at c.

14. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by setting $f(x) := x$ if x is rational, and $f(x) = 0$ if x is irrational.
    (a) Show that f has a limit at $x = 0$.
    (b) Use a sequential argument to show that if $c \ne 0$, then f does not have a limit at c.

15. Let $f : \mathbb{R} \to \mathbb{R}$, let I be an *open* interval in $\mathbb{R}$, and let $c \in I$. If f_1 is the restriction of f to I, show that f_1 has a limit at c if and only if f has a limit at c, and that the limits are equal.

16. Let $f : \mathbb{R} \to \mathbb{R}$, let J be a *closed* interval in $\mathbb{R}$, and let $c \in J$. If f_2 is the restriction of f to J, show that if f has a limit at c then f_2 has a limit at c. Show by example that it does *not* follow that if f_2 has a limit at c, then f has a limit at c.

Section 4.2 Limit Theorems

We shall now obtain results that are useful in calculating limits of functions. These results are parallel to the limit theorems established in Section 3.2 for sequences. In fact, in most cases these results can be proved by using Theorem 4.1.8 and results from Section 3.2. Alternatively, the results in this section can be proved by using ε-δ arguments that are very similar to the ones employed in Section 3.2.

4.2.1 Definition Let $A \subseteq \mathbb{R}$, let $f: A \to \mathbb{R}$, and let $c \in \mathbb{R}$ be a cluster point of A. We say that f is **bounded on a neighborhood of** c if there exists a δ-neighborhood $V_\delta(c)$ of c and a constant $M > 0$ such that we have $|f(x)| \leq M$ for all $x \in A \cap V_\delta(c)$.

4.2.2 Theorem *If $A \subseteq \mathbb{R}$ and $f: A \to \mathbb{R}$ has a limit at $c \in \mathbb{R}$, then f is bounded on some neighborhood of c.*

Proof. If $L := \lim_{x\to c} f$, then for $\varepsilon = 1$, there exists $\delta > 0$ such that if $0 < |x - c| < \delta$, then $|f(x) - L| < 1$; hence (by Corollary 2.2.4(a)),

$$|f(x)| - |L| \leq |f(x) - L| < 1.$$

Therefore, if $x \in A \cap V_\delta(c)$, $x \neq c$, then $|f(x)| \leq |L| + 1$. If $c \notin A$, we take $M = |L| + 1$, while if $c \in A$ we take $M := \sup\{|f(c)|, |L| + 1\}$. It follows that if $x \in A \cap V_\delta(c)$, then $|f(x)| \leq M$. This shows that f is bounded on the neighborhood $V_\delta(c)$ of c. Q.E.D.

The next definition is similar to the definition for sums, differences, products, and quotients of sequences given in Section 3.2.

4.2.3 Definition Let $A \subseteq \mathbb{R}$ and let f and g be functions defined on A to $\mathbb{R}$. We define the **sum** $f + g$, the **difference** $f - g$, and the **product** fg on A to $\mathbb{R}$ to be the functions given by

$$(f+g)(x) := f(x) + g(x), \qquad (f-g)(x) := f(x) - g(x),$$
$$(fg)(x) := f(x)g(x)$$

for all $x \in A$. Further, if $b \in \mathbb{R}$, we define the **multiple** bf to be the function given by

$$(bf)(x) := bf(x) \qquad \text{for all} \quad x \in A.$$

Finally, if $h(x) \neq 0$ for $x \in A$, we define the **quotient** f/h to be the function given by

$$\left(\frac{f}{h}\right)(x) := \frac{f(x)}{h(x)} \qquad \text{for all} \quad x \in A.$$

4.2.4 Theorem *Let $A \subseteq \mathbb{R}$, let f and g be functions on A to $\mathbb{R}$, and let $c \in \mathbb{R}$ be a cluster point of A. Further, let $b \in \mathbb{R}$.*

(a) *If $\lim_{x\to c} f = L$ and $\lim_{x\to c} g = M$, then:*

$$\lim_{x\to c}(f+g) = L + M, \qquad \lim_{x\to c}(f-g) = L - M,$$
$$\lim_{x\to c}(fg) = LM, \qquad \lim_{x\to c}(bf) = bL.$$

(b) *If* $h: A \to \mathbb{R}$, *if* $h(x) \neq 0$ *for all* $x \in A$, *and if* $\lim_{x \to c} h = H \neq 0$, *then*

$$\lim_{x \to c} \left(\frac{f}{h} \right) = \frac{L}{H}.$$

Proof. One proof of this theorem is exactly similar to that of Theorem 3.2.3. Alternatively, it can be proved by making use of Theorems 3.2.3 and 4.1.8. For example, let (x_n) be any sequence in A such that $x_n \neq c$ for $n \in \mathbb{N}$, and $c = \lim(x_n)$. It follows from Theorem 4.1.8 that

$$\lim \big(f(x_n)\big) = L, \qquad \lim \big(g(x_n)\big) = M.$$

On the other hand, Definition 4.2.3 implies that

$$(fg)(x_n) = f(x_n)g(x_n) \qquad \text{for} \quad n \in \mathbb{N}.$$

Therefore an application of Theorem 3.2.3 yields

$$\begin{aligned} \lim \big((fg)(x_n)\big) &= \lim \big(f(x_n)g(x_n)\big) \\ &= \big[\lim \big(f(x_n)\big)\big] \big[\lim \big(g(x_n)\big)\big] = LM. \end{aligned}$$

Consequently, it follows from Theorem 4.1.8 that

$$\lim_{x \to c} (fg) = \lim \big((fg)(x_n)\big) = LM.$$

The other parts of this theorem are proved in a similar manner. We leave the details to the reader. Q.E.D.

Remarks (1) We note that, in part (b), the additional assumption that $H = \lim_{x \to c} h \neq 0$ is made. If this assumption is not satisfied, then the limit

$$\lim_{x \to c} \frac{f(x)}{h(x)}$$

may or may not exist. But even if this limit does exist, we *cannot* use Theorem 4.2.4(b) to evaluate it.

(2) Let $A \subseteq \mathbb{R}$, and let $f_1, f_2, \cdots, f_n$ be functions on A to $\mathbb{R}$, and let c be a cluster point of A. If

$$L_k := \lim_{x \to c} f_k \qquad \text{for} \quad k = 1, \cdots, n,$$

then it follows from Theorem 4.2.4 by an Induction argument that

$$L_1 + L_2 + \cdots + L_n = \lim_{x \to c} (f_1 + f_2 + \cdots + f_n),$$

and

$$L_1 \cdot L_2 \cdots L_n = \lim (f_1 \cdot f_2 \cdots f_n).$$

In particular, we deduce that if $L = \lim_{x \to c} f$ and $n \in \mathbb{N}$, then

$$L^n = \lim_{x \to c} \big(f(x)\big)^n.$$

4.2.5 Examples (a) Some of the limits that were established in Section 4.1 can be proved by using Theorem 4.2.4. For example, it follows from this result that since $\lim_{x \to c} x = c$, then $\lim_{x \to c} x^2 = c^2$, and that if $c > 0$, then

$$\lim_{x \to c} \frac{1}{x} = \frac{1}{\lim_{x \to c} x} = \frac{1}{c}.$$

(b) $\lim_{x \to 2} (x^2 + 1)(x^3 - 4) = 20.$

It follows from Theorem 4.2.4 that

$$\begin{aligned} \lim_{x \to 2} (x^2 + 1)(x^3 - 4) &= \left(\lim_{x \to 2} (x^2 + 1) \right) \left(\lim_{x \to 2} (x^3 - 4) \right) \\ &= 5 \cdot 4 = 20. \end{aligned}$$

(c) $\lim_{x \to 2} \left(\dfrac{x^3 - 4}{x^2 + 1} \right) = \dfrac{4}{5}.$

If we apply Theorem 4.2.4(b), we have

$$\lim_{x \to 2} \frac{x^3 - 4}{x^2 + 1} = \frac{\lim_{x \to 2} (x^3 - 4)}{\lim_{x \to 2} (x^2 + 1)} = \frac{4}{5}.$$

Note that since the limit in the denominator [i.e., $\lim_{x \to 2} (x^2 + 1) = 5$] is not equal to 0, then Theorem 4.2.4(b) is applicable.

(d) $\lim_{x \to 2} = \dfrac{x^2 - 4}{3x - 6} = \dfrac{4}{3}.$

If we let $f(x) := x^2 - 4$ and $h(x) := 3x - 6$ for $x \in \mathbb{R}$, then we *cannot* use Theorem 4.2.4(b) to evaluate $\lim_{x \to 2} (f(x)/h(x))$ because

$$\begin{aligned} H = \lim_{x \to 2} h(x) &= \lim_{x \to 2} (3x - 6) \\ &= 3 \lim_{x \to 2} x - 6 = 3 \cdot 2 - 6 = 0. \end{aligned}$$

However, if $x \neq 2$, then it follows that

$$\frac{x^2 - 4}{3x - 6} = \frac{(x + 2)(x - 2)}{3(x - 2)} = \frac{1}{3}(x + 2).$$

Therefore we have

$$\lim_{x \to 2} \frac{x^2 - 4}{3x - 6} = \lim_{x \to 2} \frac{1}{3} (x + 2) = \frac{1}{3} \left(\lim_{x \to 2} x + 2 \right) = \frac{4}{3}.$$

Note that the function $g(x) = (x^2 - 4)/(3x - 6)$ has a limit at $x = 2$ *even though it is not defined there*.

(e) $\lim_{x \to 0} \dfrac{1}{x}$ does not exist in $\mathbb{R}$.

Of course $\lim_{x \to 0} 1 = 1$ and $H := \lim_{x \to 0} x = 0$. However, since $H = 0$, we *cannot* use Theorem 4.2.4(b) to evaluate $\lim_{x \to 0} (1/x)$. In fact, as was seen in Example 4.1.10(a), the function $\varphi(x) = 1/x$ does not have a limit at $x = 0$. This conclusion also follows from Theorem 4.2.2 since the function $\varphi(x) = 1/x$ is not bounded on a neighborhood of $x = 0$. (Why?)

(f) If p is a polynomial function, then $\lim_{x\to c} p(x) = p(c)$.

Let p be a polynomial function on $\mathbb{R}$ so that $p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ for all $x \in \mathbb{R}$. It follows from Theorem 4.2.4 and the fact that $\lim_{x\to c} x^k = c^k$, that

$$\begin{aligned}\lim_{x\to c} p(x) &= \lim_{x\to c}\left[a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0\right]\\ &= \lim_{x\to c}\left(a_n x^n\right) + \lim_{x\to c}\left(a_{n-1}x^{n-1}\right) + \cdots + \lim_{x\to c}\left(a_1 x\right) + \lim_{x\to c} a_0\\ &= a_n c^n + a_{n-1}c^{n-1} + \cdots + a_1 c + a_0\\ &= p(c).\end{aligned}$$

Hence $\lim_{x\to c} p(x) = p(c)$ for any polynomial function p.

(g) If p and q are polynomial functions on $\mathbb{R}$ and if $q(c) \neq 0$, then

$$\lim_{x\to c}\frac{p(x)}{q(x)} = \frac{p(c)}{q(c)}.$$

Since $q(x)$ is a polynomial function, it follows from a theorem in algebra that there are at most a finite number of real numbers $\alpha_1, \cdots, \alpha_m$ [the real zeroes of $q(x)$] such that $q(\alpha_j) = 0$ and such that if $x \notin \{\alpha_1, \cdots, \alpha_m\}$, then $q(x) \neq 0$. Hence, if $x \notin \{\alpha_1, \cdots, \alpha_m\}$, we can define

$$r(x) := \frac{p(x)}{q(x)}.$$

If c is not a zero of $q(x)$, then $q(c) \neq 0$, and it follows from part (f) that $\lim_{x\to c} q(x) = q(c) \neq 0$. Therefore we can apply Theorem 4.2.4(b) to conclude that

$$\lim_{x\to c}\frac{p(x)}{q(x)} = \frac{\lim_{x\to c} p(x)}{\lim_{x\to c} q(x)} = \frac{p(c)}{q(c)}. \qquad \square$$

The next result is a direct analogue of Theorem 3.2.6.

4.2.6 Theorem *Let $A \subseteq \mathbb{R}$, let $f: A \to \mathbb{R}$ and let $c \in \mathbb{R}$ be a cluster point of A. If*

$$a \le f(x) \le b \qquad \text{for all} \quad x \in A, x \neq c,$$

and if $\lim_{x\to c} f$ exists, then $a \le \lim_{x\to c} f \le b$.

Proof. Indeed, if $L = \lim_{x\to c} f$, then it follows from Theorem 4.1.8 that if (x_n) is any sequence of real numbers such that $c \neq x_n \in A$ for all $n \in \mathbb{N}$ and if the sequence (x_n) converges to c, then the sequence $\left(f(x_n)\right)$ converges to L. Since $a \le f(x_n) \le b$ for all $n \in \mathbb{N}$, it follows from Theorem 3.2.6 that $a \le L \le b$. Q.E.D.

We now state an analogue of the Squeeze Theorem 3.2.7. We leave its proof to the reader.

4.2.7 Squeeze Theorem *Let $A \subseteq \mathbb{R}$, let $f, g, h: A \to \mathbb{R}$, and let $c \in \mathbb{R}$ be a cluster point of A. If*

$$f(x) \le g(x) \le h(x) \qquad \text{for all} \quad x \in A, x \neq c,$$

and if $\lim_{x\to c} f = L = \lim_{x\to c} h$, then $\lim_{x\to c} g = L$.

4.2.8 Examples (a) $\lim_{x\to 0} x^{3/2} = 0 \ (x > 0)$.

Let $f(x) := x^{3/2}$ for $x > 0$. Since the inequality $x < x^{1/2} \le 1$ holds for $0 < x \le 1$ (why?), it follows that $x^2 \le f(x) = x^{3/2} \le x$ for $0 < x \le 1$. Since

$$\lim_{x\to 0} x^2 = 0 \qquad \text{and} \qquad \lim_{x\to 0} x = 0,$$

it follows from the Squeeze Theorem 4.2.7 that $\lim_{x\to 0} x^{3/2} = 0$.

(b) $\lim_{x\to 0} \sin x = 0$.

It will be proved later (see Theorem 8.4.8), that

$$-x \le \sin x \le x \qquad \text{for all} \quad x \ge 0.$$

Since $\lim_{x\to 0}(\pm x) = 0$, it follows from the Squeeze Theorem that $\lim_{x\to 0} \sin x = 0$.

(c) $\lim_{x\to 0} \cos x = 1$.

It will be proved later (see Theorem 8.4.8) that

$$(1) \qquad 1 - \tfrac{1}{2}x^2 \le \cos x \le 1 \qquad \text{for all} \quad x \in \mathbb{R}.$$

Since $\lim_{x\to 0}\left(1 - \frac{1}{2}x^2\right) = 1$, it follows from the Squeeze Theorem that $\lim_{x\to 0} \cos x = 1$.

(d) $\lim_{x\to 0}\left(\dfrac{\cos x - 1}{x}\right) = 0$.

We cannot use Theorem 4.2.4(b) to evaluate this limit. (Why not?) However, it follows from the inequality (1) in part (c) that

$$-\tfrac{1}{2}x \le (\cos x - 1)/x \le 0 \qquad \text{for} \quad x > 0$$

and that

$$0 \le (\cos x - 1)/x \le -\tfrac{1}{2}x \qquad \text{for} \quad x < 0.$$

Now let $f(x) := -x/2$ for $x \ge 0$ and $f(x) := 0$ for $x < 0$, and let $h(x) := 0$ for $x \ge 0$ and $h(x) := -x/2$ for $x < 0$. Then we have

$$f(x) \le (\cos x - 1)/x \le h(x) \qquad \text{for} \quad x \ne 0.$$

Since it is readily seen that $\lim_{x\to 0} f = 0 = \lim_{x\to 0} h$, it follows from the Squeeze Theorem that $\lim_{x\to 0}(\cos x - 1)/x = 0$.

(e) $\lim_{x\to 0}\left(\dfrac{\sin x}{x}\right) = 1$.

Again we cannot use Theorem 4.2.4(b) to evaluate this limit. However, it will be proved later (see Theorem 8.4.8) that

$$x - \tfrac{1}{6}x^3 \le \sin x \le x \qquad \text{for} \quad x \ge 0$$

and that

$$x \le \sin x \le x - \tfrac{1}{6}x^3 \qquad \text{for} \quad x \le 0.$$

Therefore it follows (why?) that

$$1 - \tfrac{1}{6}x^2 \le (\sin x)/x \le 1 \qquad \text{for all} \quad x \ne 0.$$

But since $\lim_{x\to 0}(1 - \frac{1}{6}x^2) = 1 - \frac{1}{6}\cdot \lim_{x\to 0} x^2 = 1$, we infer from the Squeeze Theorem that $\lim_{x\to 0}(\sin x)/x = 1$.

(f) $\lim_{x\to 0}(x\sin(1/x)) = 0$.

Let $f(x) = x\sin(1/x)$ for $x \neq 0$. Since $-1 \le \sin z \le 1$ for all $z \in \mathbb{R}$, we have the inequality

$$-|x| \le f(x) = x\sin(1/x) \le |x|$$

for all $x \in \mathbb{R}$, $x \neq 0$. Since $\lim_{x\to 0}|x| = 0$, it follows from the Squeeze Theorem that $\lim_{x\to 0} f = 0$. For a graph, see Figure 5.1.3 or the cover of this book. □

There are results that are parallel to Theorems 3.2.9 and 3.2.10; however, we will leave them as exercises. We conclude this section with a result that is, in some sense, a partial converse to Theorem 4.2.6.

4.2.9 Theorem *Let $A \subseteq \mathbb{R}$, let $f: A \to \mathbb{R}$ and let $c \in \mathbb{R}$ be a cluster point of A. If*

$$\lim_{x\to c} f > 0 \qquad \left[\textit{respectively, } \lim_{x\to c} f < 0\right],$$

then there exists a neighborhood $V_\delta(c)$ of c such that $f(x) > 0$ [respectively, $f(x) < 0$] for all $x \in A \cap V_\delta(c)$, $x \neq c$.

Proof. Let $L := \lim_{x\to c} f$ and suppose that $L > 0$. We take $\varepsilon = \frac{1}{2}L > 0$ in Definition 4.1.4, and obtain a number $\delta > 0$ such that if $0 < |x - c| < \delta$ and $x \in A$, then $|f(x) - L| < \frac{1}{2}L$. Therefore (why?) it follows that if $x \in A \cap V_\delta(c)$, $x \neq c$, then $f(x) > \frac{1}{2}L > 0$.

If $L < 0$, a similar argument applies. Q.E.D.

Exercises for Section 4.2

1. Apply Theorem 4.2.4 to determine the following limits:

   (a) $\lim_{x\to 1}(x+1)(2x+3) \quad (x \in \mathbb{R})$, (b) $\lim_{x\to 1}\dfrac{x^2+2}{x^2-2} \quad (x > 0)$,

   (c) $\lim_{x\to 2}\left(\dfrac{1}{x+1} - \dfrac{1}{2x}\right) \quad (x > 0)$, (d) $\lim_{x\to 0}\dfrac{x+1}{x^2+2} \quad (x \in \mathbb{R})$.

2. Determine the following limits and state which theorems are used in each case. (You may wish to use Exercise 14 below.)

   (a) $\lim_{x\to 2}\sqrt{\dfrac{2x+1}{x+3}} \quad (x > 0)$, (b) $\lim_{x\to 2}\dfrac{x^2-4}{x-2} \quad (x > 0)$,

   (c) $\lim_{x\to 0}\dfrac{(x+1)^2-1}{x} \quad (x > 0)$, (d) $\lim_{x\to 1}\dfrac{\sqrt{x}-1}{x-1} \quad (x > 0)$.

3. Find $\lim_{x\to 0}\dfrac{\sqrt{1+2x}-\sqrt{1+3x}}{x+2x^2}$ where $x > 0$.

4. Prove that $\lim_{x\to 0}\cos(1/x)$ does not exist but that $\lim_{x\to 0} x\cos(1/x) = 0$.

5. Let f, g be defined on $A \subseteq \mathbb{R}$ to $\mathbb{R}$, and let c be a cluster point of A. Suppose that f is bounded on a neighborhood of c and that $\lim_{x\to c} g = 0$. Prove that $\lim_{x\to c} fg = 0$.

6. Use the definition of the limit to prove the first assertion in Theorem 4.2.4(a).

7. Use the sequential formulation of the limit to prove Theorem 4.2.4(b).

8. Let $n \in \mathbb{N}$ be such that $n \geq 3$. Derive the inequality $-x^2 \leq x^n \leq x^2$ for $-1 < x < 1$. Then use the fact that $\lim_{x \to 0} x^2 = 0$ to show that $\lim_{x \to 0} x^n = 0$.

9. Let f, g be defined on A to $\mathbb{R}$ and let c be a cluster point of A.
   (a) Show that if both $\lim_{x \to c} f$ and $\lim_{x \to c}(f + g)$ exist, then $\lim_{x \to c} g$ exists.
   (b) If $\lim_{x \to c} f$ and $\lim_{x \to c} fg$ exist, does it follow that $\lim_{x \to c} g$ exists?

10. Give examples of functions f and g such that f and g do not have limits at a point c, but such that both $f + g$ and fg have limits at c.

11. Determine whether the following limits exist in $\mathbb{R}$.
   (a) $\lim_{x \to 0} \sin(1/x^2) \quad (x \neq 0)$,
   (b) $\lim_{x \to 0} x \sin(1/x^2) \quad (x \neq 0)$,
   (c) $\lim_{x \to 0} \operatorname{sgn} \sin(1/x) \quad (x \neq 0)$,
   (d) $\lim_{x \to 0} \sqrt{x} \sin(1/x^2) \quad (x > 0)$.

12. Let $f: \mathbb{R} \to \mathbb{R}$ be such that $f(x + y) = f(x) + f(y)$ for all x, y in $\mathbb{R}$. Assume that $\lim_{x \to 0} f = L$ exists. Prove that $L = 0$, and then prove that f has a limit at every point $c \in \mathbb{R}$. [*Hint*: First note that $f(2x) = f(x) + f(x) = 2f(x)$ for $x \in \mathbb{R}$. Also note that $f(x) = f(x - c) + f(c)$ for x, c in $\mathbb{R}$.]

13. Let $A \subseteq \mathbb{R}$, let $f: A \to \mathbb{R}$ and let $c \in \mathbb{R}$ be a cluster point of A. If $\lim_{x \to c} f$ exists, and if $|f|$ denotes the function defined for $x \in A$ by $|f|(x) := |f(x)|$, prove that $\lim_{x \to c} |f| = |\lim_{x \to c} f|$

14. Let $A \subseteq \mathbb{R}$, let $f: A \to \mathbb{R}$, and let $c \in \mathbb{R}$ be a cluster point of A. In addition, suppose that $f(x) \geq 0$ for all $x \in A$, and let $\sqrt{f}$ be the function defined for $x \in A$ by $\left(\sqrt{f}\right)(x) := \sqrt{f(x)}$. If $\lim_{x \to c} f$ exists, prove that $\lim_{x \to c} \sqrt{f} = \sqrt{\lim_{x \to c} f}$.

Section 4.3 Some Extensions of the Limit Concept†

In this section, we shall present three types of extensions of the notion of a limit of a function that often occur. Since all the ideas here are closely parallel to ones we have already encountered, this section can be read easily.

One-sided Limits

There are times when a function f may not possess a limit at a point c, yet a limit does exist when the function is restricted to an interval on one side of the cluster point c.

For example, the signum function considered in Example 4.1.10(b), and illustrated in Figure 4.1.2, has no limit at $c = 0$. However, if we restrict the signum function to the interval $(0, \infty)$, the resulting function has a limit of 1 at $c = 0$. Similarly, if we restrict the signum function to the interval $(-\infty, 0)$, the resulting function has a limit of -1 at $c = 0$. These are elementary examples of right-hand and left-hand limits at $c = 0$.

4.3.1 Definition Let $A \in \mathbb{R}$ and let $f: A \to \mathbb{R}$.

(i) If $c \in \mathbb{R}$ is a cluster point of the set $A \cap (c, \infty) = \{x \in A: x > c\}$, then we say that $L \in \mathbb{R}$ is a **right-hand limit of f at c** and we write

$$\lim_{x \to c+} f = L \qquad \text{or} \qquad \lim_{x \to c+} f(x) = L$$

†This section can be largely omitted on a first reading of this chapter.

if given any $\varepsilon > 0$ there exists a $\delta = \delta(\varepsilon) > 0$ such that for all $x \in A$ with $0 < x - c < \delta$, then $|f(x) - L| < \varepsilon$.

(ii) If $c \in \mathbb{R}$ is a cluster point of the set $A \cap (-\infty, c) = \{x \in A : x < c\}$, then we say that $L \in \mathbb{R}$ is a **left-hand limit of f at c** and we write

$$\lim_{x \to c-} f = L \qquad \text{or} \qquad \lim_{x \to c-} f(x) = L$$

if given any $\varepsilon > 0$ there exists a $\delta > 0$ such that for all $x \in A$ with $0 < c - x < \delta$, then $|f(x) - L| < \varepsilon$.

Notes (1) The limits $\lim_{x \to c+} f$ and $\lim_{x \to c-} f$ are called **one-sided limits of f at c**. It is possible that neither one-sided limit may exist. Also, one of them may exist without the other existing. Similarly, as is the case for $f(x) := \operatorname{sgn}(x)$ at $c = 0$, they may both exist and be different.

(2) If A is an interval with left endpoint c, then it is readily seen that $f: A \to \mathbb{R}$ has a limit at c if and only if it has a right-hand limit at c. Moreover, in this case the limit $\lim_{x \to c} f$ and the right-hand limit $\lim_{x \to c+} f$ are equal. (A similar situation occurs for the left-hand limit when A is an interval with right endpoint c.)

The reader can show that f can have only one right-hand (respectively, left-hand) limit at a point. There are results analogous to those established in Sections 4.1 and 4.2 for two-sided limits. In particular, the existence of one-sided limits can be reduced to sequential considerations.

4.3.2 Theorem *Let $A \subseteq \mathbb{R}$, let $f : A \to \mathbb{R}$, and let $c \in \mathbb{R}$ be a cluster point of $A \cap (c, \infty)$. Then the following statements are equivalent:*

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

(ii) *For every sequence (x_n) that converges to c such that $x_n \in A$ and $x_n > c$ for all $n \in \mathbb{N}$, the sequence $(f(x_n))$ converges to L.*

We leave the proof of this result (and the formulation and proof of the analogous result for left-hand limits) to the reader. We will not take the space to write out the formulations of the one-sided version of the other results in Sections 4.1 and 4.2.

The following result relates the notion of the limit of a function to one-sided limits. We leave its proof as an exercise.

4.3.3 Theorem *Let $A \subseteq \mathbb{R}$, let $f: A \to \mathbb{R}$, and let $c \in \mathbb{R}$ be a cluster point of both of the sets $A \cap (c, \infty)$ and $A \cap (-\infty, c)$. Then $\lim_{x \to c} f = L$ if and only if $\lim_{x \to c+} f = L = \lim_{x \to c-} f$.*

4.3.4 Examples **(a)** Let $f(x) := \operatorname{sgn}(x)$.

We have seen in Example 4.1.10(b) that sgn does not have a limit at 0. It is clear that $\lim_{x \to 0+} \operatorname{sgn}(x) = +1$ and that $\lim_{x \to 0-} \operatorname{sgn}(x) = -1$. Since these one-sided limits are different, it also follows from Theorem 4.3.3 that $\operatorname{sgn}(x)$ does not have a limit at 0.

(b) Let $g(x) := e^{1/x}$ for $x \neq 0$. (See Figure 4.3.1.)

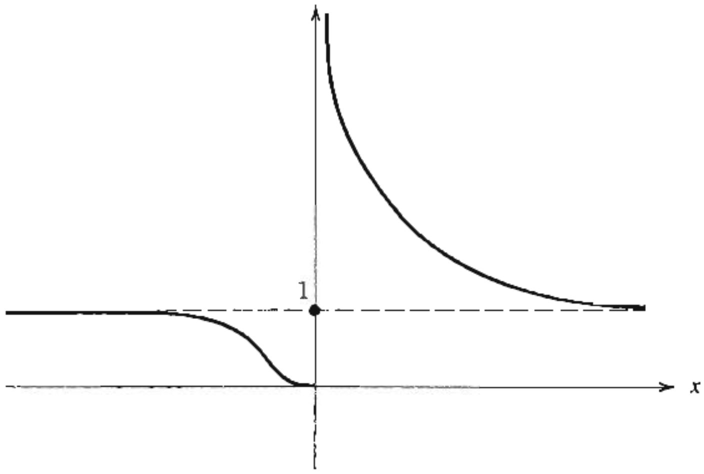


Figure 4.3.1 Graph of $g(x) = e^{1/x}$ $(x \neq 0)$.

We first show that g does not have a finite right-hand limit at $c = 0$ since it is not bounded on any right-hand neighborhood $(0, \delta)$ of 0. We shall make use of the inequality

(1) $$0 < t < e^t \quad \text{for} \quad t > 0,$$

which will be proved later (see Corollary 8.3.3). It follows from (1) that if $x > 0$, then $0 < 1/x < e^{1/x}$. Hence, if we take $x_n = 1/n$, then $g(x_n) > n$ for all $n \in \mathbb{N}$. Therefore $\lim\limits_{x \to 0+} e^{1/x}$ does not exist in $\mathbb{R}$.

However, $\lim\limits_{x \to 0-} e^{1/x} = 0$. Indeed, if $x < 0$ and we take $t = -1/x$ in (1) we obtain $0 < -1/x < e^{-1/x}$. Since $x < 0$, this implies that $0 < e^{1/x} < -x$ for all $x < 0$. It follows from this inequality that $\lim\limits_{x \to 0-} e^{1/x} = 0$.

(c) Let $h(x) := 1/(e^{1/x} + 1)$ for $x \neq 0$. (See Figure 4.3.2.)

We have seen in part (b) that $0 < 1/x < e^{1/x}$ for $x > 0$, whence

$$0 < \frac{1}{e^{1/x} + 1} < \frac{1}{e^{1/x}} < x,$$

which implies that $\lim\limits_{x \to 0+} h = 0$.

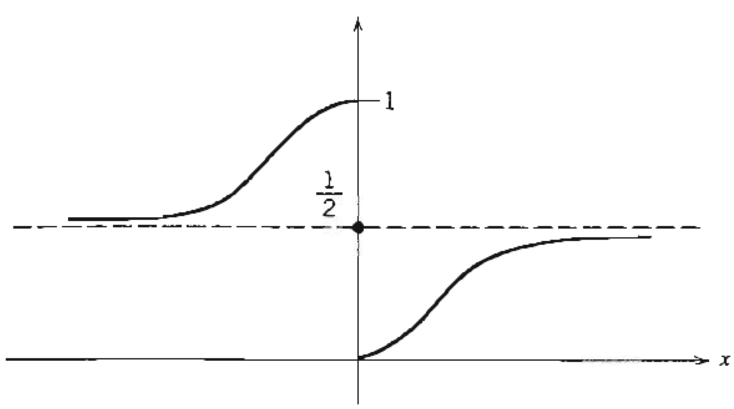


Figure 4.3.2 Graph of $h(x) = 1/(e^{1/x} + 1)$ $(x \neq 0)$.

Since we have seen in part (b) that $\lim_{x\to 0-} e^{1/x} = 0$, it follows from the analogue of Theorem 4.2.4(b) for left-hand limits that

$$\lim_{x\to 0-}\left(\frac{1}{e^{1/x}+1}\right) = \frac{1}{\lim_{x\to 0-} e^{1/x}+1} = \frac{1}{0+1} = 1.$$

Note that for this function, both one-sided limits exist in $\mathbb{R}$, but they are unequal. □

Infinite Limits

The function $f(x) := 1/x^2$ for $x \neq 0$ (see Figure 4.3.3) is not bounded on a neighborhood of 0, so it cannot have a limit in the sense of Definition 4.1.4. While the symbols ∞ $(= +\infty)$ and $-\infty$ do not represent real numbers, it is sometimes useful to be able to say that "$f(x) = 1/x^2$ tends to ∞ as $x \to 0$". This use of $\pm\infty$ will not cause any difficulties, provided we exercise caution and *never* interpret ∞ or $-\infty$ as being real numbers.

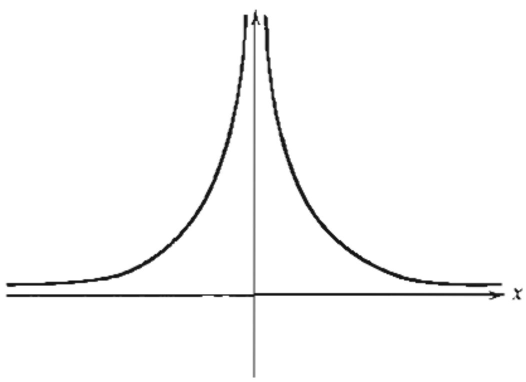


Figure 4.3.3 Graph of $f(x) = 1/x^2$ $(x \neq 0)$

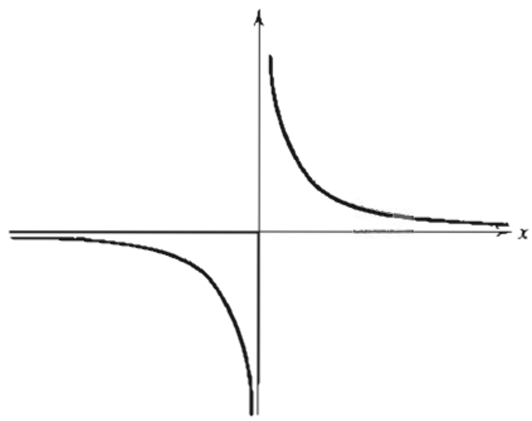


Figure 4.3.4 Graph of $g(x) = 1/x$ $(x \neq 0)$

4.3.5 Definition Let $A \subseteq \mathbb{R}$, let $f\colon A \to \mathbb{R}$, and let $c \in \mathbb{R}$ be a cluster point of A.

(i) We say that f **tends to ∞ as $x \to c$**, and write

$$\lim_{x\to c} f = \infty,$$

if for every $\alpha \in \mathbb{R}$ there exists $\delta = \delta(\alpha) > 0$ such that for all $x \in A$ with $0 < |x - c| < \delta$, then $f(x) > \alpha$.

(ii) We say that f **tends to $-\infty$ as $x \to c$**, and write

$$\lim_{x\to c} f = -\infty,$$

if for every $\beta \in \mathbb{R}$ there exists $\delta = \delta(\beta) > 0$ such that for all $x \in A$ with $0 < |x - c| < \delta$, then $f(x) < \beta$.

4.3.6 Examples **(a)** $\lim_{x\to 0}(1/x^2) = \infty$.

For, if $\alpha > 0$ is given, let $\delta := 1/\sqrt{\alpha}$. It follows that if $0 < |x| < \delta$, then $x^2 < 1/\alpha$ so that $1/x^2 > \alpha$.

(b) Let $g(x) := 1/x$ for $x \neq 0$. (See Figure 4.3.4.)

The function g does *not* tend to either ∞ or $-\infty$ as $x \to 0$. For, if $\alpha > 0$ then $g(x) < \alpha$ for all $x < 0$, so that g does not tend to ∞ as $x \to 0$. Similarly, if $\beta < 0$ then $g(x) > \beta$ for all $x > 0$, so that g does not tend to $-\infty$ as $x \to 0$. □

While many of the results in Sections 4.1 and 4.2 have extensions to this limiting notion, not all of them do since $\pm\infty$ are not real numbers. The following result is an analogue of the Squeeze Theorem 4.2.7. (See also Theorem 3.6.4.)

4.3.7 Theorem *Let $A \subseteq \mathbb{R}$, let $f, g\colon A \to \mathbb{R}$, and let $c \in \mathbb{R}$ be a cluster point of A. Suppose that $f(x) \le g(x)$ for all $x \in A$, $x \ne c$.*

(a) *If $\lim\limits_{x\to c} f = \infty$, then $\lim\limits_{x\to c} g = \infty$.*

(b) *If $\lim\limits_{x\to c} g = -\infty$, then $\lim\limits_{x\to c} f = -\infty$.*

Proof. (a) If $\lim\limits_{x\to c} f = \infty$ and $\alpha \in \mathbb{R}$ is given, then there exists $\delta(\alpha) > 0$ such that if $0 < |x - c| < \delta(\alpha)$ and $x \in A$, then $f(x) > a$. But since $f(x) \le g(x)$ for all $x \in A, x \ne c$, it follows that if $0 < |x - c| < \delta(\alpha)$ and $x \in A$, then $g(x) > \alpha$. Therefore $\lim\limits_{x\to c} g = \infty$.

The proof of (b) is similar. Q.E.D.

The function $g(x) = 1/x$ considered in Example 4.3.6(b) suggests that it might be useful to consider one-sided infinite limits. We will define only right-hand infinite limits.

4.3.8 Definition Let $A \subseteq \mathbb{R}$ and let $f : A \to \mathbb{R}$. If $c \in \mathbb{R}$ is a cluster point of the set $A \cap (c, \infty) = \{x \in A\colon x > c\}$, then we say that f **tends to** ∞ [respectively, $-\infty$] **as** $x \to c+$, and we write

$$\lim_{x\to c+} f = \infty \qquad \left[\text{respectively, } \lim_{x\to c+} f = -\infty\right],$$

if for every $\alpha \in \mathbb{R}$ there is $\delta = \delta(\alpha) > 0$ such that for all $x \in A$ with $0 < x - c < \delta$, then $f(x) > \alpha$ [respectively, $f(x) < \alpha$].

4.3.9 Examples **(a)** Let $g(x) := 1/x$ for $x \ne 0$. We have noted in Example 4.3.6(b) that $\lim\limits_{x\to 0} g$ does not exist. However, it is an easy exercise to show that

$$\lim_{x\to 0+} (1/x) = \infty \qquad \text{and} \qquad \lim_{x\to 0-} (1/x) = -\infty.$$

(b) It was seen in Example 4.3.4(b) that the function $g(x) := e^{1/x}$ for $x \ne 0$ is not bounded on any interval $(0, \delta)$, $\delta > 0$. Hence the right-hand limit of $e^{1/x}$ as $x \to 0+$ does not exist in the sense of Definition 4.3.1(i). However, since

$$1/x < e^{1/x} \qquad \text{for} \quad x > 0,$$

it is readily seen that $\lim\limits_{x\to 0+} e^{1/x} = \infty$ in the sense of Definition 4.3.8. □

Limits at Infinity

It is also desirable to define the notion of the limit of a function as $x \to \infty$. The definition as $x \to -\infty$ is similar.

4.3.10 Definition Let $A \subseteq \mathbb{R}$ and let $f\colon A \to \mathbb{R}$. Suppose that $(a, \infty) \subseteq A$ for some $a \in \mathbb{R}$. We say that $L \in \mathbb{R}$ is **a limit of** f **as** $x \to \infty$, and write

$$\lim_{x\to\infty} f = L \qquad \text{or} \qquad \lim_{x\to\infty} f(x) = L,$$

if given any $\varepsilon > 0$ there exists $K = K(\varepsilon) > a$ such that for any $x > K$, then $|f(x) - L| < \varepsilon$.

The reader should note the close resemblance between 4.3.10 and the definition of a limit of a sequence.

We leave it to the reader to show that the limits of f as $x \to \pm\infty$ are unique whenever they exist. We also have sequential criteria for these limits; we shall only state the criterion as $x \to \infty$. This uses the notion of the limit of a properly divergent sequence (see Definition 3.6.1).

4.3.11 Theorem *Let $A \subseteq \mathbb{R}$, let $f : A \to \mathbb{R}$, and suppose that $(a, \infty) \subseteq A$ for some $a \in \mathbb{R}$. Then the following statements are equivalent:*

(i) $L = \lim\limits_{x\to\infty} f$.

(ii) *For every sequence (x_n) in $A \cap (a, \infty)$ such that $\lim(x_n) = \infty$, the sequence $\big(f(x_n)\big)$ converges to L.*

We leave it to the reader to prove this theorem and to formulate and prove the companion result concerning the limit as $x \to -\infty$.

4.3.12 Examples (a) Let $g(x) := 1/x$ for $x \neq 0$.

It is an elementary exercise to show that $\lim\limits_{x\to\infty} (1/x) = 0 = \lim\limits_{x\to-\infty} (1/x)$. (See Figure 4.3.4.)

(b) Let $f(x) := 1/x^2$ for $x \neq 0$.

The reader may show that $\lim\limits_{x\to\infty} (1/x^2) = 0 = \lim\limits_{x\to-\infty} (1/x^2)$. (See Figure 4.3.3.) One way to do this is to show that if $x \geq 1$ then $0 \leq 1/x^2 \leq 1/x$. In view of part (a), this implies that $\lim\limits_{x\to\infty} (1/x^2) = 0$. □

Just as it is convenient to be able to say that $f(x) \to \pm\infty$ as $x \to c$ for $c \in \mathbb{R}$, it is convenient to have the corresponding notion as $x \to \pm\infty$. We will treat the case where $x \to \infty$.

4.3.13 Definition Let $A \subseteq \mathbb{R}$ and let $f\colon A \to \mathbb{R}$. Suppose that $(a, \infty) \subseteq A$ for some $a \in A$. We say that f **tends to** ∞ [respectively, $-\infty$] **as** $x \to \infty$, and write

$$\lim_{x\to\infty} f = \infty \qquad \left[\text{respectively, } \lim_{x\to\infty} f = -\infty\right]$$

if given any $\alpha \in \mathbb{R}$ there exists $K = K(\alpha) > a$ such that for any $x > K$, then $f(x) > \alpha$ [respectively, $f(x) < \alpha$].

As before there is a sequential criterion for this limit.

4.3.14 Theorem *Let $A \in \mathbb{R}$, let $f\colon A \to \mathbb{R}$, and suppose that $(a, \infty) \subseteq A$ for some $a \in \mathbb{R}$. Then the following statements are equivalent:*

(i) $\lim_{x\to\infty} f = \infty$ [*respectively,* $\lim_{x\to\infty} f = -\infty$].
(ii) *For every sequence* (x_n) *in* (a, ∞) *such that* $\lim(x_n) = \infty$, *then* $\lim\big(f(x_n)\big) = \infty$ [*respectively,* $\lim(f(x_n)) = -\infty$].

The next result is an analogue of Theorem 3.6.5.

4.3.15 Theorem *Let* $A \subseteq \mathbb{R}$, *let* $f, g\colon A \to \mathbb{R}$, *and suppose that* $(a, \infty) \subseteq A$ *for some* $a \in \mathbb{R}$. *Suppose further that* $g(x) > 0$ *for all* $x > a$ *and that for some* $L \in \mathbb{R}$, $L \neq 0$, *we have*

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = L.$$

(i) *If* $L > 0$, *then* $\lim_{x\to\infty} f = \infty$ *if and only if* $\lim_{x\to\infty} g = \infty$.
(ii) *If* $L < 0$, *then* $\lim_{x\to\infty} f = -\infty$ *if and only if* $\lim_{x\to\infty} g = \infty$.

Proof. (i) Since $L > 0$, the hypothesis implies that there exists $a_1 > a$ such that

$$0 < \tfrac{1}{2}L \le \frac{f(x)}{g(x)} < \tfrac{3}{2}L \qquad \text{for} \quad x > a_1.$$

Therefore we have $\left(\frac{1}{2}L\right) g(x) < f(x) < \left(\frac{3}{2}L\right) g(x)$ for all $x > a_1$, from which the conclusion follows readily.

The proof of (ii) is similar. Q.E.D.

We leave it to the reader to formulate the analogous result as $x \to -\infty$.

4.3.16 Examples (a) $\lim_{x\to\infty} x^n = \infty$ for $n \in \mathbb{N}$.

Let $g(x) := x^n$ for $x \in (0, \infty)$. Given $\alpha \in \mathbb{R}$, let $K := \sup\{1, \alpha\}$. Then for all $x > K$, we have $g(x) = x^n \ge x > \alpha$. Since $\alpha \in \mathbb{R}$ is arbitrary, it follows that $\lim_{x\to\infty} g = \infty$.

(b) $\lim_{x\to-\infty} x^n = \infty$ for $n \in \mathbb{N}$, n even, and $\lim_{x\to-\infty} x^n = -\infty$ for $n \in \mathbb{N}$, n odd.

We will treat the case n odd, say $n = 2k + 1$ with $k = 0, 1, \cdots$. Given $\alpha \in \mathbb{R}$, let $K := \inf\{\alpha, -1\}$. For any $x < K$, then since $(x^2)^k \ge 1$, we have $x^n = (x^2)^k x \le x < \alpha$. Since $\alpha \in \mathbb{R}$ is arbitrary, it follows that $\lim_{x\to-\infty} x^n = -\infty$.

(c) Let $p\colon \mathbb{R} \to \mathbb{R}$ be the polynomial function

$$p(x) := a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0.$$

Then $\lim_{x\to\infty} p = \infty$ if $a_n > 0$, and $\lim_{x\to\infty} p = -\infty$ if $a_n < 0$.

Indeed, let $g(x) := x^n$ and apply Theorem 4.3.15. Since

$$\frac{p(x)}{g(x)} = a_n + a_{n-1}\left(\frac{1}{x}\right) + \cdots + a_1\left(\frac{1}{x^{n-1}}\right) + a_0\left(\frac{1}{x^n}\right),$$

it follows that $\lim_{x\to\infty} (p(x)/g(x)) = a_n$. Since $\lim_{x\to\infty} g = \infty$, the assertion follows from Theorem 4.3.15.

(d) Let p be the polynomial function in part (c). Then $\lim_{x\to-\infty} p = \infty$ [respectively, $-\infty$] if n is even [respectively, odd] and $a_n > 0$.

We leave the details to the reader. □

Exercises for Section 4.3

1. Prove Theorem 4.3.2.
2. Give an example of a function that has a right-hand limit but not a left-hand limit at a point.
3. Let $f(x) := |x|^{-1/2}$ for $x \neq 0$. Show that $\lim_{x\to 0+} f(x) = \lim_{x\to 0-} f(x) = +\infty$.
4. Let $c \in \mathbb{R}$ and let f be defined for $x \in (c, \infty)$ and $f(x) > 0$ for all $x \in (c, \infty)$. Show that $\lim_{x\to c} f = \infty$ if and only if $\lim_{x\to c} 1/f = 0$.
5. Evaluate the following limits, or show that they do not exist.
   (a) $\lim_{x\to 1+} \dfrac{x}{x-1}$ $(x \neq 1)$,
   (b) $\lim_{x\to 1} \dfrac{x}{x-1}$ $(x \neq 1)$,
   (c) $\lim_{x\to 0+} (x+2)/\sqrt{x}$ $(x > 0)$,
   (d) $\lim_{x\to \infty} (x+2)/\sqrt{x}$ $(x > 0)$,
   (e) $\lim_{x\to 0} \left(\sqrt{x+1}\right)/x$ $(x > -1)$,
   (f) $\lim_{x\to \infty} \left(\sqrt{x+1}\right)/x$ $(x > 0)$,
   (g) $\lim_{x\to \infty} \dfrac{\sqrt{x}-5}{\sqrt{x}+3}$ $(x > 0)$,
   (h) $\lim_{x\to \infty} \dfrac{\sqrt{x}-x}{\sqrt{x}+x}$ $(x > 0)$.
6. Prove Theorem 4.3.11.
7. Suppose that f and g have limits in $\mathbb{R}$ as $x \to \infty$ and that $f(x) \leq g(x)$ for all $x \in (a, \infty)$. Prove that $\lim_{x\to\infty} f \leq \lim_{x\to\infty} g$.
8. Let f be defined on $(0, \infty)$ to $\mathbb{R}$. Prove that $\lim_{x\to\infty} f(x) = L$ if and only if $\lim_{x\to 0+} f(1/x) = L$.
9. Show that if $f: (a, \infty) \to \mathbb{R}$ is such that $\lim_{x\to\infty} xf(x) = L$ where $L \in \mathbb{R}$, then $\lim_{x\to\infty} f(x) = 0$.
10. Prove Theorem 4.3.14.
11. Suppose that $\lim_{x\to c} f(x) = L$ where $L > 0$, and that $\lim_{x\to c} g(x) = \infty$. Show that $\lim_{x\to c} f(x)g(x) = \infty$. If $L = 0$, show by example that this conclusion may fail.
12. Find functions f and g defined on $(0, \infty)$ such that $\lim_{x\to\infty} f = \infty$ and $\lim_{x\to\infty} g = \infty$, and $\lim_{x\to\infty} (f - g) = 0$. Can you find such functions, with $g(x) > 0$ for all $x \in (0, \infty)$, such that $\lim_{x\to\infty} f/g = 0$?
13. Let f and g be defined on (a, ∞) and suppose $\lim_{x\to\infty} f = L$ and $\lim_{x\to\infty} g = \infty$. Prove that $\lim_{x\to\infty} f \circ g = L$.

CHAPTER 5

CONTINUOUS FUNCTIONS

We now begin the study of the most important class of functions that arises in real analysis: the class of continuous functions. The term "continuous" has been used since the time of Newton to refer to the motion of bodies or to describe an unbroken curve, but it was not made precise until the nineteenth century. Work of Bernhard Bolzano in 1817 and Augustin-Louis Cauchy in 1821 identified continuity as a very significant property of functions and proposed definitions, but since the concept is tied to that of limit, it was the careful work of Karl Weierstrass in the 1870s that brought proper understanding to the idea of continuity.

We will first define the notions of continuity at a point and continuity on a set, and then show that various combinations of continuous functions give rise to continuous functions. Then in Section 5.3 we establish the fundamental properties that make continuous functions so important. For instance, we will prove that a continuous function on a closed bounded interval must attain a maximum and a minimum value. We also prove that a continuous function must take on every value intermediate to any two values it attains. These properties and others are not possessed by general functions, as various examples illustrate, and thus they distinguish continuous functions as a very special class of functions.

In Section 5.4 we introduce the very important notion of uniform continuity. The distinction between continuity and uniform continuity is somewhat subtle and was not fully appreciated until the work of Weierstrass and the mathematicians of his era, but it proved to

Karl Weierstrass

Karl Weierstrass (=Weierstraß) (1815–1897) was born in Westphalia, Germany. His father, a customs officer in a salt works, insisted that he study law and public finance at the University of Bonn, but he had more interest in drinking and fencing, and left Bonn without receiving a diploma. He then enrolled in the Academy of Münster where he studied mathematics with Christoph Gudermann. From 1841–1854 he taught at various *gymnasia* in Prussia. Despite the fact that he had no contact with the mathematical world during this time, he worked hard on mathematical research and was able to publish a few papers, one of which attracted considerable attention. Indeed, the University of Königsberg gave him an honorary doctoral degree for this work in 1855. The next year, he secured positions at the Industrial Institute of Berlin and the University of Berlin. He remained at Berlin until his death.

A methodical and painstaking scholar, Weierstrass distrusted intuition and worked to put everything on a firm and logical foundation. He did fundamental work on the foundations of arithmetic and analysis, on complex analysis, the calculus of variations, and algebraic geometry. Due to his meticulous preparation, he was an extremely popular lecturer; it was not unusual for him to speak about advanced mathematical topics to audiences of more than 250. Among his auditors are counted Georg Cantor, Sonya Kovalevsky, Gösta Mittag-Leffler, Max Planck, Otto Hölder, David Hilbert, and Oskar Bolza (who had many American doctoral students). Through his writings and his lectures, Weierstrass had a profound influence on contemporary mathematics.

be very significant in applications. We present one application to the idea of approximating continuous functions by more elementary functions (such as polynomials).

The notion of a "gauge" is introduced in Section 5.5 and is used to provide an alternative method of proving the fundamental properties of continuous functions. The main significance of this concept, however, is in the area of integration theory where gauges are essential in defining the generalized Riemann integral. This will be discussed in Chapter 10.

Monotone functions are an important class of functions with strong continuity properties and they are discussed in Section 5.6.

Section 5.1 Continuous Functions

In this section, which is very similar to Section 4.1, we will define what it means to say that a function is continuous at a point, or on a set. This notion of continuity is one of the central concepts of mathematical analysis, and it will be used in almost all of the following material in this book. Consequently, it is essential that the reader master it.

5.1.1 Definition Let $A \subseteq \mathbb{R}$, let $f : A \to \mathbb{R}$, and let $c \in A$. We say that f is **continuous at** c if, given any number $\varepsilon > 0$ there exists $\delta > 0$ such that if x is any point of A satisfying $|x - c| < \delta$, then $|f(x) - f(c)| < \varepsilon$.

If f fails to be continuous at c, then we say that f is **discontinuous at** c.

As with the definition of limit, the definition of continuity at a point can be formulated very nicely in terms of neighborhoods. This is done in the next result. We leave the verification as an important exercise for the reader. See Figure 5.1.1.

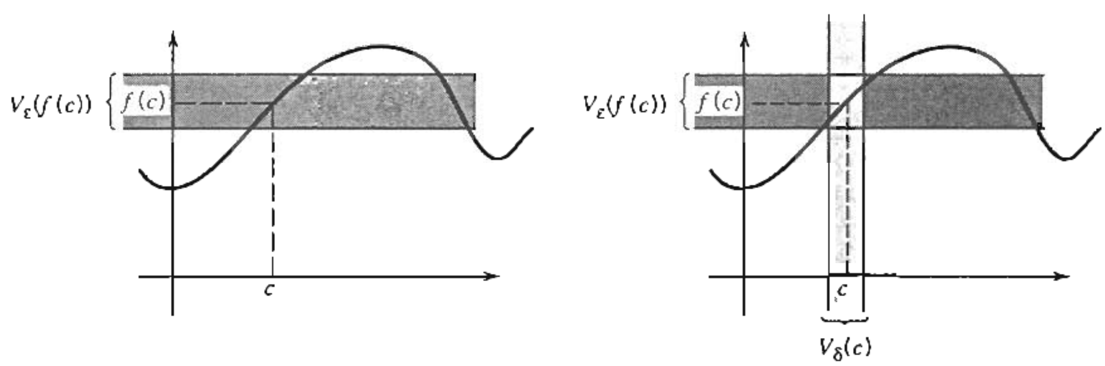


Figure 5.1.1 Given $V_\varepsilon(f(c))$, a neighborhood $V_\delta(c)$ is to be determined.

5.1.2 Theorem *A function $f : A \to \mathbb{R}$ is continuous at a point $c \in A$ if and only if given any ε-neighborhood $V_\varepsilon(f(c))$ of $f(c)$ there exists a δ-neighborhood $V_\delta(c)$ of c such that if x is any point of $A \cap V_\delta(c)$, then $f(x)$ belongs to $V_\varepsilon(f(c))$, that is,*

$$f(A \cap V_\delta(c)) \subseteq V_\varepsilon(f(c)).$$

Remark **(1)** If $c \in A$ is a cluster point of A, then a comparison of Definitions 4.1.4 and 5.1.1 show that f is continuous at c if and only if

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

Thus, if c is a cluster point of A, then three conditions must hold for f to be continuous at c:

(i) f must be defined at c (so that $f(c)$ makes sense),

(ii) the limit of f at c must exist in $\mathbb{R}$ (so that $\lim\limits_{x \to c} f(x)$ makes sense), and

(iii) these two values must be equal.

(2) If $c \in A$ is not a cluster point of A, then there exists a neighborhood $V_\delta(c)$ of c such that $A \cap V_\delta(c) = \{c\}$. Thus we conclude that a function f is automatically continuous at a point $c \in A$ that is not a cluster point of A. Such points are often called "isolated points" of A. They are of little practical interest to us, since they have no relation to a limiting process. Since continuity is automatic for such points, we generally test for continuity only at cluster points. Thus we regard condition (1) as being characteristic for continuity at c.

A slight modification of the proof of Theorem 4.1.8 for limits yields the following sequential version of continuity at a point.

5.1.3 Sequential Criterion for Continuity *A function $f : A \to \mathbb{R}$ is continuous at the point $c \in A$ if and only if for every sequence (x_n) in A that converges to c, the sequence $(f(x_n))$ converges to $f(c)$.*

The following Discontinuity Criterion is a consequence of the last theorem. It should be compared with the Divergence Criterion 4.1.9(a) with $L = f(c)$. Its proof should be written out in detail by the reader.

5.1.4 Discontinuity Criterion *Let $A \subseteq \mathbb{R}$, let $f : A \to \mathbb{R}$, and let $c \in A$. Then f is discontinuous at c if and only if there exists a sequence (x_n) in A such that (x_n) converges to c, but the sequence $(f(x_n))$ does not converge to $f(c)$.*

So far we have discussed continuity at a *point*. To talk about the continuity of a function on a *set*, we will simply require that the function be continuous at each point of the set. We state this formally in the next definition.

5.1.5 Definition Let $A \subseteq \mathbb{R}$ and let $f : A \to \mathbb{R}$. If B is a subset of A, we say that f is **continuous on the set** B if f is continuous at every point of B.

5.1.6 Examples **(a)** The constant function $f(x) := b$ is continuous on $\mathbb{R}$.

It was seen in Example 4.1.7(a) that if $c \in \mathbb{R}$, then $\lim\limits_{x \to c} f(x) = b$. Since $f(c) = b$, we have $\lim\limits_{x \to c} f(x) = f(c)$, and thus f is continuous at every point $c \in \mathbb{R}$. Therefore f is continuous on $\mathbb{R}$.

(b) $g(x) := x$ is continuous on $\mathbb{R}$.

It was seen in Example 4.1.7(b) that if $c \in \mathbb{R}$, then we have $\lim\limits_{x \to c} g = c$. Since $g(c) = c$, then g is continuous at every point $c \in \mathbb{R}$. Thus g is continuous on $\mathbb{R}$.

(c) $h(x) := x^2$ is continuous on $\mathbb{R}$.

It was seen in Example 4.1.7(c) that if $c \in \mathbb{R}$, then we have $\lim\limits_{x \to c} h = c^2$. Since $h(c) = c^2$, then h is continuous at every point $c \in \mathbb{R}$. Thus h is continuous on $\mathbb{R}$.

(d) $\varphi(x) := 1/x$ is continuous on $A := \{x \in \mathbb{R}: x > 0\}$.

It was seen in Example 4.1.7(d) that if $c \in A$, then we have $\lim_{x\to c} \varphi = 1/c$. Since $\varphi(c) = 1/c$, this shows that φ is continuous at every point $c \in A$. Thus φ is continuous on A.

(e) $\varphi(x) := 1/x$ is not continuous at $x = 0$.

Indeed, if $\varphi(x) = 1/x$ for $x > 0$, then φ is not defined for $x = 0$, so it cannot be continuous there. Alternatively, it was seen in Example 4.1.10(a) that $\lim_{x\to 0} \varphi$ does not exist in $\mathbb{R}$, so φ cannot be continuous at $x = 0$.

(f) The signum function sgn is not continuous at 0.

The signum function was defined in Example 4.1.10(b), where it was also shown that $\lim_{x\to 0} \operatorname{sgn}(x)$ does not exist in $\mathbb{R}$. Therefore sgn is not continuous at $x = 0$ (even though sgn 0 is defined).

It is an exercise to show that sgn is continuous at every point $c \neq 0$.

(g) Let $A := \mathbb{R}$ and let f be Dirichlet's "discontinuous function" defined by

$$f(x) := \begin{cases} 1 & \text{if} \quad x \text{ is rational,} \\ 0 & \text{if} \quad x \text{ is irrational.} \end{cases}$$

We claim that f is *not continuous at any point of* $\mathbb{R}$. (This function was introduced in 1829 by P. G. L. Dirichlet.)

Indeed, if c is a rational number, let (x_n) be a sequence of irrational numbers that converges to c. (Corollary 2.4.9 to the Density Theorem 2.4.8 assures us that such a sequence does exist.) Since $f(x_n) = 0$ for all $n \in \mathbb{N}$, we have $\lim\left(f(x_n)\right) = 0$, while $f(c) = 1$. Therefore f is not continuous at the rational number c.

On the other hand, if b is an irrational number, let (y_n) be a sequence of rational numbers that converge to b. (The Density Theorem 2.4.8 assures us that such a sequence does exist.) Since $f(y_n) = 1$ for all $n \in \mathbb{N}$, we have $\lim\left(f(y_n)\right) = 1$, while $f(b) = 0$. Therefore f is not continuous at the irrational number b.

Since every real number is either rational or irrational, we deduce that f is not continuous at any point in $\mathbb{R}$.

(h) Let $A := \{x \in \mathbb{R} : x > 0\}$. For any irrational number $x > 0$ we define $h(x) = 0$. For a rational number in A of the form m/n, with natural numbers m, n having no common factors except 1, we define $h(m/n) := 1/n$. (See Figure 5.1.2.)

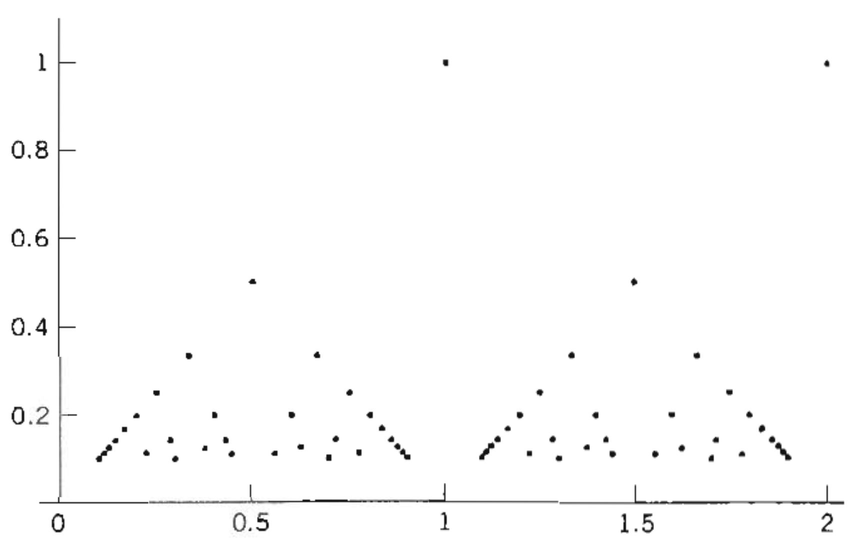


Figure 5.1.2 Thomae's function.

We claim that h *is continuous at every irrational number in* A, *and is discontinuous at every rational number in* A. (This function was introduced in 1875 by K. J. Thomae.)

Indeed, if $a > 0$ is rational, let (x_n) be a sequence of irrational numbers in A that converges to a. Then $\lim(h(x_n)) = 0$, while $h(a) > 0$. Hence h is discontinuous at a.

On the other hand, if b is an irrational number and $\varepsilon > 0$, then (by the Archimedean Property) there is a natural number n_0 such that $1/n_0 < \varepsilon$. There are only a finite number of rationals with denominator less than n_0 in the interval $(b-1, b+1)$. (Why?) Hence $\delta > 0$ can be chosen so small that the neighborhood $(b-\delta, b+\delta)$ contains no rational numbers with denominator less than n_0. It then follows that for $|x-b| < \delta$, $x \in A$, we have $|h(x) - h(b)| = |h(x)| \le 1/n_0 < \varepsilon$. Thus h is continuous at the irrational number b.

Consequently, we deduce that Thomae's function h is continuous precisely at the irrational points in A. □

5.1.7 Remarks (a) Sometimes a function $f\colon A \to \mathbb{R}$ is not continuous at a point c because it is not defined at this point. However, if the function f has a limit L at the point c and if we define F on $A \cup \{c\} \to \mathbb{R}$ by

$$F(x) := \begin{cases} L & \text{for} \quad x = c, \\ f(x) & \text{for} \quad x \in A, \end{cases}$$

then F is continuous at c. To see this, one needs to check that $\lim_{x\to c} F = L$, but this follows (why?), since $\lim_{x\to c} f = L$.

(b) If a function $g : A \to \mathbb{R}$ does not have a limit at c, then there is no way that we can obtain a function $G : A \cup \{c\} \to \mathbb{R}$ that is continuous at c by defining

$$G(x) := \begin{cases} C & \text{for} \quad x = c, \\ g(x) & \text{for} \quad x \in A. \end{cases}$$

To see this, observe that if $\lim_{x\to c} G$ exists and equals C, then $\lim_{x\to c} g$ must also exist and equal C.

5.1.8 Examples (a) The function $g(x) := \sin(1/x)$ for $x \neq 0$ (see Figure 4.1.3) does not have a limit at $x = 0$ (see Example 4.1.10(c)). Thus there is no value that we can assign at $x = 0$ to obtain a continuous extension of g at $x = 0$.

(b) Let $f(x) = x\sin(1/x)$ for $x \neq 0$. (See Figure 5.1.3.) Since f is not defined at $x = 0$, the function f cannot be continuous at this point. However, it was seen in Example 4.2.8(f) that $\lim_{x\to 0}(x\sin(1/x)) = 0$. Therefore it follows from Remark 5.1.7(a) that if we define $F : \mathbb{R} \to \mathbb{R}$ by

$$F(x) := \begin{cases} 0 & \text{for} \quad x = 0, \\ x\sin(1/x) & \text{for} \quad x \neq 0, \end{cases}$$

then F is continuous at $x = 0$. □

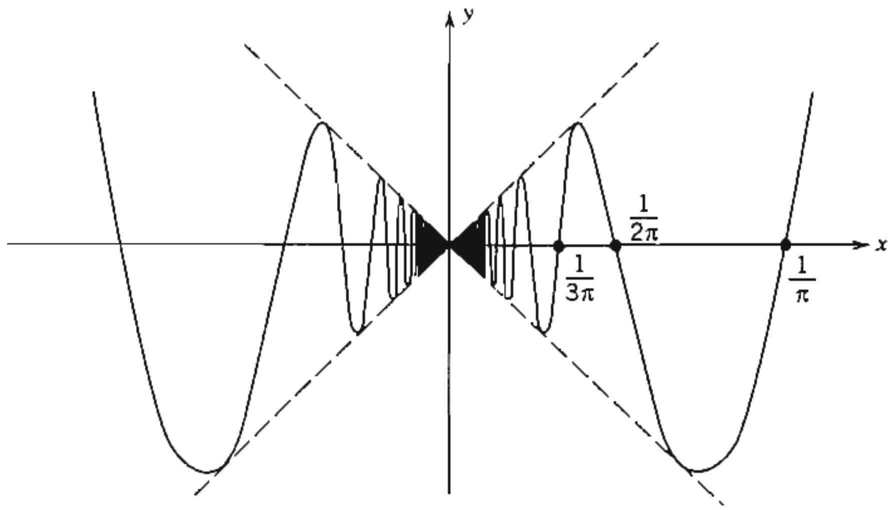


Figure 5.1.3 Graph of $f(x) = x\sin(1/x)$ $(x \neq 0)$.

Exercises for Section 5.1

1. Prove the Sequential Criterion 5.1.3.
2. Establish the Discontinuity Criterion 5.1.4.
3. Let $a < b < c$. Suppose that f is continuous on $[a, b]$, that g is continuous on $[b, c]$, and that $f(b) = g(b)$. Define h on $[a, c]$ by $h(x) := f(x)$ for $x \in [a, b]$ and $h(x) := g(x)$ for $x \in (b, c]$. Prove that h is continuous on $[a, c]$.
4. If $x \in \mathbb{R}$, we define $[\![x]\!]$ to be the greatest integer $n \in \mathbb{Z}$ such that $n \leq x$. (Thus, for example, $[\![8.3]\!] = 8$, $[\![\pi]\!] = 3$, $[\![-\pi]\!] = -4$.) The function $x \mapsto [\![x]\!]$ is called the **greatest integer function**. Determine the points of continuity of the following functions:
   (a) $f(x) := [\![x]\!]$ (b) $g(x) := x[\![x]\!]$,
   (c) $h(x) := [\![\sin x]\!]$, (d) $k(x) := [\![1/x]\!]$ $(x \neq 0)$.
5. Let f be defined for all $x \in \mathbb{R}$, $x \neq 2$, by $f(x) = (x^2 + x - 6)/(x - 2)$. Can f be defined at $x = 2$ in such a way that f is continuous at this point?
6. Let $A \subseteq \mathbb{R}$ and let $f : A \to \mathbb{R}$ be continuous at a point $c \in A$. Show that for any $\varepsilon > 0$, there exists a neighborhood $V_\delta(c)$ of c such that if $x, y \in A \cap V_\delta(c)$, then $|f(x) - f(y)| < \varepsilon$.
7. Let $f : \mathbb{R} \to \mathbb{R}$ be continuous at c and let $f(c) > 0$. Show that there exists a neighborhood $V_\delta(c)$ of c such that if $x \in V_\delta(c)$, then $f(x) > 0$.
8. Let $f : \mathbb{R} \to \mathbb{R}$ be continuous on $\mathbb{R}$ and let $S := \{x \in \mathbb{R} : f(x) = 0\}$ be the "zero set" of f. If (x_n) is in S and $x = \lim(x_n)$, show that $x \in S$.
9. Let $A \subseteq B \subseteq \mathbb{R}$, let $f: B \to \mathbb{R}$ and let g be the restriction of f to A (that is, $g(x) = f(x)$ for $x \in A$).
   (a) If f is continuous at $c \in A$, show that g is continuous at c.
   (b) Show by example that if g is continuous at c, it need not follow that f is continuous at c.
10. Show that the absolute value function $f(x) := |x|$ is continuous at every point $c \in \mathbb{R}$.
11. Let $K > 0$ and let $f : \mathbb{R} \to \mathbb{R}$ satisfy the condition $|f(x) - f(y)| \leq K|x - y|$ for all $x, y \in \mathbb{R}$. Show that f is continuous at every point $c \in \mathbb{R}$.
12. Suppose that $f : \mathbb{R} \to \mathbb{R}$ is continuous on $\mathbb{R}$ and that $f(r) = 0$ for every rational number r. Prove that $f(x) = 0$ for all $x \in \mathbb{R}$.
13. Define $g : \mathbb{R} \to \mathbb{R}$ by $g(x) := 2x$ for x rational, and $g(x) := x + 3$ for x irrational. Find all points at which g is continuous.

14. Let $A := (0, \infty)$ and let $k : A \to \mathbb{R}$ be defined as follows. For $x \in A$, x irrational, we define $k(x) = 0$; for $x \in A$ rational and of the form $x = m/n$ with natural numbers m, n having no common factors except 1, we define $k(x) := n$. Prove that k is unbounded on every open interval in A. Conclude that k is not continuous at any point of A. (See Example 5.1.6(h).)

15. Let $f : (0, 1) \to \mathbb{R}$ be bounded but such that $\lim_{x\to 0} f$ does not exist. Show that there are two sequences (x_n) and (y_n) in $(0, 1)$ with $\lim(x_n) = 0 = \lim(y_n)$, but such that $\lim\big(f(x_n)\big)$ and $\lim\big(f(y_n)\big)$ exist but are not equal.

Section 5.2 Combinations of Continuous Functions

Let $A \subseteq \mathbb{R}$ and let f and g be functions that are defined on A to $\mathbb{R}$ and let $b \in \mathbb{R}$. In Definition 4.2.3 we defined the sum, difference, product, and multiple functions denoted by $f + g$, $f - g$, fg, bf. In addition, if $h : A \to \mathbb{R}$ is such that $h(x) \neq 0$ for all $x \in A$, then we defined the quotient function denoted by f/h.

The next result is similar to Theorem 4.2.4, from which it follows.

5.2.1 Theorem *Let $A \subseteq \mathbb{R}$, let f and g be functions on A to $\mathbb{R}$, and let $b \in \mathbb{R}$. Suppose that $c \in A$ and that f and g are continuous at c.*

(a) *Then $f + g$, $f - g$, fg, and bf are continuous at c.*

(b) *If $h : A \to \mathbb{R}$ is continuous at $c \in A$ and if $h(x) \neq 0$ for all $x \in A$, then the quotient f/h is continuous at c.*

Proof. If $c \in A$ is not a cluster point of A, then the conclusion is automatic. Hence we assume that c is a cluster point of A.

(a) Since f and g are continuous at c, then

$$f(c) = \lim_{x\to c} f \qquad \text{and} \qquad g(c) = \lim_{x\to c} g.$$

Hence it follows from Theorem 4.2.4(a) that

$$(f + g)(c) = f(c) + g(c) = \lim_{x\to c}(f + g).$$

Therefore $f + g$ is continuous at c. The remaining assertions in part (a) are proved in a similar fashion.

(b) Since $c \in A$, then $h(c) \neq 0$. But since $h(c) = \lim_{x\to c} h$, it follows from Theorem 4.2.4(b) that

$$\frac{f}{h}(c) = \frac{f(c)}{h(c)} = \frac{\lim\limits_{x\to c} f}{\lim\limits_{x\to c} h} = \lim_{x\to c}\left(\frac{f}{h}\right).$$

Therefore f/h is continuous at c. Q.E.D.

The next result is an immediate consequence of Theorem 5.2.1, applied to every point of A. However, since it is an extremely important result, we shall state it formally.

5.2.2 Theorem *Let $A \subseteq \mathbb{R}$, let f and g be continuous on A to $\mathbb{R}$, and let $b \in \mathbb{R}$.*

(a) *The functions $f + g$, $f - g$, fg, and bf are continuous on A.*

(b) *If $h : A \to \mathbb{R}$ is continuous on A and $h(x) \neq 0$ for $x \in A$, then the quotient f/h is continuous on A.*

Remark To define quotients, it is sometimes more convenient to proceed as follows. If $\varphi : A \to \mathbb{R}$, let $A_1 := \{x \in A : \varphi(x) \neq 0\}$. We can define the quotient f/φ on the set A_1 by

(1) $$\left(\frac{f}{\varphi}\right)(x) := \frac{f(x)}{\varphi(x)} \quad \text{for} \quad x \in A_1.$$

If φ is continuous at a point $c \in A_1$, it is clear that the restriction φ_1 of φ to A_1 is also continuous at c. Therefore it follows from Theorem 5.2.1(b) applied to φ_1 that f/φ_1 is continuous at $c \in A$. Since $(f/\varphi)(x) = (f/\varphi_1)(x)$ for $x \in A_1$ it follows that f/φ is continuous at $c \in A_1$. Similarly, if f and φ are continuous on A, then the function f/φ, defined on A_1 by (1), is continuous on A_1.

5.2.3 Examples (a) Polynomial functions.

If p is a polynomial function, so that $p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ for all $x \in \mathbb{R}$, then it follows from Example 4.2.5(f) that $p(c) = \lim_{x \to c} p$ for any $c \in \mathbb{R}$. Thus *a polynomial function is continuous on $\mathbb{R}$.*

(b) Rational functions.

If p and q are polynomial functions on $\mathbb{R}$, then there are at most a finite number $\alpha_1, \cdots, \alpha_m$ of real roots of q. If $x \notin \{\alpha_1, \cdots, \alpha_m\}$ then $q(x) \neq 0$ so that we can define the rational function r by

$$r(x) := \frac{p(x)}{q(x)} \qquad \text{for} \quad x \notin \{\alpha_1, \cdots, \alpha_m\}.$$

It was seen in Example 4.2.5(g) that if $q(c) \neq 0$, then

$$r(c) = \frac{p(c)}{q(c)} = \lim_{x \to c} \frac{p(x)}{q(x)} = \lim_{x \to c} r(x).$$

In other words, r is continuous at c. Since c is any real number that is not a root of q, we infer that *a rational function is continuous at every real number for which it is defined.*

(c) We shall show that the sine function sin is continuous on $\mathbb{R}$.

To do so we make use of the following properties of the sine and cosine functions. (See Section 8.4.) For all $x, y, z \in \mathbb{R}$ we have:

$$|\sin z| \leq |z|, \qquad |\cos z| \leq 1,$$
$$\sin x - \sin y = 2 \sin\left[\tfrac{1}{2}(x - y)\right] \cos\left[\tfrac{1}{2}(x + y)\right].$$

Hence if $c \in \mathbb{R}$, then we have

$$|\sin x - \sin c| \leq 2 \cdot \tfrac{1}{2}|x - c| \cdot 1 = |x - c|.$$

Therefore sin is continuous at c. Since $c \in \mathbb{R}$ is arbitrary, it follows that sin is continuous on $\mathbb{R}$.

(d) The cosine function is continuous on $\mathbb{R}$.

We make use of the following properties of the sine and cosine functions. For all $x, y, z \in \mathbb{R}$ we have:

$$|\sin z| \le |z|, \qquad |\sin z| \le 1,$$
$$\cos x - \cos y = -2\sin[\tfrac{1}{2}(x+y)]\sin[\tfrac{1}{2}(x-y)].$$

Hence if $c \in \mathbb{R}$, then we have

$$|\cos x - \cos c| \le 2 \cdot 1 \cdot \tfrac{1}{2}|c - x| = |x - c|.$$

Therefore cos is continuous at c. Since $c \in \mathbb{R}$ is arbitrary, it follows that cos is continuous on $\mathbb{R}$. (Alternatively, we could use the relation $\cos x = \sin(x + \pi/2)$.)

(e) The functions tan, cot, sec, csc are continuous where they are defined.
For example, the cotangent function is defined by

$$\cot x := \frac{\cos x}{\sin x}$$

provided $\sin x \neq 0$ (that is, provided $x \neq n\pi, n \in \mathbb{Z}$). Since sin and cos are continuous on $\mathbb{R}$, it follows (see the Remark before Example 5.2.3) that the function cot is continuous on its domain. The other trigonometric functions are treated similarly. □

5.2.4 Theorem *Let $A \subseteq \mathbb{R}$, let $f : A \to \mathbb{R}$, and let $|f|$ be defined by $|f|(x) := |f(x)|$ for $x \in A$.*

(a) *If f is continuous at a point $c \in A$, then $|f|$ is continuous at c.*
(b) *If f is continuous on A, then $|f|$ is continuous on A.*

Proof. This is an immediate consequence of Exercise 4.2.13. Q.E.D.

5.2.5 Theorem *Let $A \subseteq \mathbb{R}$, let $f : A \to \mathbb{R}$, and let $f(x) \ge 0$ for all $x \in A$. We let $\sqrt{f}$ be defined for $x \in A$ by $\left(\sqrt{f}\right)(x) := \sqrt{f(x)}$.*

(a) *If f is continuous at a point $c \in A$, then $\sqrt{f}$ is continuous at c.*
(b) *If f is continuous on A, then $\sqrt{f}$ is continuous on A.*

Proof. This is an immediate consequence of Exercise 4.2.14. Q.E.D.

Composition of Continuous Functions

We now show that if the function $f : A \to \mathbb{R}$ is continuous at a point c and if $g : B \to \mathbb{R}$ is continuous at $b = f(c)$, then the composition $g \circ f$ is continuous at c. In order to assure that $g \circ f$ is defined on all of A, we also need to assume that $f(A) \subseteq B$.

5.2.6 Theorem *Let $A, B \subseteq \mathbb{R}$ and let $f : A \to \mathbb{R}$ and $g : B \to \mathbb{R}$ be functions such that $f(A) \subseteq B$. If f is continuous at a point $c \in A$ and g is continuous at $b = f(c) \in B$, then the composition $g \circ f : A \to \mathbb{R}$ is continuous at c.*

Proof. Let W be an ε-neighborhood of $g(b)$. Since g is continuous at b, there is a δ-neighborhood V of $b = f(c)$ such that if $y \in B \cap V$ then $g(y) \in W$. Since f is continuous at c, there is a γ-neighborhood U of c such that if $x \in A \cap U$, then $f(x) \in V$. (See Figure 5.2.1.) Since $f(A) \subseteq B$, it follows that if $x \in A \cap U$, then $f(x) \in B \cap V$ so that $g \circ f(x) = g(f(x)) \in W$. But since W is an arbitrary ε-neighborhood of $g(b)$, this implies that $g \circ f$ is continuous at c. Q.E.D.

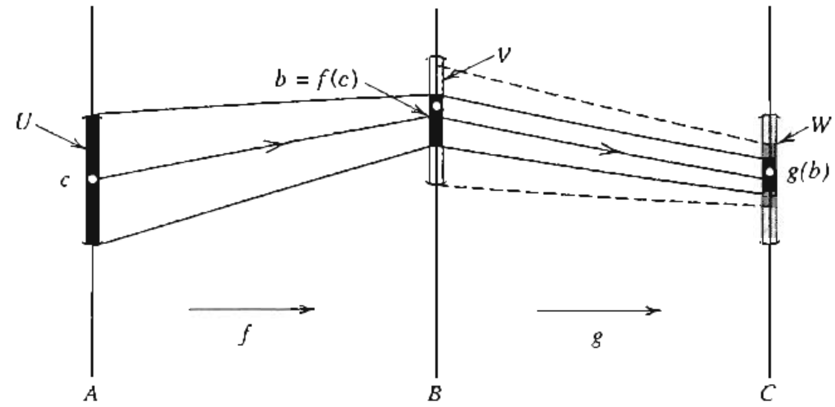


Figure 5.2.1 The composition of f and g.

5.2.7 Theorem *Let $A, B \subseteq \mathbb{R}$, let $f : A \to \mathbb{R}$ be continuous on A, and let $g: B \to \mathbb{R}$ be continuous on B. If $f(A) \subseteq B$, then the composite function $g \circ f : A \to \mathbb{R}$ is continuous on A.*

Proof. The theorem follows immediately from the preceding result, if f and g are continuous at every point of A and B, respectively. Q.E.D

Theorems 5.2.6 and 5.2.7 are very useful in establishing that certain functions are continuous. They can be used in many situations where it would be difficult to apply the definition of continuity directly.

5.2.8 Examples (a) Let $g_1(x) := |x|$ for $x \in \mathbb{R}$. It follows from the Triangle Inequality that

$$\left|g_1(x) - g_1(c)\right| \leq |x - c|$$

for all $x, c \in \mathbb{R}$. Hence g_1 is continuous at $c \in \mathbb{R}$. If $f: A \to \mathbb{R}$ is any function that is continuous on A, then Theorem 5.2.7 implies that $g_1 \circ f = |f|$ is continuous on A. This gives another proof of Theorem 5.2.4.

(b) Let $g_2(x) := \sqrt{x}$ for $x \geq 0$. It follows from Theorems 3.2.10 and 5.1.3 that g_2 is continuous at any number $c \geq 0$. If $f: A \to \mathbb{R}$ is continuous on A and if $f(x) \geq 0$ for all $x \in A$, then it follows from Theorem 5.2.7 that $g_2 \circ f = \sqrt{f}$ is continuous on A. This gives another proof of Theorem 5.2.5.

(c) Let $g_3(x) := \sin x$ for $x \in \mathbb{R}$. We have seen in Example 5.2.3(c) that g_3 is continuous on $\mathbb{R}$. If $f : A \to \mathbb{R}$ is continuous on A, then it follows from Theorem 5.2.7 that $g_3 \circ f$ is continuous on A.

In particular, if $f(x) := 1/x$ for $x \neq 0$, then the function $g(x) := \sin(1/x)$ is continuous at every point $c \neq 0$. [We have seen, in Example 5.1.8(a), that g cannot be defined at 0 in order to become continuous at that point.] □

Exercises for Section 5.2

1. Determine the points of continuity of the following functions and state which theorems are used in each case.

   (a) $f(x) := \dfrac{x^2 + 2x + 1}{x^2 + 1} \quad (x \in \mathbb{R})$, (b) $g(x) := \sqrt{x + \sqrt{x}} \quad (x \geq 0)$,

   (c) $h(x) := \dfrac{\sqrt{1 + |\sin x|}}{x} \quad (x \neq 0)$, (d) $k(x) := \cos\sqrt{1 + x^2} \quad (x \in \mathbb{R})$.

2. Show that if $f : A \to \mathbb{R}$ is continuous on $A \subseteq \mathbb{R}$ and if $n \in \mathbb{N}$, then the function f^n defined by $f^n(x) = (f(x))^n$ for $x \in A$, is continuous on A.
3. Give an example of functions f and g that are both discontinuous at a point c in $\mathbb{R}$ such that (a) the sum $f + g$ is continuous at c, (b) the product fg is continuous at c.
4. Let $x \mapsto [\![x]\!]$ denote the greatest integer function (see Exercise 5.1.4). Determine the points of continuity of the function $f(x) := x - [\![x]\!]$, $x \in \mathbb{R}$.
5. Let g be defined on $\mathbb{R}$ by $g(1) := 0$, and $g(x) := 2$ if $x \neq 1$, and let $f(x) := x + 1$ for all $x \in \mathbb{R}$. Show that $\lim_{x \to 0} g \circ f \neq (g \circ f)(0)$. Why doesn't this contradict Theorem 5.2.6?
6. Let f, g be defined on $\mathbb{R}$ and let $c \in \mathbb{R}$. Suppose that $\lim_{x \to c} f = b$ and that g is continuous at b. Show that $\lim_{x \to c} g \circ f = g(b)$. (Compare this result with Theorem 5.2.7 and the preceding exercise.)
7. Give an example of a function $f : [0, 1] \to \mathbb{R}$ that is discontinuous at every point of $[0, 1]$ but such that $|f|$ is continuous on $[0, 1]$.
8. Let f, g be continuous from $\mathbb{R}$ to $\mathbb{R}$, and suppose that $f(r) = g(r)$ for all rational numbers r. Is it true that $f(x) = g(x)$ for all $x \in \mathbb{R}$?
9. Let $h: \mathbb{R} \to \mathbb{R}$ be continuous on $\mathbb{R}$ satisfying $h(m/2^n) = 0$ for all $m \in \mathbb{Z}$, $n \in \mathbb{N}$. Show that $h(x) = 0$ for all $x \in \mathbb{R}$.
10. Let $f: \mathbb{R} \to \mathbb{R}$ be continuous on $\mathbb{R}$, and let $P := \{x \in \mathbb{R} : f(x) > 0\}$. If $c \in P$, show that there exists a neighborhood $V_\delta(c) \subseteq P$.
11. If f and g are continuous on $\mathbb{R}$, let $S := \{x \in \mathbb{R} : f(x) \geq g(x)\}$. If $(s_n) \subseteq S$ and $\lim(s_n) = s$, show that $s \in S$.
12. A function $f : \mathbb{R} \to \mathbb{R}$ is said to be **additive** if $f(x + y) = f(x) + f(y)$ for all x, y in $\mathbb{R}$. Prove that if f is continuous at some point x_0, then it is continuous at every point of $\mathbb{R}$. (See Exercise 4.2.12.)
13. Suppose that f is a continuous additive function on $\mathbb{R}$. If $c := f(1)$, show that we have $f(x) = cx$ for all $x \in \mathbb{R}$. [*Hint:* First show that if r is a rational number, then $f(r) = cr$.]
14. Let $g : \mathbb{R} \to \mathbb{R}$ satisfy the relation $g(x + y) = g(x)g(y)$ for all x, y in $\mathbb{R}$. Show that if g is continuous at $x = 0$, then g is continuous at every point of $\mathbb{R}$. Also if we have $g(a) = 0$ for some $a \in \mathbb{R}$, then $g(x) = 0$ for all $x \in \mathbb{R}$.
15. Let $f, g : \mathbb{R} \to \mathbb{R}$ be continuous at a point c, and let $h(x) := \sup\{f(x), g(x)\}$ for $x \in \mathbb{R}$. Show that $h(x) = \frac{1}{2}\big(f(x) + g(x)\big) + \frac{1}{2}|f(x) - g(x)|$ for all $x \in \mathbb{R}$. Use this to show that h is continuous at c.

Section 5.3 Continuous Functions on Intervals

Functions that are continuous on intervals have a number of very important properties that are not possessed by general continuous functions. In this section, we will establish some deep results that are of considerable importance and that will be applied later. Alternative proofs of these results will be given in Section 5.5.

5.3.1 Definition A function $f : A \to \mathbb{R}$ is said to be **bounded on** A if there exists a constant $M > 0$ such that $|f(x)| \leq M$ for all $x \in A$.

In other words, a function is bounded on a set if its range is a bounded set in $\mathbb{R}$. To say that a function is *not* bounded on a given set is to say that no particular number can serve as a bound for its range. In exact language, a function f is not bounded on the set A if given any $M > 0$, there exists a point $x_M \in A$ such that $|f(x_M)| > M$. We often say that f is **unbounded on** A in this case.

For example, the function f defined on the interval $A := (0, \infty)$ by $f(x) := 1/x$ is not bounded on A because for any $M > 0$ we can take the point $x_M := 1/(M+1)$ in A to get $f(x_M) = 1/x_M = M + 1 > M$. This example shows that continuous functions need not be bounded. In the next theorem, however, we show that continuous functions on a certain type of interval are necessarily bounded.

5.3.2 Boundedness Theorem† *Let $I := [a, b]$ be a closed bounded interval and let $f: I \to \mathbb{R}$ be continuous on I. Then f is bounded on I.*

Proof. Suppose that f is not bounded on I. Then, for any $n \in \mathbb{N}$ there is a number $x_n \in I$ such that $|f(x_n)| > n$. Since I is bounded, the sequence $X := (x_n)$ is bounded. Therefore, the Bolzano-Weierstrass Theorem 3.4.8 implies that there is a subsequence $X' = (x_{n_r})$ of X that converges to a number x. Since I is closed and the elements of X' belong to I, it follows from Theorem 3.2.6 that $x \in I$. Then f is continuous at x, so that $(f(x_{n_r}))$ converges to $f(x)$. We then conclude from Theorem 3.2.2 that the convergent sequence $(f(x_{n_r}))$ must be bounded. But this is a contradiction since

$$\left|f(x_{n_r})\right| > n_r \geq r \qquad \text{for} \quad r \in \mathbb{N}.$$

Therefore the supposition that the continuous function f is not bounded on the closed bounded interval I leads to a contradiction. Q.E.D.

To show that each hypothesis of the Boundedness Theorem is needed, we can construct examples that show the conclusion fails if any one of the hypotheses is relaxed.

(i) The interval must be bounded. The function $f(x) := x$ for x in the unbounded, closed interval $A := [0, \infty)$ is continuous but not bounded on A.

(ii) The interval must be closed. The function $g(x) := 1/x$ for x in the half-open interval $B := (0, 1]$ is continuous but not bounded on B.

(iii) The function must be continuous. The function h defined on the closed interval $C := [0, 1]$ by $h(x) := 1/x$ for $x \in (0, 1]$ and $h(0) := 1$ is discontinuous and unbounded on C.

The Maximum-Minimum Theorem

5.3.3 Definition Let $A \subseteq \mathbb{R}$ and let $f : A \to \mathbb{R}$. We say that f **has an absolute maximum** on A if there is a point $x^* \in A$ such that

$$f(x^*) \geq f(x) \qquad \text{for all} \quad x \in A.$$

We say that f **has an absolute minimum** on A if there is a point $x_* \in A$ such that

$$f(x_*) \leq f(x) \qquad \text{for all} \quad x \in A.$$

We say that x^* is an **absolute maximum point** for f on A, and that x_* is an **absolute minimum point** for f on A, if they exist.

†This theorem, as well as 5.3.4, is true for an arbitrary closed bounded set. For these developments, see Sections 11.2 and 11.3.

We note that a continuous function on a set A does not necessarily have an absolute maximum or an absolute minimum on the set. For example, $f(x) := 1/x$ has neither an absolute maximum nor an absolute minimum on the set $A := (0, \infty)$. (See Figure 5.3.1). There can be no absolute maximum for f on A since f is not bounded above on A, and there is no point at which f attains the value $0 = \inf\{f(x) : x \in A\}$. The same function has neither an absolute maximum nor an absolute minimum when it is restricted to the set $(0, 1)$, while it has *both* an absolute maximum and an absolute minimum when it is restricted to the set $[1, 2]$. In addition, $f(x) = 1/x$ has an absolute maximum but no absolute minimum when restricted to the set $[1, \infty)$, but no absolute maximum and no absolute minimum when restricted to the set $(1, \infty)$.

It is readily seen that if a function has an absolute maximum point, then this point is not necessarily uniquely determined. For example, the function $g(x) := x^2$ defined for $x \in A := [-1, +1]$ has the two points $x = \pm 1$ giving the absolute maximum on A, and the single point $x = 0$ yielding its absolute minimum on A. (See Figure 5.3.2.) To pick an extreme example, the constant function $h(x) := 1$ for $x \in \mathbb{R}$ is such that *every point* of $\mathbb{R}$ is both an absolute maximum and an absolute minimum point for h.

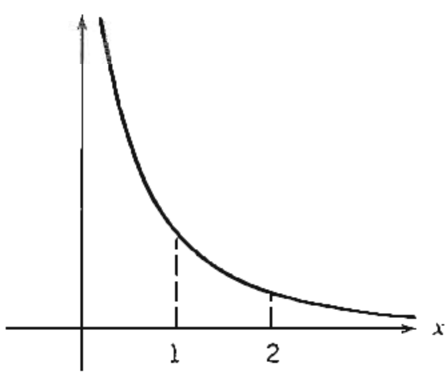


Figure 5.3.1 The function $f(x) = 1/x$ $(x > 0)$.

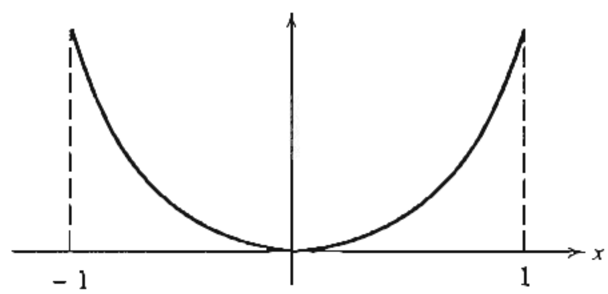


Figure 5.3.2 The function $g(x) = x^2$ $(|x| \le 1)$.

5.3.4 Maximum-Minimum Theorem *Let $I := [a, b]$ be a closed bounded interval and let $f : I \to \mathbb{R}$ be continuous on I. Then f has an absolute maximum and an absolute minimum on I.*

Proof. Consider the nonempty set $f(I) := \{f(x) : x \in I\}$ of values of f on I. In Theorem 5.3.2 it was established that $f(I)$ is a bounded subset of $\mathbb{R}$. Let $s^* := \sup f(I)$ and $s_* := \inf f(I)$. We claim that there exist points x^* and x_* in I such that $s^* = f(x^*)$ and $s_* = f(x_*)$. We will establish the existence of the point x^*, leaving the proof of the existence of x_* to the reader.

Since $s^* = \sup f(I)$, if $n \in \mathbb{N}$, then the number $s^* - 1/n$ is not an upper bound of the set $f(I)$. Consequently there exists a number $x_n \in I$ such that

$$(1) \qquad s^* - \frac{1}{n} < f(x_n) \le s^* \qquad \text{for all} \quad n \in \mathbb{N}.$$

Since I is bounded, the sequence $X := (x_n)$ is bounded. Therefore, by the Bolzano–Weierstrass Theorem 3.4.8, there is a subsequence $X' = (x_{n_r})$ of X that converges to some number x^*. Since the elements of X' belong to $I = [a, b]$, it follows from Theorem 3.2.6

that $x^* \in I$. Therefore f is continuous at x^* so that $\left(\lim f(x_{n_r})\right) = f(x^*)$. Since it follows from (1) that

$$s^* - \frac{1}{n_r} < f(x_{n_r}) \le s^* \qquad \text{for all} \quad r \in \mathbb{N},$$

we conclude from the Squeeze Theorem 3.2.7 that $\lim(f(x_{n_r})) = s^*$. Therefore we have

$$f(x^*) = \lim\left(f(x_{n_r})\right) = s^* = \sup f(I).$$

We conclude that x^* is an absolute maximum point of f on I. Q.E.D.

The next result is the theoretical basis for locating roots of a continuous function by means of sign changes of the function. The proof also provides an algorithm, known as the **Bisection Method,** for the calculation of roots to a specified degree of accuracy and can be readily programmed for a computer. It is a standard tool for finding solutions of equations of the form $f(x) = 0$, where f is a continuous function. An alternative proof of the theorem is indicated in Exercise 11.

5.3.5 Location of Roots Theorem *Let $I = [a, b]$ and let $f : I \to \mathbb{R}$ be continuous on I. If $f(a) < 0 < f(b)$, or if $f(a) > 0 > f(b)$, then there exists a number $c \in (a, b)$ such that $f(c) = 0$.*

Proof. We assume that $f(a) < 0 < f(b)$. We will generate a sequence of intervals by successive bisections. Let $I_1 := [a_1, b_1]$, where $a_1 := a$, $b_1 := b$, and let p_1 be the midpoint $p_1 := \frac{1}{2}(a_1 + b_1)$. If $f(p_1) = 0$, we take $c := p_1$ and we are done. If $f(p_1) \neq 0$, then either $f(p_1) > 0$ or $f(p_1) < 0$. If $f(p_1) > 0$, then we set $a_2 := a_1$, $b_2 := p_1$, while if $f(p_1) < 0$, then we set $a_2 := p_1$, $b_2 := b_1$. In either case, we let $I_2 := [a_2, b_2]$; then we have $I_2 \subset I_1$ and $f(a_2) < 0$, $f(b_2) > 0$.

We continue the bisection process. Suppose that the intervals $I_1, I_2, \cdots, I_k$ have been obtained by successive bisection in the same manner. Then we have $f(a_k) < 0$ and $f(b_k) > 0$, and we set $p_k := \frac{1}{2}(a_k + b_k)$. If $f(p_k) = 0$, we take $c := p_k$ and we are done. If $f(p_k) > 0$, we set $a_{k+1} := a_k$, $b_{k+1} := p_k$, while if $f(p_k) < 0$, we set $a_{k+1} := p_k$, $b_{k+1} := b_k$. In either case, we let $I_{k+1} := [a_{k+1}, b_{k+1}]$; then $I_{k+1} \subset I_k$ and $f(a_{k+1}) < 0$, $f(b_{k+1}) > 0$.

If the process terminates by locating a point p_n such that $f(p_n) = 0$, then we are done. If the process does not terminate, then we obtain a nested sequence of closed bounded intervals $I_n := [a_n, b_n]$ such that for every $n \in \mathbb{N}$ we have

$$f(a_n) < 0 \qquad \text{and} \qquad f(b_n) > 0.$$

Furthermore, since the intervals are obtained by repeated bisection, the length of I_n is equal to $b_n - a_n = (b - a)/2^{n-1}$. It follows from the Nested Intervals Property 2.5.2 that there exists a point c that belongs to I_n for all $n \in \mathbb{N}$. Since $a_n \le c \le b_n$ for all $n \in \mathbb{N}$, we have $0 \le c - a_n \le b_n - a_n = (b - a)/2^{n-1}$, and $0 \le b_n - c \le b_n - a_n = (b - a)/2^{n-1}$. Hence, it follows that $\lim(a_n) = c = \lim(b_n)$. Since f is continuous at c, we have

$$\lim\left(f(a_n)\right) = f(c) = \lim\left(f(b_n)\right).$$

The fact that $f(a_n) < 0$ for all $n \in \mathbb{N}$ implies that $f(c) = \lim\left(f(a_n)\right) \le 0$. Also, the fact that $f(b_n) \ge 0$ for all $n \in \mathbb{N}$ implies that $f(c) = \lim\left(f(b_n)\right) \ge 0$. Thus, we conclude that $f(c) = 0$. Consequently, c is a root of f. Q.E.D.

The following example illustrates how the Bisection Method for finding roots is applied in a systematic fashion.

5.3.6 Example The equation $f(x) = xe^x - 2 = 0$ has a root c in the interval $[0, 1]$, because f is continuous on this interval and $f(0) = -2 < 0$ and $f(1) = e - 2 > 0$. We construct the following table, where the sign of $f(p_n)$ determines the interval at the next step. The far right column is an upper bound on the error when p_n is used to approximate the root c, because we have

$$|p_n - c| \le \tfrac{1}{2}(b_n - a_n) = 1/2^n.$$

We will find an approximation p_n with error less than 10^{-2}.

n	a_n	b_n	p_n	$f(p_n)$	$\frac{1}{2}(b_n - a_n)$
1	0	1	.5	-1.176	.5
2	.5	1	.75	$-.412$	.25
3	.75	1	.875	$+.099$	.125
4	.75	.875	.8125	$-.169$	.0625
5	.8125	.875	.84375	$-.0382$	.03125
6	.84375	.875	.859375	$+.0296$	.015625
7	.84375	.859375	.8515625	—	.0078125

We have stopped at $n = 7$, obtaining $c \approx p_7 = .8515625$ with error less than .0078125. This is the first step in which the error is less than 10^{-2}. The decimal place values of p_7 past the second place cannot be taken seriously, but we can conclude that $.843 < c < .860$. □

Bolzano's Theorem

The next result is a generalization of the Location of Roots Theorem. It assures us that a continuous function on an interval takes on (at least once) any number that lies between two of its values.

5.3.7 Bolzano's Intermediate Value Theorem *Let I be an interval and let $f : I \to \mathbb{R}$ be continuous on I. If $a, b \in I$ and if $k \in \mathbb{R}$ satisfies $f(a) < k < f(b)$, then there exists a point $c \in I$ between a and b such that $f(c) = k$.*

Proof. Suppose that $a < b$ and let $g(x) := f(x) - k$; then $g(a) < 0 < g(b)$. By the Location of Roots Theorem 5.3.5 there exists a point c with $a < c < b$ such that $0 = g(c) = f(c) - k$. Therefore $f(c) = k$.

If $b < a$, let $h(x) := k - f(x)$ so that $h(b) < 0 < h(a)$. Therefore there exists a point c with $b < c < a$ such that $0 = h(c) = k - f(c)$, whence $f(c) = k$. Q.E.D.

5.3.8 Corollary *Let $I = [a, b]$ be a closed, bounded interval and let $f : I \to \mathbb{R}$ be continuous on I. If $k \in \mathbb{R}$ is any number satisfying*

$$\inf f(I) \le k \le \sup f(I),$$

then there exists a number $c \in I$ such that $f(c) = k$.

Proof. It follows from the Maximum-Minimum Theorem 5.3.4 that there are points c_* and c^* in I such that

$$\inf f(I) = f(c_*) \le k \le f(c^*) = \sup f(I).$$

The conclusion now follows from Bolzano's Theorem 5.3.7. Q.E.D.

The next theorem summarizes the main results of this section. It states that the image of a closed bounded interval under a continuous function is also a closed bounded interval. The endpoints of the **image interval** are the absolute minimum and absolute maximum values of the function, and the statement that all values between the absolute minimum and the absolute maximum values belong to the image is a way of describing Bolzano's Intermediate Value Theorem.

5.3.9 Theorem *Let I be a closed bounded interval and let $f : I \to \mathbb{R}$ be continuous on I. Then the set $f(I) := \{f(x) : x \in I\}$ is a closed bounded interval.*

Proof. If we let $m := \inf f(I)$ and $M := \sup f(I)$, then we know from the Maximum-Minimum Theorem 5.3.4 that m and M belong to $f(I)$. Moreover, we have $f(I) \subseteq [m, M]$. If k is any element of $[m, M]$, then it follows from the preceding corollary that there exists a point $c \in I$ such that $k = f(c)$. Hence, $k \in f(I)$ and we conclude that $[m, M] \subseteq f(I)$. Therefore, $f(I)$ is the interval $[m, M]$. Q.E.D.

Warning If $I := [a, b]$ is an interval and $f : I \to \mathbb{R}$ is continuous on I, we have proved that $f(I)$ is the interval $[m, M]$. We have *not* proved (and it is not always true) that $f(I)$ is the interval $[f(a), f(b)]$. (See Figure 5.3.3.)

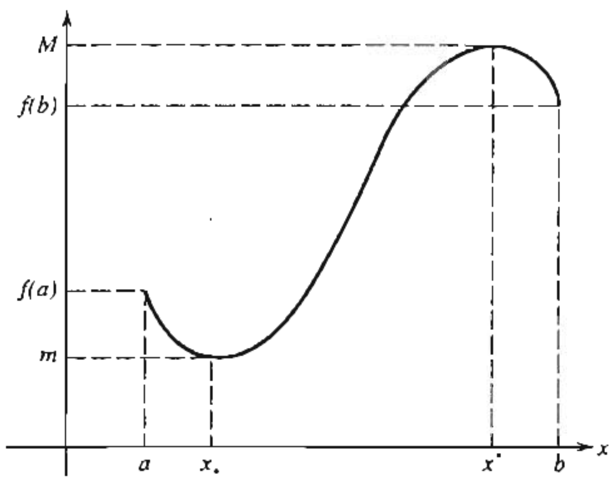


Figure 5.3.3 $f(I) = [m, M]$.

The preceding theorem is a "preservation" theorem in the sense that it states that the continuous image of a closed bounded interval is a set of the same type. The next theorem extends this result to general intervals. However, it should be noted that although the continuous image of an interval is shown to be an interval, it is *not* true that the image interval necessarily has the *same form* as the domain interval. For example, the continuous image of an open interval need not be an open interval, and the continuous image of an unbounded closed interval need not be a closed interval. Indeed, if $f(x) := 1/(x^2 + 1)$ for $x \in \mathbb{R}$, then f is continuous on $\mathbb{R}$ [see Example 5.2.3(b)]. It is easy to see that if $I_1 := (-1, 1)$, then $f(I_1) = \left(\frac{1}{2}, 1\right]$, which is not an open interval. Also, if $I_2 := [0, \infty)$, then $f(I_2) = (0, 1]$, which is not a closed interval. (See Figure 5.3.4.)

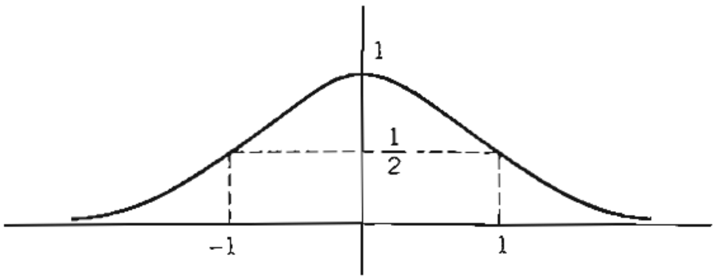


Figure 5.3.4 Graph of $f(x) = 1/(x^2+1)$ $(x \in \mathbb{R})$.

To prove the Preservation of Intervals Theorem 5.3.10, we will use Theorem 2.5.1 characterizing intervals.

5.3.10 Preservation of Intervals Theorem *Let I be an interval and let $f : I \to \mathbb{R}$ be continuous on I. Then the set $f(I)$ is an interval.*

Proof. Let $\alpha, \beta \in f(I)$ with $\alpha < \beta$; then there exist points $a, b \in I$ such that $\alpha = f(a)$ and $\beta = f(b)$. Further, it follows from Bolzano's Intermediate Value Theorem 5.3.7 that if $k \in (\alpha, \beta)$ then there exists a number $c \in I$ with $k = f(c) \in f(I)$. Therefore $[\alpha, \beta] \subseteq f(I)$, showing that $f(I)$ possesses property (1) of Theorem 2.5.1. Therefore $f(I)$ is an interval. Q.E.D.

Exercises for Section 5.3

1. Let $I := [a, b]$ and let $f : I \to \mathbb{R}$ be a continuous function such that $f(x) > 0$ for each x in I. Prove that there exists a number $\alpha > 0$ such that $f(x) \geq \alpha$ for all $x \in I$.
2. Let $I := [a, b]$ and let $f : I \to \mathbb{R}$ and $g : I \to \mathbb{R}$ be continuous functions on I. Show that the set $E := \{x \in I : f(x) = g(x)\}$ has the property that if $(x_n) \subseteq E$ and $x_n \to x_0$, then $x_0 \in E$.
3. Let $I := [a, b]$ and let $f : I \to \mathbb{R}$ be a continuous function on I such that for each x in I there exists y in I such that $|f(y)| \leq \frac{1}{2}|f(x)|$. Prove there exists a point c in I such that $f(c) = 0$.
4. Show that every polynomial of odd degree with real coefficients has at least one real root.
5. Show that the polynomial $p(x) := x^4 + 7x^3 - 9$ has at least two real roots. Use a calculator to locate these roots to within two decimal places.
6. Let f be continuous on the interval $[0, 1]$ to $\mathbb{R}$ and such that $f(0) = f(1)$. Prove that there exists a point c in $[0, \frac{1}{2}]$ such that $f(c) = f\left(c + \frac{1}{2}\right)$. [*Hint:* Consider $g(x) = f(x) - f\left(x + \frac{1}{2}\right)$.] Conclude that there are, at any time, antipodal points on the earth's equator that have the same temperature.
7. Show that the equation $x = \cos x$ has a solution in the interval $[0, \pi/2]$. Use the Bisection Method and a calculator to find an approximate solution of this equation, with error less than 10^{-3}.
8. Show that the function $f(x) := 2\ln x + \sqrt{x} - 2$ has root in the interval $[1, 2]$. Use the Bisection Method and a calculator to find the root with error less than 10^{-2}.
9. (a) The function $f(x) := (x-1)(x-2)(x-3)(x-4)(x-5)$ has five roots in the interval $[0, 7]$. If the Bisection Method is applied on this interval, which of the roots is located?
   (b) Same question for $g(x) := (x-2)(x-3)(x-4)(x-5)(x-6)$ on the interval $[0, 7]$.
10. If the Bisection Method is used on an interval of length 1 to find p_n with error $|p_n - c| < 10^{-5}$, determine the least value of n that will assure this accuracy.

11. Let $I := [a, b]$, let $f : I \to \mathbb{R}$ be continuous on I, and assume that $f(a) < 0$, $f(b) > 0$. Let $W := \{x \in I : f(x) < 0\}$, and let $w := \sup W$. Prove that $f(w) = 0$. (This provides an alternative proof of Theorem 5.3.5.)
12. Let $I := [0, \pi/2]$ and let $f : I \to \mathbb{R}$ be defined by $f(x) := \sup\{x^2, \cos x\}$ for $x \in I$. Show there exists an absolute minimum point $x_0 \in I$ for f on I. Show that x_0 is a solution to the equation $\cos x = x^2$.
13. Suppose that $f : \mathbb{R} \to \mathbb{R}$ is continuous on $\mathbb{R}$ and that $\lim\limits_{x\to-\infty} f = 0$ and $\lim\limits_{x\to\infty} f = 0$. Prove that f is bounded on $\mathbb{R}$ and attains either a maximum or minimum on $\mathbb{R}$. Give an example to show that both a maximum and a minimum need not be attained.
14. Let $f : \mathbb{R} \to \mathbb{R}$ be continuous on $\mathbb{R}$ and let $\beta \in \mathbb{R}$. Show that if $x_0 \in \mathbb{R}$ is such that $f(x_0) < \beta$, then there exists a δ-neighborhood U of x_0 such that $f(x) < \beta$ for all $x \in U$.
15. Examine which open [respectively, closed] intervals are mapped by $f(x) := x^2$ for $x \in \mathbb{R}$ onto open [respectively, closed] intervals.
16. Examine the mapping of open [respectively, closed] intervals under the functions $g(x) := 1/(x^2 + 1)$ and $h(x) := x^3$ for $x \in \mathbb{R}$.
17. If $f : [0, 1] \to \mathbb{R}$ is continuous and has only rational [respectively, irrational] values, must f be constant? Prove your assertion.
18. Let $I := [a, b]$ and let $f : I \to \mathbb{R}$ be a (not necessarily continuous) function with the property that for every $x \in I$, the function f is bounded on a neighborhood $V_{\delta_x}(x)$ of x (in the sense of Definition 4.2.1). Prove that f is bounded on I.
19. Let $J := (a, b)$ and let $g : J \to \mathbb{R}$ be a continuous function with the property that for every $x \in J$, the function g is bounded on a neighborhood $V_{\delta_x}(x)$ of x. Show by example that g is not necessarily bounded on J.

Section 5.4 Uniform Continuity

Let $A \subseteq \mathbb{R}$ and let $f : A \to \mathbb{R}$. Definition 5.1.1 states that the following statements are equivalent:

(i) f is continuous at every point $u \in A$;

(ii) given $\varepsilon > 0$ and $u \in A$, there is a $\delta(\varepsilon, u) > 0$ such that for all x such that $x \in A$ and $|x - u| < \delta(\varepsilon, u)$, then $|f(x) - f(u)| < \varepsilon$.

The point we wish to emphasize here is that δ depends, in general, on *both* $\varepsilon > 0$ and $u \in A$. The fact that δ depends on u is a reflection of the fact that the function f may change its values rapidly near certain points and slowly near other points. [For example, consider $f(x) := \sin(1/x)$ for $x > 0$; see Figure 4.1.3.]

Now it often happens that the function f is such that the number δ can be chosen to be independent of the point $u \in A$ and to depend only on ε. For example, if $f(x) := 2x$ for all $x \in \mathbb{R}$, then

$$|f(x) - f(u)| = 2\,|x - u|,$$

and so we can choose $\delta(\varepsilon, u) := \varepsilon/2$ for all $\varepsilon > 0$, $u \in \mathbb{R}$. (Why?)

On the other hand if $g(x) := 1/x$ for $x \in A := \{x \in \mathbb{R}: x > 0\}$, then

$$g(x) - g(u) = \frac{u - x}{ux}. \tag{1}$$

If $u \in A$ is given and if we take

$$\delta(\varepsilon, u) := \inf\left\{\tfrac{1}{2}u, \tfrac{1}{2}u^2\varepsilon\right\}, \tag{2}$$

then if $|x - u| < \delta(\varepsilon, u)$, we have $|x - u| < \frac{1}{2}u$ so that $\frac{1}{2}u < x < \frac{3}{2}u$, whence it follows that $1/x < 2/u$. Thus, if $|x - u| < \frac{1}{2}u$, the equality (1) yields the inequality

$$|g(x) - g(u)| \le (2/u^2)\,|x - u|. \tag{3}$$

Consequently, if $|x - u| < \delta(\varepsilon, u)$, then (2) and (3) imply that

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

We have seen that the selection of $\delta(\varepsilon, u)$ by the formula (2) "works" in the sense that it enables us to give a value of δ that will ensure that $|g(x) - g(u)| < \varepsilon$ when $|x - u| < \delta$ and $x, u \in A$. We note that the value of $\delta(\varepsilon, u)$ given in (2) certainly depends on the point $u \in A$. If we wish to consider *all* $u \in A$, formula (2) does not lead to one value $\delta(\varepsilon) > 0$ that will "work" simultaneously for all $u > 0$, since $\inf\{\delta(\varepsilon, u) : u > 0\} = 0$.

An alert reader will have observed that there are other selections that can be made for δ. (For example we could also take $\delta_1(\varepsilon, u) := \inf\left\{\frac{1}{3}u, \frac{2}{3}u^2\varepsilon\right\}$, as the reader can show; however, we still have $\inf\left\{\delta_1(\varepsilon, u) : u > 0\right\} = 0$.) In fact, there is no way of choosing one value of δ that will "work" for all $u > 0$ for the function $g(x) = 1/x$, as we shall see.

The situation is exhibited graphically in Figures 5.4.1 and 5.4.2 where, for a given ε-neighborhood $V_\varepsilon(\frac{1}{2})$ about $\frac{1}{2} = f(2)$ and $V_\varepsilon(2)$ about $2 = f(\frac{1}{2})$, the corresponding maximum values of δ are seen to be considerably different. As u tends to 0, the permissible values of δ tend to 0.

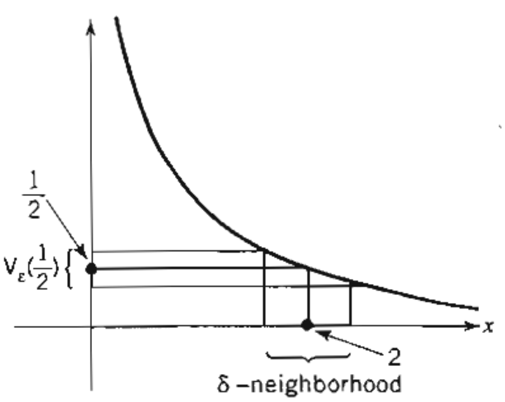


Figure 5.4.1 $g(x) = 1/x \quad (x > 0)$.

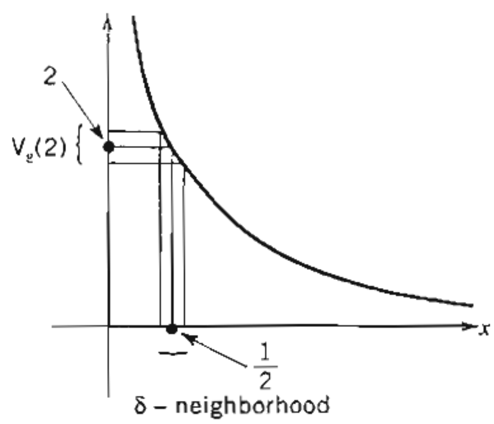


Figure 5.4.2 $g(x) = 1/x \quad (x > 0)$.

5.4.1 Definition Let $A \subseteq \mathbb{R}$ and let $f : A \to \mathbb{R}$. We say that f is **uniformly continuous** on A if for each $\varepsilon > 0$ there is a $\delta(\varepsilon) > 0$ such that if $x, u \in A$ are any numbers satisfying $|x - u| < \delta(\varepsilon)$, then $|f(x) - f(u)| < \varepsilon$.

It is clear that if f is uniformly continuous on A, then it is continuous at every point of A. In general, however, the converse does not hold, as is shown by the function $g(x) = 1/x$ on the set $A := \{x \in \mathbb{R} : x > 0\}$.

It is useful to formulate a condition equivalent to saying that f is *not* uniformly continuous on A. We give such criteria in the next result, leaving the proof to the reader as an exercise.

5.4.2 Nonuniform Continuity Criteria *Let $A \subseteq \mathbb{R}$ and let $f : A \to \mathbb{R}$. Then the following statements are equivalent:*

(i) *f is not uniformly continuous on A.*

(ii) *There exists an $\varepsilon_0 > 0$ such that for every $\delta > 0$ there are points x_δ, u_δ in A such that $|x_\delta - u_\delta| < \delta$ and $|f(x_\delta) - f(u_\delta)| \geq \varepsilon_0$.*

(iii) *There exists an $\varepsilon_0 > 0$ and two sequences (x_n) and (u_n) in A such that $\lim(x_n - u_n) = 0$ and $|f(x_n) - f(u_n)| \geq \varepsilon_0$ for all $n \in \mathbb{N}$.*

We can apply this result to show that $g(x) := 1/x$ is not uniformly continuous on $A := \{x \in \mathbb{R} : x > 0\}$. For, if $x_n := 1/n$ and $u_n := 1/(n+1)$, then we have $\lim(x_n - u_n) = 0$, but $|g(x_n) - g(u_n)| = 1$ for all $n \in \mathbb{N}$.

We now present an important result that assures that a continuous function on a closed bounded interval I is uniformly continuous on I. Other proofs of this theorem are given in Sections 5.5 and 11.3.

5.4.3 Uniform Continuity Theorem *Let I be a closed bounded interval and let $f : I \to \mathbb{R}$ be continuous on I. Then f is uniformly continuous on I.*

Proof. If f is not uniformly continuous on I then, by the preceding result, there exists $\varepsilon_0 > 0$ and two sequences (x_n) and (u_n) in I such that $|x_n - u_n| < 1/n$ and $|f(x_n) - f(u_n)| \geq \varepsilon_0$ for all $n \in \mathbb{N}$. Since I is bounded, the sequence (x_n) is bounded; by the Bolzano-Weierstrass Theorem 3.4.8 there is a subsequence (x_{n_k}) of (x_n) that converges to an element z. Since I is closed, the limit z belongs to I, by Theorem 3.2.6. It is clear that the corresponding subsequence (u_{n_k}) also converges to z, since

$$|u_{n_k} - z| \leq |u_{n_k} - x_{n_k}| + |x_{n_k} - z|.$$

Now if f is continuous at the point z, then both of the sequences $\big(f(x_{n_k})\big)$ and $\big(f(u_{n_k})\big)$ must converge to $f(z)$. But this is not possible since

$$|f(x_n) - f(u_n)| \geq \varepsilon_0$$

for all $n \in \mathbb{N}$. Thus the hypothesis that f is not uniformly continuous on the closed bounded interval I implies that f is not continuous at some point $z \in I$. Consequently, if f is continuous at every point of I, then f is uniformly continuous on I. Q.E.D.

Lipschitz Functions

If a uniformly continuous function is given on a set that is not a closed bounded interval, then it is sometimes difficult to establish its uniform continuity. However, there is a condition that frequently occurs that is sufficient to guarantee uniform continuity.

5.4.4 Definition Let $A \subseteq \mathbb{R}$ and let $f : A \to \mathbb{R}$. If there exists a constant $K > 0$ such that

$$|f(x) - f(u)| \leq K\,|x - u| \tag{4}$$

for all $x, u \in A$, then f is said to be a **Lipschitz function** (or to satisfy a **Lipschitz condition**) on A.

The condition (4) that a function $f : I \to \mathbb{R}$ on an interval I is a Lipschitz function an be interpreted geometrically as follows. If we write the condition as

$$\left|\frac{f(x) - f(u)}{x - u}\right| \le K, \qquad x, u \in I, x \ne u,$$

n the quantity inside the absolute values is the slope of a line segment joining the points $x, f(x))$ and $(u, f(u))$. Thus a function f satisfies a Lipschitz condition if and only if the opes of all line segments joining two points on the graph of $y = f(x)$ over I are bounded some number K.

.5 **Theorem** *If $f : A \to \mathbb{R}$ is a Lipschitz function, then f is uniformly continuous* *n A.*

oof. If condition (4) is satisfied, then given $\varepsilon > 0$, we can take $\delta := \varepsilon/K$. If $x, u \in A$ atisfy $|x - u| < \delta$, then

$$|f(x) - f(u)| < K \cdot \frac{\varepsilon}{K} = \varepsilon.$$

erefore f is uniformly continuous on A. Q.E.D.

.6 **Examples** (a) If $f(x) := x^2$ on $A := [0, b]$, where $b > 0$, then

$$|f(x) - f(u)| = |x + u|\,|x - u| \le 2b\,|x - u|$$

all x, u in $[0, b]$. Thus f satisfies (4) with $K := 2b$ on A, and therefore f is uniformly ontinuous on A. Of course, since f is continuous and A is a closed bounded interval, this n also be deduced from the Uniform Continuity Theorem. (Note that f does *not* satisfy Lipschitz condition on the interval $[0, \infty)$.)

) Not every uniformly continuous function is a Lipschitz function.

Let $g(x) := \sqrt{x}$ for x in the closed bounded interval $I := [0, 2]$. Since g is continuous on I, it follows from the Uniform Continuity Theorem 5.4.3 that g is uniformly continuous I. However, there is no number $K > 0$ such that $|g(x)| \le K|x|$ for all $x \in I$. (Why not?) Therefore, g is not a Lipschitz function on I.

) The Uniform Continuity Theorem and Theorem 5.4.5 can sometimes be combined to establish the uniform continuity of a function on a set.

We consider $g(x) := \sqrt{x}$ on the set $A := [0, \infty)$. The uniform continuity of g on e interval $I := [0, 2]$ follows from the Uniform Continuity Theorem as noted in (b). If $J := [1, \infty)$, then if both x, u are in J, we have

$$|g(x) - g(u)| = |\sqrt{x} - \sqrt{u}| = \frac{|x - u|}{\sqrt{x} + \sqrt{u}} \le \tfrac{1}{2}\,|x - u|.$$

hus g is a Lipschitz function on J with constant $K = \frac{1}{2}$, and hence by Theorem 5.4.5, g is uniformly continuous on $[1, \infty)$. Since $A = I \cup J$, it follows [by taking $\delta(\varepsilon) :=$ f$\{1, \delta_I(\varepsilon), \delta_J(\varepsilon)\}$] that g is uniformly continuous on A. We leave the details to the reader. □

he Continuous Extension Theorem

e have seen examples of functions that are continuous but not uniformly continuous on open intervals; for example, the function $f(x) = 1/x$ on the interval $(0, 1)$. On the other nd, by the Uniform Continuity Theorem, a function that is continuous on a closed bounded terval is always uniformly continuous. So the question arises: Under what conditions is a function uniformly continuous on a bounded *open* interval? The answer reveals the strength of uniform continuity, for it will be shown that a function on (a, b) is uniformly continuous if and only if it can be defined at the endpoints to produce a function that is continuous on the closed interval. We first establish a result that is of interest in itself.

5.4.7 Theorem *If $f : A \to \mathbb{R}$ is uniformly continuous on a subset A of $\mathbb{R}$ and if (x_n) is a Cauchy sequence in A, then $(f(x_n))$ is a Cauchy sequence in $\mathbb{R}$.*

Proof. Let (x_n) be a Cauchy sequence in A, and let $\varepsilon > 0$ be given. First choose $\delta > 0$ such that if x, u in A satisfy $|x - u| < \delta$, then $|f(x) - f(u)| < \varepsilon$. Since (x_n) is a Cauchy sequence, there exists $H(\delta)$ such that $|x_n - x_m| < \delta$ for all $n, m > H(\delta)$. By the choice of δ, this implies that for $n, m > H(\delta)$, we have $|f(x_n) - f(x_m)| < \varepsilon$. Therefore the sequence $(f(x_n))$ is a Cauchy sequence. Q.E.D

The preceding result gives us an alternative way of seeing that $f(x) := 1/x$ is not uniformly continuous on $(0, 1)$. We note that the sequence given by $x_n := 1/n$ in $(0, 1)$ is a Cauchy sequence, but the image sequence, where $f(x_n) = n$, is not a Cauchy sequence.

5.4.8 Continuous Extension Theorem *A function f is uniformly continuous on the interval (a, b) if and only if it can be defined at the endpoints a and b such that the extended function is continuous on $[a, b]$.*

Proof. $(\Leftarrow)$ This direction is trivial.

$(\Rightarrow)$ Suppose f is uniformly continuous on (a, b). We shall show how to extend f to a; the argument for b is similar. This is done by showing that $\lim_{x \to c} f(x) = L$ exists, and this is accomplished by using the sequential criterion for limits. If (x_n) is a sequence in (a, b) with $\lim(x_n) = a$, then it is a Cauchy sequence, and by the preceding theorem, the sequence $(f(x_n))$ is also a Cauchy sequence, and so is convergent by Theorem 3.5.5. Thus the limit $\lim(f(x_n)) = L$ exists. If (u_n) is any other sequence in (a, b) that converges to a, then $\lim(u_n - x_n) = a - a = 0$, so by the uniform continuity of f we have

$$\begin{aligned}\lim(f(u_n)) &= \lim(f(u_n) - f(x_n)) + \lim(f(x_n)) \\ &= 0 + L = L.\end{aligned}$$

Since we get the same value L for every sequence converging to a, we infer from the sequential criterion for limits that f has limit L at a. If we define $f(a) := L$, then f is continuous at a. The same argument applies to b, so we conclude that f has a continuous extension to the interval $[a, b]$. Q.E.D.

Since the limit of $f(x) := \sin(1/x)$ at 0 does not exist, we infer from the Continuous Extension Theorem that the function is not uniformly continuous on $(0, b]$ for any $b > 0$. On the other hand, since $\lim_{x \to 0} x \sin(1/x) = 0$ exists, the function $g(x) := x \sin(1/x)$ is uniformly continuous on $(0, b]$ for all $b > 0$.

Approximation†

In many applications it is important to be able to approximate continuous functions by functions of an elementary nature. Although there are a variety of definitions that can be used to make the word "approximate" more precise, one of the most natural (as well as one of

†The rest of this section can be omitted on a first reading of this chapter.

the most important) is to require that, at every point of the given domain, the approximating function shall not differ from the given function by more than the preassigned error.

5.4.9 Definition Let $I \subseteq \mathbb{R}$ be an interval and let $s : I \to \mathbb{R}$. Then s is called a **step function** if it has only a finite number of distinct values, each value being assumed on one or more intervals in I.

For example, the function $s : [-2, 4] \to \mathbb{R}$ defined by

$$s(x) := \begin{cases} 0, & -2 \le x < -1, \\ 1, & -1 \le x \le 0, \\ \frac{1}{2}, & 0 < x < \frac{1}{2}, \\ 3, & \frac{1}{2} \le x < 1, \\ -2, & 1 \le x \le 3, \\ 2, & 3 < x \le 4, \end{cases}$$

is a step function. (See Figure 5.4.3.)

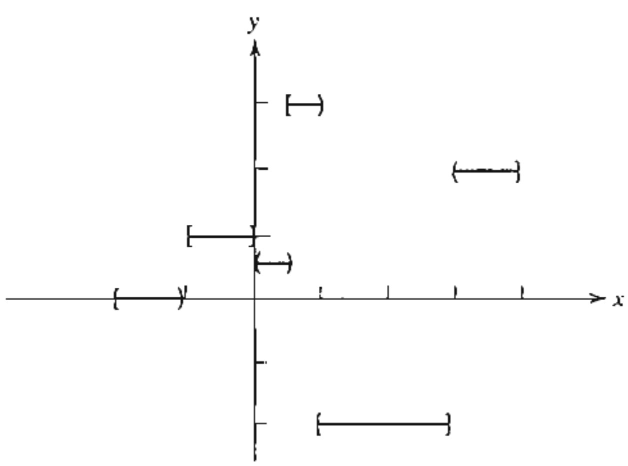


Figure 5.4.3 Graph of $y = s(x)$.

We will now show that a continuous function on a closed bounded interval I can be approximated arbitrarily closely by step functions.

5.4.10 Theorem *Let I be a closed bounded interval and let $f : I \to \mathbb{R}$ be continuous on I. If $\varepsilon > 0$, then there exists a step function $s_\varepsilon : I \to \mathbb{R}$ such that $|f(x) - s_\varepsilon(x)| < \varepsilon$ for all $x \in I$.*

Proof. Since (by the Uniform Continuity Theorem 5.4.3) the function f is uniformly continuous, it follows that given $\varepsilon > 0$ there is a number $\delta(\varepsilon) > 0$ such that if $x, y \in I$ and $|x - y| < \delta(\varepsilon)$, then $|f(x) - f(y)| < \varepsilon$. Let $I := [a, b]$ and let $m \in \mathbb{N}$ be sufficiently large so that $h := (b - a)/m < \delta(\varepsilon)$. We now divide $I = [a, b]$ into m disjoint intervals of length h; namely, $I_1 := [a, a + h]$, and $I_k := \big(a + (k-1)h, a + kh\big]$ for $k = 2, \cdots, m$. Since the length of each subinterval I_k is $h < \delta(\varepsilon)$, the difference between any two values of f in I_k is less than ε. We now define

(5) $$s_\varepsilon(x) := f(a + kh) \qquad \text{for} \quad x \in I_k, \quad k = 1, \cdots, m,$$

so that s_ε is constant on each interval I_k. (In fact the value of s_ε on I_k is the value of f at the right endpoint of I_k. See Figure 5.4.4.) Consequently if $x \in I_k$, then

$$|f(x) - s_\varepsilon(x)| = |f(x) - f(a + kh)| < \varepsilon.$$

Therefore we have $|f(x) - s_\varepsilon(x)| < \varepsilon$ for all $x \in I$. Q.E.D.

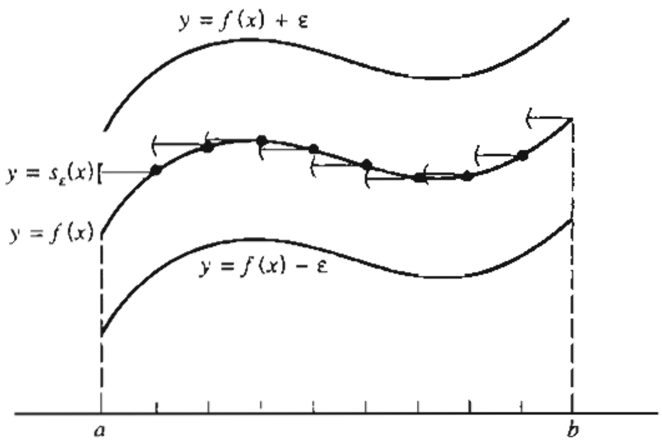


Figure 5.4.4 Approximation by step functions.

Note that the proof of the preceding theorem establishes somewhat more than was announced in the statement of the theorem. In fact, we have proved the following, more precise, assertion.

5.4.11 Corollary *Let $I := [a, b]$ be a closed bounded interval and let $f : I \to \mathbb{R}$ be continuous on I. If $\varepsilon > 0$, there exists a natural number m such that if we divide I into m disjoint intervals I_k having length $h := (b - a)/m$, then the step function s_ε defined in equation (5) satisfies $|f(x) - s_\varepsilon(x)| < \varepsilon$ for all $x \in I$.*

Step functions are extremely elementary in character, but they are not continuous (except in trivial cases). Since it is often desirable to approximate continuous functions by elementary continuous functions, we now shall show that we can approximate continuous functions by continuous piecewise linear functions.

5.4.12 Definition Let $I := [a, b]$ be an interval. Then a function $g : I \to \mathbb{R}$ is said to be **piecewise linear** on I if I is the union of a finite number of disjoint intervals $I_1, \cdots, I_m$, such that the restriction of g to each interval I_k is a linear function.

Remark It is evident that in order for a piecewise linear function g to be continuous on I, the line segments that form the graph of g must meet at the endpoints of adjacent subintervals I_k, I_{k+1} $(k = 1, \cdots, m - 1)$.

5.4.13 Theorem *Let I be a closed bounded interval and let $f : I \to \mathbb{R}$ be continuous on I. If $\varepsilon > 0$, then there exists a continuous piecewise linear function $g_\varepsilon : I \to \mathbb{R}$ such that $|f(x) - g_\varepsilon(x)| < \varepsilon$ for all $x \in I$.*

Proof. Since f is uniformly continuous on $I := [a, b]$, there is a number $\delta(\varepsilon) > 0$ such that if $x, y \in I$ and $|x - y| < \delta(\varepsilon)$, then $|f(x) - f(y)| < \varepsilon$. Let $m \in \mathbb{N}$ be sufficiently large so that $h := (b - a)/m < \delta(\varepsilon)$. Divide $I = [a, b]$ into m disjoint intervals of length h; namely let $I_1 = [a, a + h]$, and let $I_k = \big(a + (k - 1)h, a + kh\big]$ for $k = 2, \cdots, m$. On each interval I_k we define g_ε to be the linear function joining the points

$$\big(a + (k - 1)h, f(a + (k - 1)h)\big) \qquad \text{and} \qquad \big(a + kh, f(a + kh)\big).$$

Then g_ε is a continuous piecewise linear function on I. Since, for $x \in I_k$ the value $f(x)$ is within ε of $f(a + (k - 1)h)$ and $f(a + kh)$, it is an exercise to show that $|f(x) - g_\varepsilon(x)| < \varepsilon$ for all $x \in I_k$; therefore this inequality holds for all $x \in I$. (See Figure 5.4.5.) Q.E.D.

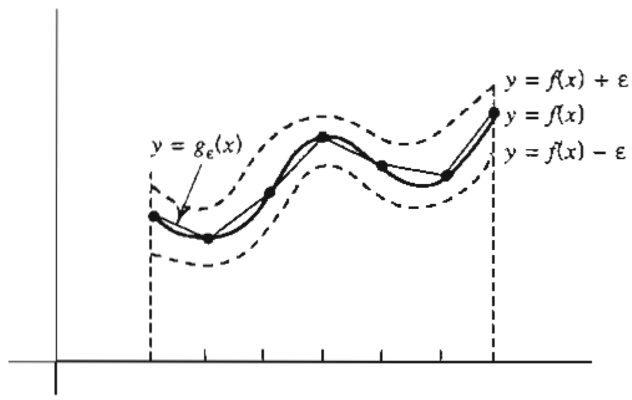


Figure 5.4.5 Approximation by piecewise linear function.

We shall close this section by stating the important theorem of Weierstrass concerning the approximation of continuous functions by polynomial functions. As would be expected, in order to obtain an approximation within an arbitrarily preassigned $\varepsilon > 0$, we must be prepared to use polynomials of arbitrarily high degree.

5.4.14 Weierstrass Approximation Theorem *Let $I = [a, b]$ and let $f : I \to \mathbb{R}$ be a continuous function. If $\varepsilon > 0$ is given, then there exists a polynomial function p_ε such that $|f(x) - p_\varepsilon(x)| < \varepsilon$ for all $x \in I$.*

There are a number of proofs of this result. Unfortunately, all of them are rather intricate, or employ results that are not yet at our disposal. One of the most elementary proofs is based on the following theorem, due to Serge Bernsteĭn, for continuous functions on $[0, 1]$. Given $f : [0, 1] \to \mathbb{R}$, Bernsteĭn defined the sequence of polynomials:

(6)
$$B_n(x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right)\binom{n}{k} x^k (1 - x)^{n-k}.$$

The polynomial function B_n is called the ***n*th Bernsteĭn polynomial for** f; it is a polynomial of degree at most n and its coefficients depend on the values of the function f at the $n + 1$ equally spaced points $0, 1/n, 2/n, \cdots, k/n, \cdots, 1$, and on the binomial coefficients

$$\binom{n}{k} = \frac{n!}{k!(n-k)} = \frac{n(n-1)\cdots(n-k+1)}{1 \cdot 2 \cdots k}.$$

5.4.15 Bernsteĭn's Approximation Theorem *Let $f : [0, 1] \to \mathbb{R}$ be continuous and let $\varepsilon > 0$. There exists an $n_\varepsilon \in \mathbb{N}$ such that if $n \geq n_\varepsilon$, then we have $|f(x) - B_n(x)| < \varepsilon$ for all $x \in [0, 1]$.*

The proof of Bernsteĭn's Approximation Theorem is given in [ERA, pp. 169–172].

The Weierstrass Approximation Theorem 5.4.14 can be derived from the Bernsteĭn Approximation Theorem 5.4.15 by a change of variable. Specifically, we replace $f : [a, b] \to \mathbb{R}$ by a function $F : [0, 1] \to \mathbb{R}$, defined by

$$F(t) := f\big(a + (b - a)t\big) \qquad \text{for} \quad t \in [0, 1].$$

The function F can be approximated by Bernsteĭn polynomials for F on the interval $[0, 1]$, which can then yield polynomials on $[a, b]$ that approximate f.

Exercises for Section 5.4

1. Show that the function $f(x) := 1/x$ is uniformly continuous on the set $A := [a, \infty)$, where a is a positive constant.
2. Show that the function $f(x) := 1/x^2$ is uniformly continuous on $A := [1, \infty)$, but that it is not uniformly continuous on $B := (0, \infty)$.
3. Use the Nonuniform Continuity Criterion 5.4.2 to show that the following functions are not uniformly continuous on the given sets.
   (a) $f(x) := x^2$, $A := [0, \infty)$.
   (b) $g(x) := \sin(1/x)$, $B := (0, \infty)$.
4. Show that the function $f(x) := 1/(1 + x^2)$ for $x \in \mathbb{R}$ is uniformly continuous on $\mathbb{R}$.
5. Show that if f and g are uniformly continuous on a subset A of $\mathbb{R}$, then $f + g$ is uniformly continuous on A.
6. Show that if f and g are uniformly continuous on $A \subseteq \mathbb{R}$ and if they are *both* bounded on A, then their product fg is uniformly continuous on A.
7. If $f(x) := x$ and $g(x) := \sin x$, show that both f and g are uniformly continuous on $\mathbb{R}$, but that their product fg is not uniformly continuous on $\mathbb{R}$.
8. Prove that if f and g are each uniformly continuous on $\mathbb{R}$, then the composite function $f \circ g$ is uniformly continuous on $\mathbb{R}$.
9. If f is uniformly continuous on $A \subseteq \mathbb{R}$, and $|f(x)| \geq k > 0$ for all $x \in A$, show that $1/f$ is uniformly continuous on A.
10. Prove that if f is uniformly continuous on a bounded subset A of $\mathbb{R}$, then f is bounded on A.
11. If $g(x) := \sqrt{x}$ for $x \in [0, 1]$, show that there does not exist a constant K such that $|g(x)| \leq K|x|$ for all $x \in [0, 1]$. Conclude that the uniformly continuous g is not a Lipschitz function on $[0, 1]$.
12. Show that if f is continuous on $[0, \infty)$ and uniformly continuous on $[a, \infty)$ for some positive constant a, then f is uniformly continuous on $[0, \infty)$.
13. Let $A \subseteq \mathbb{R}$ and suppose that $f : A \to \mathbb{R}$ has the following property: for each $\varepsilon > 0$ there exists a function $g_\varepsilon : A \to \mathbb{R}$ such that g_ε is uniformly continuous on A and $|f(x) - g_\varepsilon(x)| < \varepsilon$ for all $x \in A$. Prove that f is uniformly continuous on A.
14. A function $f\colon \mathbb{R} \to \mathbb{R}$ is said to be **periodic** on $\mathbb{R}$ if there exists a number $p > 0$ such that $f(x + p) = f(x)$ for all $x \in \mathbb{R}$. Prove that a continuous periodic function on $\mathbb{R}$ is bounded and uniformly continuous on $\mathbb{R}$.

15. If $f_0(x) := 1$ for $x \in [0, 1]$, calculate the first few Bernsteĭn polynomials for f_0. Show that they coincide with f_0. [*Hint*: The Binomial Theorem asserts that

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}.]$$

16. If $f_1(x) := x$ for $x \in [0, 1]$, calculate the first few Bernsteĭn polynomials for f_1. Show that they coincide with f_1.

17. If $f_2(x) := x^2$ for $x \in [0, 1]$, calculate the first few Bernsteĭn polynomials for f_2. Show that $B_n(x) = (1 - 1/n)x^2 + (1/n)x$.

Section 5.5 Continuity and Gauges

We will now introduce some concepts that will be used later—especially in Chapters 7 and 10 on integration theory. However, we wish to introduce the notion of a "gauge" now because of its connection with the study of continuous functions. We first define the notion of a tagged partition of an interval.

5.5.1 Definition A **partition** of an interval $I := [a, b]$ is a collection $\mathcal{P} = \{I_1, \cdots, I_n\}$ of non-overlapping closed intervals whose union is $[a, b]$. We ordinarily denote the intervals by $I_i := [x_{i-1}, x_i]$, where

$$a = x_0 < \cdots < x_{i-1} < x_i < \cdots < x_n = b.$$

The points x_i $(i = 0, \cdots, n)$ are called the **partition points** of $\mathcal{P}$. If a point t_i has been chosen from each interval I_i, for $i = 1, \cdots, n$, then the points t_i are called the **tags** and the set of ordered pairs

$$\dot{\mathcal{P}} = \{(I_1, t_1), \cdots, (I_n, t_n)\}$$

is called a **tagged partition** of I. (The dot signifies that the partition is tagged.)

The "fineness" of a partition $\mathcal{P}$ refers to the lengths of the subintervals in $\mathcal{P}$. Instead of requiring that all subintervals have length less than some specific quantity, it is often useful to allow varying degrees of fineness for different subintervals I_i in $\mathcal{P}$. This is accomplished by the use of a "gauge", which we now define.

5.5.2 Definition A **gauge** on I is a strictly positive function defined on I. If δ is a gauge on I, then a (tagged) partition $\dot{\mathcal{P}}$ is said to be **δ-fine** if

(1) $$t_i \in I_i \subseteq [t_i - \delta(t_i), t_i + \delta(t_i)] \quad \text{for} \quad i = 1, \cdots, n.$$

We note that the notion of δ-fineness requires that the partition be tagged, so we do not need to say "tagged partition" in this case.

Figure 5.5.1 Inclusion (1).

A gauge δ on an interval I assigns an interval $[t - \delta(t), t + \delta(t)]$ to each point $t \in I$. The δ-fineness of a partition $\dot{\mathcal{P}}$ requires that each subinterval I_i of $\dot{\mathcal{P}}$ is contained in the interval determined by the gauge δ and the tag t_i for that subinterval. This is indicated by the inclusions in (1); see Figure 5.5.1. Note that the length of the subintervals is also controlled by the gauge and the tags; the next lemma reflects that control.

5.5.3 Lemma *If a partition $\dot{\mathcal{P}}$ of $I := [a, b]$ is δ-fine and $x \in I$, then there exists a tag t_i in $\dot{\mathcal{P}}$ such that $|x - t_i| \leq \delta(t_i)$.*

Proof. If $x \in I$, there exists a subinterval $[x_{i-1}, x_i]$ from $\dot{\mathcal{P}}$ that contains x. Since $\dot{\mathcal{P}}$ is δ-fine, then

$$t_i - \delta(t_i) \leq x_{i-1} \leq x \leq x_i \leq t_i + \delta(t_i), \tag{2}$$

whence it follows that $|x - t_i| \leq \delta(t_i)$. Q.E.D.

In the theory of Riemann integration, we will use gauges δ that are constant functions to control the fineness of the partition; in the theory of the *generalized* Riemann integral, the use of nonconstant gauges is essential. But nonconstant gauge functions arise quite naturally in connection with continuous functions. For, let $f : I \to \mathbb{R}$ be continuous on I and let $\varepsilon > 0$ be given. Then, for each point $t \in I$ there exists $\delta_\varepsilon(t) > 0$ such that if $|x - t| < \delta_\varepsilon(t)$ and $x \in I$, then $|f(x) - f(t)| < \varepsilon$. Since δ_ε is defined and is strictly positive on I, the function δ_ε is a gauge on I. Later in this section, we will use the relations between gauges and continuity to give alternative proofs of the fundamental properties of continuous functions discussed in Sections 5.3 and 5.4.

5.5.4 Examples (a) If δ and γ are gauges on $I := [a, b]$ and if $0 < \delta(x) \leq \gamma(x)$ for all $x \in I$, then every partition $\dot{\mathcal{P}}$ that is δ-fine is also γ-fine. This follows immediately from the inequalities

$$t_i - \gamma(t_i) \leq t_i - \delta(t_i) \quad \text{and} \quad t_i + \delta(t_i) \leq t_i + \gamma(t_i)$$

which imply that

$$t_i \in [t_i - \delta(t_i), t_i + \delta(t_i)] \subseteq [t_i - \gamma(t_i), t_i + \gamma(t_i)] \quad \text{for} \quad i = 1, \cdots, n.$$

(b) If δ_1 and δ_2 are gauges on $I := [a, b]$ and if

$$\delta(x) := \min\{\delta_1(x), \delta_2(x)\} \quad \text{for all} \quad x \in I,$$

then δ is also a gauge on I. Moreover, since $\delta(x) \leq \delta_1(x)$, then every δ-fine partition is δ_1-fine. Similarly, every δ-fine partition is also δ_2-fine.

(c) Suppose that δ is defined on $I := [0, 1]$ by

$$\delta(x) := \begin{cases} \frac{1}{10} & \text{if} \quad x = 0, \\ \frac{1}{2}x & \text{if} \quad 0 < x \leq 1. \end{cases}$$

Then δ is a gauge on $[0, 1]$. If $0 < t \leq 1$, then $[t - \delta(t), t + \delta(t)] = \left[\frac{1}{2}t, \frac{3}{2}t\right]$, which does not contain the point 0. Thus, if $\dot{\mathcal{P}}$ is a δ-fine partition of I, then the only subinterval in $\dot{\mathcal{P}}$ that contains 0 must have the point 0 as its tag.

(d) Let γ be defined on $I := [0, 1]$ by

$$\gamma(x) := \begin{cases} \frac{1}{10} & \text{if} \quad x = 0 \text{ or } x = 1, \\ \frac{1}{2}x & \text{if} \quad 0 < x \leq \frac{1}{2}, \\ \frac{1}{2}(1 - x) & \text{if} \quad \frac{1}{2} < x < 1. \end{cases}$$

Then γ is a gauge on I, and it is an exercise to show that the subintervals in any γ-fine partition that contain the points 0 or 1 must have these points as tags. □

Existence of δ-Fine Partitions

In view of the above examples, it is not obvious that an arbitrary gauge δ admits a δ-fine partition. We now use the Supremum Property of $\mathbb{R}$ to establish the existence of δ-fine partitions. In the Exercises, we will sketch a proof based on the Nested Intervals Theorem 2.5.2.

5.5.5 Theorem *If δ is a gauge defined on the interval $[a, b]$, then there exists a δ-fine partition of $[a, b]$.*

Proof. Let E denote the set of all points $x \in [a, b]$ such that there exists a δ-fine partition of the subinterval $[a, x]$. The set E is not empty, since the pair $([a, x], a)$ is a δ-fine partition of the interval $[a, x]$ when $x \in [a, a + \delta(a)]$ and $x \le b$. Since $E \subseteq [a, b]$, the set E is also bounded. Let $u := \sup E$ so that $a < u \le b$. We will show that $u \in E$ and that $u = b$.

We claim that $u \in E$. Since $u - \delta(u) < u = \sup E$, there exists $v \in E$ such that $u - \delta(u) < v < u$. Let $\dot{\mathcal{P}}_1$ be a δ-fine partition of $[a, v]$ and let $\dot{\mathcal{P}}_2 := \dot{\mathcal{P}}_1 \cup ([v, u], u)$. Then $\dot{\mathcal{P}}_2$ is a δ-fine partition of $[a, u]$, so that $u \in E$.

If $u < b$, let $w \in [a, b]$ be such that $u < w < u + \delta(u)$. If $\dot{\mathcal{Q}}_1$ is a δ-fine partition of $[a, u]$, we let $\dot{\mathcal{Q}}_2 := \dot{\mathcal{Q}}_1 \cup ([u, w], u)$. Then $\dot{\mathcal{Q}}_2$ is a δ-fine partition of $[a, w]$, whence $w \in E$. But this contradicts the supposition that u is an upper bound of E. Therefore $u = b$.

Q.E.D.

Some Applications

Following R. A. Gordon (see his *Monthly* article), we will now show that some of the major theorems in the two preceding sections can be proved by using gauges.

Alternate Proof of Theorem 5.3.2: Boundedness Theorem. Since f is continuous on I, then for each $t \in I$ there exists $\delta(t) > 0$ such that if $x \in I$ and $|x - t| \le \delta(t)$, then $|f(x) - f(t)| \le 1$. Thus δ is a gauge on I. Let $\{(I_i, t_i)\}_{i=1}^n$ be a δ-fine partition of I and let $K := \max\{|f(t_i)| : i = 1, \cdots, n\}$. By Lemma 5.5.3, given any $x \in I$ there exists i with $|x - t_i| \le \delta(t_i)$, whence

$$|f(x)| \le |f(x) - f(t_i)| + |f(t_i)| \le 1 + K.$$

Since $x \in I$ is arbitrary, then f is bounded by $1 + K$ on I. Q.E.D.

Alternate Proof of Theorem 5.3.4: Maximum-Minimum Theorem. We will prove the existence of x^*. Let $M := \sup\{f(x) : x \in I\}$ and suppose that $f(x) < M$ for all $x \in I$. Since f is continuous on I, for each $t \in I$ there exists $\delta(t) > 0$ such that if $x \in I$ and $|x - t| \le \delta(t)$, then $f(x) < \frac{1}{2}(M + f(t))$. Thus δ is a gauge on I, and if $\{(I_i, t_i)\}_{i=1}^n$ is a δ-fine partition of I, we let

$$\tilde{M} := \tfrac{1}{2}\max\{M + f(t_1), \cdots, M + f(t_n)\}.$$

By Lemma 5.5.3, given any $x \in I$, there exists i with $|x - t_i| \le \delta(t_i)$, whence

$$f(x) < \tfrac{1}{2}\left(M + f(t_i)\right) \le \tilde{M}.$$

Since $x \in I$ is arbitrary, then $\tilde{M}$ $(< M)$ is an upper bound for f on I, contrary to the definition of M as the supremum of f. Q.E.D.

Alternate Proof of Theorem 5.3.5: Location of Roots Theorem. We assume that $f(t) \neq 0$ for all $t \in I$. Since f is continuous at t, Exercise 5.1.7 implies that there exists $\delta(t) > 0$ such that if $x \in I$ and $|x - t| \le \delta(t)$, then $f(x) < 0$ if $f(t) < 0$, and $f(x) > 0$ if $f(t) > 0$. Then δ is a gauge on I and we let $\{(I_i, t_i)\}_{i=1}^n$ be a δ-fine partition. Note that for each i, either $f(x) < 0$ for all $x \in [x_{i-1}, x_i]$ or $f(x) > 0$ for all such x. Since $f(x_0) = f(a) < 0$, this implies that $f(x_1) < 0$, which in turn implies that $f(x_2) < 0$. Continuing in this way, we have $f(b) = f(x_n) < 0$, contrary to the hypothesis that $f(b) > 0$. Q.E.D.

Alternate Proof of Theorem 5.4.3: Uniform Continuity Theorem. Let $\varepsilon > 0$ be given. Since f is continuous at $t \in I$, there exists $\delta(t) > 0$ such that if $x \in I$ and $|x - t| \le 2\delta(t)$, then $|f(x) - f(t)| \le \frac{1}{2}\varepsilon$. Thus δ is a gauge on I. If $\{(I_i, t_i)\}_{i=1}^n$ is a δ-fine partition of I, let $\delta_\varepsilon := \min\{\delta(t_1), \cdots, \delta(t_n)\}$. Now suppose that $x, u \in I$ and $|x - u| \le \delta_\varepsilon$, and choose i with $|x - t_i| \le \delta(t_i)$. Since

$$|u - t_i| \le |u - x| + |x - t_i| \le \delta_\varepsilon + \delta(t_i) \le 2\delta(t_i),$$

then it follows that

$$|f(x) - f(u)| \le |f(x) - f(t_i)| + |f(t_i) - f(u)| \le \tfrac{1}{2}\varepsilon + \tfrac{1}{2}\varepsilon = \varepsilon.$$

Therefore, f is uniformly continuous on I. Q.E.D.

Exercises for Section 5.5

1. Let δ be the gauge on $[0, 1]$ defined by $\delta(0) := \frac{1}{4}$ and $\delta(t) := \frac{1}{2}t$ for $t \in (0, 1]$.
   (a) Show that $\dot{\mathcal{P}}_1 := \{([0, \frac{1}{4}], 0), ([\frac{1}{4}, \frac{1}{2}], \frac{1}{2}), ([\frac{1}{2}, 1], \frac{3}{4})\}$ is δ-fine.
   (b) Show that $\dot{\mathcal{P}}_2 := \{([0, \frac{1}{4}], 0), ([\frac{1}{4}, \frac{1}{2}], \frac{1}{2}), ([\frac{1}{2}, 1], \frac{3}{5})\}$ is not δ-fine.
2. Suppose that δ_1 is the gauge defined by $\delta_1(0) := \frac{1}{4}$, $\delta_1(t) := \frac{3}{4}t$ for $t \in (0, 1]$. Are the partitions given in Exercise 1 δ_1-fine? Note that $\delta(t) \le \delta_1(t)$ for all $t \in [0, 1]$.
3. Suppose that δ_2 is the gauge defined by $\delta_2(0) := \frac{1}{10}$ and $\delta_2(t) := \frac{9}{10}t$ for $t \in (0, 1]$. Are the partitions given in Exercise 1 δ_2-fine?
4. Let γ be the gauge in Example 5.5.4(d).
   (a) If $t \in (0, \frac{1}{2}]$ show that $[t - \gamma(t), t + \gamma(t)] = [\frac{1}{2}t, \frac{3}{2}t] \subseteq (0, \frac{3}{4}]$.
   (b) If $t \in (\frac{1}{2}, 1)$ show that $[t - \gamma(t), t + \gamma(t)] \subseteq (\frac{1}{4}, 1)$.
5. Let $a < c < b$ and let δ be a gauge on $[a, b]$. If $\dot{\mathcal{P}}'$ is a δ-fine partition of $[a, c]$ and if $\dot{\mathcal{P}}''$ is a δ-fine partition of $[c, b]$, show that $\dot{\mathcal{P}}' \cup \dot{\mathcal{P}}''$ is δ-fine partition of $[a, b]$ having c as a partition point.
6. Let $a < c < b$ and let δ' and δ'' be gauges on $[a, c]$ and $[c, b]$, respectively. If δ is defined on $[a, b]$ by
$$\delta(t) := \begin{cases} \delta'(t) & \text{if } t \in [a, c), \\ \min\{\delta'(c), \delta''(c)\} & \text{if } t = c, \\ \delta''(t) & \text{if } t \in (c, b], \end{cases}$$
   then δ is a gauge on $[a, b]$. Moreover, if $\dot{\mathcal{P}}'$ is a δ'-fine partition of $[a, c]$ and $\dot{\mathcal{P}}''$ is a δ''-fine partition of $[c, b]$, then $\dot{\mathcal{P}}' \cup \dot{\mathcal{P}}''$ is a tagged partition of $[a, b]$ having c as a partition point. Explain why $\dot{\mathcal{P}}' \cup \dot{\mathcal{P}}''$ may not be δ-fine. Give an example.
7. Let δ' and δ'' be as in the preceding exercise and let δ^* be defined by
$$\delta^*(t) := \begin{cases} \min\{\delta'(t), \frac{1}{2}(c - t)\} & \text{if } t \in [a, c), \\ \min\{\delta'(c), \delta''(c)\} & \text{if } t = c, \\ \min\{\delta''(t), \frac{1}{2}(t - c)\} & \text{if } t \in (c, b]. \end{cases}$$
   Show that δ^* is a gauge on $[a, b]$ and that every δ^*-fine partition $\dot{\mathcal{P}}$ of $[a, b]$ having c as a partition point gives rise to a δ'-fine partition $\dot{\mathcal{P}}'$ of $[a, c]$ and a δ''-fine partition $\dot{\mathcal{P}}''$ of $[c, b]$ such that $\dot{\mathcal{P}} = \dot{\mathcal{P}}' \cup \dot{\mathcal{P}}''$.

8. Let δ be a gauge on $I := [a, b]$ and suppose that I *does not* have a δ-fine partition.
   (a) Let $c := \frac{1}{2}(a + b)$. Show that at least one of the intervals $[a, c]$ and $[c, b]$ does not have a δ-fine partition.
   (b) Construct a nested sequence (I_n) of subintervals with the length of I_n equal to $(b - a)/2^n$ such that I_n does not have a δ-fine partition.
   (c) Let $\xi \in \bigcap_{n=1}^{\infty} I_n$ and let $p \in \mathbb{N}$ be such that $(b - a)/2^p < \delta(\xi)$. Show that $I_p \subseteq [\xi - \delta(\xi), \xi + \delta(\xi)]$, so the pair (I_p, ξ) is a δ-fine partition of I_p.
9. Let $I := [a, b]$ and let $f : I \to \mathbb{R}$ be a (not necessarily continuous) function. We say that f is "locally bounded" at $c \in I$ if there exists $\delta(c) > 0$ such that f is bounded on $I \cap [c - \delta(c), c + \delta(c)]$. Prove that if f is locally bounded at every point of I, then f is bounded on I.
10. Let $I := [a, b]$ and $f : I \to \mathbb{R}$. We say that f is "locally increasing" at $c \in I$ if there exists $\delta(c) > 0$ such that f is increasing on $I \cap [c - \delta(c), c + \delta(c)]$. Prove that if f is locally increasing at every point of I, then f is increasing on I.

Section 5.6 Monotone and Inverse Functions

Recall that if $A \subseteq \mathbb{R}$, then a function $f : A \to \mathbb{R}$ is said to be **increasing on** A if whenever $x_1, x_2 \in A$ and $x_1 \le x_2$, then $f(x_1) \le f(x_2)$. The function f is said to be **strictly increasing on** A if whenever $x_1, x_2 \in A$ and $x_1 < x_2$, then $f(x_1) < f(x_2)$. Similarly, $g : A \to \mathbb{R}$ is said to be **decreasing on** A if whenever $x_1, x_2 \in A$ and $x_1 \le x_2$ then $g(x_1) \ge g(x_2)$. The function g is said to be **strictly decreasing on** A if whenever $x_1, x_2 \in A$ and $x_1 < x_2$ then $g(x_1) > g(x_2)$.

If a function is either increasing or decreasing on A, we say that it is **monotone** on A. If f is either strictly increasing or strictly decreasing on A, we say that f is **strictly monotone** on A.

We note that if $f : A \to \mathbb{R}$ is increasing on A then $g := -f$ is decreasing on A; similarly if $\varphi : A \to \mathbb{R}$ is decreasing on A then $\psi := -\varphi$ is increasing on A.

In this section, we will be concerned with monotone functions that are defined on an interval $I \subseteq \mathbb{R}$. We will discuss increasing functions explicitly, but it is clear that there are corresponding results for decreasing functions. These results can either be obtained directly from the results for increasing functions or proved by similar arguments.

Monotone functions are not necessarily continuous. For example, if $f(x) := 0$ for $x \in [0, 1]$ and $f(x) := 1$ for $x \in (1, 2]$, then f is increasing on $[0, 2]$, but fails to be continuous at $x = 1$. However, the next result shows that a monotone function always has both one-sided limits (see Definition 4.3.1) in $\mathbb{R}$ at every point that is not an endpoint of its domain.

5.6.1 Theorem *Let $I \subseteq \mathbb{R}$ be an interval and let $f : I \to \mathbb{R}$ be increasing on I. Suppose that $c \in I$ is not an endpoint of I. Then*

(i) $\lim\limits_{x \to c-} f = \sup\{f(x) : x \in I, x < c\}$,

(ii) $\lim\limits_{x \to c+} f = \inf\{f(x) : x \in I, x > c\}$.

Proof. (i) First note that if $x \in I$ and $x < c$, then $f(x) \le f(c)$. Hence the set $\{f(x) : x \in I, x < c\}$, which is nonvoid since c is not an endpoint of I, is bounded above by $f(c)$. Thus the indicated supremum exists; we denote it by L. If $\varepsilon > 0$ is given, then $L - \varepsilon$ is not

Since f is increasing, we deduce that if $\delta_\varepsilon := c - y_\varepsilon$ and if $0 < c - y < \delta_\varepsilon$, then $y_\varepsilon < y < c$ so that

$$L - \varepsilon < f(y_\varepsilon) \le f(y) \le L.$$

Therefore $|f(y) - L| < \varepsilon$ when $0 < c - y < \delta_\varepsilon$. Since $\varepsilon > 0$ is arbitrary we infer that (i) holds.

The proof of (ii) is similar. Q.E.D.

The next result gives criteria for the continuity of an increasing function f at a point c that is not an endpoint of the interval on which f is defined.

5.6.2 Corollary *Let $I \subseteq \mathbb{R}$ be an interval and let $f : I \to \mathbb{R}$ be increasing on I. Suppose that $c \in I$ is not an endpoint of I. Then the following statements are equivalent.*

(a) *f is continuous at c.*

(b) $\lim\limits_{x\to c-} f = f(c) = \lim\limits_{x\to c+} f$.

(c) $\sup\{f(x) : x \in I, x < c\} = f(c) = \inf\{f(x) : x \in I, x > c\}$.

This follows easily from Theorems 5.6.1 and 4.3.3. We leave the details to the reader.

Let I be an interval and let $f : I \to \mathbb{R}$ be an increasing function. If a is the left endpoint of I, it is an exercise to show that f is continuous at a if and only if

$$f(a) = \inf\{f(x) : x \in I, a < x\}$$

or if and only if $f(a) = \lim\limits_{x\to a+} f$. Similar conditions apply at a right endpoint, and for decreasing functions.

If $f : I \to \mathbb{R}$ is increasing on I and if c is not an endpoint of I, we define the **jump of f at c** to be $j_f(c) := \lim\limits_{x\to c+} f - \lim\limits_{x\to c-} f$. (See Figure 5.6.1.) It follows from Theorem 5.5.1 that

$$j_f(c) = \inf\{f(x) : x \in I, x > c\} - \sup\{f(x) : x \in I, x < c\}$$

for an increasing function. If the left endpoint a of I belongs to I, we define the **jump of f at a** to be $j_f(a) := \lim\limits_{x\to a+} f - f(a)$. If the right endpoint b of I belongs to I, we define the **jump of f at b** to be $j_f(b) := f(b) - \lim\limits_{x\to b-} f$.

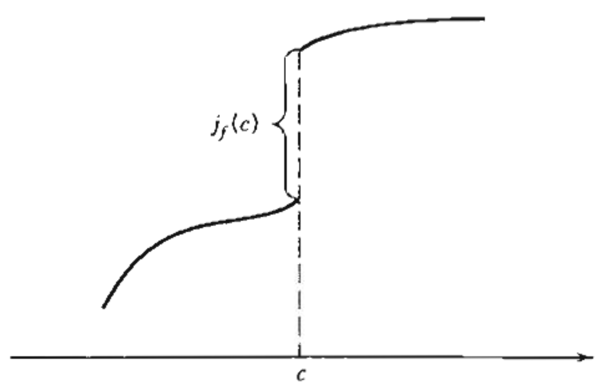


Figure 5.6.1 The jump of f at c.

5.6.3 Theorem *Let $I \subseteq \mathbb{R}$ be an interval and let $f : I \to \mathbb{R}$ be increasing on I. If $c \in I$, then f is continuous at c if and only if $j_f(c) = 0$.*

Proof. If c is not an endpoint, this follows immediately from Corollary 5.6.2. If $c \in I$ is the left endpoint of I, then f is continuous at c if and only if $f(c) = \lim_{x \to c+} f$, which is equivalent to $j_f(c) = 0$. Similar remarks apply to the case of a right endpoint. Q.E.D.

We now show that there can be at most a countable set of points at which a monotone function is discontinuous.

5.6.4 Theorem *Let $I \subseteq \mathbb{R}$ be an interval and let $f : I \to \mathbb{R}$ be monotone on I. Then the set of points $D \subseteq I$ at which f is discontinuous is a countable set.*

Proof. We shall suppose that f is increasing on I. It follows from Theorem 5.6.3 that $D = \{x \in I : j_f(x) \neq 0\}$. We shall consider the case that $I := [a, b]$ is a closed bounded interval, leaving the case of an arbitrary interval to the reader.

We first note that since f is increasing, then $j_f(c) \geq 0$ for all $c \in I$. Moreover, if $a \leq x_1 < \cdots < x_n \leq b$, then (why?) we have

$$f(a) \leq f(a) + j_f(x_1) + \cdots + j_f(x_n) \leq f(b), \tag{1}$$

whence it follows that

$$j_f(x_1) + \cdots + j_f(x_n) \leq f(b) - f(a).$$

(See Figure 5.6.2.) Consequently there can be at most k points in $I = [a, b]$ where $j_f(x) \geq (f(b) - f(a))/k$. We conclude that there is at most one point $x \in I$ where $j_f(x) = f(b) - f(a)$; there are at most two points in I where $j_f(x) \geq (f(b) - f(a))/2$; at most three points in I where $j_f(x) \geq (f(b) - f(a))/3$, and so on. Therefore there is at most a countable set of points x where $j_f(x) > 0$. But since every point in D must be included in this set, we deduce that D is a countable set. Q.E.D.

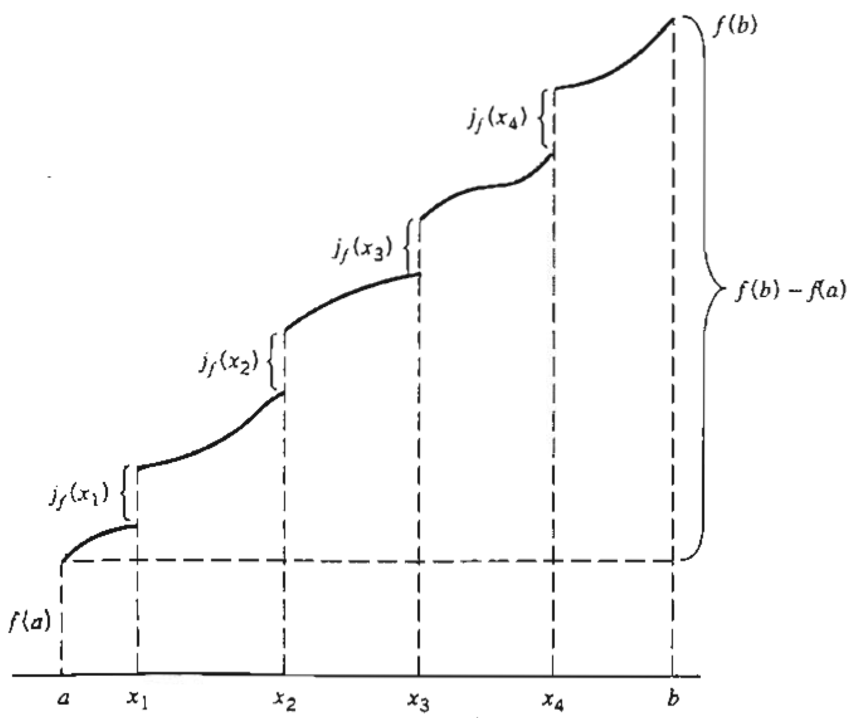


Figure 5.6.2 $j_f(x_1) + \cdots + j_f(x_n) \leq f(b) - f(a)$.

Theorem 5.6.4 has some useful applications. For example, it was seen in Exercise 5.2.12 that if $h : \mathbb{R} \to \mathbb{R}$ satisfies the identity

$$(2) \qquad h(x + y) = h(x) + h(y) \qquad \text{for all} \quad x, y \in \mathbb{R},$$

and if h is continuous at a single point x_0, then h is continuous at *every* point of $\mathbb{R}$. Thus, if h is a monotone function satisfying (2), then h must be continuous on $\mathbb{R}$. [It follows from this that $h(x) = Cx$ for all $x \in \mathbb{R}$, where $C := h(1)$.]

Inverse Functions

We shall now consider the existence of inverses for functions that are continuous on an interval $I \subseteq \mathbb{R}$. We recall (see Section 1.1) that a function $f : I \to \mathbb{R}$ has an inverse function if and only if f is injective (= one-one); that is, $x, y \in I$ and $x \neq y$ imply that $f(x) \neq f(y)$. We note that a strictly monotone function is injective and so has an inverse. In the next theorem, we show that if $f : I \to \mathbb{R}$ is a strictly monotone *continuous* function, then f has an inverse function g on $J := f(I)$ that is strictly monotone and continuous on J. In particular, if f is strictly increasing then so is g, and if f is strictly decreasing then so is g.

5.6.5 Continuous Inverse Theorem *Let $I \subseteq \mathbb{R}$ be an interval and let $f : I \to \mathbb{R}$ be strictly monotone and continuous on I. Then the function g inverse to f is strictly monotone and continuous on $J := f(I)$.*

Proof. We consider the case that f is strictly increasing, leaving the case that f is strictly decreasing to the reader.

Since f is continuous and I is an interval, it follows from the Preservation of Intervals Theorem 5.3.10 that $J := f(I)$ is an interval. Moreover, since f is strictly increasing on I, it is injective on I; therefore the function $g : J \to \mathbb{R}$ inverse to f exists. We claim that g is strictly increasing. Indeed, if $y_1, y_2 \in J$ with $y_1 < y_2$, then $y_1 = f(x_1)$ and $y_2 = f(x_2)$ for some $x_1, x_2 \in I$. We must have $x_1 < x_2$; otherwise $x_1 \geq x_2$, which implies that $y_1 = f(x_1) \geq f(x_2) = y_2$, contrary to the hypothesis that $y_1 < y_2$. Therefore we have $g(y_1) = x_1 < x_2 = g(y_2)$. Since y_1 and y_2 are arbitrary elements of J with $y_1 < y_2$, we conclude that g is strictly increasing on J.

It remains to show that g is continuous on J. However, this is a consequence of the fact that $g(J) = I$ is an interval. Indeed, if g is discontinuous at a point $c \in J$, then the jump of g at c is nonzero so that $\lim_{y\to c-} g < \lim_{y\to c+} g$. If we choose any number $x \neq g(c)$ satisfying $\lim_{x\to c-} g < x < \lim_{x\to c+} g$, then x has the property that $x \neq g(y)$ for any $y \in J$. (See Figure 5.6.3.) Hence $x \notin I$, which contradicts the fact that I is an interval. Therefore we conclude that g is continuous on J. Q.E.D.

The nth Root Function

We will apply the Continuous Inverse Theorem 5.6.5 to the nth power function. We need to distinguish two cases: (i) n even, and (ii) n odd.

(i) **n even**. In order to obtain a function that is strictly monotone, we restrict our attention to the interval $I := [0, \infty)$. Thus, let $f(x) := x^n$ for $x \in I$. (See Figure 5.6.4.) We have seen (in Exercise 2.1.23) that if $0 \leq x < y$, then $f(x) = x^n < y^n = f(y)$; therefore f is strictly increasing on I. Moreover, it follows from Example 5.2.3(a) that f is continuous on I. Therefore, by the Preservation of Intervals Theorem 5.3.10, $J := f(I)$ is an interval.

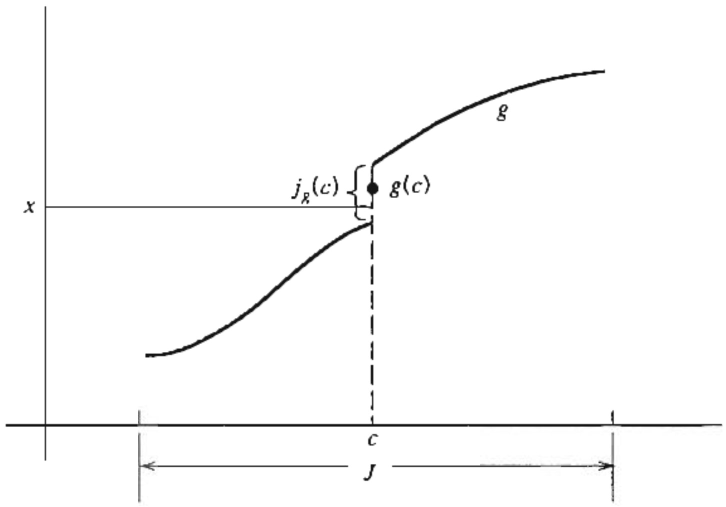


Figure 5.6.3 $g(y) \neq x$ for $y \in J$.

We will show that $J = [0, \infty)$. Let $y \geq 0$ be arbitrary; by the Archimedean Property, there exists $k \in \mathbb{N}$ such that $0 \leq y < k$. Since

$$f(0) = 0 \leq y < k \leq k^n = f(k),$$

it follows from Bolzano's Intermediate Value Theorem 5.3.7 that $y \in J$. Since $y \geq 0$ is arbitrary, we deduce that $J = [0, \infty)$.

We conclude from the Continuous Inverse Theorem 5.6.5 that the function g that is inverse to $f(x) = x^n$ on $I = [0, \infty)$ is strictly increasing and continuous on $J = [0, \infty)$. We usually write

$$g(x) = x^{1/n} \qquad \text{or} \qquad g(x) = \sqrt[n]{x}$$

for $x \geq 0$ (n even), and call $x^{1/n} = \sqrt[n]{x}$ the **nth root of** $x \geq 0$ (n even). The function g is called the **nth root function** (n even). (See Figure 5.6.5.)

Since g is inverse to f we have

$$g\big(f(x)\big) = x \qquad \text{and} \qquad f\big(g(x)\big) = x \qquad \text{for all} \quad x \in [0, \infty).$$

We can write these equations in the following form:

$$\left(x^n\right)^{1/n} = x \qquad \text{and} \qquad \left(x^{1/n}\right)^n = x$$

for all $x \in [0, \infty)$ and n even.

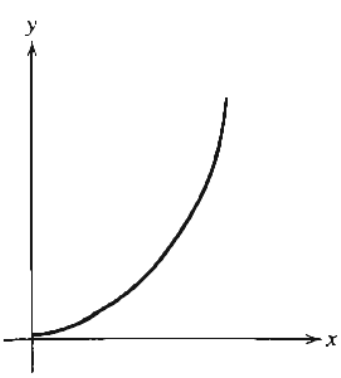


Figure 5.6.4 Graph of $f(x) = x^n$ ($x \geq 0$, n even).

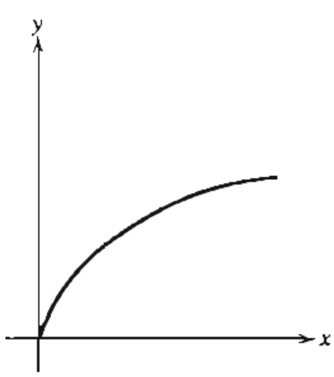


Figure 5.6.5 Graph of $g(x) = x^{1/n}$ ($x \geq 0$, n even).

(ii) n **odd**. In this case we let $F(x) := x^n$ for all $x \in \mathbb{R}$; by 5.2.3(a), F is continuous on $\mathbb{R}$. We leave it to the reader to show that F is strictly increasing on $\mathbb{R}$ and that $F(\mathbb{R}) = \mathbb{R}$. (See Figure 5.6.6.)

It follows from the Continuous Inverse Theorem 5.6.5 that the function G that is inverse to $F(x) = x^n$ for $x \in \mathbb{R}$, is strictly increasing and continuous on $\mathbb{R}$. We usually write

$$G(x) = x^{1/n} \quad \text{or} \quad G(x) = \sqrt[n]{x} \quad \text{for } x \in \mathbb{R}, n \text{ odd},$$

and call $x^{1/n}$ the **nth root of** $x \in \mathbb{R}$. The function G is called the **nth root function** (n odd). (See Figure 5.6.7.) Here we have

$$\left(x^n\right)^{1/n} = x \qquad \text{and} \qquad \left(x^{1/n}\right)^n = x$$

for all $x \in \mathbb{R}$ and n odd.

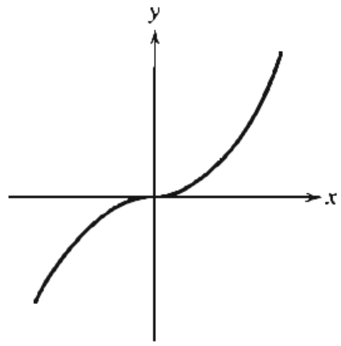


Figure 5.6.6 Graph of $F(x) = x^n$ ($x \in \mathbb{R}$, n odd).

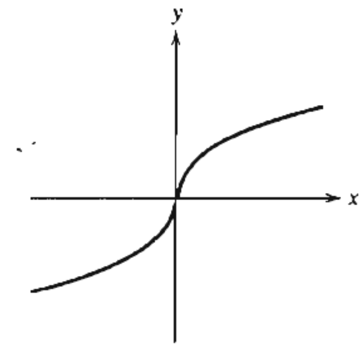


Figure 5.6.7 Graph of $G(x) = x^{1/n}$ ($x \in \mathbb{R}$, n odd).

Rational Powers

Now that the nth root functions have been defined for $n \in \mathbb{N}$, it is easy to define rational powers.

5.6.6 Definition (i) If $m, n \in \mathbb{N}$ and $x \ge 0$, we define $x^{m/n} := (x^{1/n})^m$.
(ii) If $m, n \in \mathbb{N}$ and $x > 0$, we define $x^{-m/n} := (x^{1/n})^{-m}$.

Hence we have defined x^r when r is a rational number and $x > 0$. The graphs of $x \mapsto x^r$ depend on whether $r > 1$, $r = 1$, $0 < r < 1$, $r = 0$, or $r < 0$. (See Figure 5.6.8.) Since a rational number $r \in \mathbb{Q}$ can be written in the form $r = m/n$ with $m \in \mathbb{Z}$, $n \in \mathbb{N}$, in many ways, it should be shown that Definition 5.6.6 is not ambiguous. That is if $r = m/n = p/q$ with $m, p \in \mathbb{Z}$ and $n, q \in \mathbb{N}$ and if $x > 0$, then $(x^{1/n})^m = (x^{1/q})^p$. We leave it as an exercise to the reader to establish this relation.

5.6.7 Theorem *If $m \in \mathbb{Z}$, $n \in \mathbb{N}$, and $x > 0$, then $x^{m/n} = (x^m)^{1/n}$.*

Proof. If $x > 0$ and $m, n \in \mathbb{Z}$, then $(x^m)^n = x^{mn} = (x^n)^m$. Now let $y := x^{m/n} = (x^{1/n})^m > 0$ so that $y^n = \left((x^{1/n})^m\right)^n = \left((x^{1/n})^n\right)^m = x^m$. Therefore it follows that $y = (x^m)^{1/n}$. Q.E.D.

The reader should also show, as an exercise, that if $x > 0$ and $r, s \in \mathbb{Q}$, then

$$x^r x^s = x^{r+s} = x^s x^r \quad \text{and} \quad (x^r)^s = x^{rs} = (x^s)^r.$$

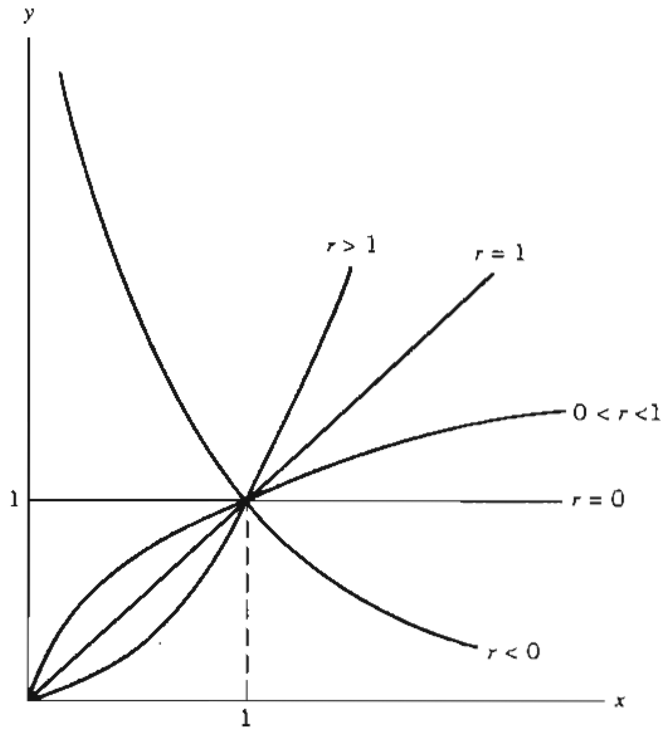


Figure 5.6.8 Graphs of $x \to x^r$ $(x \geq 0)$.

Exercises for Section 5.6

1. If $I := [a, b]$ is an interval and $f : I \to \mathbb{R}$ is an increasing function, then the point a [respectively, b] is an absolute minimum [respectively, maximum] point for f on I. If f is strictly increasing, then a is the only absolute minimum point for f on I.

2. If f and g are increasing functions on an interval $I \subseteq \mathbb{R}$, show that $f + g$ is an increasing function on I. If f is also strictly increasing on I, then $f + g$ is strictly increasing on I.

3. Show that both $f(x) := x$ and $g(x) := x - 1$ are strictly increasing on $I := [0, 1]$, but that their product fg is not increasing on I.

4. Show that if f and g are positive increasing functions on an interval I, then their product fg is increasing on I.

5. Show that if $I := [a, b]$ and $f : I \to \mathbb{R}$ is increasing on I, then f is continuous at a if and only if $f(a) = \inf\{f(x) : x \in (a, b]\}$.

6. Let $I \subseteq \mathbb{R}$ be an interval and let $f : I \to \mathbb{R}$ be increasing on I. Suppose that $c \in I$ is not an endpoint of I. Show that f is continuous at c if and only if there exists a sequence (x_n) in I such that $x_n < c$ for $n = 1, 3, 5, \cdots$; $x_n > c$ for $n = 2, 4, 6, \cdots$; and such that $c = \lim(x_n)$ and $f(c) = \lim\big(f(x_n)\big)$.

7. Let $I \subseteq \mathbb{R}$ be an interval and let $f : I \to \mathbb{R}$ be increasing on I. If c is not an endpoint of I, show that the jump $j_f(c)$ of f at c is given by $\inf\{f(y) - f(x) : x < c < y, x, y \in I\}$.

8. Let f, g be increasing on an interval $I \subseteq \mathbb{R}$ and let $f(x) > g(x)$ for all $x \in I$. If $y \in f(I) \cap g(I)$, show that $f^{-1}(y) < g^{-1}(y)$. [*Hint:* First interpret this statement geometrically.]

9. Let $I := [0, 1]$ and let $f : I \to \mathbb{R}$ be defined by $f(x) := x$ for x rational, and $f(x) := 1 - x$ for x irrational. Show that f is injective on I and that $f\,(f(x)) = x$ for all $x \in I$. (Hence f is its own inverse function!) Show that f is continuous only at the point $x = \frac{1}{2}$.

10. Let $I := [a, b]$ and let $f : I \to \mathbb{R}$ be continuous on I. If f has an absolute maximum [respectively, minimum] at an interior point c of I, show that f is not injective on I.

11. Let $f(x) := x$ for $x \in [0, 1]$, and $f(x) := 1 + x$ for $x \in (1, 2]$. Show that f and f^{-1} are strictly increasing. Are f and f^{-1} continuous at every point?

12. Let $f : [0, 1] \to \mathbb{R}$ be a continuous function that does not take on any of its values twice and with $f(0) < f(1)$. Show that f is strictly increasing on $[0, 1]$.

13. Let $h : [0, 1] \to \mathbb{R}$ be a function that takes on each of its values exactly twice. Show that h cannot be continuous at every point. [*Hint:* If $c_1 < c_2$ are the points where h attains its supremum, show that $c_1 = 0$, $c_2 = 1$. Now examine the points where h attains its infimum.]

14. Let $x \in \mathbb{R}$, $x > 0$. Show that if $m, p \in \mathbb{Z}$, $n, q \in \mathbb{N}$, and $mq = np$, then $(x^{1/n})^m = (x^{1/q})^p$.

15. If $x \in \mathbb{R}$, $x > 0$, and if $r, s \in \mathbb{Q}$, show that $x^r x^s = x^{r+s} = x^s x^r$ and $(x^r)^s = x^{rs} = (x^s)^r$.

CHAPTER 6

DIFFERENTIATION

Prior to the seventeenth century, a curve was generally described as a locus of points satisfying some geometric condition, and tangent lines were obtained through geometric construction. This viewpoint changed dramatically with the creation of analytic geometry in the 1630s by René Descartes (1596–1650) and Pierre de Fermat (1601–1665). In this new setting geometric problems were recast in terms of algebraic expressions, and new classes of curves were defined by algebraic rather than geometric conditions. The concept of derivative evolved in this new context. The problem of finding tangent lines and the seemingly unrelated problem of finding maximum or minimum values were first seen to have a connection by Fermat in the 1630s. And the relation between tangent lines to curves and the velocity of a moving particle was discovered in the late 1660s by Isaac Newton. Newton's theory of "fluxions", which was based on an intuitive idea of limit, would be familiar to any modern student of differential calculus once some changes in terminology and notation were made. But the vital observation, made by Newton and, independently, by Gottfried Leibniz in the 1680s, was that areas under curves could be calculated by reversing the differentiation process. This exciting technique, one that solved previously difficult area problems with ease, sparked enormous interest among the mathematicians of the era and led to a coherent theory that became known as the differential and integral calculus.

Isaac Newton

Isaac Newton (1642–1727) was born in Woolsthorpe, in Lincolnshire, England, on Christmas Day; his father, a farmer, had died three months earlier. His mother remarried when he was three years old and he was sent to live with his grandmother. He returned to his mother at age eleven, only to be sent to boarding school in Grantham the next year. Fortunately, a perceptive teacher noticed his mathematical talent and, in 1661, Newton entered Trinity College at Cambridge University, where he studied with Isaac Barrow.

When the bubonic plague struck in 1665–1666, leaving dead nearly 70,000 persons in London, the university closed and Newton spent two years back in Woolsthorpe. It was during this period that he formulated his basic ideas concerning optics, gravitation, and his method of "fluxions", later called "calculus". He returned to Cambridge in 1667 and was appointed Lucasian Professor in 1669. His theories of universal gravitation and planetary motion were published to world acclaim in 1687 under the title *Philosophiæ Naturalis Principia Mathematica*. However, he neglected to publish his method of inverse tangents for finding areas and other work in calculus, and this led to a controversy over priority with Leibniz.

Following an illness, he retired from Cambridge University and in 1696 was appointed Warden of the British mint. However, he maintained contact with advances in science and mathematics and served as President of the Royal Society from 1703 until his death in 1727. At his funeral, Newton was eulogized as "the greatest genius that ever existed". His place of burial in Westminster Abbey is a popular tourist site.

In this chapter we will develop the theory of differentiation. Integration theory, including the fundamental theorem that relates differentiation and integration, will be the subject of the next chapter. We will assume that the reader is already familiar with the geometrical and physical interpretations of the derivative of a function as described in introductory calculus courses. Consequently, we will concentrate on the mathematical aspects of the derivative and not go into its applications in geometry, physics, economics, and so on.

The first section is devoted to a presentation of the basic results concerning the differentiation of functions. In Section 6.2 we discuss the fundamental Mean Value Theorem and some of its applications. In Section 6.3 the important L'Hospital Rules are presented for the calculation of certain types of "indeterminate" limits.

In Section 6.4 we give a brief discussion of Taylor's Theorem and a few of its applications—for example, to convex functions and to Newton's Method for the location of roots.

Section 6.1 The Derivative

In this section we will present some of the elementary properties of the derivative. We begin with the definition of the derivative of a function.

6.1.1 Definition Let $I \subseteq \mathbb{R}$ be an interval, let $f : I \to \mathbb{R}$, and let $c \in I$. We say that a real number L is the **derivative of f at c** if given any $\varepsilon > 0$ there exists $\delta(\varepsilon) > 0$ such that if $x \in I$ satisfies $0 < |x - c| < \delta(\varepsilon)$, then

$$(1) \qquad \left| \frac{f(x) - f(c)}{x - c} - L \right| < \varepsilon.$$

In this case we say that f is **differentiable** at c, and we write $f'(c)$ for L.

In other words, the derivative of f at c is given by the limit

$$(2) \qquad f'(c) = \lim_{x \to c} \frac{f(x) - f(c)}{x - c}$$

provided this limit exists. (We allow the possibility that c may be the endpoint of the interval.)

Note It is possible to define the derivative of a function having a domain more general than an interval (since the point c need only be an element of the domain and also a cluster point of the domain) but the significance of the concept is most naturally apparent for functions defined on intervals. Consequently we shall limit our attention to such functions.

Whenever the derivative of $f : I \to \mathbb{R}$ exists at a point $c \in I$, its value is denoted by $f'(c)$. In this way we obtain a function f' whose domain is a subset of the domain of f. In working with the function f', it is convenient to regard it also as a function of x. For example, if $f(x) := x^2$ for $x \in \mathbb{R}$, then at any c in $\mathbb{R}$ we have

$$f'(c) = \lim_{x \to c} \frac{f(x) - f(c)}{x - c} = \lim_{x \to c} \frac{x^2 - c^2}{x - c} = \lim_{x \to c} (x + c) = 2c.$$

Thus, in this case, the function f' is defined on all of $\mathbb{R}$ and $f'(x) = 2x$ for $x \in \mathbb{R}$.

We now show that continuity of f at a point c is a necessary (but not sufficient) condition for the existence of the derivative at c.

6.1.2 Theorem *If $f: I \to \mathbb{R}$ has a derivative at $c \in I$, then f is continuous at c.*

Proof. For all $x \in I$, $x \neq c$, we have

$$f(x) - f(c) = \left(\frac{f(x) - f(c)}{x - c}\right)(x - c).$$

Since $f'(c)$ exists, we may apply Theorem 4.2.4 concerning the limit of a product to conclude that

$$\begin{aligned}\lim_{x \to c}(f(x) - f(c)) &= \lim_{x \to c}\left(\frac{f(x) - f(c)}{x - c}\right)\left(\lim_{x \to c}(x - c)\right) \\ &= f'(c) \cdot 0 = 0.\end{aligned}$$

Therefore, $\lim_{x \to c} f(x) = f(c)$ so that f is continuous at c. Q.E.D.

The continuity of $f: I \to \mathbb{R}$ at a point does not assure the existence of the derivative at that point. For example, if $f(x) := |x|$ for $x \in \mathbb{R}$, then for $x \neq 0$ we have $(f(x) - f(0))/(x - 0) = |x|/x$ which is equal to 1 if $x > 0$, and equal to -1 if $x < 0$. Thus the limit at 0 does not exist [see Example 4.1.10(b)], and therefore the function is not differentiable at 0. Hence, continuity at a point c is *not* a sufficient condition for the derivative to exist at c.

Remark By taking simple algebraic combinations of functions of the form $x \mapsto |x - c|$, it is not difficult to construct continuous functions that do not have a derivative at a finite (or even a countable) number of points. In 1872, Karl Weierstrass astounded the mathematical world by giving an example of a function that is *continuous at every point but whose derivative does not exist anywhere*. Such a function defied geometric intuition about curves and tangent lines, and consequently spurred much deeper investigations into the concepts of real analysis. It can be shown that the function f defined by the series

$$f(x) := \sum_{n=0}^{\infty} \frac{1}{2^n} \cos(3^n x)$$

has the stated property. A very interesting historical discussion of this and other examples of continuous, nondifferentiable functions is given in Kline, p. 955–966, and also in Hawkins, p. 44–46. A detailed proof for a slightly different example can be found in Appendix E.

There are a number of basic properties of the derivative that are very useful in the calculation of the derivatives of various combinations of functions. We now provide the justification of some of these properties, which will be familiar to the reader from earlier courses.

6.1.3 Theorem *Let $I \subseteq \mathbb{R}$ be an interval, let $c \in I$, and let $f: I \to \mathbb{R}$ and $g: I \to \mathbb{R}$ be functions that are differentiable at c. Then:*

(a) *If $\alpha \in \mathbb{R}$, then the function αf is differentiable at c, and*

(3) $$(\alpha f)'(c) = \alpha f'(c).$$

(b) *The function $f + g$ is differentiable at c, and*

(4) $$(f+g)'(c) = f'(c) + g'(c).$$

(c) *(Product Rule) The function fg is differentiable at c, and*

(5) $$(fg)'(c) = f'(c)g(c) + f(c)g'(c).$$

(d) *(Quotient Rule) If $g(c) \neq 0$, then the function f/g is differentiable at c, and*

(6) $$\left(\frac{f}{g}\right)'(c) = \frac{f'(c)g(c) - f(c)g'(c)}{\left(g(c)\right)^2}.$$

Proof. We shall prove (c) and (d), leaving (a) and (b) as exercises for the reader.

(c) Let $p := fg$; then for $x \in I$, $x \neq c$, we have

$$\begin{aligned}\frac{p(x) - p(c)}{x - c} &= \frac{f(x)g(x) - f(c)g(c)}{x - c} \\ &= \frac{f(x)g(x) - f(c)g(x) + f(c)g(x) - f(c)g(c)}{x - c} \\ &= \frac{f(x) - f(c)}{x - c} \cdot g(x) + f(c) \cdot \frac{g(x) - g(c)}{x - c}.\end{aligned}$$

Since g is continuous at c, by Theorem 6.1.2, then $\lim_{x \to c} g(x) = g(c)$. Since f and g are differentiable at c, we deduce from Theorem 4.2.4 on properties of limits that

$$\lim_{x \to c} \frac{p(x) - p(c)}{x - c} = f'(c)g(c) + f(c)g'(c).$$

Hence $p := fg$ is differentiable at c and (5) holds.

(d) Let $q := f/g$. Since g is differentiable at c, it is continuous at that point (by Theorem 6.1.2). Therefore, since $g(c) \neq 0$, we know from Theorem 4.2.9 that there exists an interval $J \subseteq I$ with $c \in J$ such that $g(x) \neq 0$ for all $x \in J$. For $x \in J$, $x \neq c$, we have

$$\begin{aligned}\frac{q(x) - q(c)}{x - c} = \frac{f(x)/g(x) - f(c)/g(c)}{x - c} &= \frac{f(x)g(c) - f(c)g(x)}{g(x)g(c)(x - c)} \\ &= \frac{f(x)g(c) - f(c)g(c) + f(c)g(c) - f(c)g(x)}{g(x)g(c)(x - c)} \\ &= \frac{1}{g(x)g(c)}\left[\frac{f(x) - f(c)}{x - c} \cdot g(c) - f(c) \cdot \frac{g(x) - g(c)}{x - c}\right].\end{aligned}$$

Using the continuity of g at c and the differentiability of f and g at c, we get

$$q'(c) = \lim_{x \to c} \frac{q(x) - q(c)}{x - c} = \frac{f'(c)g(c) - f(c)g'(c)}{\left(g(c)\right)^2}.$$

Thus, $q = f/g$ is differentiable at c and equation (6) holds. Q.E.D.

Mathematical Induction may be used to obtain the following extensions of the differentiation rules.

6.1.4 Corollary *If $f_1, f_2, \cdots, f_n$ are functions on an interval I to $\mathbb{R}$ that are differentiable at $c \in I$, then:*

(a) *The function $f_1 + f_2 + \cdots + f_n$ is differentiable at c and*

(7) $$(f_1 + f_2 + \cdots + f_n)'(c) = f_1'(c) + f_2'(c) + \cdots + f_n'(c).$$

(b) *The function $f_1 f_2 \cdots f_n$ is differentiable at c, and*

$$(f_1 f_2 \cdots f_n)'(c) = f_1'(c) f_2(c) \cdots f_n(c) + f_1(c) f_2'(c) \cdots f_n(c) + \cdots + f_1(c) f_2(c) \cdots f_n'(c). \tag{8}$$

An important special case of the extended product rule (8) occurs if the functions are equal, that is, $f_1 = f_2 = \cdots = f_n = f$. Then (8) becomes

$$(f^n)'(c) = n(f(c))^{n-1} f'(c). \tag{9}$$

In particular, if we take $f(x) := x$, then we find the derivative of $g(x) := x^n$ to be $g'(x) = nx^{n-1}$, $n \in \mathbb{N}$. The formula is extended to include negative integers by applying the Quotient Rule 6.1.3(d).

Notation If $I \subseteq \mathbb{R}$ is an interval and $f : I \to \mathbb{R}$, we have introduced the notation f' to denote the function whose domain is a subset of I and whose value at a point c is the derivative $f'(c)$ of f at c. There are other notations that are sometimes used for f'; for example, one sometimes writes Df for f'. Thus one can write formulas (4) and (5) in the form:

$$D(f + g) = Df + Dg, \qquad D(fg) = (Df) \cdot g + f \cdot (Dg).$$

When x is the "independent variable", it is common practice in elementary courses to write df/dx for f'. Thus formula (5) is sometimes written in the form

$$\frac{d}{dx}\big(f(x)g(x)\big) = \left(\frac{df}{dx}(x)\right) g(x) + f(x) \left(\frac{dg}{dx}(x)\right).$$

This last notation, due to Leibniz, has certain advantages. However, it also has certain disadvantages and must be used with some care.

The Chain Rule

We now turn to the theorem on the differentiation of composite functions known as the "Chain Rule". It provides a formula for finding the derivative of a composite function $g \circ f$ in terms of the derivatives of g and f.

We first establish the following theorem concerning the derivative of a function at a point that gives us a very nice method for proving the Chain Rule. It will also be used to derive the formula for differentiating inverse functions.

6.1.5 Carathéodory's Theorem *Let f be defined on an interval I containing the point c. Then f is differentiable at c if and only if there exists a function φ on I that is continuous at c and satisfies*

$$f(x) - f(c) = \varphi(x)(x - c) \qquad \textit{for} \quad x \in I. \tag{10}$$

In this case, we have $\varphi(c) = f'(c)$.

Proof. ($\Rightarrow$) If $f'(c)$ exists, we can define φ by

$$\varphi(x) := \begin{cases} \dfrac{f(x) - f(c)}{x - c} & \text{for} \quad x \neq c, x \in I, \\ f'(c) & \text{for} \quad x = c. \end{cases}$$

The continuity of φ follows from the fact that $\lim\limits_{x \to c} \varphi(x) = f'(c)$. If $x = c$, then both sides of (10) equal 0, while if $x \neq c$, then multiplication of $\varphi(x)$ by $x - c$ gives (10) for all other $x \in I$.

($\Leftarrow$) Now assume that a function φ that is continuous at c and satisfying (10) exists. If we divide (10) by $x - c \neq 0$, then the continuity of φ implies that

$$\varphi(c) = \lim_{x \to c} \varphi(x) = \lim_{x \to c} \frac{f(x) - f(c)}{x - c}$$

exists. Therefore f is differentiable at c and $f'(c) = \varphi(c)$. Q.E.D.

To illustrate Carathéodory's Theorem, we consider the function f defined by $f(x) := x^3$ for $x \in \mathbb{R}$. For $c \in \mathbb{R}$, we see from the factorization

$$x^3 - c^3 = (x^2 + cx + c^2)(x - c)$$

that $\varphi(x) := x^2 + cx + c^2$ satisfies the conditions of the theorem. Therefore, we conclude that f is differentiable at $c \in \mathbb{R}$ and that $f'(c) = \varphi(c) = 3c^2$.

We will now establish the Chain Rule. If f is differentiable at c and g is differentiable at $f(c)$, then the Chain Rule states that the derivative of the composite function $g \circ f$ at c is the product $(g \circ f)'(c) = g'(f(c)) \cdot f'(c)$. Note this can be written

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

One approach to the Chain Rule is the observation that the difference quotient can be written, when $f(x) \neq f(c)$, as the product

$$\frac{g\big(f(x)\big) - g\big(f(c)\big)}{x - c} = \frac{g\big(f(x)\big) - g\big(f(c)\big)}{f(x) - f(c)} \cdot \frac{f(x) - f(c)}{x - c}.$$

This suggests the correct limiting value. Unfortunately, the first factor in the product on the right is undefined if the denominator $f(x) - f(c)$ equals 0 for values of x near c, and this presents a problem. However, the use of Carathéodory's Theorem neatly avoids this difficulty.

6.1.6 Chain Rule *Let I, J be intervals in $\mathbb{R}$, let $g : I \to \mathbb{R}$ and $f : J \to \mathbb{R}$ be functions such that $f(J) \subseteq I$, and let $c \in J$. If f is differentiable at c and if g is differentiable at $f(c)$, then the composite function $g \circ f$ is differentiable at c and*

$$(g \circ f)'(c) = g'(f(c)) \cdot f'(c). \tag{11}$$

Proof. Since $f'(c)$ exists, Carathéodory's Theorem 6.1.5 implies that there exists a function φ on J such that φ is continuous at c and $f(x) - f(c) = \varphi(x)(x - c)$ for $x \in J$, and where $\varphi(c) = f'(c)$. Also, since $g'\big(f(c)\big)$ exists, there is a function ψ defined on I such that ψ is continuous at $d := f(c)$ and $g(y) - g(d) = \psi(y)(y - d)$ for $y \in I$, where $\psi(d) = g'(d)$. Substitution of $y = f(x)$ and $d = f(c)$ then produces

$$g\big(f(x)\big) - g\big(f(c)\big) = \psi\big(f(x)\big)\big(f(x) - f(c)\big) = \big[\big(\psi \circ f(x)\big) \cdot \varphi(x)\big](x - c)$$

for all $x \in J$ such that $f(x) \in I$. Since the function $(\psi \circ f) \cdot \varphi$ is continuous at c and its value at c is $g'(f(c)) \cdot f'(c)$, Carathéodory's Theorem gives (11). Q.E.D.

If g is differentiable on I, if f is differentiable on J and if $f(J) \subseteq I$, then it follows from the Chain Rule that $(g \circ f)' = (g' \circ f) \cdot f'$ which can also be written in the form $D(g \circ f) = (Dg \circ f) \cdot Df$.

6.1.7 Examples **(a)** If $f : I \to \mathbb{R}$ is differentiable on I and $g(y) := y^n$ for $y \in \mathbb{R}$ and $n \in \mathbb{N}$, then since $g'(y) = ny^{n-1}$, it follows from the Chain Rule 6.1.6 that

$$(g \circ f)'(x) = g'\big(f(x)\big) \cdot f'(x) \qquad \text{for} \quad x \in I.$$

Therefore we have $(f^n)'(x) = n\big(f(x)\big)^{n-1} f'(x)$ for all $x \in I$ as was seen in (9).

(b) Suppose that $f : I \to \mathbb{R}$ is differentiable on I and that $f(x) \neq 0$ and $f'(x) \neq 0$ for $x \in I$. If $h(y) := 1/y$ for $y \neq 0$, then it is an exercise to show that $h'(y) = -1/y^2$ for $y \in \mathbb{R}$, $y \neq 0$. Therefore we have

$$\left(\frac{1}{f}\right)'(x) = (h \circ f)'(x) = h'\big(f(x)\big)f'(x) = -\frac{f'(x)}{\big(f(x)\big)^2} \qquad \text{for} \quad x \in I.$$

(c) The absolute value function $g(x) := |x|$ is differentiable at all $x \neq 0$ and has derivative $g'(x) = \operatorname{sgn}(x)$ for $x \neq 0$. (The signum function is defined in Example 4.1.10(b).) Though sgn is defined everywhere, it is not equal to g' at $x = 0$ since $g'(0)$ does not exist.

Now if f is a differentiable function, then the Chain Rule implies that the function $g \circ f = |f|$ is also differentiable at all points x where $f(x) \neq 0$, and its derivative is given by

$$|f|'(x) = \operatorname{sgn}(f(x)) \cdot f'(x) = \begin{cases} f'(x) & \text{if} \quad f(x) > 0, \\ -f'(x) & \text{if} \quad f(x) < 0. \end{cases}$$

If f is differentiable at a point c with $f(c) = 0$, then it is an exercise to show that $|f|$ is differentiable at c if and only if $f'(c) = 0$. (See Exercise 7.)

For example, if $f(x) := x^2 - 1$ for $x \in \mathbb{R}$, then the derivative of its absolute value $|f|(x) = |x^2 - 1|$ is equal to $|f|'(x) = \operatorname{sgn}(x^2 - 1) \cdot (2x)$ for $x \neq 1, -1$. See Figure 6.1.1 for a graph of $|f|$.

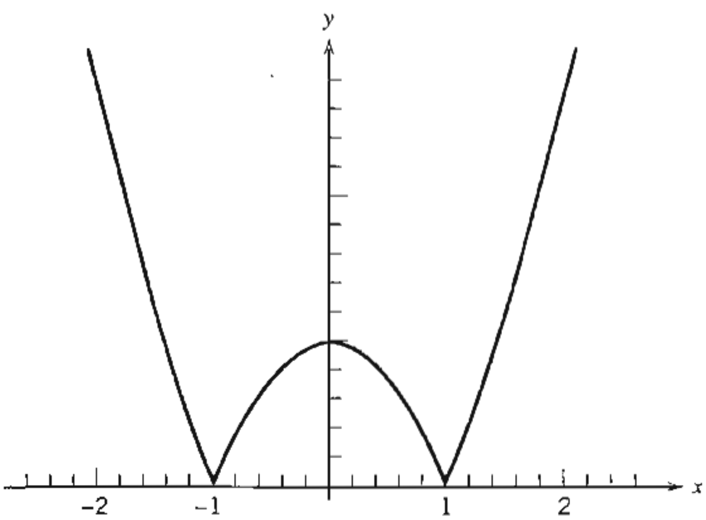


Figure 6.1.1 The function $|f|(x) = |x^2 - 1|$.

(d) It will be proved later that if $S(x) := \sin x$ and $C(x) := \cos x$ for all $x \in \mathbb{R}$, then

$$S'(x) = \cos x = C(x) \qquad \text{and} \qquad C'(x) = -\sin x = -S(x)$$

for all $x \in \mathbb{R}$. If we use these facts together with the definitions

$$\tan x := \frac{\sin x}{\cos x}, \qquad \sec x := \frac{1}{\cos x},$$

for $x \neq (2k+1)\pi/2$, $k \in \mathbb{Z}$, and apply the Quotient Rule 6.1.3(d), we obtain

$$D\tan x = \frac{(\cos x)(\cos x) - (\sin x)(-\sin x)}{(\cos x)^2} = (\sec x)^2,$$

$$D\sec x = \frac{0 - 1(-\sin x)}{(\cos x)^2} = \frac{\sin x}{(\cos x)^2} = (\sec x)(\tan x)$$

for $x \neq (2k+1)\pi/2$, $k \in \mathbb{Z}$.

Similarly, since

$$\cot x := \frac{\cos x}{\sin x}, \qquad \csc x := \frac{1}{\sin x}$$

for $x \neq k\pi$, $k \in \mathbb{Z}$, then we obtain

$$D\cot x = -(\csc x)^2 \qquad \text{and} \qquad D\csc x = -(\csc x)(\cot x)$$

for $x \neq k\pi$, $k \in \mathbb{Z}$.

(e) Suppose that f is defined by

$$f(x) := \begin{cases} x^2 \sin(1/x) & \text{for} \quad x \neq 0, \\ 0 & \text{for} \quad x = 0. \end{cases}$$

If we use the fact that $D\sin x = \cos x$ for all $x \in \mathbb{R}$ and apply the Product Rule 6.1.3(c) and the Chain Rule 6.1.6, we obtain (why?)

$$f'(x) = 2x\sin(1/x) - \cos(1/x) \qquad \text{for} \quad x \neq 0.$$

If $x = 0$, none of the calculational rules may be applied. (Why?) Consequently, the derivative of f at $x = 0$ must be found by applying the definition of derivative. We find that

$$f'(0) = \lim_{x\to 0} \frac{f(x) - f(0)}{x - 0} = \lim_{x\to 0} \frac{x^2 \sin(1/x)}{x} = \lim_{x\to 0} x\sin(1/x) = 0.$$

Hence, the derivative f' of f exists at all $x \in \mathbb{R}$. However, the function f' does not have a limit at $x = 0$ (why?), and consequently f' is discontinuous at $x = 0$. Thus, a function f that is differentiable at every point of $\mathbb{R}$ need not have a continuous derivative f'. □

Inverse Functions

We will now relate the derivative of a function to the derivative of its inverse function, when this inverse function exists. We will limit our attention to a continuous strictly monotone function and use the Continuous Inverse Theorem 5.6.5 to ensure the existence of a continuous inverse function.

If f is a continuous strictly monotone function on an interval I, then its inverse function $g = f^{-1}$ is defined on the interval $J := f(I)$ and satisfies the relation

$$g\big(f(x)\big) = x \qquad \text{for} \quad x \in I.$$

If $c \in I$ and $d := f(c)$, and if we knew that both $f'(c)$ and $g'(d)$ exist, then we could differentiate both sides of the equation and apply the Chain Rule to the left side to get $g'\big(f(c)\big) \cdot f'(c) = 1$. Thus, if $f'(c) \neq 0$, we would obtain

$$g'(d) = \frac{1}{f'(c)}.$$

However, it is necessary to deduce the differentiability of the inverse function g from the assumed differentiability of f before such a calculation can be performed. This is nicely accomplished by using Carathéodory's Theorem.

6.1.8 Theorem *Let I be an interval in $\mathbb{R}$ and let $f : I \to \mathbb{R}$ be strictly monotone and continuous on I. Let $J := f(I)$ and let $g : J \to \mathbb{R}$ be the strictly monotone and continuous function inverse to f. If f is differentiable at $c \in I$ and $f'(c) \neq 0$, then g is differentiable at $d := f(c)$ and*

$$g'(d) = \frac{1}{f'(c)} = \frac{1}{f'(g(d))}. \tag{12}$$

Proof. Given $c \in \mathbb{R}$, we obtain from Carathéodory's Theorem 6.1.5 a function φ on I with properties that φ is continuous at c, $f(x) - f(c) = \varphi(x)(x - c)$ for $x \in I$, and $\varphi(c) = f'(c)$. Since $\varphi(c) \neq 0$ by hypothesis, there exists a neighborhood $V := (c - \delta, c + \delta)$ such that $\varphi(x) \neq 0$ for all $x \in V \cap I$. (See Theorem 4.2.9.) If $U := f(V \cap I)$, then the inverse function g satisfies $f\big(g(y)\big) = y$ for all $y \in U$, so that

$$y - d = f\big(g(y)\big) - f(c) = \varphi\big(g(y)\big) \cdot \big(g(y) - g(d)\big).$$

Since $\varphi(g(y)) \neq 0$ for $y \in U$, we can divide to get

$$g(y) - g(d) = \frac{1}{\varphi\big(g(y)\big)} \cdot (y - d).$$

Since the function $1/(\varphi \circ g)$ is continuous at d, we apply Theorem 6.1.5 to conclude that $g'(d)$ exists and $g'(d) = 1/\varphi\big(g(d)\big) = 1/\varphi(c) = 1/f'(c)$. Q.E.D.

Note The hypothesis, made in Theorem 6.1.8, that $f'(c) \neq 0$ is essential. In fact, if $f'(c) = 0$, then the inverse function g is *never* differentiable at $d = f(c)$, since the assumed existence of $g'(d)$ would lead to $1 = f'(c)g'(d) = 0$, which is impossible. The function $f(x) := x^3$ with $c = 0$ is such an example.

6.1.9 Theorem *Let I be an interval and let $f : I \to \mathbb{R}$ be strictly monotone on I. Let $J := f(I)$ and let $g : J \to \mathbb{R}$ be the function inverse to f. If f is differentiable on I and $f'(x) \neq 0$ for $x \in I$, then g is differentiable on J and*

$$g' = \frac{1}{f' \circ g}. \tag{13}$$

Proof. If f is differentiable on I, then Theorem 6.1.2 implies that f is continuous on I, and by the Continuous Inverse Theorem 5.6.5, the inverse function g is continuous on J. Equation (13) now follows from Theorem 6.1.8. Q.E.D.

Remark If f and g are the functions of Theorem 6.1.9, and if $x \in I$ and $y \in J$ are related by $y = f(x)$ and $x = g(y)$, then equation (13) can be written in the form

$$g'(y) = \frac{1}{(f' \circ g)(y)}, \quad y \in J, \qquad \text{or} \qquad (g' \circ f)(x) = \frac{1}{f'(x)}, \quad x \in I.$$

It can also be written in the form $g'(y) = 1/f'(x)$, *provided* that it is kept in mind that x and y are related by $y = f(x)$ and $x = g(y)$.

6.1.10 Examples **(a)** The function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) := x^5 + 4x + 3$ is continuous and strictly monotone increasing (since it is the sum of two strictly increasing functions). Moreover, $f'(x) = 5x^4 + 4$ is never zero. Therefore, by Theorem 6.1.8, the inverse function $g = f^{-1}$ is differentiable at every point. If we take $c = 1$, then since $f(1) = 8$, we obtain $g'(8) = g'(f(1)) = 1/f'(1) = 1/9$.

(b) Let $n \in \mathbb{N}$ be even, let $I := [0, \infty)$, and let $f(x) := x^n$ for $x \in I$. It was seen at the end of Section 5.6 that f is strictly increasing and continuous on I, so that its inverse function $g(y) := y^{1/n}$ for $y \in J := [0, \infty)$ is also strictly increasing and continuous on J. Moreover, we have $f'(x) = nx^{n-1}$ for all $x \in I$. Hence it follows that if $y > 0$, then $g'(y)$ exists and

$$g'(y) = \frac{1}{f'\big(g(y)\big)} = \frac{1}{n\big(g(y)\big)^{n-1}} = \frac{1}{ny^{(n-1)/n}}.$$

Hence we deduce that

$$g'(y) = \frac{1}{n} y^{(1/n)-1} \qquad \text{for} \quad y > 0.$$

However, g is *not* differentiable at 0. (For a graph of f and g, see Figures 5.6.4 and 5.6.5.)

(c) Let $n \in \mathbb{N}, n \neq 1$, be odd, let $F(x) := x^n$ for $x \in \mathbb{R}$, and let $G(y) := y^{1/n}$ be its inverse function defined for all $y \in \mathbb{R}$. As in part (b) we find that G is differentiable for $y \neq 0$ and that $G'(y) = (1/n)y^{(1/n)-1}$ for $y \neq 0$. However, G is not differentiable at 0, even though G is differentiable for all $y \neq 0$. (For a graph of F and G, see Figures 5.6.6 and 5.6.7.)

(d) Let $r := m/n$ be a positive rational number, let $I := [0, \infty)$, and let $R(x) := x^r$ for $x \in I$. (Recall Definition 5.6.6.) Then R is the composition of the functions $f(x) := x^m$ and $g(x) := x^{1/n}$, $x \in I$. That is, $R(x) = f(g(x))$ for $x \in I$. If we apply the Chain Rule 6.1.6 and the results of (b) [or (c), depending on whether n is even or odd], then we obtain

$$\begin{aligned} R'(x) &= f'\big(g(x)\big)g'(x) = m(x^{1/n})^{m-1} \cdot \frac{1}{n} x^{(1/n)-1} \\ &= \frac{m}{n} x^{(m/n)-1} = rx^{r-1} \end{aligned}$$

for all $x > 0$. If $r > 1$, then it is an exercise to show that the derivative also exists at $x = 0$ and $R'(0) = 0$. (For a graph of R see Figure 5.6.8.)

(e) The sine function is strictly increasing on the interval $I := [-\pi/2, \pi/2]$; therefore its inverse function, which we will denote by Arcsin, exists on $J := [-1, 1]$. That is, if $x \in [-\pi/2, \pi/2]$ and $y \in [-1, 1]$ then $y = \sin x$ if and only if $\operatorname{Arcsin} y = x$. It was asserted (without proof) in Example 6.1.7(d) that sin is differentiable on I and that $D \sin x = \cos x$ for $x \in I$. Since $\cos x \neq 0$ for x in $(-\pi/2, \pi/2)$ it follows from Theorem 6.1.8 that

$$\begin{aligned} D \operatorname{Arcsin} y &= \frac{1}{D \sin x} = \frac{1}{\cos x} \\ &= \frac{1}{\sqrt{1 - (\sin x)^2}} = \frac{1}{\sqrt{1 - y^2}} \end{aligned}$$

for all $y \in (-1, 1)$. The derivative of Arcsin does *not* exist at the points -1 and 1. □

Exercises for Section 6.1

1. Use the definition to find the derivative of each of the following functions:
   (a) $f(x) := x^3$ for $x \in \mathbb{R}$,
   (b) $g(x) := 1/x$ for $x \in \mathbb{R}, x \neq 0$,
   (c) $h(x) := \sqrt{x}$ for $x > 0$,
   (d) $k(x) := 1/\sqrt{x}$ for $x > 0$.

2. Show that $f(x) := x^{1/3}$, $x \in \mathbb{R}$, is not differentiable at $x = 0$.
3. Prove Theorem 6.1.3(a), (b).
4. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) := x^2$ for x rational, $f(x) := 0$ for x irrational. Show that f is differentiable at $x = 0$, and find $f'(0)$.
5. Differentiate and simplify:
   (a) $f(x) := \dfrac{x}{1+x^2}$, (b) $g(x) := \sqrt{5 - 2x + x^2}$,
   (c) $h(x) := (\sin x^k)^m$ for $m, k \in \mathbb{N}$, (d) $k(x) := \tan(x^2)$ for $|x| < \sqrt{\pi/2}$.
6. Let $n \in \mathbb{N}$ and let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) := x^n$ for $x \geq 0$ and $f(x) := 0$ for $x < 0$. For which values of n is f' continuous at 0? For which values of n is f' differentiable at 0?
7. Suppose that $f : \mathbb{R} \to \mathbb{R}$ is differentiable at c and that $f(c) = 0$. Show that $g(x) := |f(x)|$ is differentiable at c if and only if $f'(c) = 0$.
8. Determine where each of the following functions from $\mathbb{R}$ to $\mathbb{R}$ is differentiable and find the derivative:
   (a) $f(x) := |x| + |x + 1|$, (b) $g(x) := 2x + |x|$,
   (c) $h(x) := x|x|$, (d) $k(x) := |\sin x|$.
9. Prove that if $f : \mathbb{R} \to \mathbb{R}$ is an **even function** [that is, $f(-x) = f(x)$ for all $x \in \mathbb{R}$] and has a derivative at every point, then the derivative f' is an **odd function** [that is, $f'(-x) = -f'(x)$ for all $x \in \mathbb{R}$]. Also prove that if $g : \mathbb{R} \to \mathbb{R}$ is a differentiable odd function, then g' is an even function.
10. Let $g : \mathbb{R} \to \mathbb{R}$ be defined by $g(x) := x^2 \sin(1/x^2)$ for $x \neq 0$, and $g(0) := 0$. Show that g is differentiable for all $x \in \mathbb{R}$. Also show that the derivative g' is not bounded on the interval $[-1, 1]$.
11. Assume that there exists a function $L : (0, \infty) \to \mathbb{R}$ such that $L'(x) = 1/x$ for $x > 0$. Calculate the derivatives of the following functions:
    (a) $f(x) := L(2x + 3)$ for $x > 0$, (b) $g(x) := (L(x^2))^3$ for $x > 0$,
    (c) $h(x) := L(ax)$ for $a > 0, x > 0$, (d) $k(x) := L(L(x))$ when $L(x) > 0, x > 0$.
12. If $r > 0$ is a rational number, let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) := x^r \sin(1/x)$ for $x \neq 0$, and $f(0) := 0$. Determine those values of r for which $f'(0)$ exists.
13. If $f : \mathbb{R} \to \mathbb{R}$ is differentiable at $c \in \mathbb{R}$, show that
    $$f'(c) = \lim \left(n\{f(c + 1/n) - f(c)\}\right).$$
    However, show by example that the existence of the limit of this sequence does not imply the existence of $f'(c)$.
14. Given that the function $h(x) := x^3 + 2x + 1$ for $x \in \mathbb{R}$ has an inverse h^{-1} on $\mathbb{R}$, find the value of $(h^{-1})'(y)$ at the points corresponding to $x = 0, 1, -1$.
15. Given that the restriction of the cosine function cos to $I := [0, \pi]$ is strictly decreasing and that $\cos 0 = 1$, $\cos \pi = -1$, let $J := [-1, 1]$, and let Arccos: $J \to \mathbb{R}$ be the function inverse to the restriction of cos to I. Show that Arccos is differentiable on $(-1, 1)$ and $D\text{Arccos}\, y = (-1)/(1 - y^2)^{1/2}$ for $y \in (-1, 1)$. Show that Arccos is not differentiable at -1 and 1.
16. Given that the restriction of the tangent function tan to $I := (-\pi/2, \pi/2)$ is strictly increasing and that $\tan(I) = \mathbb{R}$, let Arctan: $\mathbb{R} \to \mathbb{R}$ be the function inverse to the restriction of tan to I. Show that Arctan is differentiable on $\mathbb{R}$ and that $D\text{Arctan}(y) = (1 + y^2)^{-1}$ for $y \in \mathbb{R}$.
17. Let $f : I \to \mathbb{R}$ be differentiable at $c \in I$. Establish the **Straddle Lemma:** Given $\varepsilon > 0$ there exists $\delta(\varepsilon) > 0$ such that if $u, v \in I$ satisfy $c - \delta(\varepsilon) < u \leq c \leq v < c + \delta(\varepsilon)$, then we have $|f(v) - f(u) - (v - u)f'(c)| \leq \varepsilon(v - u)$. [*Hint:* The $\delta(\varepsilon)$ is given by Definition 6.1.1. Subtract and add the term $f(c) - cf'(c)$ on the left side and use the Triangle Inequality.]

Section 6.2 The Mean Value Theorem

The Mean Value Theorem, which relates the values of a function to values of its derivative, is one of the most useful results in real analysis. In this section we will establish this important theorem and sample some of its many consequences.

We begin by looking at the relationship between the relative extrema of a function and the values of its derivative. Recall that the function $f : I \to \mathbb{R}$ is said to have a **relative maximum** [respectively, **relative minimum**] at $c \in I$ if there exists a neighborhood $V := V_\delta(c)$ of c such that $f(x) \le f(c)$ [respectively, $f(c) \le f(x)$] for all x in $V \cap I$. We say that f has a **relative extremum** at $c \in I$ if it has either a relative maximum or a relative minimum at c.

The next result provides the theoretical justification for the familiar process of finding points at which f has relative extrema by examining the zeros of the derivative. However, it must be realized that this procedure applies only to *interior* points of the interval. For example, if $f(x) := x$ on the interval $I := [0, 1]$, then the endpoint $x = 0$ yields the unique relative minimum and the endpoint $x = 1$ yields the unique maximum of f on I, but neither point is a zero of the derivative of f.

6.2.1 Interior Extremum Theorem *Let c be an interior point of the interval I at which $f: I \to \mathbb{R}$ has a relative extremum. If the derivative of f at c exists, then $f'(c) = 0$.*

Proof. We will prove the result only for the case that f has a relative maximum at c; the proof for the case of a relative minimum is similar.

If $f'(c) > 0$, then by Theorem 4.2.9 there exists a neighborhood $V \subseteq I$ of c such that

$$\frac{f(x) - f(c)}{x - c} > 0 \qquad \text{for} \quad x \in V,\ x \ne c.$$

If $x \in V$ and $x > c$, then we have

$$f(x) - f(c) = (x - c) \cdot \frac{f(x) - f(c)}{x - c} > 0.$$

But this contradicts the hypothesis that f has a relative maximum at c. Thus we cannot have $f'(c) > 0$. Similarly (how?), we cannot have $f'(c) < 0$. Therefore we must have $f'(c) = 0$. Q.E.D.

6.2.2 Corollary *Let $f: I \to \mathbb{R}$ be continuous on an interval I and suppose that f has a relative extremum at an interior point c of I. Then either the derivative of f at c does not exist, or it is equal to zero.*

We note that if $f(x) := |x|$ on $I := [-1, 1]$, then f has an interior minimum at $x = 0$; however, the derivative of f fails to exist at $x = 0$.

6.2.3 Rolle's Theorem *Suppose that f is continuous on a closed interval $I := [a, b]$, that the derivative f' exists at every point of the open interval (a, b), and that $f(a) = f(b) = 0$. Then there exists at least one point c in (a, b) such that $f'(c) = 0$.*

Proof. If f vanishes identically on I, then any c in (a, b) will satisfy the conclusion of the theorem. Hence we suppose that f does not vanish identically; replacing f by $-f$ if necessary, we may suppose that f assumes some positive values. By the Maximum–Minimum Theorem 5.3.4, the function f attains the value $\sup\{f(x) : x \in I\} > 0$ at some point c in I. Since $f(a) = f(b) = 0$, the point c must lie in (a, b); therefore $f'(c)$ exists.

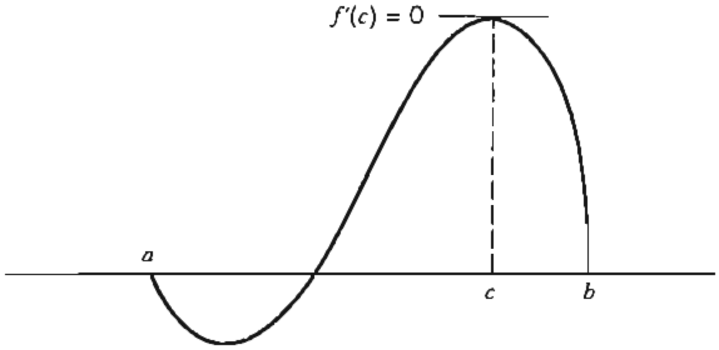


Figure 6.2.1 Rolle's Theorem

Since f has a relative maximum at c, we conclude from the Interior Extremum Theorem 6.2.1 that $f'(c) = 0$. (See Figure 6.2.1.) Q.E.D.

As a consequence of Rolle's Theorem, we obtain the fundamental Mean Value Theorem.

6.2.4 Mean Value Theorem *Suppose that f is continuous on a closed interval $I := [a, b]$, and that f has a derivative in the open interval (a, b). Then there exists at least one point c in (a, b) such that*

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

Proof. Consider the function φ defined on I by

$$\varphi(x) := f(x) - f(a) - \frac{f(b) - f(a)}{b - a}(x - a).$$

[The function φ is simply the difference of f and the function whose graph is the line segment joining the points $(a, f(a))$ and $(b, f(b))$; see Figure 6.2.2.] The hypotheses of

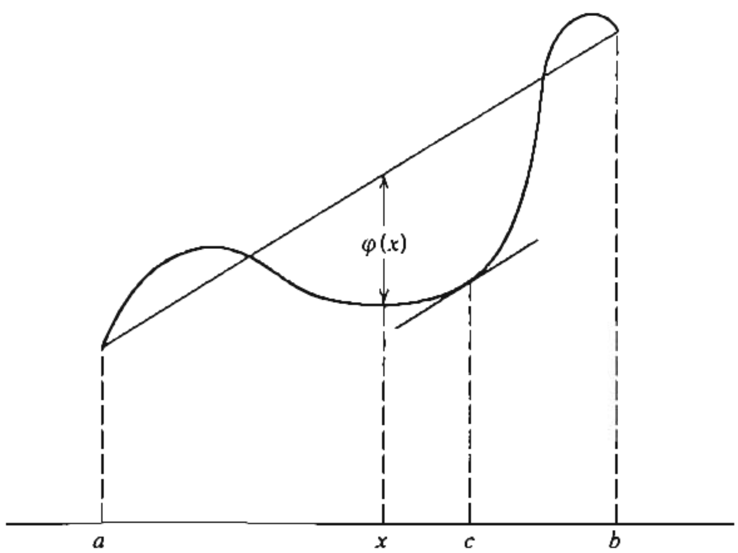


Figure 6.2.2 The Mean Value Theorem

Rolle's Theorem are satisfied by φ since φ is continuous on $[a, b]$, differentiable on (a, b), and $\varphi(a) = \varphi(b) = 0$. Therefore, there exists a point c in (a, b) such that

$$0 = \varphi'(c) = f'(c) - \frac{f(b) - f(a)}{b - a}.$$

Hence, $f(b) - f(a) = f'(c)(b - a)$. Q.E.D.

Remark The geometric view of the Mean Value Theorem is that there is some point on the curve $y = f(x)$ at which the tangent line is parallel to the line segment through the points $(a, f(a))$ and $(b, f(b))$. Thus it is easy to remember the statement of the Mean Value Theorem by drawing appropriate diagrams. While this should not be discouraged, it tends to suggest that its importance is geometrical in nature, which is quite misleading. In fact the Mean Value Theorem is a wolf in sheep's clothing and is *the* Fundamental Theorem of Differential Calculus. In the remainder of this section, we will present some of the consequences of this result. Other applications will be given later.

The Mean Value Theorem permits one to draw conclusions about the nature of a function f from information about its derivative f'. The following results are obtained in this manner.

6.2.5 Theorem *Suppose that f is continuous on the closed interval $I := [a, b]$, that f is differentiable on the open interval (a, b), and that $f'(x) = 0$ for $x \in (a, b)$. Then f is constant on I.*

Proof. We will show that $f(x) = f(a)$ for all $x \in I$. Indeed, if $x \in I$, $x > a$, is given, we apply the Mean Value Theorem to f on the closed interval $[a, x]$. We obtain a point c (depending on x) between a and x such that $f(x) - f(a) = f'(c)(x - a)$. Since $f'(c) = 0$ (by hypothesis), we deduce that $f(x) - f(a) = 0$. Hence, $f(x) = f(a)$ for any $x \in I$.
Q.E.D.

6.2.6 Corollary *Suppose that f and g are continuous on $I := [a, b]$, that they are differentiable on (a, b), and that $f'(x) = g'(x)$ for all $x \in (a, b)$. Then there exists a constant C such that $f = g + C$ on I.*

Recall that a function $f : I \to \mathbb{R}$ is said to be **increasing** on the interval I if whenever x_1, x_2 in I satisfy $x_1 < x_2$, then $f(x_1) \leq f(x_2)$. Also recall that f is **decreasing** on I if the function $-f$ is increasing on I.

6.2.7 Theorem *Let $f : I \to \mathbb{R}$ be differentiable on the interval I. Then:*

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

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

Proof. (a) Suppose that $f'(x) \geq 0$ for all $x \in I$. If x_1, x_2 in I satisfy $x_1 < x_2$, then we apply the Mean Value Theorem to f on the closed interval $J := [x_1, x_2]$ to obtain a point c in (x_1, x_2) such that

$$f(x_2) - f(x_1) = f'(c)(x_2 - x_1).$$

Since $f'(c) \geq 0$ and $x_2 - x_1 > 0$, it follows that $f(x_2) - f(x_1) \geq 0$. (Why?) Hence, $f(x_1) \leq f(x_2)$ and, since $x_1 < x_2$ are arbitrary points in I, we conclude that f is increasing on I.

For the converse assertion, we suppose that f is differentiable and increasing on I. Thus, for any point $x \neq c$ in I, we have $\big(f(x) - f(c)\big)/(x - c) \geq 0$. (Why?) Hence, by Theorem 4.2.6 we conclude that

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

(b) The proof of part (b) is similar and will be omitted. Q.E.D.

A function f is said to be **strictly increasing** on an interval I if for any points x_1, x_2 in I such that $x_1 < x_2$, we have $f(x_1) < f(x_2)$. An argument along the same lines of the proof of Theorem 6.2.7 can be made to show that a function having a strictly positive derivative on an interval is strictly increasing there. (See Exercise 13.) However, the converse assertion is not true, since a strictly increasing differentiable function may have a derivative that vanishes at certain points. For example, the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) := x^3$ is strictly increasing on $\mathbb{R}$, but $f'(0) = 0$. The situation for strictly decreasing functions is similar.

Remark It is reasonable to define a function to be **increasing at a point** if there is a neighborhood of the point on which the function is increasing. One might suppose that, if the derivative is strictly positive at a point, then the function is increasing at this point. However, this supposition is false; indeed, the differentiable function defined by

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

is such that $g'(0) = 1$, yet it can be shown that g is not increasing in any neighborhood of $x = 0$. (See Exercise 10.)

We next obtain a sufficient condition for a function to have a relative extremum at an interior point of an interval.

6.2.8 First Derivative Test for Extrema *Let f be continuous on the interval $I := [a, b]$ and let c be an interior point of I. Assume that f is differentiable on (a, c) and (c, b). Then:*

(a) *If there is a neighborhood $(c - \delta, c + \delta) \subseteq I$ such that $f'(x) \geq 0$ for $c - \delta < x < c$ and $f'(x) \leq 0$ for $c < x < c + \delta$, then f has a relative maximum at c.*

(b) *If there is a neighborhood $(c - \delta, c + \delta) \subseteq I$ such that $f'(x) \leq 0$ for $c - \delta < x < c$ and $f'(x) \geq 0$ for $c < x < c + \delta$, then f has a relative minimum at c.*

Proof. (a) If $x \in (c - \delta, c)$, then it follows from the Mean Value Theorem that there exists a point $c_x \in (x, c)$ such that $f(c) - f(x) = (c - x)f'(c_x)$. Since $f'(c_x) \geq 0$ we infer that $f(x) \leq f(c)$ for $x \in (c - \delta, c)$. Similarly, it follows (how?) that $f(x) \leq f(c)$ for $x \in (c, c + \delta)$. Therefore $f(x) \leq f(c)$ for all $x \in (c - \delta, c + \delta)$ so that f has a relative maximum at c.

(b) The proof is similar. Q.E.D.

Remark The converse of the First Derivative Test 6.2.8 is *not* true. For example, there exists a differentiable function $f : \mathbb{R} \to \mathbb{R}$ with absolute minimum at $x = 0$ but such that

f' takes on both positive and negative values on both sides of (and arbitrarily close to) $x = 0$. (See Exercise 9.)

Further Applications of the Mean Value Theorem

We will continue giving other types of applications of the Mean Value Theorem; in doing so we will draw more freely than before on the past experience of the reader and his or her knowledge concerning the derivatives of certain well-known functions.

6.2.9 Examples (a) Rolle's Theorem can be used for the location of roots of a function. For, if a function g can be identified as the derivative of a function f, then between any two roots of f there is at least one root of g. For example, let $g(x) := \cos x$, then g is known to be the derivative of $f(x) := \sin x$. Hence, between any two roots of $\sin x$ there is at least one root of $\cos x$. On the other hand, $g'(x) = -\sin x = -f(x)$, so another application of Rolle's Theorem tells us that between any two roots of cos there is at least one root of sin. Therefore, we conclude that the roots of sin and cos *interlace each other*. This conclusion is probably not news to the reader; however, the same type of argument can be applied to the *Bessel functions* J_n of order $n = 0, 1, 2, \cdots$ by using the relations

$$[x^n J_n(x)]' = x^n J_{n-1}(x), \quad [x^{n-1} J_n(x)]' = -x^{-n} J_{n+1}(x) \qquad \text{for} \quad x > 0.$$

The details of this argument should be supplied by the reader.

(b) We can apply the Mean Value Theorem for approximate calculations and to obtain error estimates. For example, suppose it is desired to evaluate $\sqrt{105}$. We employ the Mean Value Theorem with $f(x) := \sqrt{x}$, $a = 100$, $b = 105$, to obtain

$$\sqrt{105} - \sqrt{100} = \frac{5}{2\sqrt{c}}$$

for some number c with $100 < c < 105$. Since $10 < \sqrt{c} < \sqrt{105} < \sqrt{121} = 11$, we can assert that

$$\frac{5}{2(11)} < \sqrt{105} - 10 < \frac{5}{2(10)},$$

whence it follows that $10.2272 < \sqrt{105} < 10.2500$. This estimate may not be as sharp as desired. It is clear that the estimate $\sqrt{c} < \sqrt{105} < \sqrt{121}$ was wasteful and can be improved by making use of our conclusion that $\sqrt{105} < 10.2500$. Thus, $\sqrt{c} < 10.2500$ and we easily determine that

$$0.2439 < \frac{5}{2(10.2500)} < \sqrt{105} - 10.$$

Our improved estimate is $10.2439 < \sqrt{105} < 10.2500$. □

Inequalities

One very important use of the Mean Value Theorem is to obtain certain inequalities. Whenever information concerning the range of the derivative of a function is available, this information can be used to deduce certain properties of the function itself. The following examples illustrate the valuable role that the Mean Value Theorem plays in this respect.

6.2.10 Examples (a) The exponential function $f(x) := e^x$ has the derivative $f'(x) = e^x$ for all $x \in \mathbb{R}$. Thus $f'(x) > 1$ for $x > 0$, and $f'(x) < 1$ for $x < 0$. From these relationships, we will derive the inequality

$$(1) \qquad e^x \geq 1 + x \qquad \text{for} \quad x \in \mathbb{R},$$

with equality occurring if and only if $x = 0$.

If $x = 0$, we have equality with both sides equal to 1. If $x > 0$, we apply the Mean Value Theorem to the function f on the interval $[0, x]$. Then for some c with $0 < c < x$ we have

$$e^x - e^0 = e^c(x - 0).$$

Since $e^0 = 1$ and $e^c > 1$, this becomes $e^x - 1 > x$ so that we have $e^x > 1 + x$ for $x > 0$. A similar argument establishes the same strict inequality for $x < 0$. Thus the inequality (1) holds for all x, and equality occurs only if $x = 0$.

(b) The function $g(x) := \sin x$ has the derivative $g'(x) = \cos x$ for all $x \in \mathbb{R}$. On the basis of the fact that $-1 \leq \cos x \leq 1$ for all $x \in \mathbb{R}$, we will show that

$$(2) \qquad -x \leq \sin x \leq x \quad \text{for all} \quad x \geq 0.$$

Indeed, if we apply the Mean Value Theorem to g on the interval $[0, x]$, where $x > 0$, we obtain

$$\sin x - \sin 0 = (\cos c)(x - 0)$$

for some c between 0 and x. Since $\sin 0 = 0$ and $-1 \leq \cos c \leq 1$, we have $-x \leq \sin x \leq x$. Since equality holds at $x = 0$, the inequality (2) is established.

(c) (Bernoulli's inequality) If $\alpha > 1$, then

$$(3) \qquad (1 + x)^\alpha \geq 1 + \alpha x \qquad \text{for all} \quad x > -1,$$

with equality if and only if $x = 0$.

This inequality was established earlier, in Example 2.1.13(c), for positive integer values of α by using Mathematical Induction. We now derive the more general version by employing the Mean Value Theorem.

If $h(x) := (1 + x)^\alpha$ then $h'(x) = \alpha(1 + x)^{\alpha-1}$ for all $x > -1$. [For rational α this derivative was established in Example 6.1.10(c). The extension to irrational will be discussed in Section 8.3.] If $x > 0$, we infer from the Mean Value Theorem applied to h on the interval $[0, x]$ that there exists c with $0 < c < x$ such that $h(x) - h(0) = h'(c)(x - 0)$. Thus, we have

$$(1 + x)^\alpha - 1 = \alpha(1 + c)^{\alpha-1}x.$$

Since $c > 0$ and $\alpha - 1 > 0$, it follows that $(1 + c)^{\alpha-1} > 1$ and hence that $(1 + x)^\alpha > 1 + \alpha x$. If $-1 < x < 0$, a similar use of the Mean Value Theorem on the interval $[x, 0]$ leads to the same strict inequality. Since the case $x = 0$ results in equality, we conclude that (3) is valid for all $x > -1$ with equality if and only if $x = 0$.

(d) Let α be a real number satisfying $0 < \alpha < 1$ and let $g(x) = \alpha x - x^\alpha$ for $x \geq 0$. Then $g'(x) = \alpha(1 - x^{\alpha-1})$, so that $g'(x) < 0$ for $0 < x < 1$ and $g'(x) > 0$ for $x > 1$. Consequently, if $x \geq 0$, then $g(x) \geq g(1)$ and $g(x) = g(1)$ if and only if $x = 1$. Therefore, if $x \geq 0$ and $0 < \alpha < 1$, then we have

$$x^\alpha \leq \alpha x + (1 - \alpha).$$

If $a > 0$ and $b > 0$ and if we let $x = a/b$ and multiply by b, we obtain the inequality

$$a^{\alpha} b^{1-\alpha} \leq \alpha a + (1-\alpha)b,$$

where equality holds if and only if $a = b$. □

The Intermediate Value Property of Derivatives

We conclude this section with an interesting result, often referred to as Darboux's Theorem. It states that if a function f is differentiable at every point of an interval I, then the function f' has the Intermediate Value Property. This means that if f' takes on values A and B, then it also takes on all values between A and B. The reader will recognize this property as one of the important consequences of continuity as established in Theorem 5.3.7. It is remarkable that derivatives, which need not be continuous functions, also possess this property.

6.2.11 Lemma *Let $I \subseteq \mathbb{R}$ be an interval, let $f : I \to \mathbb{R}$, let $c \in I$, and assume that f has a derivative at c. Then:*

(a) *If $f'(c) > 0$, then there is a number $\delta > 0$ such that $f(x) > f(c)$ for $x \in I$ such that $c < x < c + \delta$.*

(b) *If $f'(c) < 0$, then there is a number $\delta > 0$ such that $f(x) > f(c)$ for $x \in I$ such that $c - \delta < x < c$.*

Proof. (a) Since

$$\lim_{x \to c} \frac{f(x) - f(c)}{x - c} = f'(c) > 0,$$

it follows from Theorem 4.2.9 that there is a number $\delta > 0$ such that if $x \in I$ and $0 < |x - c| < \delta$, then

$$\frac{f(x) - f(c)}{x - c} > 0.$$

If $x \in I$ also satisfies $x > c$, then we have

$$f(x) - f(c) = (x - c) \cdot \frac{f(x) - f(c)}{x - c} > 0.$$

Hence, if $x \in I$ and $c < x < c + \delta$, then $f(x) > f(c)$.

The proof of (b) is similar. Q.E.D.

6.2.12 Darboux's Theorem *If f is differentiable on $I = [a, b]$ and if k is a number between $f'(a)$ and $f'(b)$, then there is at least one point c in (a, b) such that $f'(c) = k$.*

Proof. Suppose that $f'(a) < k < f'(b)$. We define g on I by $g(x) := kx - f(x)$ for $x \in I$. Since g is continuous, it attains a maximum value on I. Since $g'(a) = k - f'(a) > 0$, it follows from Lemma 6.2.11(a) that the maximum of g does not occur at $x = a$. Similarly, since $g'(b) = k - f'(b) < 0$, it follows from Lemma 6.2.11(b) that the maximum does not occur at $x = b$. Therefore, g attains its maximum at some c in (a, b). Then from Theorem 6.2.1 we have $0 = g'(c) = k - f'(c)$. Hence, $f'(c) = k$. Q.E.D.

6.2.13 Example The function $g: [-1, 1] \to \mathbb{R}$ defined by

$$g(x) := \begin{cases} 1 & \text{for } 0 < x \leq 1, \\ 0 & \text{for } x = 0, \\ -1 & \text{for } -1 \leq x < 0, \end{cases}$$

(which is a restriction of the signum function) clearly fails to satisfy the intermediate value property on the interval $[-1, 1]$. Therefore, by Darboux's Theorem, there does not exist a function f such that $f'(x) = g(x)$ for all $x \in [-1, 1]$. In other words, g is *not* the derivative on $[-1, 1]$ of any function. □

Exercises for Section 6.2

1. For each of the following functions on $\mathbb{R}$ to $\mathbb{R}$, find points of relative extrema, the intervals on which the function is increasing, and those on which it is decreasing:
(a) $f(x) := x^2 - 3x + 5$, (b) $g(x) := 3x - 4x^2$,
(c) $h(x) := x^3 - 3x - 4$, (d) $k(x) := x^4 + 2x^2 - 4$.

2. Find the points of relative extrema, the intervals on which the following functions are increasing, and those on which they are decreasing:
(a) $f(x) := x + 1/x$ for $x \neq 0$, (b) $g(x) := x/(x^2 + 1)$ for $x \in \mathbb{R}$,
(c) $h(x) := \sqrt{x} - 2\sqrt{x + 2}$ for $x > 0$, (d) $k(x) := 2x + 1/x^2$ for $x \neq 0$.

3. Find the points of relative extrema of the following functions on the specified domain:
(a) $f(x) := |x^2 - 1|$ for $-4 \leq x \leq 4$, (b) $g(x) := 1 - (x - 1)^{2/3}$ for $0 \leq x \leq 2$,
(c) $h(x) := x|x^2 - 12|$ for $-2 \leq x \leq 3$, (d) $k(x) := x(x - 8)^{1/3}$ for $0 \leq x \leq 9$.

4. Let $a_1, a_2, \cdots, a_n$ be real numbers and let f be defined on $\mathbb{R}$ by
$$f(x) := \sum_{i=1}^{n} (a_i - x)^2 \qquad \text{for} \quad x \in \mathbb{R}.$$
Find the unique point of relative minimum for f.

5. Let $a > b > 0$ and let $n \in \mathbb{N}$ satisfy $n \geq 2$. Prove that $a^{1/n} - b^{1/n} < (a - b)^{1/n}$. [*Hint*: Show that $f(x) := x^{1/n} - (x - 1)^{1/n}$ is decreasing for $x \geq 1$, and evaluate f at 1 and a/b.]

6. Use the Mean Value Theorem to prove that $|\sin x - \sin y| \leq |x - y|$ for all x, y in $\mathbb{R}$.

7. Use the Mean Value Theorem to prove that $(x - 1)/x < \ln x < x - 1$ for $x > 1$. [*Hint*: Use the fact that $D \ln x = 1/x$ for $x > 0$.]

8. Let $f: [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable in (a, b). Show that if $\lim_{x \to a} f'(x) = A$, then $f'(a)$ exists and equals A. [*Hint*: Use the definition of $f'(a)$ and the Mean Value Theorem.]

9. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) := 2x^4 + x^4 \sin(1/x)$ for $x \neq 0$ and $f(0) := 0$. Show that f has an absolute minimum at $x = 0$, but that its derivative has both positive and negative values in every neighborhood of 0.

10. Let $g : \mathbb{R} \to \mathbb{R}$ be defined by $g(x) := x + 2x^2 \sin(1/x)$ for $x \neq 0$ and $g(0) := 0$. Show that $g'(0) = 1$, but in every neighborhood of 0 the derivative $g'(x)$ takes on both positive and negative values. Thus g is not monotonic in any neighborhood of 0.

11. Give an example of a uniformly continuous function on $[0, 1]$ that is differentiable on $(0, 1)$ but whose derivative is not bounded on $(0, 1)$.

12. If $h(x) := 0$ for $x < 0$ and $h(x) := 1$ for $x \geq 0$, prove there does not exist a function $f : \mathbb{R} \to \mathbb{R}$ such that $f'(x) = h(x)$ for all $x \in \mathbb{R}$. Give examples of two functions, not differing by a constant, whose derivatives equal $h(x)$ for all $x \neq 0$.

13. Let I be an interval and let $f : I \to \mathbb{R}$ be differentiable on I. Show that if f' is positive on I, then f is strictly increasing on I.

14. Let I be an interval and let $f : I \to \mathbb{R}$ be differentiable on I. Show that if the derivative f' is never 0 on I, then either $f'(x) > 0$ for all $x \in I$ or $f'(x) < 0$ for all $x \in I$.

15. Let I be an interval. Prove that if f is differentiable on I and if the derivative f' is bounded on I, then f satisfies a Lipschitz condition on I. (See Definition 5.4.4.)
16. Let $f\colon [0, \infty) \to \mathbb{R}$ be differentiable on $(0, \infty)$ and assume that $f'(x) \to b$ as $x \to \infty$.
    (a) Show that for any $h > 0$, we have $\lim\limits_{x\to\infty} \big(f(x+h) - f(x)\big)/h = b$.
    (b) Show that if $f(x) \to a$ as $x \to \infty$, then $b = 0$.
    (c) Show that $\lim\limits_{x\to\infty} \big(f(x)/x\big) = b$.
17. Let f, g be differentiable on $\mathbb{R}$ and suppose that $f(0) = g(0)$ and $f'(x) \le g'(x)$ for all $x \ge 0$. Show that $f(x) \le g(x)$ for all $x \ge 0$.
18. Let $I := [a, b]$ and let $f : I \to \mathbb{R}$ be differentiable at $c \in I$. Show that for every $\varepsilon > 0$ there exists $\delta > 0$ such that if $0 < |x - y| < \delta$ and $a \le x \le c \le y \le b$, then
$$\left|\frac{f(x) - f(y)}{x - y} - f'(c)\right| < \varepsilon.$$
19. A differentiable function $f : I \to \mathbb{R}$ is said to be **uniformly differentiable** on $I := [a, b]$ if for every $\varepsilon > 0$ there exists $\delta > 0$ such that if $0 < |x - y| < \delta$ and $x, y \in I$, then
$$\left|\frac{f(x) - f(y)}{x - y} - f'(x)\right| < \varepsilon.$$
Show that if f is uniformly differentiable on I, then f' is continuous on I.
20. Suppose that $f : [0, 2] \to \mathbb{R}$ is continuous on $[0, 2]$ and differentiable on $(0, 2)$, and that $f(0) = 0$, $f(1) = 1$, $f(2) = 1$.
    (a) Show that there exists $c_1 \in (0, 1)$ such that $f'(c_1) = 1$.
    (b) Show that there exists $c_2 \in (1, 2)$ such that $f'(c_2) = 0$.
    (c) Show that there exists $c \in (0, 2)$ such that $f'(c) = 1/3$.

Section 6.3 L'Hospital's Rules

The Marquis Guillame François L'Hospital (1661–1704) was the author of the first calculus book, *L'Analyse des infiniment petits*, published in 1696. He studied the then new differential calculus from Johann Bernoulli (1667–1748), first when Bernoulli visited L'Hospital's country estate and subsequently through a series of letters. The book was the result of L'Hospital's studies. The limit theorem that became known as L'Hospital's Rule first appeared in this book, though in fact it was discovered by Bernoulli.

The initial theorem was refined and extended, and the various results are collectively referred to as L'Hospital's (or L'Hôpital's) Rules. In this section we establish the most basic of these results and indicate how others can be derived.

Indeterminate Forms

In the preceding chapters we have often been concerned with methods of evaluating limits. It was shown in Theorem 4.2.4(b) that if $A := \lim\limits_{x\to c} f(x)$ and $B := \lim\limits_{x\to c} g(x)$, and if $B \ne 0$, then
$$\lim_{x\to c} \frac{f(x)}{g(x)} = \frac{A}{B}.$$
However, if $B = 0$, then no conclusion was deduced. It will be seen in Exercise 2 that if $B = 0$ and $A \ne 0$, then the limit is infinite (when it exists).

The case $A = 0$, $B = 0$ has not been covered previously. In this case, the limit of the quotient f/g is said to be "indeterminate". We will see that in this case the limit may not exist or may be any real value, depending on the particular functions f and g. The symbolism $0/0$ is used to refer to this situation. For example, if α is any real number, and if we define $f(x) := \alpha x$ and $g(x) := x$, then

$$\lim_{x\to 0} \frac{f(x)}{g(x)} = \lim_{x\to 0} \frac{\alpha x}{x} = \lim_{x\to 0} \alpha = \alpha.$$

Thus the indeterminate form $0/0$ can lead to any real number α as a limit.

Other indeterminate forms are represented by the symbols ∞/∞, $0\cdot\infty$, 0^0, 1^∞, ∞^0, and $\infty - \infty$. These notations correspond to the indicated limiting behavior and juxtaposition of the functions f and g. Our attention will be focused on the indeterminate forms $0/0$ and ∞/∞. The other indeterminate cases are usually reduced to the form $0/0$ or ∞/∞ by taking logarithms, exponentials, or algebraic manipulations.

A Preliminary Result

To show that the use of differentiation in this context is a natural and not surprising development, we first establish an elementary result that is based simply on the definition of the derivative.

6.3.1 Theorem *Let f and g be defined on $[a, b]$, let $f(a) = g(a) = 0$, and let $g(x) \neq 0$ for $a < x < b$. If f and g are differentiable at a and if $g'(a) \neq 0$, then the limit of f/g at a exists and is equal to $f'(a)/g'(a)$. Thus*

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

Proof. Since $f(a) = g(a) = 0$, we can write the quotient $f(x)/g(x)$ for $a < x < b$ as follows:

$$\frac{f(x)}{g(x)} = \frac{f(x) - f(a)}{g(x) - g(a)} = \frac{\dfrac{f(x) - f(a)}{x - a}}{\dfrac{g(x) - g(a)}{x - a}}.$$

Applying Theorem 4.2.4(b), we obtain

$$\lim_{x\to a+} \frac{f(x)}{g(x)} = \frac{\lim\limits_{x\to a+} \dfrac{f(x) - f(a)}{x - a}}{\lim\limits_{x\to a+} \dfrac{g(x) - g(a)}{x - a}} = \frac{f'(a)}{g'(a)}.$$

Q.E.D.

Warning The hypothesis that $f(a) = g(a) = 0$ is essential here. For example, if $f(x) := x + 17$ and $g(x) := 2x + 3$ for $x \in \mathbb{R}$, then

$$\lim_{x\to 0} \frac{f(x)}{g(x)} = \frac{17}{3}, \quad \text{while} \quad \frac{f'(0)}{g'(0)} = \frac{1}{2}.$$

The preceding result enables us to deal with limits such as

$$\lim_{x\to 0} \frac{x^2 + x}{\sin 2x} = \frac{2\cdot 0 + 1}{2\cos 0} = \frac{1}{2}.$$

To handle limits where f and g are not differentiable at the point a, we need a more general version of the Mean Value Theorem due to Cauchy.

6.3.2 Cauchy Mean Value Theorem *Let f and g be continuous on $[a, b]$ and differentiable on (a, b), and assume that $g'(x) \neq 0$ for all x in (a, b). Then there exists c in (a, b) such that*

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

Proof. As in the proof of the Mean Value Theorem, we introduce a function to which Rolle's Theorem will apply. First we note that since $g'(x) \neq 0$ for all x in (a, b), it follows from Rolle's Theorem that $g(a) \neq g(b)$. For x in $[a, b]$, we now define

$$h(x) := \frac{f(b) - f(a)}{g(b) - g(a)}\big(g(x) - g(a)\big) - \big(f(x) - f(a)\big).$$

Then h is continuous on $[a, b]$, differentiable on (a, b), and $h(a) = h(b) = 0$. Therefore, it follows from Rolle's Theorem 6.2.3 that there exists a point c in (a, b) such that

$$0 = h'(c) = \frac{f(b) - f(a)}{g(b) - g(a)} g'(c) - f'(c).$$

Since $g'(c) \neq 0$, we obtain the desired result by dividing by $g'(c)$. Q.E.D.

Remarks The preceding theorem has a geometric interpretation that is similar to that of the Mean Value Theorem 6.2.4. The functions f and g can be viewed as determining a curve in the plane by means of the parametric equations $x = f(t)$, $y = g(t)$ where $a \leq t \leq b$. Then the conclusion of the theorem is that there exists a point $(f(c), g(c))$ on the curve for some c in (a, b) such that the slope $g'(c)/f'(c)$ of the line tangent to the curve at that point is equal to the slope of the line segment joining the endpoints of the curve.

Note that if $g(x) = x$, then the Cauchy Mean Value Theorem reduces to the Mean Value Theorem 6.2.4.

L'Hospital's Rule, I

We will now establish the first of L'Hospital's Rules. For convenience, we will consider right-hand limits at a point a; left-hand limits, and two-sided limits are treated in exactly the same way. In fact, the theorem even allows the possibility that $a = -\infty$. The reader should observe that, in contrast with Theorem 6.3.1, the following result does not assume the differentiability of the functions at the point a. The result asserts that the limiting behavior of $f(x)/g(x)$ as $x \to a+$ is the same as the limiting behavior of $f'(x)/g'(x)$ as $x \to a+$, including the case where this limit is infinite. An important hypothesis here is that both f and g approach 0 as $x \to a+$.

6.3.3 L'Hospital's Rule, I *Let $-\infty \leq a < b \leq \infty$ and let f, g be differentiable on (a, b) such that $g'(x) \neq 0$ for all $x \in (a, b)$. Suppose that*

(1) $$\lim_{x \to a+} f(x) = 0 = \lim_{x \to a+} g(x).$$

(a) *If* $\displaystyle\lim_{x \to a+} \frac{f'(x)}{g'(x)} = L \in \mathbb{R}$, *then* $\displaystyle\lim_{x \to a+} \frac{f(x)}{g(x)} = L$.

(b) *If* $\displaystyle\lim_{x \to a+} \frac{f'(x)}{g'(x)} = L \in \{-\infty, \infty\}$, *then* $\displaystyle\lim_{x \to a+} \frac{f(x)}{g(x)} = L$.

Proof. If $a < \alpha < \beta < b$, then Rolle's Theorem implies that $g(\beta) \neq g(\alpha)$. Further, by the Cauchy Mean Value Theorem 6.3.2, there exists $u \in (\alpha, \beta)$ such that

$$\frac{f(\beta) - f(\alpha)}{g(\beta) - g(\alpha)} = \frac{f'(u)}{g'(u)}. \tag{2}$$

Case (a): If $L \in \mathbb{R}$ and if $\varepsilon > 0$ is given, there exists $c \in (a, b)$ such that

$$L - \varepsilon < \frac{f'(u)}{g'(u)} < L + \varepsilon \qquad \text{for} \quad u \in (a, c),$$

whence it follows from (2) that

$$L - \varepsilon < \frac{f(\beta) - f(\alpha)}{g(\beta) - g(\alpha)} < L + \varepsilon \qquad \text{for} \quad a < \alpha < \beta \le c. \tag{3}$$

If we take the limit in (3) as $\alpha \to a+$, we have

$$L - \varepsilon \le \frac{f(\beta)}{g(\beta)} \le L + \varepsilon \qquad \text{for} \quad \beta \in (a, c].$$

Since $\varepsilon > 0$ is arbitrary, the assertion follows.

Case (b): If $L = +\infty$ and if $M > 0$ is given, there exists $c \in (a, b)$ such that

$$\frac{f'(u)}{g'(u)} > M \qquad \text{for} \quad u \in (a, c),$$

whence it follows from (2) that

$$\frac{f(\beta) - f(\alpha)}{g(\beta) - g(\alpha)} > M \qquad \text{for} \quad a < \alpha < \beta < c. \tag{4}$$

If we take the limit in (4) as $\alpha \to a+$, we have

$$\frac{f(\beta)}{g(\beta)} \ge M \qquad \text{for} \quad \beta \in (a, c).$$

Since $M > 0$ is arbitrary, the assertion follows.

If $L = -\infty$, the argument is similar. Q.E.D.

6.3.4 Examples **(a)** We have

$$\lim_{x \to 0+} \frac{\sin x}{\sqrt{x}} = \lim_{x \to 0+} \left[\frac{\cos x}{1/(2\sqrt{x})} \right] = \lim_{x \to 0+} 2\sqrt{x} \cos x = 0.$$

Observe that the denominator is not differentiable at $x = 0$ so that Theorem 6.3.1 cannot be applied. However $f(x) := \sin x$ and $g(x) := \sqrt{x}$ are differentiable on $(0, \infty)$ and both approach 0 as $x \to 0+$. Moreover, $g'(x) \neq 0$ on $(0, \infty)$, so that 6.3.3 is applicable.

(b) We have $\lim\limits_{x \to 0} \left[\dfrac{1 - \cos x}{x^2} \right] = \lim\limits_{x \to 0} \dfrac{\sin x}{2x}$.

We need to consider both left and right hand limits here. The quotient in the second limit is again indeterminate in the form $0/0$. However, the hypotheses of 6.3.3 are again satisfied so that a second application of L'Hospital's Rule is permissible. Hence, we obtain

$$\lim_{x \to 0} \left[\frac{1 - \cos x}{x^2} \right] = \lim_{x \to 0} \frac{\sin x}{2x} = \lim_{x \to 0} \frac{\cos x}{2} = \frac{1}{2}.$$

(c) We have $\lim_{x\to 0} \dfrac{e^x - 1}{x} = \lim_{x\to 0} \dfrac{e^x}{1} = 1.$
Again, both left- and right-hand limits need to be considered. Similarly, we have

$$\lim_{x\to 0}\left[\frac{e^x - 1 - x}{x^2}\right] = \lim_{x\to 0}\frac{e^x - 1}{2x} = \lim_{x\to 0}\frac{e^x}{2} = \frac{1}{2}.$$

(d) We have $\lim_{x\to 1}\left[\dfrac{\ln x}{x-1}\right] = \lim_{x\to 1}\dfrac{(1/x)}{1} = 1.$ □

L'Hospital's Rule, II

This Rule is very similar to the first one, except that it treats the case where the denominator becomes infinite as $x \to a+$. Again we will consider only right-hand limits, but it is possible that $a = -\infty$. Left-hand limits and two-sided limits are handled similarly.

6.3.5 L'Hospital's Rule, II *Let $-\infty \le a < b \le \infty$ and let f, g be differentiable on (a, b) such that $g'(x) \ne 0$ for all $x \in (a, b)$. Suppose that*

$$\lim_{x\to a+} g(x) = \pm\infty. \tag{5}$$

(a) *If $\lim_{x\to a+} \dfrac{f'(x)}{g'(x)} = L \in \mathbb{R}$, then $\lim_{x\to a+}\dfrac{f(x)}{g(x)} = L$.*

(b) *If $\lim_{x\to a+} \dfrac{f'(x)}{g'(x)} = L \in \{-\infty, \infty\}$, then $\lim_{x\to a+}\dfrac{f(x)}{g(x)} = L$.*

Proof. We will suppose that (5) holds with limit ∞.

As before, we have $g(\beta) \ne g(\alpha)$ for $\alpha, \beta \in (a, b)$, $\alpha < \beta$. Further, equation (2) in the proof of 6.3.3 holds for some $u \in (\alpha, \beta)$.

Case (a): If $L \in \mathbb{R}$ with $L > 0$ and $\varepsilon > 0$ is given, there is $c \in (a, b)$ such that (3) in the proof of 6.3.3 holds when $a < \alpha < \beta \le c$. Since $g(x) \to \infty$, we may also assume that $g(c) > 0$. Taking $\beta = c$ in (3), we have

$$L - \varepsilon < \frac{f(c) - f(\alpha)}{g(c) - g(\alpha)} < L + \varepsilon \qquad \text{for} \quad \alpha \in (a, c). \tag{6}$$

Since $g(c)/g(\alpha) \to 0$ as $\alpha \to a+$, we may assume that $0 < g(c)/g(\alpha) < 1$ for all $\alpha \in (a, c)$, whence it follows that

$$\frac{g(\alpha) - g(c)}{g(\alpha)} = 1 - \frac{g(c)}{g(\alpha)} > 0 \qquad \text{for} \quad \alpha \in (a, c).$$

If we multiply (6) by $(g(\alpha) - g(c))/g(\alpha) > 0$, we have

$$(L - \varepsilon)\left(1 - \frac{g(c)}{g(\alpha)}\right) < \frac{f(\alpha)}{g(\alpha)} - \frac{f(c)}{g(\alpha)} < (L + \varepsilon)\left(1 - \frac{g(c)}{g(\alpha)}\right). \tag{7}$$

Now, since $g(c)/g(\alpha) \to 0$ and $f(c)/g(\alpha) \to 0$ as $\alpha \to a+$, then for any δ with $0 < \delta < 1$ there exists $d \in (a, c)$ such that $0 < g(c)/g(\alpha) < \delta$ and $|f(c)|/g(\alpha) < \delta$ for all $\alpha \in (a, d)$, whence (7) gives

$$(L - \varepsilon)(1 - \delta) - \delta < \frac{f(\alpha)}{g(\alpha)} < (L + \varepsilon) + \delta. \tag{8}$$

If we take $\delta := \min\{1, \varepsilon, \varepsilon/(|L| + 1)\}$, it is an exercise to show that

$$L - 2\varepsilon \le \frac{f(\alpha)}{g(\alpha)} \le L + 2\varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, this yields the assertion. The cases $L = 0$ and $L < 0$ are handled similarly.

Case (b): If $L = +\infty$, let $M > 1$ be given and $c \in (a, b)$ be such that $f'(u)/g'(u) > M$ for all $u \in (a, c)$. Then it follows as before that

$$\frac{f(\beta) - f(\alpha)}{g(\beta) - g(\alpha)} > M \qquad \text{for} \quad a < \alpha < \beta \le c. \tag{9}$$

Since $g(x) \to \infty$ as $x \to a+$, we may suppose that c also satisfies $g(c) > 0$, that $|f(c)|/g(\alpha) < \frac{1}{2}$, and that $0 < g(c)/g(\alpha) < \frac{1}{2}$ for all $\alpha \in (a, c)$. If we take $\beta = c$ in (9) and multiply by $1 - g(c)/g(\alpha) > \frac{1}{2}$, we get

$$\frac{f(\alpha) - f(c)}{g(\alpha)} > M\left(1 - \frac{g(c)}{g(\alpha)}\right) > \tfrac{1}{2}M,$$

so that

$$\frac{f(\alpha)}{g(\alpha)} > \tfrac{1}{2}M + \frac{f(c)}{g(\alpha)} > \tfrac{1}{2}(M - 1) \qquad \text{for} \quad \alpha \in (a, c).$$

Since $M > 1$ is arbitrary, it follows that $\lim\limits_{\alpha \to a+} f(\alpha)/g(\alpha) = \infty$.

If $L = -\infty$, the argument is similar. Q.E.D.

6.3.6 Examples **(a)** We consider $\lim\limits_{x \to \infty} \dfrac{\ln x}{x}$.

Here $f(x) := \ln x$ and $g(x) := x$ on the interval $(0, \infty)$. If we apply the left-hand version of 6.3.5, we obtain $\lim\limits_{x \to \infty} \dfrac{\ln x}{x} = \lim\limits_{x \to \infty} \dfrac{1/x}{1} = 0$.

(b) We consider $\lim\limits_{x \to \infty} e^{-x}x^2$.

Here we take $f(x) := x^2$ and $g(x) := e^x$ on $\mathbb{R}$. We obtain

$$\lim_{x \to \infty} \frac{x^2}{e^x} = \lim_{x \to \infty} \frac{2x}{e^x} = \lim_{x \to \infty} \frac{2}{e^x} = 0.$$

(c) We consider $\lim\limits_{x \to 0+} \dfrac{\ln \sin x}{\ln x}$.

Here we take $f(x) := \ln \sin x$ and $g(x) := \ln x$ on $(0, \pi)$. If we apply 6.3.5, we obtain

$$\lim_{x \to 0+} \frac{\ln \sin x}{\ln x} = \lim_{x \to 0+} \frac{\cos x / \sin x}{1/x} = \lim_{x \to 0+} \left[\frac{x}{\sin x}\right] \cdot [\cos x].$$

Since $\lim\limits_{x \to 0+} [x/\sin x] = 1$ and $\lim\limits_{x \to 0+} \cos x = 1$, we conclude that the limit under consideration equals 1. □

Other Indeterminate Forms

Indeterminate forms such as $\infty - \infty$, $0 \cdot \infty$, 1^∞, 0^0, ∞^0 can be reduced to the previously considered cases by algebraic manipulations and the use of the logarithmic and exponential functions. Instead of formulating these variations as theorems, we illustrate the pertinent techniques by means of examples.

6.3.7 Examples **(a)** Let $I := (0, \pi/2)$ and consider

$$\lim_{x \to 0+} \left(\frac{1}{x} - \frac{1}{\sin x}\right),$$

which has the indeterminate form $\infty - \infty$. We have

$$\lim_{x\to 0+}\left(\frac{1}{x}-\frac{1}{\sin x}\right) = \lim_{x\to 0+}\frac{\sin x - x}{x\sin x} = \lim_{x\to 0+}\frac{\cos x - 1}{\sin x + x\cos x}$$

$$= \lim_{x\to 0+}\frac{-\sin x}{2\cos x - x\sin x} = \frac{0}{2} = 0.$$

(b) Let $I := (0, \infty)$ and consider $\lim_{x\to 0+} x\ln x$, which has the indeterminate form $0\cdot(-\infty)$. We have

$$\lim_{x\to 0+} x\ln x = \lim_{x\to 0+}\frac{\ln x}{1/x} = \lim_{x\to 0+}\frac{1/x}{-1/x^2} = \lim_{x\to 0+}(-x) = 0.$$

(c) Let $I := (0, \infty)$ and consider $\lim_{x\to 0+} x^x$, which has the indeterminate form 0^0.

We recall from calculus (see also Section 8.3) that $x^x = e^{x\ln x}$. It follows from part (b) and the continuity of the function $y \mapsto e^y$ at $y = 0$ that $\lim_{x\to 0+} x^x = e^0 = 1$.

(d) Let $I := (1, \infty)$ and consider $\lim_{x\to\infty}(1 + 1/x)^x$, which has the indeterminate form 1^∞. We note that

$$(1 + 1/x)^x = e^{x\ln(1+1/x)}. \tag{10}$$

Moreover, we have

$$\lim_{x\to\infty} x\ln(1 + 1/x) = \lim_{x\to\infty}\frac{\ln(1 + 1/x)}{1/x}$$

$$= \lim_{x\to\infty}\frac{(1 + 1/x)^{-1}(-x^{-2})}{-x^{-2}} = \lim_{x\to\infty}\frac{1}{1 + 1/x} = 1.$$

Since $y \mapsto e^y$ is continuous at $y = 1$, we infer that $\lim_{x\to\infty}(1 + 1/x)^x = e$.

(e) Let $I := (0, \infty)$ and consider $\lim_{x\to 0+}(1 + 1/x)^x$, which has the indeterminate form ∞^0. In view of formula (10), we consider

$$\lim_{x\to 0+} x\ln(1 + 1/x) = \lim_{x\to 0+}\frac{\ln(1 + 1/x)}{1/x} = \lim_{x\to 0+}\frac{1}{1 + 1/x} = 0.$$

Therefore we have $\lim_{x\to 0+}(1 + 1/x)^x = e^0 = 1$. □

Exercises for Section 6.3

1. Suppose that f and g are continuous on $[a, b]$, differentiable on (a, b), that $c \in [a, b]$ and that $g(x) \neq 0$ for $x \in [a, b]$, $x \neq c$. Let $A := \lim_{x\to c} f$ and $B := \lim_{x\to c} g$. If $B = 0$, and if $\lim_{x\to c} f(x)/g(x)$ exists in $\mathbb{R}$, show that we must have $A = 0$. [*Hint:* $f(x) = \{f(x)/g(x)\}g(x)$.]

2. In addition to the suppositions of the preceding exercise, let $g(x) > 0$ for $x \in [a, b]$, $x \neq c$. If $A > 0$ and $B = 0$, prove that we must have $\lim_{x\to c} f(x)/g(x) = \infty$. If $A < 0$ and $B = 0$, prove that we must have $\lim_{x\to c} f(x)/g(x) = -\infty$.

3. Let $f(x) := x^2\sin(1/x)$ for $0 < x \le 1$ and $f(0) := 0$, and let $g(x) := x^2$ for $x \in [0, 1]$. Then both f and g are differentiable on $[0, 1]$ and $g(x) > 0$ for $x \neq 0$. Show that $\lim_{x\to 0} f(x) = 0 = \lim_{x\to 0} g(x)$ and that $\lim_{x\to 0} f(x)/g(x)$ does not exist.

4. Let $f(x) := x^2$ for x rational, let $f(x) := 0$ for x irrational, and let $g(x) := \sin x$ for $x \in \mathbb{R}$. Use Theorem 6.3.1 to show that $\lim_{x\to 0} f(x)/g(x) = 0$. Explain why Theorem 6.3.3 cannot be used.

5. Let $f(x) := x^2 \sin(1/x)$ for $x \neq 0$, let $f(0) := 0$, and let $g(x) := \sin x$ for $x \in \mathbb{R}$. Show that $\lim_{x\to 0} f(x)/g(x) = 0$ but that $\lim_{x\to 0} f'(x)/g'(x)$ does not exist.

6. Evaluate the following limits, where the domain of the quotient is as indicated.
   (a) $\lim_{x\to 0+} \dfrac{\ln(x+1)}{\sin x}$ $(0, \pi/2)$,
   (b) $\lim_{x\to 0+} \dfrac{\tan x}{x}$ $(0, \pi/2)$,
   (c) $\lim_{x\to 0+} \dfrac{\ln \cos x}{x}$ $(0, \pi/2)$,
   (d) $\lim_{x\to 0+} \dfrac{\tan x - x}{x^3}$ $(0, \pi/2)$.

7. Evaluate the following limits:
   (a) $\lim_{x\to 0} \dfrac{\operatorname{Arctan} x}{x}$ $(-\infty, \infty)$,
   (b) $\lim_{x\to 0} \dfrac{1}{x(\ln x)^2}$ $(0, 1)$,
   (c) $\lim_{x\to 0+} x^3 \ln x$ $(0, \infty)$,
   (d) $\lim_{x\to\infty} \dfrac{x^3}{e^x}$ $(0, \infty)$.

8. Evaluate the following limits:
   (a) $\lim_{x\to\infty} \dfrac{\ln x}{x^2}$ $(0, \infty)$,
   (b) $\lim_{x\to\infty} \dfrac{\ln x}{\sqrt{x}}$ $(0, \infty)$,
   (c) $\lim_{x\to 0} x \ln \sin x$ $(0, \pi)$,
   (d) $\lim_{x\to\infty} \dfrac{x + \ln x}{x \ln x}$ $(0, \infty)$.

9. Evaluate the following limits:
   (a) $\lim_{x\to 0+} x^{2x}$ $(0, \infty)$,
   (b) $\lim_{x\to 0} (1 + 3/x)^x$ $(0, \infty)$,
   (c) $\lim_{x\to\infty} (1 + 3/x)^x$ $(0, \infty)$,
   (d) $\lim_{x\to 0+} \left(\dfrac{1}{x} - \dfrac{1}{\operatorname{Arctan} x}\right)$ $(0, \infty)$.

10. Evaluate the following limits:
    (a) $\lim_{x\to\infty} x^{1/x}$ $(0, \infty)$,
    (b) $\lim_{x\to 0+} (\sin x)^x$ $(0, \pi)$,
    (c) $\lim_{x\to 0+} x^{\sin x}$ $(0, \infty)$,
    (d) $\lim_{x\to\pi/2-} (\sec x - \tan x)$ $(0, \pi/2)$.

11. Let f be differentiable on $(0, \infty)$ and suppose that $\lim_{x\to\infty} \big(f(x) + f'(x)\big) = L$. Show that $\lim_{x\to\infty} f(x) = L$ and $\lim_{x\to\infty} f'(x) = 0$. [*Hint*: $f(x) = e^x f(x)/e^x$.]

12. Try to use L'Hospital's Rule to find the limit of $\dfrac{\tan x}{\sec x}$ as $x \to (\pi/2)-$. Then evaluate directly by changing to sines and cosines.

Section 6.4 Taylor's Theorem

A very useful technique in the analysis of real functions is the approximation of functions by polynomials. In this section we will prove a fundamental theorem in this area which goes back to Brook Taylor (1685–1731), although the remainder term was not provided until much later by Joseph-Louis Lagrange (1736–1813). Taylor's Theorem is a powerful result that has many applications. We will illustrate the versatility of Taylor's Theorem by briefly discussing some of its applications to numerical estimation, inequalities, extreme values of a function, and convex functions.

Taylor's Theorem can be regarded as an extension of the Mean Value Theorem to "higher order" derivatives. Whereas the Mean Value Theorem relates the values of a function and its first derivative, Taylor's Theorem provides a relation between the values of a function and its higher order derivatives.

Derivatives of order greater than one are obtained by a natural extension of the differentiation process. If the derivative $f'(x)$ of a function f exists at every point x in an interval I containing a point c, then we can consider the existence of the derivative of the function f' at the point c. In case f' has a derivative at the point c, we refer to the resulting number as the **second derivative** of f at c, and we denote this number by $f''(c)$ or by $f^{(2)}(c)$. In similar fashion we define the third derivative $f'''(c) = f^{(3)}(c), \cdots$, and the ***n*th derivative** $f^{(n)}(c)$, whenever these derivatives exist. It is noted that the existence of the nth derivative at c presumes the existence of the $(n-1)$st derivative in an interval containing c, but we do allow the possibility that c might be an endpoint of such an interval.

If a function f has an nth derivative at a point x_0, it is not difficult to construct an nth degree polynomial P_n such that $P_n(x_0) = f(x_0)$ and $P_n^{(k)}(x_0) = f^{(k)}(x_0)$ for $k = 1, 2, \cdots, n$. In fact, the polynomial

$$(1) \qquad P_n(x) := f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$

has the property that it and its derivatives up to order n agree with the function f and its derivatives up to order n, at the specified point x_0. This polynomial P_n is called the ***n*th Taylor polynomial for f at x_0**. It is natural to expect this polynomial to provide a reasonable approximation to f for points near x_0, but to gauge the quality of the approximation, it is necessary to have information concerning the remainder $R_n := f - P_n$. The following fundamental result provides such information.

6.4.1 Taylor's Theorem *Let $n \in \mathbb{N}$, let $I := [a, b]$, and let $f : I \to \mathbb{R}$ be such that f and its derivatives $f', f'', \cdots, f^{(n)}$ are continuous on I and that $f^{(n+1)}$ exists on (a, b). If $x_0 \in I$, then for any x in I there exists a point c between x and x_0 such that*

$$(2) \qquad f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + \frac{f^{(n+1)}(c)}{(n+1)!}(x - x_0)^{n+1}.$$

Proof. Let x_0 and x be given and let J denote the closed interval with endpoints x_0 and x. We define the function F on J by

$$F(t) := f(x) - f(t) - (x - t)f'(t) - \cdots - \frac{(x - t)^n}{n!} f^{(n)}(t)$$

for $t \in J$. Then an easy calculation shows that we have

$$F'(t) = -\frac{(x - t)^n}{n!} f^{(n+1)}(t).$$

If we define G on J by

$$G(t) := F(t) - \left(\frac{x - t}{x - x_0}\right)^{n+1} F(x_0)$$

for $t \in J$, then $G(x_0) = G(x) = 0$. An application of Rolle's Theorem 6.2.3 yields a point c between x and x_0 such that

$$0 = G'(c) = F'(c) + (n + 1)\frac{(x - c)^n}{(x - x_0)^{n+1}} F(x_0).$$

Hence, we obtain

$$\begin{aligned} F(x_0) &= -\frac{1}{n+1}\frac{(x-x_0)^{n+1}}{(x-c)^n}F'(c) \\ &= \frac{1}{n+1}\frac{(x-x_0)^{n+1}}{(x-c)^n}\frac{(x-c)^n}{n!}f^{(n+1)}(c) = \frac{f^{(n+1)}(c)}{(n+1)!}(x-x_0)^{n+1}, \end{aligned}$$

which implies the stated result. Q.E.D.

We shall use the notation P_n for the nth Taylor polynomial (1) of f, and R_n for the remainder. Thus we may write the conclusion of Taylor's Theorem as $f(x) = P_n(x) + R_n(x)$ where R_n is given by

$$(3) \qquad R_n(x) := \frac{f^{(n+1)}(c)}{(n+1)!}(x-x_0)^{n+1}$$

for some point c between x and x_0. This formula for R_n is referred to as the **Lagrange form** (or the **derivative form**) of the remainder. Many other expressions for R_n are known; one is in terms of integration and will be given later. (See Theorem 7.3.18.)

Applications of Taylor's Theorem

The remainder term R_n in Taylor's Theorem can be used to estimate the error in approximating a function by its Taylor polynomial P_n. If the number n is prescribed, then the question of the accuracy of the approximation arises. On the other hand, if a certain accuracy is specified, then the question of finding a suitable value of n is germane. The following examples illustrate how one responds to these questions.

6.4.2 Examples (a) Use Taylor's Theorem with $n = 2$ to approximate $\sqrt[3]{1+x}$, $x > -1$.

We take the function $f(x) := (1+x)^{1/3}$, the point $x_0 = 0$, and $n = 2$. Since $f'(x) = \frac{1}{3}(1+x)^{-2/3}$ and $f''(x) = \frac{1}{3}\left(-\frac{2}{3}\right)(1+x)^{-5/3}$, we have $f'(0) = \frac{1}{3}$ and $f''(0) = -2/9$. Thus we obtain

$$f(x) = P_2(x) + R_2(x) = 1 + \tfrac{1}{3}x - \tfrac{1}{9}x^2 + R_2(x),$$

where $R_2(x) = \frac{1}{3!}f'''(c)x^3 = \frac{5}{81}(1+c)^{-8/3}x^3$ for some point c between 0 and x.

For example, if we let $x = 0.3$, we get the approximation $P_2(0.3) = 1.09$ for $\sqrt[3]{1.3}$. Moreover, since $c > 0$ in this case, then $(1+c)^{-8/3} < 1$ and so the error is at most

$$R_2(0.3) \le \frac{5}{81}\left(\frac{3}{10}\right)^3 = \frac{1}{600} < 0.17 \times 10^{-2}.$$

Hence, we have $|\sqrt[3]{1.3} - 1.09| < 0.5 \times 10^{-2}$, so that two decimal place accuracy is assured.

(b) Approximate the number e with error less than 10^{-5}.

We shall consider the function $g(x) := e^x$ and take $x_0 = 0$ and $x = 1$ in Taylor's Theorem. We need to determine n so that $|R_n(1)| < 10^{-5}$. To do so, we shall use the fact that $g'(x) = e^x$ and the initial bound of $e^x \le 3$ for $0 \le x \le 1$.

Since $g'(x) = e^x$, it follows that $g^{(k)}(x) = e^x$ for all $k \in \mathbb{N}$, and therefore $g^{(k)}(0) = 1$ for all $k \in \mathbb{N}$. Consequently the nth Taylor polynomial is given by

$$P_n(x) := 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!}$$

and the remainder for $x = 1$ is given by $R_n(1) = e^c/(n+1)!$ for some c satisfying $0 < c < 1$. Since $e^c < 3$, we seek a value of n such that $3/(n+1)! < 10^{-5}$. A calculation reveals that $9! = 362,880 > 3 \times 10^5$ so that the value $n = 8$ will provide the desired accuracy; moreover, since $8! = 40,320$, no smaller value of n will be certain to suffice. Thus, we obtain

$$e \approx P_8(1) = 1 + 1 + \frac{1}{2!} + \cdots + \frac{1}{8!} = 2.718\,28$$

with error less than 10^{-5}. □

Taylor's Theorem can also be used to derive inequalities.

6.4.3 Examples **(a)** $1 - \frac{1}{2}x^2 \le \cos x$ for all $x \in \mathbb{R}$.

Use $f(x) := \cos x$ and $x_0 = 0$ in Taylor's Theorem, to obtain

$$\cos x = 1 - \frac{1}{2}x^2 + R_2(x),$$

where for some c between 0 and x we have

$$R_2(x) = \frac{f'''(c)}{3!}x^3 = \frac{\sin c}{6}x^3.$$

If $0 \le x \le \pi$, then $0 \le c < \pi$; since c and x^3 are both positive, we have $R_2(x) \ge 0$. Also, if $-\pi \le x \le 0$, then $-\pi \le c \le 0$; since $\sin c$ and x^3 are both negative, we again have $R_2(x) \ge 0$. Therefore, we see that $1 - \frac{1}{2}x^2 \le \cos x$ for $|x| \le \pi$. If $|x| \ge \pi$, then we have $1 - \frac{1}{2}x^2 < -3 \le \cos x$ and the inequality is trivially valid. Hence, the inequality holds for all $x \in \mathbb{R}$.

(b) For any $k \in \mathbb{N}$, and for all $x > 0$, we have

$$x - \frac{1}{2}x^2 + \cdots - \frac{1}{2k}x^{2k} < \ln(1+x) < x - \frac{1}{2}x^2 + \cdots + \frac{1}{2k+1}x^{2k+1}.$$

Using the fact that the derivative of $\ln(1+x)$ is $1/(1+x)$ for $x > 0$, we see that the nth Taylor polynomial for $\ln(1+x)$ with $x_0 = 0$ is

$$P_n(x) = x - \frac{1}{2}x^2 + \cdots + (-1)^{n-1}\frac{1}{n}x^n$$

and the remainder is given by

$$R_n(x) = \frac{(-1)^n c^{n+1}}{n+1}x^{n+1}$$

for some c satisfying $0 < c < x$. Thus for any $x > 0$, if $n = 2k$ is even, then we have $R_{2k}(x) > 0$; and if $n = 2k+1$ is odd, then we have $R_{2k+1}(x) < 0$. The stated inequality then follows immediately. □

Relative Extrema

It was established in Theorem 6.2.1 that if a function $f : I \to \mathbb{R}$ is differentiable at a point c interior to the interval I, then a necessary condition for f to have a relative extremum at c is that $f'(c) = 0$. One way to determine whether f has a relative maximum or relative minimum [or neither] at c, is to use the First Derivative Test 6.2.8. Higher order derivatives, if they exist, can also be used in this determination, as we now show.

6.4.4 Theorem *Let I be an interval, let x_0 be an interior point of I, and let $n \geq 2$. Suppose that the derivatives $f', f'', \cdots, f^{(n)}$ exist and are continuous in a neighborhood of x_0 and that $f'(x_0) = \cdots = f^{(n-1)}(x_0) = 0$, but $f^{(n)}(x_0) \neq 0$.*

(i) *If n is even and $f^{(n)}(x_0) > 0$, then f has a relative minimum at x_0.*

(ii) *If n is even and $f^{(n)}(x_0) < 0$, then f has a relative maximum at x_0.*

(iii) *If n is odd, then f has neither a relative minimum nor relative maximum at x_0.*

Proof. Applying Taylor's Theorem at x_0, we find that for $x \in I$ we have

$$f(x) = P_{n-1}(x) + R_{n-1}(x) = f(x_0) + \frac{f^{(n)}(c)}{n!}(x - x_0)^n,$$

where c is some point between x_0 and x. Since $f^{(n)}$ is continuous, if $f^{(n)}(x_0) \neq 0$, then there exists an interval U containing x_0 such that $f^{(n)}(x)$ will have the same sign as $f^{(n)}(x_0)$ for $x \in U$. If $x \in U$, then the point c also belongs to U and consequently $f^{(n)}(c)$ and $f^{(n)}(x_0)$ will have the same sign.

(i) If n is even and $f^{(n)}(x_0) > 0$, then for $x \in U$ we have $f^{(n)}(c) > 0$ and $(x - x_0)^n \geq 0$ so that $R_{n-1}(x) \geq 0$. Hence, $f(x) \geq f(x_0)$ for $x \in U$, and therefore f has a relative minimum at x_0.

(ii) If n is even and $f^{(n)}(x_0) < 0$, then it follows that $R_{n-1}(x) \leq 0$ for $x \in U$, so that $f(x) \leq f(x_0)$ for $x \in U$. Therefore, f has a relative maximum at x_0.

(iii) If n is odd, then $(x - x_0)^n$ is positive if $x > x_0$ and negative if $x < x_0$. Consequently, if $x \in U$, then $R_{n-1}(x)$ will have opposite signs to the left and to the right of x_0. Therefore, f has neither a relative minimum nor a relative maximum at x_0. Q.E.D.

Convex Functions

The notion of convexity plays an important role in a number of areas, particularly in the modern theory of optimization. We shall briefly look at convex functions of one real variable and their relation to differentiation. The basic results, when appropriately modified, can be extended to higher dimensional spaces.

6.4.5 Definition Let $I \subseteq \mathbb{R}$ be an interval. A function $f : I \to \mathbb{R}$ is said to be **convex** on I if for any t satisfying $0 \leq t \leq 1$ and any points x_1, x_2 in I, we have

$$f\left((1-t)x_1 + tx_2\right) \leq (1-t)f(x_1) + tf(x_2).$$

Note that if $x_1 < x_2$, then as t ranges from 0 to 1, the point $(1-t)x_1 + tx_2$ traverses the interval from x_1 to x_2. Thus if f is convex on I and if $x_1, x_2 \in I$, then the chord joining any two points $(x_1, f(x_1))$ and $(x_2, f(x_2))$ on the graph of f lies above the graph of f. (See Figure 6.4.1.)

A convex function need not be differentiable at every point, as the example $f(x) := |x|$, $x \in \mathbb{R}$, reveals. However, it can be shown that if I is an open interval and if $f : I \to \mathbb{R}$ is convex on I, then the left and right derivatives of f exist at every point of I. As a consequence, it follows that a convex function on an open interval is necessarily continuous. We will not verify the preceding assertions, nor will we develop many other interesting properties of convex functions. Rather, we will restrict ourselves to establishing the connection between a convex function f and its second derivative f'', assuming that f'' exists.

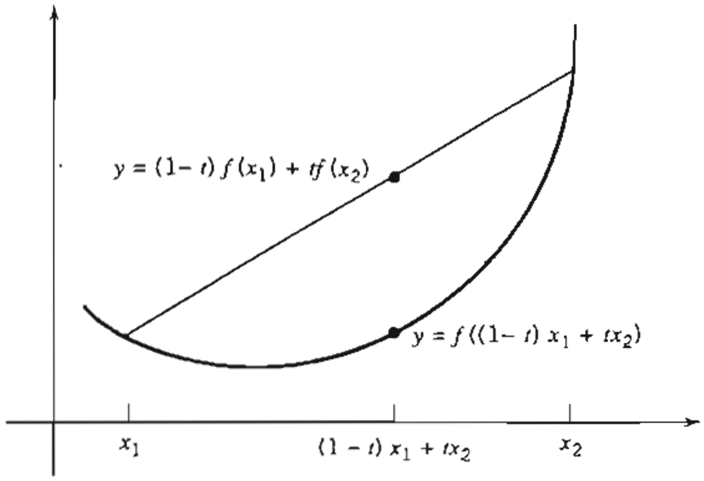


Figure 6.4.1 A convex function.

6.4.6 Theorem *Let I be an open interval and let $f : I \to \mathbb{R}$ have a second derivative on I. Then f is a convex function on I if and only if $f''(x) \geq 0$ for all $x \in I$.*

Proof. ($\Rightarrow$) We will make use of the fact that the second derivative is given by the limit

$$f''(a) = \lim_{h\to 0} \frac{f(a+h) - 2f(a) + f(a-h)}{h^2} \tag{4}$$

for each $a \in I$. (See Exercise 16.) Given $a \in I$, let h be such that $a+h$ and $a-h$ belong to I. Then $a = \frac{1}{2}\big((a+h) + (a-h)\big)$, and since f is convex on I, we have

$$f(a) = f\left(\tfrac{1}{2}(a+h) + \tfrac{1}{2}(a-h)\right) \leq \tfrac{1}{2}f(a+h) + \tfrac{1}{2}f(a-h).$$

Therefore, we have $f(a+h) - 2f(a) + f(a-h) \geq 0$. Since $h^2 > 0$ for all $h \neq 0$, we see that the limit in (4) must be nonnegative. Hence, we obtain $f''(a) \geq 0$ for any $a \in I$.

($\Leftarrow$) We will use Taylor's Theorem. Let x_1, x_2 be any two points of I, let $0 < t < 1$, and let $x_0 := (1-t)x_1 + tx_2$. Applying Taylor's Theorem to f at x_0 we obtain a point c_1 between x_0 and x_1 such that

$$f(x_1) = f(x_0) + f'(x_0)(x_1 - x_0) + \tfrac{1}{2}f''(c_1)(x_1 - x_0)^2,$$

and a point c_2 between x_0 and x_2 such that

$$f(x_2) = f(x_0) + f'(x_0)(x_2 - x_0) + \tfrac{1}{2}f''(c_2)(x_2 - x_0)^2.$$

If f'' is nonnegative on I, then the term

$$R := \tfrac{1}{2}(1-t)f''(c_1)(x_1 - x_0)^2 + \tfrac{1}{2}tf''(c_2)(x_2 - x_0)^2$$

is also nonnegative. Thus we obtain

$$\begin{aligned}(1-t)f(x_1) + tf(x_2) &= f(x_0) + f'(x_0)\big((1-t)x_1 + tx_2 - x_0\big)\\ &\quad + \tfrac{1}{2}(1-t)f''(c_1)(x_1 - x_0)^2 + \tfrac{1}{2}tf''(c_2)(x_2 - x_0)^2\\ &= f(x_0) + R\\ &\geq f(x_0) = f\big((1-t)x_1 + tx_2\big).\end{aligned}$$

Hence, f is a convex function on I. Q.E.D.