Undergraduate Texts in Mathematics

Undergraduate Texts in Mathematics

Sudhir R. Ghorpade
Balmohan V. Limaye

A Course in Calculus and Real Analysis

With 71 Figures

Sudhir R. Ghorpade
Department of Mathematics
Indian Institute of Technology Bombay
Powai, Mumbai 400076
INDIA
srg@math.iitb.ac.in

Balmohan V. Limaye
Department of Mathematics
Indian Institute of Technology Bombay
Powai, Mumbai 400076
INDIA
bvl@math.iitb.ac.in

Series Editors:
S. Axler
Mathematics Department
San Francisco State University
San Francisco, CA 94132
USA
axler@sfsu.edu

K. A. Ribet
Department of Mathematics
University of California at Berkeley
Berkeley, CA 94720-3840
USA
ribet@math.berkeley.edu

Preface

Calculus is one of the triumphs of the human mind. It emerged from investigations into such basic questions as finding areas, lengths and volumes. In the third century B.C., Archimedes determined the area under the arc of a parabola. In the early seventeenth century, Fermat and Descartes studied the problem of finding tangents to curves. But the subject really came to life in the hands of Newton and Leibniz in the late seventeenth century. In particular, they showed that the geometric problems of finding the areas of planar regions and of finding the tangents to plane curves are intimately related to one another. In subsequent decades, the subject developed further through the work of several mathematicians, most notably Euler, Cauchy, Riemann, and Weierstrass.

Today, calculus occupies a central place in mathematics and is an essential component of undergraduate education. It has an immense number of applications both within and outside mathematics. Judged by the sheer variety of the concepts and results it has generated, calculus can be rightly viewed as a fountainhead of ideas and disciplines in mathematics.

Real analysis, often called mathematical analysis or simply analysis, may be regarded as a formidable counterpart of calculus. It is a subject where one revisits notions encountered in calculus, but with greater rigor and sometimes with greater generality. Nonetheless, the basic objects of study remain the same, namely, real-valued functions of one or several real variables.

This book attempts to give a self-contained and rigorous introduction to calculus of functions of one variable. The presentation and sequencing of topics emphasizes the structural development of calculus. At the same time, due importance is given to computational techniques and applications. In the course of our exposition, we highlight the fact that calculus provides a firm foundation to several concepts and results that are generally encountered in high school and accepted on faith. For instance, this book can help students get a clear understanding of (i) the definitions of the logarithmic, exponential and trigonometric functions and a proof of the fact that these are not algebraic functions, (ii) the definition of an angle and (iii) the result that the ratio of

the circumference of a circle to its diameter is the same for all circles. It is our experience that a majority of students are unable to absorb these concepts and results without getting into vicious circles. This may partly be due to the division of calculus and real analysis in compartmentalized courses. Calculus is often taught as a service course and as such there is little time to dwell on subtleties and gain perspective. On the other hand, real analysis courses may start at once with metric spaces and devote more time to pathological examples than to consolidating students' knowledge of calculus. A host of topics such as L'Hôpital's rule, points of inflection, convergence criteria for Newton's method, solids of revolution, and quadrature rules, which may have been inadequately covered in calculus courses, become passé when one studies real analysis. Trigonometric, exponential, and logarithmic functions are defined, if at all, in terms of infinite series, thereby missing out on purely algebraic motivations for introducing these functions. The ubiquitous role of π as a ratio of various geometric quantities and as a constant that can be defined independently using calculus is often not well understood. A possible remedy would be to avoid the separation of calculus and real analysis into seemingly disjoint courses and textbooks. Attempts along these lines have been made in the past as in the excellent books of Hardy and of Courant and John. Ours is another attempt to give a unified exposition of calculus and real analysis and address the concerns expressed above. While this book deals with functions of one variable, we intend to treat functions of several variables in another book.

The genesis of this book lies in the notes we prepared for an undergraduate course at the Indian Institute of Technology Bombay in 1997. Encouraged by the feedback from students and colleagues, the notes and problem sets were put together in March 1998 into a booklet that has been in private circulation. Initially, it seemed that it would be relatively easy to convert that booklet into a book. Seven years have passed since then and we now know a little better! While that booklet was certainly helpful, this book has evolved to acquire a form and philosophy of its own and is quite distinct from the original notes.

A glance at the table of contents should give the reader an idea of the topics covered. For the most part, these are standard topics and novelty, if any, lies in how we approach them. Throughout this text we have sought to make a distinction between the intrinsic definition of a geometric notion and the analytic characterizations or criteria that are normally employed in studying it. In many cases we have used articles such as those in *A Century of Calculus* to simplify the treatment. Usually each important result is followed by two kinds of examples: one to illustrate the result and the other to show that a hypothesis cannot be dropped.

When a concept is defined it appears in boldface. Definitions are not numbered but can be located using the index. Everything else (propositions, examples, remarks, etc.) is numbered serially in each chapter. The end of a proof is marked by the symbol □, while the symbol ◇ marks the end of an example or a remark. Bibliographic details about the books and articles mentioned in the text and in this preface can be found in the list of references. Citations

within the text appear in square brackets. A list of symbols and abbreviations used in the text appears after the list of references.

The *Notes and Comments* that appear at the end of each chapter are an important part of the book. Distinctive features of the exposition are mentioned here and often pointers to some relevant literature and further developments are provided. We hope that these may inspire many fruitful visits to the library—not when a quiz or the final is around the corner, but perhaps after it is over. The *Notes and Comments* are followed by a fairly large collection of exercises. These are divided into two parts. Exercises in Part A are relatively routine and should be attempted by all students. Part B contains problems that are of a theoretical nature or are particularly challenging. These may be done at leisure. Besides the two sets of exercises in every chapter, there is a separate collection of problems, called Revision Exercises which appear at the end of Chapter 7. It is in Chapter 7 that the logarithmic, exponential, and trigonometric functions are formally introduced. Their use is strictly avoided in the preceding chapters. This meant that standard examples and counterexamples such as $x \sin(1/x)$ could not be discussed earlier. The Revision Exercises provide an opportunity to revisit the material covered in Chapters 1–7 and to work out problems that involve the use of elementary transcendental functions.

The formal prerequisites for this course do not go beyond what is normally covered in high school. No knowledge of trigonometry is assumed and exposure to linear algebra is not taken for granted. However, we do expect some mathematical maturity and an ability to understand and appreciate proofs. This book can be used as a textbook for a serious undergraduate course in calculus. Parts of the book could be useful for advanced undergraduate and graduate courses in real analysis. Further, this book can also be used for self-study by students who wish to consolidate their knowledge of calculus and real analysis. For teachers and researchers this may be a useful reference for topics that are usually not covered in standard texts.

Apart from the first paragraph of this preface, we have not discussed the history of the subject or placed each result in historical context. However, we do not doubt that a knowledge of the historical development of concepts and results is important as well as interesting. Indeed, it can greatly enrich one's understanding and appreciation of the subject. For those interested, we encourage looking on the Internet, where a wealth of information about the history of mathematics and mathematicians can be readily found. Among the various sources available, we particularly recommend the MacTutor History of Mathematics archive `http://www-groups.dcs.st-and.ac.uk/history/` at the University of St. Andrews. The books of Boyer, Edwards, and Stillwell are also excellent sources for the history of mathematics, especially calculus.

In preparing this book we have received generous assistance from various organizations and individuals. First, we thank our parent institution IIT Bombay and in particular its Department of Mathematics for providing excellent infrastructure and granting a sabbatical leave for each of us to work

We would appreciate receiving comments, suggestions, and corrections. These can be sent by e-mail to `acicara@gmail.com` or by writing to either of us. Corrections, modifications, and relevant information will be posted at `http://www.math.iitb.ac.in/`∼`srg/acicara` and we encourage interested readers to visit this website to learn about updates concerning the book.

Mumbai, India   *Sudhir Ghorpade*
July 2005   *Balmohan Limaye*

Contents

1

Numbers and Functions

Let us begin at the beginning. When we learn the script of a language, such as the English language, we begin with the letters of the alphabet A, B, C, ...; when we learn the sounds of music, such as those of western classical music, we begin with the notes Do, Re, Mi, ... Likewise, in mathematics, one begins with $1, 2, 3, \ldots$; these are the **positive integers** or the **natural numbers**. We shall denote the set of positive integers by $\mathbb{N}$. Thus,

$$\mathbb{N} = \{1, 2, 3, \ldots\} .$$

These numbers have been known since antiquity. Over the years, the number 0 was conceived[1] and subsequently, the negative integers. Together, these form the set $\mathbb{Z}$ of integers.[2] Thus,

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

Quotients of integers are called **rational numbers**. We shall denote the set of all rational numbers by $\mathbb{Q}$. Thus,

$$\mathbb{Q} = \left\{ \frac{m}{n} : m, n \in \mathbb{Z}, \ n \neq 0 \right\} .$$

Geometrically, the integers can be represented by points on a line by fixing a base point (signifying the number 0) and a unit distance. Such a line is called the **number line** and it may be drawn as in Figure 1.1. By suitably subdividing the segment between 0 and 1, we can also represent rational numbers such as $1/n$, where $n \in \mathbb{N}$, and this can, in turn, be used to represent any

---

[1] The invention of 'zero', which also paves the way for the place value system of enumeration, is widely credited to the Indians. Great psychological barriers had to be overcome when 'zero' was being given the status of a legitimate number. For more on this, see the books of Kaplan [39] and Kline [41].

[2] The notation $\mathbb{Z}$ for the set of integers is inspired by the German word *Zahlen* for numbers.
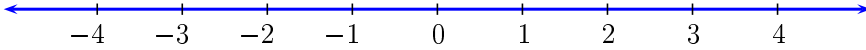
Fig. 1.1. The number line

rational number by a unique point on the number line. It is seen that the rational numbers spread themselves rather densely on this line. Nevertheless, several gaps do remain. For example, the 'number' $\sqrt{2}$ can be represented by a unique point between 1 and 2 on the number line using simple geometric constructions, but as we shall see later, this is not a rational number. We are, therefore, forced to reckon with the so-called *irrational numbers*, which are precisely the 'numbers' needed to fill the gaps left on the number line after marking all the rational numbers. The rational numbers and the irrational numbers together constitute the set $\mathbb{R}$, called the set of *real numbers*. The geometric representation of the real numbers as points on the number line naturally implies that there is an *order* among the real numbers. In particular, those real numbers that are greater than 0, that is, which correspond to points to the right of 0, are called *positive*.

1.1 Properties of Real Numbers

To be sure, we haven't precisely defined what real numbers are and what it means for them to be positive. For that matter, we haven't even defined the positive integers $1, 2, 3, \ldots$ or the rational numbers.[3] But at least we are familiar with the latter. We are also familiar with the addition and the multiplication of rational numbers. As for the real numbers, which are not easy to define, it is better to at least specify the properties that we shall take for granted. We shall take adequate care that in the subsequent development, we use only these properties or the consequences derived from them. In this way, we don't end up taking too many things on faith. So let us specify our assumptions.

We assume that there is a set $\mathbb{R}$ (whose elements are called real numbers), which contains the set $\mathbb{Q}$ of all rational numbers (and, in particular, the numbers 0 and 1) such that the following three types of properties are satisfied.

_____

[3] To a purist, this may appear unsatisfactory. A conscientious beginner in calculus may also become uncomfortable at some point of time that the basic notion of a (real) number is undefined. Such persons are first recommended to read the 'Notes and Comments' at the end of this chapter and then look up some of the references mentioned therein.

Algebraic Properties

We have the operations of addition (denoted by $+$) and multiplication (denoted by $\cdot$ or by juxtaposition) on $\mathbb{R}$, which extend the usual addition and multiplication of rational numbers and satisfy the following properties:

A1 $a + (b + c) = (a + b) + c$ and $a(bc) = (ab)c$ for all $a, b, c \in \mathbb{R}$.

A2 $a + b = b + a$ and $ab = ba$ for all $a, b \in \mathbb{R}$.

A3 $a + 0 = a$ and $a \cdot 1 = a$ for all $a \in \mathbb{R}$.

A4 Given any $a \in \mathbb{R}$, there is $a' \in \mathbb{R}$ such that $a + a' = 0$. Further, if $a \neq 0$, then there is $a^* \in \mathbb{R}$ such that $aa^* = 1$.

A5 $a(b + c) = ab + ac$ for all $a, b, c \in \mathbb{R}$.

It is interesting to note that several simple properties of real numbers that one is tempted to take for granted can be derived as consequences of the above properties. For example, let us prove that $a \cdot 0 = 0$ for all $a \in \mathbb{R}$. First, by A3, we have $0 + 0 = 0$. So, by A5, $a \cdot 0 = a(0 + 0) = a \cdot 0 + a \cdot 0$. Now, by A4, there is a $b' \in \mathbb{R}$ such that $a \cdot 0 + b' = 0$. Thus,

$$0 = a \cdot 0 + b' = (a \cdot 0 + a \cdot 0) + b' = a \cdot 0 + (a \cdot 0 + b') = a \cdot 0 + 0 = a \cdot 0,$$

where the third equality follows from A1 and the last equality follows from A3. This completes the proof! A number of similar properties are listed in the exercises and we invite the reader to supply the proofs. These show, in particular, that given any $a \in \mathbb{R}$, an element $a' \in \mathbb{R}$ such that $a + a' = 0$ is unique; this element will be called the **negative** or the **additive inverse** of a and denoted by $-a$. Likewise, if $a \in \mathbb{R}$ and $a \neq 0$, then an element $a^* \in \mathbb{R}$ such that $aa^* = 1$ is unique; this element is called the **reciprocal** or the **multiplicative inverse** of a and is denoted by a^{-1} or by $1/a$. Once all these formalities are understood, we will be free to replace expressions such as

$$a\,(1/b), \ a + a, \ aa, \ (a + b) + c, \ (ab)c, \ a + (-b),$$

by the corresponding simpler expressions, namely,

$$a/b, \ 2a, \ a^2, \ a + b + c, \ abc, \ a - b.$$

Here, for instance, it is meaningful and unambiguous to write $a + b + c$, thanks to A1. More generally, given finitely many real numbers $a_1, \ldots, a_n$, the sum $a_1 + \cdots + a_n$ has an unambiguous meaning. To represent such sums, the "sigma notation" can be quite useful. Thus, $a_1 + \cdots + a_n$ is often denoted by $\sum_{i=1}^{n} a_i$ or sometimes simply by $\sum_i a_i$ or $\sum a_i$. Likewise, the product $a_1 \cdots a_n$ of the real numbers $a_1, \ldots, a_n$ has an unambiguous meaning and it is often denoted by $\prod_{i=1}^{n} a_i$ or sometimes simply by $\prod_i a_i$ or $\prod a_i$. We remark that as a convention, the empty sum is defined to be zero, whereas an empty product is defined to be one. Thus, if $n = 0$, then $\sum_{i=1}^{n} a_i := 0$, whereas $\prod_{i=1}^{n} a_i := 1$.

Order Properties

The set $\mathbb{R}$ contains a subset $\mathbb{R}^+$, called the set of all positive real numbers, satisfying the following properties:

O1 *Given any $a \in \mathbb{R}$, exactly one of the following statements is true:*

$$a \in \mathbb{R}^+; \qquad a = 0; \qquad -a \in \mathbb{R}^+.$$

O2 *If $a, b \in \mathbb{R}^+$, then $a + b \in \mathbb{R}^+$ and $ab \in \mathbb{R}^+$.*

Given the existence of $\mathbb{R}^+$, we can define an *order relation* on $\mathbb{R}$ as follows. For $a, b \in \mathbb{R}$, define a to be **less than** b, and write $a < b$, if $b - a \in \mathbb{R}^+$. Sometimes, we write $b > a$ in place of $a < b$ and say that b is **greater than** a. With this notation, it follows from the algebraic properties that $\mathbb{R}^+ = \{x \in \mathbb{R} : x > 0\}$. Moreover, the following properties are easy consequences of A1–A5 and O1–O2:

(i) Given any $a, b \in \mathbb{R}$, exactly one of the following statements is true.

$$a < b; \qquad a = b; \qquad b < a.$$

(ii) If $a, b, c \in \mathbb{R}$ with $a < b$ and $b < c$, then $a < c$.
(iii) If $a, b, c \in \mathbb{R}$, with $a < b$, then $a + c < b + c$. Further, if $c > 0$, then $ac < bc$, whereas if $c < 0$, then $ac > bc$.

Note that it is also a consequence of the properties above that $1 > 0$. Indeed, by (i), we have either $1 > 0$ or $1 < 0$. If we had $1 < 0$, then we must have $-1 > 0$ and hence by (iii), $1 = (-1)(-1) > 0$, which is a contradiction. Therefore, $1 > 0$. A similar argument shows that $a^2 > 0$ for any $a \in \mathbb{R}$, $a \neq 0$.

The notation $a \leq b$ is often used to mean that either $a < b$ or $a = b$. Likewise, $a \geq b$ means that $a > b$ or $a = b$.

Let S be a subset of $\mathbb{R}$. We say that S is **bounded above** if there exists $\alpha \in \mathbb{R}$ such that $x \leq \alpha$ for all $x \in S$. Any such α is called an **upper bound** of S. We say that S is **bounded below** if there exists $\beta \in \mathbb{R}$ such that $x \geq \beta$ for all $x \in S$. Any such β is called a **lower bound** of S. The set S is said to be **bounded** if it is bounded above as well as bounded below; otherwise, S is said to be **unbounded**. Note that if $S = \emptyset$, that is, if S is the empty set, then every real number is an upper bound as well as a lower bound of S.

Examples 1.1. (i) The set $\mathbb{N}$ of positive integers is bounded below, and any real number $\beta \leq 1$ is a lower bound of $\mathbb{N}$. However, as we shall see later in Proposition 1.3, the set $\mathbb{N}$ is not bounded above.
(ii) The set S of reciprocals of positive integers, that is,

$$S := \left\{ 1, \frac{1}{2}, \frac{1}{3}, \ldots \right\}$$

is bounded. Any real number $\alpha \geq 1$ is an upper bound of S, whereas any real number $\beta \leq 0$ is a lower bound of S.

(iii) The set $S := \{x \in \mathbb{Q} : x^2 < 2\}$ is bounded. Here, for example, 2 is an upper bound, while -2 is a lower bound. $\diamond$

Let S be a subset of $\mathbb{R}$. An element $M \in \mathbb{R}$ is called a **supremum** or a **least upper bound** of the set S if

(i) M is an upper bound of S, that is, $x \leq M$ for all $x \in S$, and
(ii) $M \leq \alpha$ for any upper bound α of S.

It is easy to see from the definition that if S has a supremum, then it must be unique; we denote it by $\sup S$. Note that $\emptyset$ does not have a supremum.

An element $m \in \mathbb{R}$ is called an **infimum** or a **greatest lower bound** of the set S if

(i) m is a lower bound of S, that is, $m \leq x$ for all $x \in S$, and
(ii) $\beta \leq m$ for any lower bound β of S.

Again, it is easy to see from the definition that if S has an infimum, then it must be unique; we denote it by $\inf S$. Note that $\emptyset$ does not have an infimum.

For example, if $S = \{x \in \mathbb{R} : 0 < x \leq 1\}$, then $\inf S = 0$ and $\sup S = 1$. In this example, $\inf S$ is not an element of S, but $\sup S$ is an element of S.

If the supremum of a set S is an element of S, then it is called the **maximum** of S, and denoted by $\max S$; likewise, if the infimum of S is in S, then it is called the **minimum** of S, and denoted by $\min S$.

The last, and perhaps the most important, property of $\mathbb{R}$ that we shall assume is the following.

Completeness Property

Every nonempty subset of $\mathbb{R}$ that is bounded above has a supremum.

The significance of the Completeness Property (which is also known as the Least Upper Bound Property) will become clearer from the results proved in this as well as the subsequent chapters.

Proposition 1.2. *Let S be a nonempty subset of $\mathbb{R}$ that is bounded below. Then S has an infimum.*

Proof. Let $T = \{\beta \in \mathbb{R} : \beta \leq a \text{ for all } a \in S\}$. Since S is bounded below, T is nonempty, and since S is nonempty, T is bounded above. Hence T has a supremum. It is easily seen that $\sup T$ is the infimum of S. $\square$

Proposition 1.3. *Given any $x \in \mathbb{R}$, there is some $n \in \mathbb{N}$ such that $n > x$. Consequently, there is also an $m \in \mathbb{N}$ such that $-m < x$.*

Proof. Assume the contrary. Then x is an upper bound of $\mathbb{N}$. Therefore, $\mathbb{N}$ has a supremum. Let $M = \sup \mathbb{N}$. Then $M - 1 < M$ and hence $M - 1$ is not an upper bound of $\mathbb{N}$. So, there is $n \in \mathbb{N}$ such that $M - 1 < n$. But then $n + 1 \in \mathbb{N}$ and $M < n + 1$, which is a contradiction since M is an upper bound of $\mathbb{N}$. The second assertion about the existence of $m \in \mathbb{N}$ with $-m < x$ follows by applying the first assertion to $-x$. $\square$

The first assertion in the proposition above is sometimes referred to as the **Archimedean property** of $\mathbb{R}$. Observe that for any positive real number ϵ, by applying the Proposition 1.3 to $x = 1/\epsilon$, we see that there exists $n \in \mathbb{N}$ such that $(1/n) < \epsilon$. Note also that thanks to Proposition 1.3, for any $x \in \mathbb{R}$, there are $m, n \in \mathbb{N}$ such that $-m < x < n$. The largest among the finitely many integers k satisfying $-m \leq k \leq n$ and also $k \leq x$ is called the **integer part** of x and is denoted by $[x]$. Note that the integer part of x is characterized by the following properties:

$$[x] \in \mathbb{Z} \quad \text{and} \quad [x] \leq x < [x] + 1.$$

Sometimes, the integer part of x is called the **floor** of x and is denoted by $\lfloor x \rfloor$. In the same vein, the smallest integer $\geq x$ is called the **ceiling** of x and is denoted by $\lceil x \rceil$. For example, $\lfloor \frac{3}{2} \rfloor = \lfloor 1 \rfloor = 1$, whereas $\lceil \frac{3}{2} \rceil = \lceil 2 \rceil = 2$.

Given any $a \in \mathbb{R}$ and $n \in \mathbb{N}$, we define the nth **power** a^n of a to be the product $a \cdots a$ of a with itself taken n times. Further, we define $a^0 = 1$ and $a^{-n} = (1/a)^n$ provided $a \neq 0$. In this way integral powers of all nonzero real numbers are defined. The following elementary properties are immediate consequences of the algebraic properties and the order properties of $\mathbb{R}$.

(i) $(a_1 a_2)^n = a_1^n a_2^n$ for all $n \in \mathbb{Z}$ and $a_1, a_2 \in \mathbb{R}$ (with $a_1 a_2 \neq 0$ if $n \leq 0$).
(ii) $(a^m)^n = a^{mn}$ and $a^{m+n} = a^m a^n$ for all $m, n \in \mathbb{Z}$ and $a \in \mathbb{R}$ (with $a \neq 0$ if $m \leq 0$ or $n \leq 0$).
(iii) If $n \in \mathbb{N}$ and $b_1, b_2 \in \mathbb{R}$ with $0 \leq b_1 < b_2$, then $b_1^n < b_2^n$.

The first two properties above are sometimes called the **laws of exponents** or the **laws of indices** (for integral powers).

Proposition 1.4. *Given any $n \in \mathbb{N}$ and $a \in \mathbb{R}$ with $a \geq 0$, there exists a unique $b \in \mathbb{R}$ such that $b \geq 0$ and $b^n = a$.*

Proof. Uniqueness is clear since $b_1, b_2 \in \mathbb{R}$ with $0 \leq b_1 < b_2$ implies that $b_1^n < b_2^n$. To prove the existence of $b \in \mathbb{R}$ with $b \geq 0$ and $b^n = a$, note that the case of $a = 0$ is trivial, and moreover, the case of $0 < a < 1$ follows from the case of $a > 1$ by taking reciprocals. Thus we will assume that $a \geq 1$. Let

$$S_a = \{x \in \mathbb{R} : x^n \leq a\}.$$

Then S_a is a subset of $\mathbb{R}$, which is nonempty (since $1 \in S_a$) and bounded above (by a, for example). Define $b = \sup S_a$. Note that since $1 \in S_a$, we have $b \geq 1 > 0$. We will show that $b^n = a$ by showing that each of the inequalities $b^n < a$ and $b^n > a$ leads to a contradiction.

Note that by Binomial Theorem, for any $\delta \in \mathbb{R}$, we have

$$(b + \delta)^n = b^n + \binom{n}{1} b^{n-1} \delta + \binom{n}{2} b^{n-2} \delta^2 + \cdots + \delta^n.$$

Now, suppose $b^n < a$. Let us define

$$\epsilon := a - b^n, \quad M := \max\left\{ \binom{n}{k} b^{n-k} : k = 1, \ldots, n \right\} \quad \text{and} \quad \delta := \min\left\{ 1, \frac{\epsilon}{nM} \right\}.$$

Then $M \geq 1$ and $0 < \delta^k \leq \delta$ for $k = 1, 2, \ldots, n$. Therefore,

$$(b + \delta)^n \leq b^n + M\delta + M\delta^2 + \cdots + M\delta^n \leq b^n + nM\delta \leq b^n + \epsilon = a.$$

Hence, $b + \delta \in S_a$. But this is a contradiction since b is an upper bound of S_a.

Next, suppose $b^n > a$. This time, take $\epsilon = b^n - a$ and define M and δ as before. Similar arguments will show that

$$(b - \delta)^n \geq b^n - nM\delta \geq b^n - \epsilon = a.$$

But $b - \delta < b$, and hence $b - \delta$ cannot be an upper bound of S_a. This means that there is some $x \in S_a$ such that $b - \delta < x$. Therefore, $(b - \delta)^n < x^n \leq a$, which is a contradiction. Thus $b^n = a$. $\qquad\qquad\square$

Thanks to Proposition 1.4, we define, for any $n \in \mathbb{N}$ and $a \in \mathbb{R}$ with $a \geq 0$, the nth **root** of a to be the unique real number b such that $b \geq 0$ and $b^n = a$; we denote this real number by $\sqrt[n]{a}$ or by $a^{1/n}$. In case $n = 2$, we simply write $\sqrt{a}$ instead of $\sqrt[2]{a}$. From the uniqueness of the nth root, the analogues of the properties (i), (ii), and (iii) stated just before Proposition 1.4 can be easily proved for nth roots instead of the nth powers. More generally, given any $r \in \mathbb{Q}$, we write $r = m/n$, where $m, n \in \mathbb{Z}$ with $n > 0$, and define $a^r = (a^m)^{1/n}$ for any $a \in \mathbb{R}$ with $a > 0$. Note that if also $r = p/q$, for some $p, q \in \mathbb{Z}$ with $q > 0$, then for any $a \in \mathbb{R}$ with $a > 0$, we have $(a^m)^{1/n} = (a^p)^{1/q}$. This can be seen, for example, by raising both sides to the nqth power, using laws of exponents for integral powers and the uniqueness of roots. Thus, rational powers of positive real numbers are unambiguously defined. In general, for negative real numbers, nonintegral rational powers are not defined in $\mathbb{R}$. For example, $(-1)^{1/2}$ cannot equal any $b \in \mathbb{R}$ since $b^2 \geq 0$. However, in a special case, rational powers of negative real numbers can be defined. More precisely, if $n \in \mathbb{N}$ is odd and $a \in \mathbb{R}$ is positive, then we define

$$(-a)^{1/n} = -\left(a^{1/n}\right).$$

It is easily seen that this is well defined, and as a result, for any $x \in \mathbb{R}$, $x \neq 0$, the rth **power** x^r is defined whenever $r \in \mathbb{Q}$ has an odd denominator, that is, when $r = m/n$ for some $m \in \mathbb{Z}$ and $n \in \mathbb{N}$ with n odd. Finally, if r is any positive rational number, then we set $0^r = 0$. For rational powers, wherever they are defined, analogues of the properties (i), (ii), and (iii) stated just before Proposition 1.4 are valid. These analogues can be easily proved by raising both sides of the desired equality or inequality to sufficiently high integral powers so as to reduce to the corresponding properties of integral powers, and using the uniqueness of roots.

Real numbers that are not rational numbers are called **irrational numbers**. The possibility of taking nth roots provides a useful method to construct several examples of irrational numbers. For instance, we prove below a classical result that $\sqrt{2}$ is an irrational number. The proof here is such that it can easily be adapted to prove that several such numbers, for example, $\sqrt{3}, \sqrt{15}, \sqrt[3]{2}, \sqrt[5]{16}$, are not rational. [See Exercises 10 and 44.] We recall first the familiar notion of divisibility in the set $\mathbb{Z}$ of integers. Given $m, n \in \mathbb{Z}$, we say that m **divides** n or that m is a **factor** of n (and write $m \mid n$) if $n = \ell m$ for some $\ell \in \mathbb{Z}$. Sometimes, we write $m \nmid n$ if m does not divide n. Two integers m and n are said to be **relatively prime** if the only integers that divide both m and n are 1 and -1. It can be shown that if $m, n, n' \in \mathbb{Z}$ are such that m, n are relatively prime and $m \mid nn'$, then $m \mid n'$. It can also be shown that any rational number r can be written as

$$r = \frac{p}{q}, \quad \text{where } p, q \in \mathbb{Z}, \ q > 0, \text{ and } p, q \text{ are relatively prime.}$$

The above representation of r is called the **reduced form** of r. The numerator (namely, p) and the denominator (namely, q) in the case of a reduced form representation are uniquely determined by r.

Proposition 1.5. *No rational number has a square equal to* 2. *In other words,* $\sqrt{2}$ *is an irrational number.*

Proof. Suppose $\sqrt{2}$ is rational. Write $\sqrt{2}$ in the reduced form as p/q, where $p, q \in \mathbb{Z}$, $q > 0$, and p, q are relatively prime. Then $p^2 = 2q^2$. Hence q divides p^2. This implies that q divides p, and so p/q is an integer. But there is no integer whose square is 2 because $(\pm 1)^2 = 1$ and the square of any integer other than 1 or -1 is ≥ 4. Hence $\sqrt{2}$ is not rational. $\qquad \square$

The following result shows that the rational numbers as well as the irrational numbers spread themselves rather densely on the number line.

Proposition 1.6. *Given any* $a, b \in \mathbb{R}$ *with* $a < b$, *there exists a rational number as well as an irrational number between* a *and* b.

Proof. By Proposition 1.3, we can find $n \in \mathbb{N}$ such that $n > 1/(b - a)$. Let $m = [na] + 1$. Then $m - 1 \leq na < m$, and hence

$$a < \frac{m}{n} \leq \frac{na + 1}{n} = a + \frac{1}{n} < a + (b - a) = b.$$

Thus we have found a rational number (namely, m/n) between a and b. Now, $a + \sqrt{2} < b + \sqrt{2}$, and if r is a rational number between $a + \sqrt{2}$ and $b + \sqrt{2}$, then $r - \sqrt{2}$ is an irrational number between a and b. $\qquad \square$

We shall now introduce some basic terminology that is useful in dealing with real numbers. Given any $a, b \in \mathbb{R}$, we define the **open interval** from a to b to be the set

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

and the **closed interval** from a to b to be the set

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

The **semiopen** or the **semiclosed** intervals from a to b are defined by

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

In other words, $(a, b] := [a, b] \setminus \{a\}$ and $[a, b) := [a, b] \setminus \{b\}$. Note that if $a > b$, then each of these intervals is empty, whereas if $a = b$, then $[a, b] = \{a\}$ while the other intervals from a to b are empty. If I is a subset of $\mathbb{R}$ of the form $[a, b]$, (a, b), $[a, b)$ or $(a, b]$, where $a, b \in \mathbb{R}$ with $a < b$, then a is called the **left (hand) endpoint** of I while b is called the **right (hand) endpoint** of I. Collectively, a and b are called the **endpoints** of I.

It is often useful to consider the symbols ∞ (called **infinity**) and $-\infty$ (called **minus infinity**), which may be thought as the fictional (right and left) endpoints of the number line. Thus

$$-\infty < a < \infty \quad \text{for all } a \in \mathbb{R}.$$

The set $\mathbb{R}$ together with the additional symbols ∞ and $-\infty$ is sometimes called the set of **extended real numbers**. We use the symbols ∞ and $-\infty$ to define, for any $a \in \mathbb{R}$, the following **semi-infinite intervals**:

$$(-\infty, a) := \{x \in \mathbb{R} : x < a\}, \quad (-\infty, a] := \{x \in \mathbb{R} : x \leq a\}$$

and

$$(a, \infty) := \{x \in \mathbb{R} : x > a\}, \quad [a, \infty) := \{x \in \mathbb{R} : x \geq a\}.$$

The set $\mathbb{R}$ can also be thought of as the doubly infinite interval $(-\infty, \infty)$, and as such we may sometimes use this interval notation for the set of all real numbers.

It may be noted that each of the above types of intervals has a basic property in common. We state this in the form of the following definition.

Let $I \subseteq \mathbb{R}$, that is, let I be a subset of $\mathbb{R}$. We say that I is an **interval** if

$$a, b \in I \text{ and } a < b \Longrightarrow [a, b] \subseteq I.$$

In other words, the line segment connecting any two points of I is in I. This is sometimes expressed by saying that an interval is a 'connected set'.

Proposition 1.7. *If $I \subseteq \mathbb{R}$ is an interval, then I is either an open interval or a closed interval or a semiopen interval or a semi-infinite interval or the doubly infinite interval.*

Proof. If $I = \emptyset$, then $I = (a, a)$ for any $a \in \mathbb{R}$. Suppose $I \neq \emptyset$. Define

$$a := \begin{cases} \inf I & \text{if } I \text{ is bounded below,} \\ -\infty & \text{otherwise,} \end{cases} \quad \text{and } b := \begin{cases} \sup I & \text{if } I \text{ is bounded above,} \\ \infty & \text{otherwise.} \end{cases}$$

Note that by the Completeness Property and Proposition 1.2, both a and b are well defined and $a \leq b$. Since I is an interval, it follows that

(i) $I = (a, b)$,   or   (ii) $I = [a, b]$,   or   (iii) $I = [a, b)$,   or   (iv) $I = (a, b]$,

according as (i) $a \notin I$ and $b \notin I$, or (ii) $a \in I$ and $b \in I$, or (iii) $a \in I$ and $b \notin I$, or (iv) $a \notin I$ and $b \in I$. This proves the proposition. $\qquad \square$

In the proof of the above proposition, we have considered intervals that can reduce to the empty set or to a set containing only one point. However, to avoid trivialities, we shall usually refrain from doing so in the sequel. Henceforth, when we write $[a, b]$, (a, b), $[a, b)$ or $(a, b]$, it will be tacitly assumed that a and b are real numbers and $a < b$.

Given any real number a, the **absolute value** or the **modulus** of a is denoted by $|a|$ and is defined by

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

Note that $|a| \geq 0$, $|a| = |-a|$, and $|ab| = |a|\,|b|$ for any $a, b \in \mathbb{R}$. The notion of absolute value can sometimes be useful in describing certain intervals that are symmetric about a point. For example, if $a \in \mathbb{R}$ and ϵ is a positive real number, then

$$(a - \epsilon, a + \epsilon) = \{x \in \mathbb{R} : |x - a| < \epsilon\}.$$

1.2 Inequalities

In this section, we describe and prove some inequalities that will be useful to us in the sequel.

Proposition 1.8 (Basic Inequalities for Absolute Values). *Given any* $a, b \in \mathbb{R}$*, we have*

(i) $|a + b| \leq |a| + |b|$,
(ii) $|\,|a| - |b|\,| \leq |a - b|$.

Proof. It is clear that $a \leq |a|$ and $b \leq |b|$. Thus, $a + b \leq |a| + |b|$. Likewise, $-(a + b) \leq |a| + |b|$. This implies (i). To prove (ii), note that by (i), we have $|a - b| \geq |(a - b) + b| - |b| = |a| - |b|$ and also $|a - b| = |b - a| \geq |b| - |a|$. $\qquad \square$

The first inequality in the proposition above is sometimes referred to as the **triangle inequality**. An immediate consequence of this is that if $a_1, \ldots, a_n$ are any real numbers, then

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

Proposition 1.9 (Basic Inequalities for Powers and Roots). *Given any* $a, b \in \mathbb{R}$ *and* $n \in \mathbb{N}$, *we have*

(i) $|a^n - b^n| \le nM^{n-1}|a - b|$, *where* $M = \max\{|a|, |b|\}$,
(ii) $|a^{1/n} - b^{1/n}| \le |a - b|^{1/n}$, *provided* $a \ge 0$ *and* $b \ge 0$.

Proof. (i) Consider the identity

$$a^n - b^n = (a - b)(a^{n-1}b + a^{n-2}b^2 + \cdots + a^2b^{n-2} + ab^{n-1}).$$

Take the absolute value of both sides and use Proposition 1.8. The absolute value of the second factor on the right is bounded above by nM^{n-1}. This implies the inequality in (i).

(ii) We may assume, without loss of generality, that $a \ge b$. Let $c = a^{1/n}$ and $d = b^{1/n}$. Then $c - d \ge 0$ and by the Binomial Theorem,

$$c^n = [(c - d) + d]^n = (c - d)^n + \cdots + d^n \ge (c - d)^n + d^n.$$

Therefore,

$$a - b = c^n - d^n \ge (c - d)^n = [a^{1/n} - b^{1/n}]^n.$$

This implies the inequality in (ii). □

We remark that the basic inequality for powers in part (i) of Proposition 1.9 is valid, more generally, for rational powers. [See Exercise 54 (i).] As for part (ii), a slightly weaker inequality holds if instead of nth roots, we consider rational roots. [See Exercise 54 (ii).]

Proposition 1.10 (Binomial Inequalities). *Given any* $a \in \mathbb{R}$ *such that* $1 + a \ge 0$, *we have*

$$(1 + a)^n \ge 1 + na \quad \text{for all } n \in \mathbb{N}.$$

More generally, given any $n \in \mathbb{N}$ *and* $a_1, \ldots, a_n \in \mathbb{R}$ *such that* $1 + a_i \ge 0$ *for* $i = 1, \ldots, n$ *and* $a_1, \ldots, a_n$ *all have the same sign, we have*

$$(1 + a_1)(1 + a_2) \cdots (1 + a_n) \ge 1 + (a_1 + \cdots + a_n).$$

Proof. Clearly, the first inequality follows from the second by substituting $a_1 = \cdots = a_n = a$. To prove the second inequality, we use induction on n. The case of $n = 1$ is obvious. If $n > 1$ and the result holds for $n - 1$, then

$$(1 + a_1)(1 + a_2) \cdots (1 + a_n) \ge (1 + b_n)(1 + a_n),$$

where $b_n = a_1 + \cdots + a_{n-1}$. Now, b_n and a_n have the same sign, and hence

$$(1 + b_n)(1 + a_n) = 1 + b_n + a_n + b_n a_n \ge 1 + b_n + a_n.$$

This proves that $(1 + a_1)(1 + a_2) \cdots (1 + a_n) \ge 1 + (a_1 + \cdots + a_n).$ □

Note that the first inequality in the proposition above is an immediate consequence of the Binomial Theorem when $a \geq 0$, although we have proved it in the more general case of $a \geq -1$. We shall refer to the first inequality in Proposition 1.10 as the **binomial inequality**. On the other hand, we shall refer to the second inequality in Proposition 1.10 as the **generalized binomial inequality**. We remark that the binomial inequality is valid, more generally, for rational powers. [See Exercise 54 (iii).]

Proposition 1.11 (A.M.-G.M. Inequality). *Let $n \in \mathbb{N}$ and let $a_1, \ldots, a_n$ be nonnegative real numbers. Then the arithmetic mean of $a_1, \ldots, a_n$ is greater than or equal to their geometric mean, that is,*

$$\frac{a_1 + \cdots + a_n}{n} \geq \sqrt[n]{a_1 \cdots a_n}.$$

Moreover, equality holds if and only if $a_1 = \cdots = a_n$.

Proof. If some $a_i = 0$, then the result is obvious. Hence we shall assume that $a_i > 0$ for $i = 1, \ldots, n$. Let $g = (a_1 \cdots a_n)^{1/n}$ and $b_i = a_i/g$ for $i = 1, \ldots, n$. Then $b_1, \ldots, b_n$ are positive and $b_1 \cdots b_n = 1$. We shall now show, using induction on n, that $b_1 + \cdots + b_n \geq n$. This is clear if $n = 1$ or if each of $b_1, \ldots, b_n$ equals 1. Suppose $n > 1$ and not every b_i equals 1. Then $b_1 \cdots b_n = 1$ implies that among $b_1, \ldots, b_n$ there is a number < 1 as well as a number > 1. Relabeling $b_1, \ldots, b_n$ if necessary, we may assume that $b_1 < 1$ and $b_n > 1$. Let $c_1 = b_1 b_n$. Then $c_1 b_2 \cdots b_{n-1} = 1$, and hence by the induction hypothesis $c_1 + b_2 + \cdots + b_{n-1} \geq n - 1$. Now observe that

$$\begin{aligned}
b_1 + \cdots + b_n &= (c_1 + b_2 + \cdots + b_{n-1}) + b_1 + b_n - c_1 \\
&\geq (n-1) + b_1 + b_n - b_1 b_n \\
&= n + (1 - b_1)(b_n - 1) \\
&> n,
\end{aligned}$$

where the last inequality follows since $b_1 < 1$ and $b_n > 1$. This proves that $b_1 + \cdots + b_n \geq n$, and moreover the inequality is strict unless $b_1 = \cdots = b_n = 1$. Substituting $b_i = a_i/g$, we obtain the desired result. $\square$

Proposition 1.12 (Cauchy–Schwarz Inequality). *Let $n \in \mathbb{N}$ and let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be any real numbers. Then*

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

Moreover, equality holds if and only if $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ are proportional to each other, that is, if $a_i b_j = a_j b_i$ for all $i, j = 1, \ldots, n$.

Proof. Observe that

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

Now for any $\alpha, \beta \in \mathbb{R}$, we have $2\alpha\beta \le \alpha^2 + \beta^2$ and equality holds if and only if $\alpha = \beta$. (This follows by considering $(\alpha - \beta)^2$.) If we apply this to each of the terms in the second summation above, then we obtain

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

and moreover, equality holds if and only if $a_i b_j = a_j b_i$ for all $i, j = 1, \ldots, n$. This proves the desired result. $\qquad\square$

Remark 1.13. Analyzing the argument in the above proof of the Cauchy–Schwarz inequality, we obtain, in fact, the following identity, which is easy to verify directly:

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

This is known as **Lagrange's Identity** and it may be viewed as a one-line proof of Proposition 1.12. See also Exercise 16 for yet another proof. $\qquad\diamond$

1.3 Functions and Their Geometric Properties

The concept of a function is of basic importance in calculus and real analysis. In this section, we begin with an informal description of this concept followed by a precise definition. Next, we outline some basic terminology associated with functions. Later, we give basic examples of functions, including polynomial functions, rational functions, and algebraic functions. Finally, we discuss a number of geometric properties of functions and state some results concerning them. These results are proved here without invoking any of the notions of calculus that are encountered in the subsequent chapters.

Typically, a function is described with the help of an expression in a single parameter (say x), which varies over a stipulated set; this set is called the *domain* of that function. For example, each of the expressions

(i) $f(x) := 2x + 1, \ x \in \mathbb{R},$ (ii) $f(x) := x^2, \ x \in \mathbb{R},$
(iii) $f(x) := 1/x, \ x \in \mathbb{R}, \ x \ne 0,$ (iv) $f(x) := x^3, \ x \in \mathbb{R},$

defines a function f. In (i), (ii), and (iv), the domain is the set $\mathbb{R}$ of all real numbers whereas in (iii), the domain is the set $\mathbb{R} \setminus \{0\}$ of all nonzero real numbers. Note that each of the functions in (i)–(iv) takes its 'values' in the

set $\mathbb{R}$; to indicate this, we say that $\mathbb{R}$ is the *codomain* of these functions or that these are *real-valued functions*.

Given a real-valued function f having a subset D of $\mathbb{R}$ as its domain, it is often useful to consider the **graph** of f, which is defined as the subset $\{(x, f(x)) : x \in D\}$ of the plane $\mathbb{R}^2$. In other words, this is the set of points on the *curve* given by $y = f(x)$, $x \in D$, in the xy-plane. For example, the graphs of the functions in (i) and (ii) are shown in Figure 1.2, while the graphs of the functions in (iii) and (iv) above are shown in Figure 1.3.
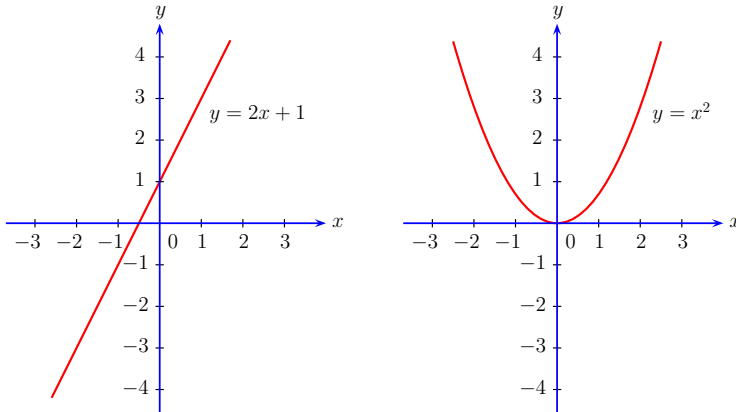
Fig. 1.2. Graphs of $f(x) = 2x + 1$ and $f(x) = x^2$

In general, we can talk about a function from any set D to any set E, and this associates to each point of D a unique element of E. A formal definition of a function is given below. It may be seen that this, in essence, identifies a function with its graph!

Definitions and Terminology

Let D and E be any sets. We denote by $D \times E$ the set of all ordered pairs (x, y) where x varies over elements of D and y varies over elements of E. A **function** from D to E is a subset f of $D \times E$ with the property that for each $x \in D$, there is a unique $y \in E$ such that $(x, y) \in f$. The set D is called the **domain** or the **source** of f and E the **codomain** or the **target** of f.

Usually, we write $f : D \to E$ to indicate that f is a function from D to E. Also, instead of $(x, y) \in f$, we usually write $y = f(x)$, and call $f(x)$ the **value** of f at x. This may also be indicated by writing $x \mapsto f(x)$, and saying that f **maps** x to $f(x)$. Functions $f : D \to E$ and $g : D \to E$ are said to be **equal** and we write $f = g$ if $f(x) = g(x)$ for all $x \in D$.

If $f : D \to E$ is a function, then the subset $f(D) := \{f(x) : x \in D\}$ of E is called the **range** of f. We say that f is **onto** or **surjective** if $f(D) = E$.

On the other hand, if f maps distinct points to distinct points, that is, if

$$x_1, x_2 \in D, \ f(x_1) = f(x_2) \Longrightarrow x_1 = x_2$$

then f is said to be **one-one** or **injective**. If f is both one-one and onto, then it is said to be **bijective** or a **one-to-one correspondence**.

The notion of a bijective function can be used to define a basic terminology concerning sets as follows. Given any nonnegative integer n, consider the set $\{1, \ldots, n\}$ of the first n positive integers. Note that if $n = 0$, then $\{1, \ldots, n\}$ is the empty set. A set D is said to be **finite** if there is a bijective map from $\{1, \ldots, n\}$ onto D, for some nonnegative integer n. In this case the nonnegative integer n is unique (Exercise 18) and it is called the **cardinality** of D or the **number of elements** in D. A set that is not finite is said to be **infinite**.
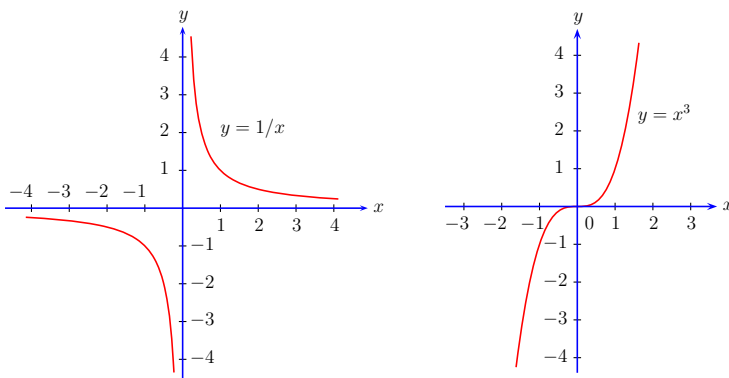
Fig. 1.3. Graphs of $f(x) = 1/x$ and $f(x) = x^3$

The simplest examples of functions defined on arbitrary sets are an identity function and a constant function. Given any set D, the **identity function** on D is the function $\mathrm{id}_D : D \to D$ defined by $\mathrm{id}_D(x) = x$ for all $x \in D$. Given any sets D and E, a function $f : D \to E$ defined by $f(x) = c$ for all $x \in D$, where c is a fixed element of E, is called a **constant function**. Note that id_D is always bijective, whereas a constant function is neither one-one (unless D is a singleton set!) nor onto (unless E is a singleton set!). To look at more specific examples, note that $f : \mathbb{R} \to \mathbb{R}$ defined by (i) or by (iv) above is bijective, while $f : \mathbb{R} \to [0, \infty)$ defined by (ii) is onto but not one-one, and $f : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ defined by (iii) is one-one but not onto.

If $f : D \to E$ and $g : D' \to E'$ are functions with $f(D) \subseteq D'$, then the function $h : D \to E'$ defined by $h(x) = g(f(x))$, $x \in D$, is called the **composite** of g with f, and is denoted by $g \circ f$ [read as g composed with f, or as f followed by g].

Note that any function $f : D \to E$ can be made an onto function by replacing the codomain E with its range $f(D)$; more formally, this may be

done by looking at the function $\tilde{f} : D \to f(D)$ defined by $\tilde{f}(x) = f(x)$, $x \in D$. In particular, if $f : D \to E$ is one-one, then for every $y \in f(D)$, there exists a unique $x \in D$ such that $f(x) = y$. In this case, we write $x = f^{-1}(y)$. We thus obtain a function $f^{-1} : f(D) \to D$ such that $f^{-1} \circ f = \mathrm{id}_D$ and $f \circ f^{-1} = \mathrm{id}_{f(D)}$. We call f^{-1} the **inverse function** of f.

For example, the inverse of $f : \mathbb{R} \to \mathbb{R}$ defined by (i) above is the function $f^{-1} : \mathbb{R} \to \mathbb{R}$ given by $f^{-1}(y) = (y-1)/2$ for $y \in \mathbb{R}$, whereas the inverse of $f : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ defined by (iii) above is the function $f^{-1} : \mathbb{R} \setminus \{0\} \to \mathbb{R} \setminus \{0\}$ given by $f^{-1}(y) = 1/y$ for $y \in \mathbb{R} \setminus \{0\}$.

In general, if a function $f : D \to E$ is not one-one, then we cannot talk about its inverse. However, sometimes it is possible to restrict the domain of a function to a smaller set and then a 'restriction' of f may become injective. For any subset C of D, the **restriction** of f to C is the function $f_{|C} : C \to E$, defined by $f_{|C}(x) = f(x)$ for $x \in C$. For example, if $f : \mathbb{R} \to \mathbb{R}$ is the function defined by (ii), then f is not one-one but its restriction $f|_{[0,\infty)}$ is one-one and its inverse $g = \left(f_{|[0,\infty)}\right)^{-1}$ is given by $g(y) = \sqrt{y}$ for $y \in [0, \infty)$.

Suppose $D \subseteq \mathbb{R}$ is **symmetric** about the origin, that is, we have $-x \in D$ whenever $x \in D$. For example, D can be the whole real line $\mathbb{R}$ or an interval of the form $[-a, a]$ or the punctured real line $\mathbb{R} \setminus \{0\}$. A function $f : D \to \mathbb{R}$ is said to be an **even function** if $f(-x) = f(x)$ for all $x \in D$, whereas f is said to be an **odd function** if $f(-x) = -f(x)$ for all $x \in D$. For example, $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$ is an even function, whereas $f : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ defined by $f(x) = 1/x$ and $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^3$ are both odd functions. On the other hand, $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = 2x + 1$ is neither even nor odd.

Geometrically speaking, given $D \subseteq \mathbb{R}$ and $f : D \to \mathbb{R}$, the fact that f is a function corresponds to the property that for every $x_0 \in D$, the vertical line $x = x_0$ in the xy-plane meets the graph of f in exactly one point. Further, the property that f is one-one corresponds to requiring, in addition, that for any $y_0 \in \mathbb{R}$, the horizontal line $y = y_0$ meet the graph of f in at most one point. On the other hand, the property that a point $y_0 \in \mathbb{R}$ is in the range $f(D)$ of f corresponds to requiring, in addition, that the horizontal line $y = y_0$ meet the graph of f in at least one point. In case the inverse function $f^{-1} : f(D) \to \mathbb{R}$ exists, then its graph is obtained from that of f by reflecting along the diagonal line $y = x$. Assuming that D is symmetric, to say that f is an even function corresponds to saying that the graph of f is symmetric with respect to the y-axis, whereas to say that f is an odd function corresponds to saying that the graph of f is symmetric with respect to the origin. Notice that if f is odd and one-one, then its range $f(D)$ is also symmetric, and $f^{-1} : f(D) \to \mathbb{R}$ is an odd function.

Given any real-valued functions f, $g : D \to \mathbb{R}$, we can associate new functions $f + g : D \to \mathbb{R}$ and $fg : D \to \mathbb{R}$, called respectively the **sum** and the **product** of f and g, which are defined componentwise, that is, by

$$(f+g)(x) = f(x) + g(x) \quad \text{and} \quad (fg)(x) = f(x)g(x) \quad \text{for } x \in D.$$

In case f is the constant function given by $f(x) = c$ for all $x \in D$, then fg is often denoted by cg and called the **multiple** of g (by c). We often write $f - g$ in place of $f + (-1)g$. In case $g(x) \neq 0$ for all $x \in D$, the **quotient** f/g is defined and this is a function from D to $\mathbb{R}$ given by $(f/g)(x) = f(x)/g(x)$ for $x \in D$. Sometimes, we write $f \leq g$ to mean that $f(x) \leq g(x)$ for all $x \in D$.

Basic Examples of Functions

Among the most basic functions are those that are obtained from polynomials. Let us first review some relevant algebraic facts about polynomials.

A **polynomial** (in one variable x) with real coefficients is an expression[4] of the form

$$c_n x^n + c_{n-1} x^{n-1} + \cdots + c_1 x + c_0,$$

where n is a nonnegative integer and $c_0, c_1, \ldots, c_n$ are real numbers. We call $c_0, c_1, \ldots, c_n$ the **coefficients** of the above polynomial and more specifically, c_i as the **coefficient** of x^i for $i = 0, 1, \ldots, n$. In case $c_n \neq 0$, the polynomial is said to have **degree** n, and c_n is said to be its **leading coefficient**. A polynomial (in x) whose leading coefficient is 1 is said to be **monic** (in x). Two polynomials are said to be equal if the corresponding coefficients are equal. In particular, $c_n x^n + \cdots + c_1 x + c_0$ is the **zero polynomial** if and only if $c_0 = c_1 = \cdots = c_n = 0$. The degree of the zero polynomial is not defined. If $p(x)$ is a nonzero polynomial, then its degree is denoted by $\deg p(x)$. Polynomials of degrees 1, 2, and 3 are often referred to as **linear**, **quadratic**, and **cubic** polynomials, respectively. Polynomials of degree zero as well as the zero polynomial are called **constant polynomials**. The set of all polynomials in x with real coefficients is denoted by $\mathbb{R}[x]$. Addition and multiplication of polynomials is defined in a natural manner. For example,

$$(x^2 + 2x + 3) + (x^3 + 2x^2 + 5) = x^3 + 3x^2 + 2x + 8$$

and

$$(x^2 + 2x + 3)(x^3 + 2x^2 + 5) = x^5 + 4x^4 + 7x^3 + 11x^2 + 10x + 15.$$

---

[4] For those who consider 'expression' a vague term and wonder what x really is, a formal and pedantic definition of a polynomial (in one variable) can be given as follows. A polynomial with real coefficients is a function from the set $\{0, 1, 2, \ldots\}$ of nonnegative integers into $\mathbb{R}$ such that all except finitely many nonnegative integers are mapped to zero. Thus, the expression $c_n x^n + \cdots + c_1 x + c_0$ corresponds to the function which sends 0 to c_0, 1 to c_1, $\ldots$, n to c_n and m to 0 for all $m \in \mathbb{N}$ with $m > n$. In this set up, one can *define* x to be the unique function that maps 1 to 1, and all other nonnegative integers to 0. More generally, we may define x^n to be the function that maps n to 1, and all other integer to 0. We may also identify a real number a with the function that maps 0 to a and all the positive integers to 0. Now, with componentwise addition of functions, $c_n x^n + \cdots + c_1 x + c_0$ has a formal meaning, which is in accord with our intuition!

In general, for any $p(x), q(x) \in \mathbb{R}[x]$, the sum $p(x) + q(x)$ and the product $p(x)q(x)$ are polynomials in $\mathbb{R}[x]$. Moreover, if $p(x)$ and $q(x)$ are nonzero, then so is $p(x)q(x)$ and $\deg(p(x)q(x)) = \deg p(x) + \deg q(x)$, whereas $p(x) + q(x)$ is either the zero polynomial or $\deg(p(x) + q(x)) \leq \max\{\deg p(x), \deg q(x)\}$. We say that $q(x)$ **divides** $p(x)$ and write $q(x) \mid p(x)$ if $p(x) = q(x)r(x)$ for some $r(x) \in \mathbb{R}[x]$. We may write $q(x) \nmid p(x)$ if $q(x)$ does not divide $p(x)$.

If $p(x) = c_n x^n + \cdots + c_1 x + c_0 \in \mathbb{R}[x]$ and $\alpha \in \mathbb{R}$, then we denote by $p(\alpha)$ the real number $c_n \alpha^n + \cdots + c_1 \alpha + c_0$ and call it the **evaluation** of $p(x)$ at α. In case $p(\alpha) = 0$, we say that α is a (real) **root** of $p(x)$. There do exist polynomials with no real roots. For example, the quadratic polynomial $x^2 + 1$ has no real root since $\alpha^2 + 1 \geq 1 > 0$ for all $\alpha \in \mathbb{R}$. More generally, if $q(x) = ax^2 + bx + c$ is any quadratic polynomial (so that $a \neq 0$), then we have

$$4aq(x) = (2ax + b)^2 - (b^2 - 4ac).$$

Consequently, $q(x)$ has a real root if and only if $b^2 - 4ac \geq 0$; indeed, if $b^2 - 4ac \geq 0$, then $\left(-b \pm \sqrt{b^2 - 4ac}\right)/2a$ are the roots of $q(x)$. We call $b^2 - 4ac$ the **discriminant** of the quadratic polynomial $q(x) = ax^2 + bx + c$.

Quotients of polynomials, that is, expressions of the form $p(x)/q(x)$, where $p(x)$ is a polynomial and $q(x)$ is a nonzero polynomial, are called **rational functions**. Two rational functions $p_1(x)/q_1(x)$ and $p_2(x)/q_2(x)$ are regarded as equal if upon cross-multiplying, the corresponding polynomials are equal, that is, if $p_1(x)q_2(x) = p_2(x)q_1(x)$. Sums and products of rational functions are defined in a natural manner. Basic facts about polynomials and rational functions are as follows:

 (i) If a nonzero polynomial has degree n, then it has at most n roots. Consequently, if $p(x)$ is a polynomial with real coefficients such that $p(\alpha) = 0$ for all α in an infinite subset D of $\mathbb{R}$, then $p(x)$ is the zero polynomial.
 (ii) [**Real Fundamental Theorem of Algebra**] Every nonzero polynomial with real coefficients can be factored as a finite product of linear polynomials and quadratic polynomials with negative discriminants.
(iii) [**Partial Fraction Decomposition**] Every rational function can be decomposed as the sum of a polynomial and finitely many rational functions of the form
$$\frac{A}{(x - \alpha)^i} \quad \text{or} \quad \frac{Bx + C}{(x^2 + \beta x + \gamma)^j},$$
where A, B, C and α, β, γ are real numbers and i, j are positive integers.

The factorization in (ii) is, in fact, unique up to a rearrangement of terms. In (iii), we can choose $(x - \alpha)^i$ and $(x^2 + \beta x + \gamma)^j$ to be among the factors of the denominator of the given rational function and in that case the partial fraction decomposition is also unique up to a rearrangement of terms. See Exercises 60 and 67 (and some of the preceding exercises) for a proof of (i) and (iii) above. See also Exercise 69 for more on (ii) above. A simple and useful example of partial fraction decomposition is obtained by taking any distinct real numbers α, β and noting that

$$\frac{1}{(x-\alpha)(x-\beta)} = \frac{A_1}{x-\alpha} + \frac{A_2}{x-\beta}, \text{ where } A_1 = \frac{1}{\alpha-\beta} \text{ and } A_2 = \frac{1}{\beta-\alpha}.$$

More generally, if $p(x), q(x)$ are polynomials with $\deg p(x) < \deg q(x)$ and $q(x) = (x-\alpha_1)\cdots(x-\alpha_k)$ where $\alpha_1, \ldots, \alpha_k$ are distinct real numbers, then

$$\frac{p(x)}{q(x)} = \frac{A_1}{x-\alpha_1} + \cdots + \frac{A_r}{x-\alpha_k} \quad \text{where } A_i = \frac{p(\alpha_i)}{\prod_{j\neq i}(\alpha_i - \alpha_j)} \text{ for } i = 1, \ldots, r.$$

This, then, is the partial fraction decomposition of $p(x)/q(x)$. In general, the partial fraction decomposition of a rational function can be more complicated. A typical example is the following:

$$\frac{x^5 - 4x^4 + 8x^3 - 13x^2 + 3x - 7}{x^4 - 3x^3 + x^2 + 4} = (x-1) + \frac{2}{(x-2)} - \frac{3}{(x-2)^2} + \frac{2x+1}{(x^2+x+1)}.$$

Now let us revert to functions. Evaluating polynomials at real numbers, we obtain functions known as polynomial functions. Thus, if $D \subseteq \mathbb{R}$, then a **polynomial function** on D is a function $f : D \to \mathbb{R}$ given by

$$f(x) = c_n x^n + c_{n-1} x^{n-1} + \cdots + c_1 x + c_0 \quad \text{for } x \in D,$$

where n is a nonnegative integer and $c_0, c_1, \ldots, c_n$ are real numbers. Alternatively, we can view the polynomial functions on D as the class of functions obtained from the identity function on D and the constant functions from D to $\mathbb{R}$ by the construction of forming sums and products of functions. If D is an infinite set, then it follows from (i) above that a polynomial function on D and the corresponding polynomial determine each other uniquely. In this case, it is possible to identify them with each other, and permit polynomial functions to inherit some of the terminology applicable to polynomials. For example, a polynomial function is said to have **degree** n if the corresponding polynomial has degree n.

Rational functions give rise to real-valued functions on subsets D of $\mathbb{R}$ provided their denominators do not vanish at any point of D. Thus, a **rational function** on D is a function $f : D \to \mathbb{R}$ such that $f(x) = p(x)/q(x)$ for $x \in D$, where p and q are polynomial functions on D with $q(x) \neq 0$ for all $x \in D$.

Polynomial functions and rational functions (on $D \subseteq \mathbb{R}$) are special cases of *algebraic functions* (on D), which are defined as follows. A function $f : D \to \mathbb{R}$ is said to be an **algebraic function** if $y = f(x)$ satisfies an equation whose coefficients are polynomials, that is,

$$p_n(x)y^n + p_{n-1}(x)y^{n-1} + \cdots + p_1(x)y + p_0(x) = 0 \quad \text{for } x \in D,$$

where $n \in \mathbb{N}$ and $p_0(x), p_1(x), \ldots, p_n(x)$ are polynomials such that $p_n(x)$ is a nonzero polynomial. For example, the function $f : [0, \infty) \to \mathbb{R}$ defined by $f(x) := \sqrt[n]{x}$ is an algebraic function since $y = f(x)$ satisfies the equation

$y^n - x = 0$ for $x \in [0, \infty)$. It can be shown[5] that sums, products, and quotients of algebraic functions are algebraic. Here is a simple example that illustrates why such a property is true. Consider the sum $y = \sqrt{x} + \sqrt{x+1}$ of functions that are clearly algebraic. To show that this sum is algebraic, write $y - \sqrt{x} = \sqrt{x+1}$, square both sides, and simplify to get $y^2 - 1 = 2y\sqrt{x}$; now squaring once again we obtain the equation $y^4 - 2(1 + 2x)y^2 + 1 = 0$, which is of the desired type. Algebraic functions also have the property that their radicals are algebraic. More precisely, if $f : D \to \mathbb{R}$ is algebraic and $f(x) \geq 0$ for all $x \in D$, then any root of f is algebraic, that is, for any $d \in \mathbb{N}$ the function $g : D \to \mathbb{R}$ defined by $g(x) := f(x)^{1/d}$ is algebraic. This follows simply by changing y to y^d in the algebraic equation satisfied by $y = f(x)$, and noting that the resulting equation is satisfied by $y = g(x)$. It is seen, therefore, that algebraic functions constitute a fairly large class of functions, which is *closed* under the basic operations of algebra. This class may be viewed as a basic stockpile of functions from which various examples can be drawn. A real-valued function that is not algebraic is called a **transcendental function**. The transcendental functions are also important in calculus and we will discuss them in greater detail in Chapter 7.
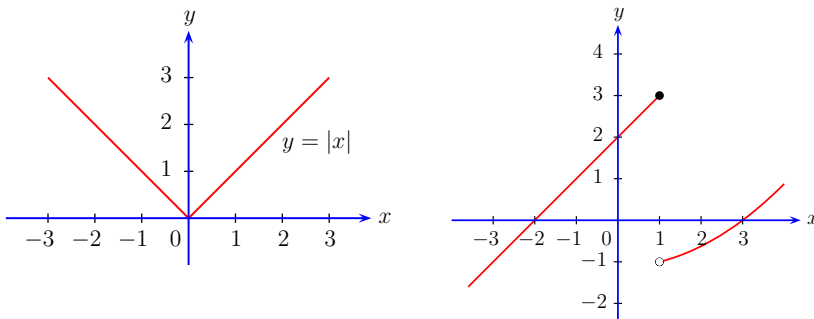
Fig. 1.4. Graphs of $f(x) := |x|$ and $f(x) := \begin{cases} x + 2 & \text{if } x \leq 1, \\ (x^2 - 9)/8 & \text{if } x > 1 \end{cases}$

Apart from algebra, a fruitful way to construct new functions is by piecing together known functions. For example, consider $f : \mathbb{R} \to \mathbb{R}$ defined by either of the following.

(i) $f(x) := |x| = \begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x < 0; \end{cases}$ (ii) $f(x) := \begin{cases} x + 2 & \text{if } x \leq 1, \\ (x^2 - 9)/8 & \text{if } x > 1. \end{cases}$

The graphs of these functions may be drawn as in Figure 1.4. Taking the integer part or the floor of a real number gives rise to a function $f : \mathbb{R} \to \mathbb{R}$

---

[5] A general proof of this requires some ideas from algebra. The interested reader is referred to [16] or [30].

defined by $f(x) := [x]$, which we refer to as the **integer part function** or the **floor function**. Likewise, $g : \mathbb{R} \to \mathbb{R}$ given by $g(x) := \lceil x \rceil$ is called the **ceiling function**. These two functions may also be viewed as examples of functions obtained by piecing together known functions, and their graphs are shown in Figure 1.5. As seen in Figures 1.4 and 1.5, it is often the case that the graphs of functions defined by piecing together different functions look broken or have beak-like edges. Also, in general, such functions are not algebraic. Nevertheless, such functions can be quite useful in constructing examples of certain 'wild behavior'.
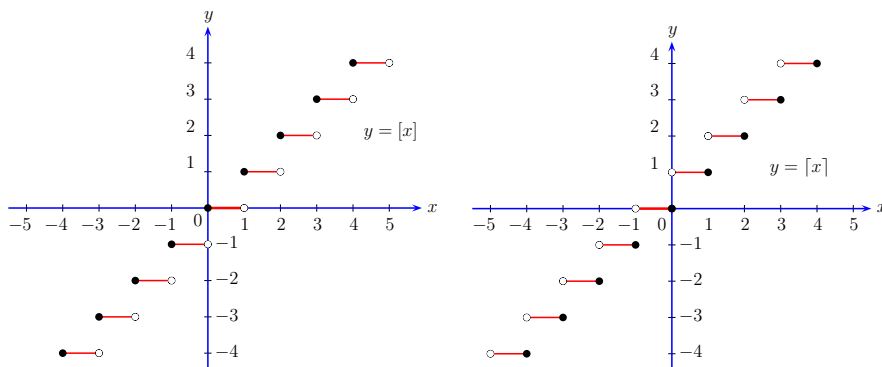
Fig. 1.5. Graphs of the integer part function $[x]$ and the ceiling function $\lceil x \rceil$

Remark 1.14. Polynomials (in one variable) are analogous to integers. Likewise, rational functions are analogous to rational numbers. Algebraic functions and transcendental functions also have analogues in arithmetic, which are defined as follows. A real number α is called an **algebraic number** if it satisfies a nonzero polynomial with integer coefficients. Numbers that are not algebraic are called **transcendental numbers**. For example, it can be easily seen that $\sqrt{2}, \sqrt{3}, \sqrt[5]{7}, \sqrt{2} + \sqrt{3}$ are algebraic numbers. Also, every rational number is an algebraic number. On the other hand, it is not easy to give concrete examples of transcendental numbers. Those interested are referred to the book of Baker [7] for the proof of transcendence of several well-known numbers. $\diamond$

We shall now discuss a number of geometric properties of real-valued functions defined on certain subsets of $\mathbb{R}$.

Bounded Functions

The notion of a bounded set has an analogue in the case of functions. In effect, we use for functions the terminology that is applicable to their range. More precisely, we make the following definitions.

Let $D \subseteq \mathbb{R}$ and $f : D \to \mathbb{R}$ be a function.

1. f is said to be **bounded above** on D if there is $\alpha \in \mathbb{R}$ such that $f(x) \leq \alpha$ for all $x \in D$. Any such α is called an **upper bound** for f.
2. f is said to be **bounded below** on D if there is $\beta \in \mathbb{R}$ such that $f(x) \geq \beta$ for all $x \in D$. Any such β is called a **lower bound** for f.
3. f is said to be **bounded** on D if it is bounded above on D and also bounded below on D.

Notice that f is bounded on D if and only if there is $\gamma \in \mathbb{R}$ such that $|f(x)| \leq \gamma$ for all $x \in D$. Any such γ is called a **bound** for the absolute value of f. Geometrically speaking, f is bounded above means that the graph of f lies below some horizontal line, while f is bounded below means that its graph lies above some horizontal line.

For example, $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) := -x^2$ is bounded above on $\mathbb{R}$, while $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) := x^2$ is bounded below on $\mathbb{R}$. However, neither of these functions is bounded on $\mathbb{R}$. On the other hand, $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) := x^2/(x^2+1)$ gives an example of a function that is bounded on $\mathbb{R}$. For this function, we see readily that $0 \leq f(x) < 1$ for all $x \in \mathbb{R}$. The bounds 0 and 1 are, in fact, optimal in the sense that

$$\inf\{f(x) : x \in \mathbb{R}\} = 0 \quad \text{and} \quad \sup\{f(x) : x \in \mathbb{R}\} = 1.$$

Of these, the first equality is obvious since $f(x) \geq 0$ for all $x \in \mathbb{R}$ and $f(0) = 0$. To see the second equality, let α be an upper bound such that $\alpha < 1$. Then $1 - \alpha > 0$ and so we can find $n \in \mathbb{N}$ such that

$$\frac{1}{n} < 1 - \alpha \quad \text{and hence} \quad f(\sqrt{n-1}) = \frac{n-1}{n} = 1 - \frac{1}{n} > \alpha,$$

which is a contradiction. This shows that $\sup\{f(x) : x \in \mathbb{R}\} = 1$. Thus there is a qualitative difference between the infimum of (the range of) f, which is attained, and the supremum, which is not attained. This suggests the following general definition.

Let $D \subseteq \mathbb{R}$ and $f : D \to \mathbb{R}$ be a function. We say that

1. f **attains its upper bound** on D if there is $c \in D$ such that

$$\sup\{f(x) : x \in D\} = f(c),$$

2. f **attains its lower bound** on D if there is $d \in D$ such that

$$\inf\{f(x) : x \in D\} = f(d),$$

3. f **attains its bounds** on D if it attains its upper bound on D and also attains its lower bound on D.

In case f attains its upper bound, we may write $\max\{f(x) : x \in D\}$ in place of $\sup\{f(x) : x \in D\}$. Likewise, if f attains its lower bound, then "inf" may be replaced by "min".

Monotonicity, Convexity, and Concavity

Monotonicity is a geometric property of a real-valued function defined on a subset of $\mathbb{R}$ that corresponds to its graph being increasing or decreasing. For example, consider Figure 1.6, where the graph on the left is increasing while that on the right is decreasing.
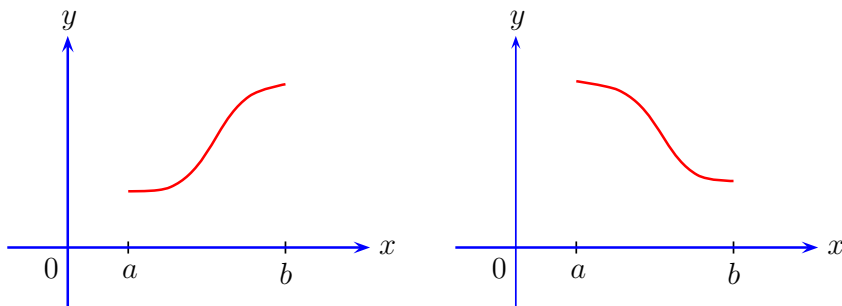
Fig. 1.6. Typical graphs of increasing and decreasing functions on $I = [a, b]$

A formal definition is as follows. Let $D \subseteq \mathbb{R}$ be such that D contains an interval I and $f : D \to \mathbb{R}$ be a function. We say that

1. f is (**monotonically**) **increasing** on I if

$$x_1, x_2 \in I, \ x_1 < x_2 \Longrightarrow f(x_1) \le f(x_2),$$

2. f is (**monotonically**) **decreasing** on I if

$$x_1, x_2 \in I, \ x_1 < x_2 \Longrightarrow f(x_1) \ge f(x_2),$$

3. f is **monotonic** on I if f is monotonically increasing on I or f is monotonically decreasing on I.

Next, we discuss more subtle properties of a function, known as convexity and concavity. Geometrically, these notions are easily described. A function is convex if the line segment joining any two points on its graph lies on or above the graph. A function is concave if any such line segment lies on or below the graph. An illustration is given in Figure 1.7. To formulate a more precise definition, one should first note that convexity or concavity can be defined relative to an interval I contained in the domain of a function f, and also that given any $x_1, x_2 \in I$ with $x_1 < x_2$, the equation of the line joining the corresponding points $(x_1, f(x_1))$ and $(x_2, f(x_2))$ on the graph of f is given by

$$y - f(x_1) = m(x - x_1), \quad \text{where} \quad m = \frac{f(x_2) - f(x_1)}{x_2 - x_1}.$$

So, once again let $D \subseteq \mathbb{R}$ be such that D contains an interval I and $f : D \to \mathbb{R}$ be a function. We say that

1. f is **convex on** I or **concave upward on** I if

$$x_1, x_2, x \in I, \ x_1 < x < x_2 \implies f(x) - f(x_1) \le \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_1),$$

2. f is **concave on** I or **concave downward on** I if

$$x_1, x_2, x \in I, \ x_1 < x < x_2 \implies f(x) - f(x_1) \ge \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_1).$$
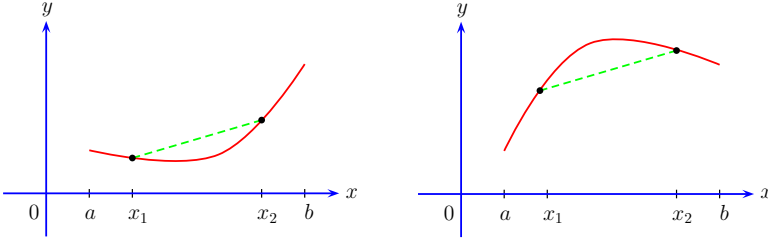
Fig. 1.7. Typical graphs of convex and concave functions on $I = [a, b]$

An alternative way to formulate the definitions of convexity and concavity is as follows. First, note that for any $x_1, x_2 \in \mathbb{R}$ with $x_1 < x_2$, the points x between x_1 and x_2 are of the form $(1 - t)x_1 + tx_2$ for some $t \in (0, 1)$; in fact, t and x determine each other uniquely since

$$x = (1 - t)x_1 + tx_2 \iff t = \frac{x - x_1}{x_2 - x_1}.$$

Substituting this in the definition above, we see that f is convex on I if (and only if) for any $x_1, x_2 \in I$ with $x_1 < x_2$ and any $t \in (0, 1)$ we have $f((1 - t)x_1 + tx_2) \le (1 - t)f(x_1) + tf(x_2)$. Of course, the roles of t and $1 - t$ can be readily reversed, and with this in view, one need not assume that $x_1 < x_2$. Thus, f is convex on I if (and only if)

$$f(tx_1 + (1 - t)x_2) \le tf(x_1) + (1 - t)f(x_2) \quad \text{for all } x_1, x_2 \in I \text{ and } t \in (0, 1).$$

Similarly, f is concave on I if (and only if)

$$f(tx_1 + (1 - t)x_2) \ge tf(x_1) + (1 - t)f(x_2) \quad \text{for all } x_1, x_2 \in I \text{ and } t \in (0, 1).$$

Examples 1.15. (i) The function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) := x^2$ is increasing on $[0, \infty)$ and decreasing on $(-\infty, 0]$. Indeed, if $x_1, x_2 \in \mathbb{R}$ with $x_1 < x_2$, then $(x_2^2 - x_1^2) = (x_2 - x_1)(x_2 + x_1)$ is positive if $x_1, x_2 \in [0, \infty)$ and negative if $x_1, x_2 \in (-\infty, 0]$. Further, f is convex on $\mathbb{R}$. To see this, note that if $x_1, x_2, x \in \mathbb{R}$ with $x_1 < x < x_2$, then $(x - x_1) > 0$ and

$$x^2 - x_1^2 = (x + x_1)(x - x_1) < (x_2 + x_1)(x - x_1) = \frac{(x_2^2 - x_1^2)}{(x_2 - x_1)}(x - x_1).$$

(ii) The function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) := x^3$ is increasing on $(-\infty, \infty)$. Indeed, if $x_1, x_2 \in \mathbb{R}$ with $x_1 < 0 < x_2$, then clearly $x_1^3 < 0 < x_2^3$, whereas if $x_1, x_2 \in [0, \infty)$ or $x_1, x_2 \in (-\infty, 0]$ with $x_1 < x_2$, then $(x_2^3 - x_1^3) = (x_2 - x_1)(x_2^2 + x_2 x_1 + x_1^2)$ is positive. Further, f is concave on $(-\infty, 0]$ and convex on $[0, \infty)$. To see this, first note that if $x_1 < x < x_2 \le 0$, then $(x - x_1) > 0$, $x^2 > x_2^2$, and $x_1 x > x_1 x_2$, and so $x^3 - x_1^3 = (x^2 + x_1 x + x_1^2)(x - x_1)$ satisfies

$$x^3 - x_1^3 > (x_2^2 + x_1 x_2 + x_1^2)(x - x_1) = \frac{(x_2^3 - x_1^3)}{(x_2 - x_1)}(x - x_1).$$

Also, if $0 \le x_1 < x < x_2$, then $(x - x_1) > 0$, $x^2 < x_2^2$, and $x_1 x < x_1 x_2$, and so in this case $x^3 - x_1^3 = (x^2 + x_1 x + x_1^2)(x - x_1)$ satisfies

$$x^3 - x_1^3 < (x_2^2 + x_1 x_2 + x_1^2)(x - x_1) = \frac{(x_2^3 - x_1^3)}{(x_2 - x_1)}(x - x_1).$$

(iii) The function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) := |x|$ is decreasing on $(-\infty, 0]$, increasing on $[0, \infty)$, and convex on $\mathbb{R} = (-\infty, \infty)$. Indeed, the first two assertions about the monotonicity of f are obvious. The convexity of f is easily verified from the definition by considering separately various cases depending on the signs of x_1, x, and x_2. $\diamond$

Remark 1.16. In each of the examples above, we have in fact obtained a stronger conclusion than was needed to satisfy the definitions of increasing/decreasing and convex/concave functions. Namely, instead of the inequalities "$\le$" and "$\ge$", we obtained the corresponding strict inequalities "$<$" and "$>$". If one wants to emphasize this, the terminology of **strictly increasing**, **strictly decreasing**, **strictly convex**, or **strictly concave**, is employed. The definitions of these concepts are obtained by changing the inequality "$\le$" or "$\ge$" appearing on the right in 1, 2, 4, and 5 above by the corresponding strict inequality "$<$" or "$>$", respectively. Also, we say that a function is **strictly monotonic** if it is strictly increasing or strictly decreasing. $\diamond$

Local Extrema and Points of Inflection

Points where the graph of a function has peaks or dips, or where the convexity changes to concavity (or vice versa), are of great interest in calculus and its applications. We shall now formally introduce the terminology used in describing this type of behavior.

Let $D \subseteq \mathbb{R}$ and $c \in D$ be such that D contains an interval $(c - r, c + r)$ for some $r > 0$. Given $f : D \to \mathbb{R}$, we say that

1. f has a **local maximum** at c if there is $\delta > 0$ with $\delta \le r$ such that $f(x) \le f(c)$ for all $x \in (c - \delta, c + \delta)$,
2. f has a **local minimum** at c if there is $\delta > 0$ with $\delta \le r$ such that $f(x) \ge f(c)$ for all $x \in (c - \delta, c + \delta)$.

3. f has a **local extremum** at c if f has a local maximum at c or a local
   minimum at c,
4. c is a **point of inflection** for f if there is $\delta > 0$ with $\delta \leq r$ such that f
   is convex in $(c - \delta, c)$, while f is concave in $(c, c + \delta)$, or vice versa, that
   is, f is concave in $(c - \delta, c)$, while f is convex in $(c, c + \delta)$.

It may be noted that the terms **local maxima**, **local minima**, and **local
extrema** are often used as plural forms of **local maximum**, **local mini-
mum**, and **local extremum**, respectively.

Examples 1.17. (i) The function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) := -x^2$ has a
   local maximum at the origin, that is, at 0.
(ii) The function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) := |x|$ has a local minimum at
   the origin, that is, at 0. [See Figure 1.4.]
(iii) For the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) := x^3$, the origin, that is, 0,
   is a point of inflection. [See Figure 1.3.] $\diamond$

It is easy to see that if $D \subseteq \mathbb{R}$ contains an open interval of the form
$(c - r, c + r)$ for some $r > 0$ and $f : D \to \mathbb{R}$ is a function such that f is
decreasing on $(c - \delta, c]$ and increasing on $[c, c + \delta)$, for some $0 < \delta \leq r$, then
f must have a local minimum at c. But as the following example shows, the
converse of this need not be true.

Example 1.18. Consider the function $f : [-1, 1] \to \mathbb{R}$, which is obtained by
piecing together infinitely many zigzags as follows. On $[1/(n+1), 1/n]$, we
define f to be such that its graph is formed by the line segments PM and
MQ, where P, Q are the points on the line $y = x$ whose x-coordinates are
$1/n + 1$ and $1/n$, respectively, while M is the point on the line $y = 2x$ whose
x-coordinate is the midpoint of the x-coordinates of P and Q. More precisely,
for $n \in \mathbb{N}$, we define

$$
f(x) :=
\begin{cases}
2(n+1)x - \dfrac{2n+1}{n+1} & \text{if } \dfrac{1}{n+1} \leq x \leq \dfrac{2n+1}{2n(n+1)}, \\[3mm]
-2nx + \dfrac{2n+1}{n} & \text{if } \dfrac{2n+1}{2n(n+1)} \leq x \leq \dfrac{1}{n}.
\end{cases}
$$

Further, let $f(0) := 0$ and $f(x) := f(-x)$ for $x \in [-1, 0)$. The graph of this
piecewise linear function can be drawn as in Figure 1.8. It is clear that f has
a local minimum at 0. However, there is no $\delta > 0$ such that f is decreasing
on $(-\delta, 0]$ and f is increasing on $[0, \delta)$.
   A similar comment holds for the notion of local maximum. $\diamond$

Remark 1.19. As before, in each of the examples above, the given function
satisfies the property mentioned in a strong sense. For example, for $f : \mathbb{R} \to \mathbb{R}$
defined by (i), we not only have $f(x) \leq f(0)$ in an interval around 0 but in fact,
$f(x) < f(0)$ for each point x, except 0, in an interval around 0. To indicate
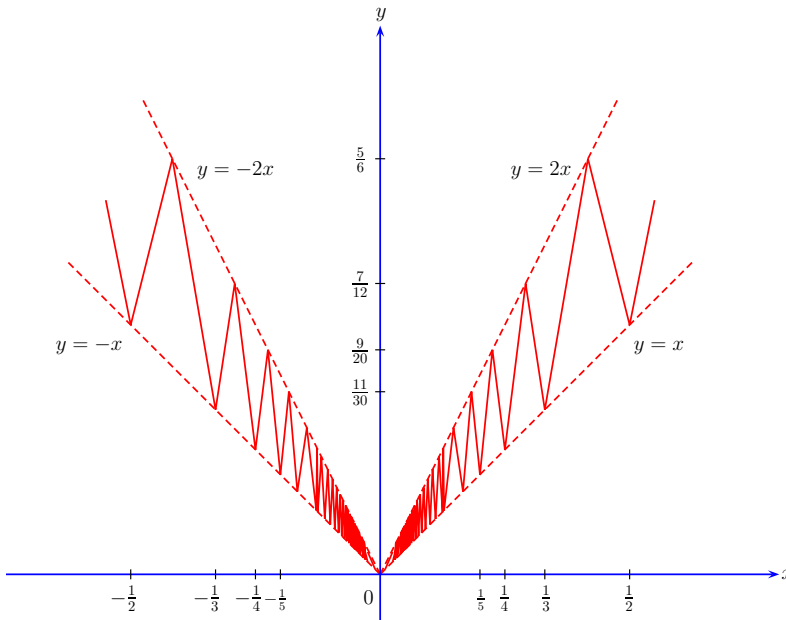
Fig. 1.8. Graph of the piecewise linear zigzag function in Example 1.18

this, the terminology **strict local maximum**, **strict local minimum**, **strict local extremum**, and **strict point of inflection** can be employed. The first two of these notions are defined by changing in 1 and 2 above the inequalities "$\leq$" and "$\geq$" by the corresponding strict inequalities "$<$" and "$>$", and the condition "$x \in (c - \delta, c + \delta)$" by the condition "$x \in (c - \delta, c + \delta)$, $x \neq c$". To say that f has a strict local extremum at c just means it has a strict local maximum or a strict local minimum at c. Finally, the notion of a strict point of inflection is defined by adding "strictly" before the words "convex" and "concave" in the above definition of a point of inflection. $\diamond$

In Examples 1.15 and 1.17, which illustrate the geometric phenomena of increasing/decreasing functions, convexity/concavity, local maxima/minima, and points of inflection, the verification of the corresponding property has been fairly easy. In fact, we have looked at what are possibly the simplest functions that are prototypes of the above phenomena. But even here, the proofs of convexity or concavity in the case of functions given by x^2 and x^3 did require some effort. As one considers functions that are more complicated, the verification of all these geometric properties can become increasingly difficult. Later in this book, we shall describe some results from calculus that can make such verification significantly simpler for a large class of functions. It is, nevertheless, useful to remember that the definition as well as the intuitive idea behind these properties is geometric, and as such, it is independent of the notions from calculus that we shall encounter in the subsequent chapters.

Intermediate Value Property

We now consider a geometric property of a function that corresponds, intuitively, to the idea that the graph of a function has no "breaks" or "disconnections". For example, if $f : \mathbb{R} \to \mathbb{R}$ is defined by $f(x) := 2x + 1$ or by $f(x) := x^2$ or by $f(x) := |x|$, then the graph of f has apparently no "breaks". [See Figures 1.2 and 1.4.] But if $f : \mathbb{R} \to \mathbb{R}$ is defined by

$$f(x) := \begin{cases} x + 2 & \text{if } x \leq 1, \\ (x^2 - 9)/8 & \text{if } x > 1, \end{cases}$$

then the graph of f does seem to have a "break". [See Figure 1.4.] This intuitive condition on the graph of a real-valued function f can be formulated by stating that every intermediate value of f is attained by f. More precisely, we make the following definition.

Let I be an interval and $f : I \to \mathbb{R}$ be a function. We say that f has the **Intermediate Value Property**, or in short, f has the **IVP**, on I if for any $a, b \in I$ with $a < b$ and $r \in \mathbb{R}$,

$$r \text{ lies between } f(a) \text{ and } f(b) \implies r = f(x) \text{ for some } x \in [a, b].$$

Note that if $f : I \to \mathbb{R}$ has the IVP on I, and J is a subinterval of I, then f has the IVP on J.

Proposition 1.20. *Let I be an interval and $f : I \to \mathbb{R}$ be any function. Then*

$$f \text{ has the IVP on } I \implies f(I) \text{ is an interval.}$$

Proof. Let $c, d \in f(I)$ with $c < d$. Then $c = f(a)$ and $d = f(b)$ for some $a, b \in I$. If $r \in (c, d)$, then by the IVP for f on I, there is $x \in I$ between a and b such that $f(x) = r$. Hence $r \in f(I)$. It follows that $f(I)$ is an interval.   □

Remark 1.21. The converse of the above result is true for monotonic functions. To see this, suppose I is an interval and $f : I \to \mathbb{R}$ is a monotonic function such that $f(I)$ is an interval. Let $x_1, x_2 \in I$ be such that $x_1 < x_2$ and r be a real number between $f(x_1)$ and $f(x_2)$. Since $f(I)$ is an interval, there is $x \in I$ such that $r = f(x)$. Now, if f is monotonically increasing on I, then we must have $f(x_1) \leq f(x_2)$; thus, $f(x_1) \leq f(x) \leq f(x_2)$, and consequently, $x_1 \leq x \leq x_2$. Likewise, if f is monotonically decreasing on I, then we have $f(x_1) \geq f(x) \geq f(x_2)$, and consequently, $x_1 \leq x \leq x_2$. This shows that f has the IVP on I.

However, in general, the converse of the result in Proposition 1.20 is not true. For example, if $I = [0, 2]$ and $f : I \to \mathbb{R}$ is defined by

$$f(x) = \begin{cases} x & \text{if } 0 \leq x \leq 1, \\ 3 - x & \text{if } 1 < x \leq 2, \end{cases}$$

then $f(I) = I$ is an interval but f does not have the IVP on I. The latter follows, for example, since $\frac{5}{4}$ lies between $1 = f(1)$ and $\frac{3}{2} = f\left(\frac{3}{2}\right)$, but $\frac{5}{4} \neq f(x)$

for any $x \in [1, \frac{3}{2}]$. It may be noted in this example that f is one-one but is not monotonic on I. Also, $f\left(\left[\frac{1}{2}, \frac{3}{2}\right]\right) = \left[\frac{1}{2}, 1\right] \cup \left[\frac{3}{2}, 2\right]$ is not an interval. $\diamond$

Proposition 1.22. *Let I be an interval and $f : I \to \mathbb{R}$ be any function. Then*

$$f \text{ has the IVP on } I \iff f(J) \text{ is an interval for every subinterval } J \text{ of } I.$$

Proof. The implication $\Longrightarrow$ follows from applying Proposition 1.20 to restrictions of f to subintervals of I. Conversely, suppose $f(J)$ is an interval for every subinterval J of I. Let $a, b \in I$ with $a < b$ and $r \in \mathbb{R}$ lie between $f(a)$ and $f(b)$. Consider $J = [a, b]$. Then J is a subinterval of I and hence $f(J)$ is an interval containing $f(a)$ and $f(b)$. Therefore, $r = f(x)$ for some $x \in J$. Thus, f has the IVP on I. $\square$

The relation between (strict) monotonicity and the IVP is made clearer by the following result.

Proposition 1.23. *Let I be an interval and $f : I \to \mathbb{R}$ be a function. Then f is one-one and has the IVP on I if and only if f is strictly monotonic and $f(I)$ is an interval. In this case, $f^{-1} : f(I) \to \mathbb{R}$ is strictly monotonic and has the IVP on $f(I)$.*

Proof. Assume that f is one-one and has the IVP on I. By Proposition 1.20, $f(I)$ is an interval. Suppose f is not strictly monotonic on I. Then there are $x_1, x_2 \in I$ and $y_1, y_2 \in I$ such that

$$x_1 < x_2 \text{ but } f(x_1) \geq f(x_2) \quad \text{and} \quad y_1 < y_2 \text{ but } f(y_1) \leq f(y_2).$$

Let $a := \min\{x_1, y_1\}$ and $b := \max\{x_2, y_2\}$. Note that $a < b$. Now, suppose $f(a) \leq f(b)$. Then we must have $f(x_1) \leq f(b)$ because otherwise, $f(x_1) > f(b) \geq f(a)$ and hence by the IVP of f on I, there is $z_1 \in [a, x_1]$ such that $f(z_1) = f(b)$. But since $z_1 \leq x_1 < x_2 \leq b$, this contradicts the assumption that f is one-one. Thus, we have $f(x_2) \leq f(x_1) \leq f(b)$. Again, by the IVP of f on I, there is $w_1 \in [x_2, b]$ such that $f(w_1) = f(x_1)$. But since $x_1 < x_2 \leq w_1$, this contradicts the assumption that f is one-one. Next, suppose $f(b) < f(a)$. Here, we must have $f(y_2) \leq f(a)$ because otherwise, $f(y_2) > f(a) > f(b)$ and hence by the IVP of f on I, there is $z_2 \in [y_2, b]$ such that $f(z_2) = f(a)$. But since $a \leq y_1 < y_2 \leq z_2$, this contradicts the assumption that f is one-one. Thus, we have $f(y_1) \leq f(y_2) \leq f(a)$. Again, by the IVP of f on I, there is $w_2 \in [a, y_1]$ such that $f(w_2) = f(y_2)$. But since $w_2 \leq y_1 < y_2$, this contradicts the assumption that f is one-one. It follows that f is strictly monotonic on I.

To prove the converse, assume that f is strictly monotonic on I and $f(I)$ is an interval. Then we have seen in Remark 1.21 above that f has the IVP on I. Also, strict monotonicity obviously implies that f is one-one.

Finally, suppose f is one-one and has the IVP on I. Then as seen above, f is strictly monotonic on I. This implies readily that f^{-1} is strictly monotonic on $f(I)$. Also, $f(I)$ is an interval and so is $I = f^{-1}(f(I))$. Hence by the equivalence proved above, f^{-1} has the IVP on $f(I)$. $\square$

Notes and Comments

It is often said, and believed, that mathematics is an exact science, and all the terms one uses in mathematics are always precisely defined. This is of course true to a large extent. But one should realize that it is impossible to precisely define everything. Indeed, to define one term, we would have to use another, to define which we would have to use yet another, and so on. Since our vocabulary is finite (check!), we would soon land in a vicious circle! Mathematicians find a way out of this dilemma by agreeing to regard certain terms as undefined or primitive. Further, one also stipulates certain axioms or postulates that describe some 'natural' properties that the primitive terms possess. Once this is done, every other term is defined using the primitive terms or the ones defined earlier. Also, a result is not accepted unless it is precisely proved using the axioms or the results proved before.

The terms that are usually considered primitive or undefined in mathematics are "set" and "is an element of" (a set). One has a small number of axioms that postulate certain basic and seemingly obvious 'facts' about sets. Taking these for granted, we can define just about everything else that one encounters in mathematics. The formal definition of a function given in this chapter is a good illustration of this phenomenon.

Good references for the nitty-gritty about sets, or rather, the subject of axiomatic set theory, are the books by Enderton [24] and Halmos [29]. To define the real numbers, one begins with the set $\mathbb{Q}$ of rational numbers and constructs a set that satisfies the properties we postulated for $\mathbb{R}$. There are two standard approaches for the construction of $\mathbb{R}$ from $\mathbb{Q}$, one due to Dedekind and the other due to Cantor. An old-fashioned but thorough discussion of both the approaches can be found in the book of Hobson [37]. A sleek presentation of Dedekind's approach can be found in the appendix to Chapter 1 of Rudin [53]. For a precise account of Cantor's construction, see Section 5 of the book of Hewitt and Stromberg [36]. A classic reference for the construction of real numbers and more generally the foundation of calculus is the charming book of Landau [44].

The topic of inequalities, which we briefly discussed in Section 1.2 is now a subject in itself, and to get a glimpse of it, one can see the book of Hardy, Littlewood, and Polya [33] or of Beckenbach and Bellman [9]. There are also specialized books like that of Bullen, Mitrinoviv, and Vasic [15], which has, for example, more than 50 proofs of the A.M.-G.M. inequality! A more elementary and accessible introduction is the little booklet [43] of Korovkin.

The notion of a function is of basic importance not only in calculus and analysis, but in all of mathematics. Not surprisingly, it has evolved over the years and the formal definition is a distilled form of various ideas one has about this notion. Classically, the notion of a function was supple enough to admit y as an (implicit) function of x if the two are related by an equation $F(x, y) = 0$, even though for a given value of x, there could be multiple values of y satisfying $F(x, y) = 0$. But in modern parlance, there is no such thing

*as a multivalued function! Nonetheless, the so-called "multivalued functions"
have played an important role in the development of calculus and other parts
of mathematics. With this in view, we have included a discussion of algebraic
functions, remaining within the confines of modern definitions and the sub-
ject of calculus. For a glimpse of the classical viewpoint, see the two-volume
textbook of Chrystal [16], which is also an excellent reference for algebra in
general. A relatively modern and accessible book on algebra, which includes
a discussion of partial fraction decomposition, is the survey of Birkhoff and
Mac Lane [11]. A variety of proofs of the Fundamental Theorem of Algbra,
which implies the Real Fundamental Theorem of Algebra stated in this chap-
ter, can be found in the book of Fine and Rosenberger [25].*

*For real-valued functions defined on intervals, we have discussed a number
of geometric properties such as monotonicity, convexity, local extrema, and
the Intermediate Value Property. Typically, these appear in calculus books in
conjunction with the notions of differentiability and continuity. The reason to
include these in the first chapter is to stress the fact that these are geometric
notions and should not be confused with various criteria one has, involving
differentiability or continuity, to check them.*

Exercises

Part A

1. Using only the algebraic properties A1–A5 on page 3, prove the following.
   (i) 0 is the unique real number such that $a + 0 = a$ for all $a \in \mathbb{R}$. In other
       words, if some $z \in \mathbb{R}$ is such that $a + z = a$ for all $a \in \mathbb{R}$, then $z = 0$.
   (ii) 1 is the unique real number such that $a \cdot 1 = a$ for all $a \in \mathbb{R}$.
   (iii) Given any $a \in \mathbb{R}$, an element $a' \in \mathbb{R}$ such that $a + a' = 0$ is unique.
       [As noted before, this unique real number a' is denoted by $-a$.]
   (iv) Given any $a \in \mathbb{R}$ with $a \neq 0$, an element $a^* \in \mathbb{R}$ such that $a \cdot a^* = 1$
       is unique. [As noted before, this unique real number a^* is denoted by
       a^{-1} or by $1/a$.]
   (v) Given any $a \in \mathbb{R}$, we have $-(-a) = a$. Further if $a \neq 0$, then
       $\left(a^{-1}\right)^{-1} = a$.
   (vi) Given any $a, b \in \mathbb{R}$, we have $a(-b) = -(ab)$ and $(-a)(-b) = ab$.
2. Given any $a \in \mathbb{R}$ and $k \in \mathbb{Z}$, the **binomial coefficient** associated with a
   and k is defined by

$$
\binom{a}{k} = \begin{cases} \dfrac{a(a-1)\cdots(a-k+1)}{k!} & \text{if } k \geq 0, \\ 0 & \text{if } k < 0, \end{cases}
$$

   where for $k \in \mathbb{N}$, $k!$ (read as k **factorial**) denotes the product of the first
   k positive integers. Note that $0! = 1$ and $\binom{a}{0} = 1$ for any $a \in \mathbb{R}$.

(i) Show that if $a, k \in \mathbb{Z}$ with $0 \leq k \leq a$, then

$$\binom{a}{k} = \frac{a!}{k!(a-k)!} = \binom{a}{a-k}.$$

(ii) If $a \in \mathbb{R}$ and $k \in \mathbb{Z}$, then show that

$$\binom{a}{k} = \binom{a-1}{k} + \binom{a}{k-1}.$$

[Note: This identity is sometimes called the **Pascal triangle identity**. If we compute the values of the binomial coefficients $\binom{n}{k}$ for $n \in \mathbb{N}$ and $0 \leq k \leq n$, and write them in a triangular array such that the nth row consists of the numbers $\binom{n}{0}, \binom{n}{1}, \ldots, \binom{n}{n}$, then this array is called the **Pascal triangle**. It may be instructive to write the first few rows of the Pascal triangle and *see* what the identity means pictorially.]

(iii) Use the identity in (ii) and induction to prove the Binomial Theorem (for positive integral exponents). In other words, prove that for any $n \in \mathbb{N}$ and $x, y \in \mathbb{R}$, we have

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

[Note: Proving a statement defined for $n \in \mathbb{N}$ such as the above identity by **induction** means that we should prove it for the initial value $n = 1$, and further prove it for an arbitrary value of $n \in \mathbb{N}$, $n > 1$, by assuming either that it holds for $n-1$ or that it holds for values of n smaller than the given one. The technique of induction also works when $\mathbb{N}$ is replaced by any subset S of $\mathbb{Z}$ that is bounded below; the only difference would be that the initial value 1 would have to be changed to the least element of S.]

3. Use induction to prove the following statements for each $n \in \mathbb{N}$:

(i) $\displaystyle\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$,

(ii) $\displaystyle\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$,

(iii) $\displaystyle\sum_{i=1}^{n} i^3 = \frac{n^2(n+1)^2}{4} = \left(\sum_{i=1}^{n} i\right)^2$.

4. Use the algebraic properties and the order properties of $\mathbb{R}$ to prove that

(i) $a^2 > 0$ for any $a \in \mathbb{R}$, $a \neq 0$.

(ii) Given $a, b \in \mathbb{R}$ with $0 < a < b$, we have $0 < (1/b) < (1/a)$.

5. Let S be a nonempty subset of $\mathbb{R}$. If S is bounded above, then show that the set $U_S = \{\alpha \in \mathbb{R} : \alpha \text{ is an upper bound of } S\}$ is bounded below, $\min U_S$ exists, and $\sup S = \min U_S$. Likewise, if S is bounded below, then show that the set $L_S = \{\beta \in \mathbb{R} : \beta \text{ is a lower bound of } S\}$ is bounded above, $\max L_S$ exists, and $\inf S = \max L_S$.

6. Let S be a nonempty subset of $\mathbb{R}$. If S is bounded above and $M \in \mathbb{R}$, then show that $M = \sup S$ if and only if M is an upper bound of S and for every $\epsilon > 0$, there exists $x \in S$ such that $M - \epsilon < x \leq M$. Likewise, if S is bounded below and $m \in \mathbb{R}$, then show that $m = \inf S$ if and only if m is a lower bound of S and for every $\epsilon > 0$, there exists $x \in S$ such that $m \leq x < m + \epsilon$.

7. Let S be a nonempty subset of $\mathbb{R}$ and $c \in \mathbb{R}$. Define the additive translate $c + S$ and the multiplicative translate cS of S as follows:

$$c + S = \{c + x : x \in S\} \quad \text{and} \quad cS = \{cx : x \in S\}.$$

If S is bounded, then show that $c + S$ and cS are bounded. Also show that

$$\sup(c + S) = c + \sup S \quad \text{and} \quad \inf(c + S) = c + \inf S,$$

whereas

$$\sup(cS) = \begin{cases} c \sup S & \text{if } c \geq 0, \\ c \inf S & \text{if } c \leq 0, \end{cases} \quad \text{and} \quad \inf(cS) = \begin{cases} c \inf S & \text{if } c \geq 0, \\ c \sup S & \text{if } c \leq 0. \end{cases}$$

8. Given any $x, y \in \mathbb{R}$ with $y > 0$, show that there exists $n \in \mathbb{N}$ such that $ny > x$.
   [Note: The above property is equivalent to the Archimedean property, which was stated in Proposition 1.3.]

9. Given any $x, y \in \mathbb{R}$ with $x \neq y$, show that there exists $\delta > 0$ such that the intervals $(x - \delta, x + \delta)$ and $(y - \delta, y + \delta)$ have no point in common.
   [Note: The above property is sometimes called the **Hausdorff property**.]

10. Prove that the following numbers are irrational:

   (i) $\sqrt{3}$,   (ii) $\sqrt{15}$,   (iii) $\sqrt[3]{2}$,   (iv) $\sqrt[4]{11}$,   (v) $\sqrt[5]{16}$,   (vi) $\sqrt{2} + \sqrt{3}$.

11. If $a, b \in \mathbb{R}$ with $a < b$, then show that there exist infinitely many rational numbers as well as infinitely many irrational numbers between a and b.

12. Show that $n! \leq 2^{-n} (n + 1)^n$ for every $n \in \mathbb{N}$, and that equality holds if and only if $n = 1$.

13. Let $n \in \mathbb{N}$ and $a_1, \ldots, a_n$ be positive real numbers. Prove that

$$\sqrt[n]{a_1 \cdots a_n} \geq \frac{n}{r} \quad \text{where} \quad r := \frac{1}{a_1} + \cdots + \frac{1}{a_n}$$

and that equality holds if and only if $a_1 = \cdots = a_n$.
   [Note: The above result is sometimes called the **G.M.-H.M. inequality** and n/r is called the **harmonic mean** of $a_1, \ldots, a_n$.]

14. Let $a_1, \ldots, a_n, b_1, \ldots, b_n$ be real numbers such that $a_1 > \cdots > a_n > 0$.
   (i) Let $m = \min\{B_1, \ldots, B_n\}$ and $M = \max\{B_1, \ldots, B_n\}$, where $B_i = b_1 + \cdots + b_i$ for $i = 1, \ldots, n$. Show that $ma_1 \leq a_1 b_1 + \cdots + a_n b_n \leq Ma_1$.
   [Note: The above inequality is sometimes called **Abel's inequality**.]

(ii) Show that the alternating sum $a_1 - a_2 + \cdots + (-1)^{n+1}a_n$ is always between 0 and a_1.

15. Let $n, m \in \mathbb{N}$ be such that $m \geq n$.
   (i) Let $a_1, \ldots, a_n, \ldots, a_m$ be real numbers, and let A_n denote the arithmetic mean of $a_1, \ldots, a_n$, and A_m denote the arithmetic mean of $a_1, \ldots, a_m$. Show that

$$A_n \leq A_m \quad \text{if} \quad a_1 \leq \cdots \leq a_m \quad \text{and} \quad A_n \geq A_m \quad \text{if} \quad a_1 \geq \cdots \geq a_m.$$

   Further, show that if $m > n$, then equality holds if and only if $a_1 = \cdots = a_m$. (Hint: Induct on m.)
   (ii) Use (i) to show that for any $x \in \mathbb{R}$ with $x \geq 0$, we have

$$\frac{x^m - 1}{m} \geq \frac{x^n - 1}{n}$$

   and further, if $m > n$, then equality holds if and only if $x = 1$.
   [Note: Exercise 52 gives an alternative approach to this inequality.]
   (iii) Use (ii) to show that for any $a \in \mathbb{R}$ and $r \in \mathbb{Q}$ with $1 + a \geq 0$ and $r \geq 1$, we have $(1 + a)^r \geq 1 + ra$. Further, show that if $r > 1$, then equality holds if and only if $a = 0$.
   [Note: Exercise 54 (iii) gives an alternative approach to this inequality.]

16. Let $n \in \mathbb{N}$ and let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be any real numbers. Assume that not all $a_1, \ldots, a_n$ are zero. Consider the quadratic polynomial

$$q(x) = \sum_{i=1}^{n} (xa_i + b_i)^2.$$

   Show that the discriminant Δ of $q(x)$ is nonnegative, and $\Delta = 0$ if and only if there is $c \in \mathbb{R}$ such that $b_i = ca_i$ for all $i = 1, \ldots, n$. Use this to give an alternative proof of Proposition 1.12.

17. Show that if $n \in \mathbb{N}$ and $a_1, \ldots, a_n$ are nonnegative real numbers, then $(a_1 + \cdots + a_n)^2 \leq n(a_1^2 + \cdots + a_n^2)$. (Hint: Write $(a_1 + \cdots + a_n)^2$ as $t_1 + \cdots + t_n$, where $t_k := a_1a_k + a_2a_{k+1} + \cdots + a_{n-k+1}a_n + a_{n-k+2}a_1 + \cdots + a_na_{k-1}$.)
   [Note: Exercise 35 gives an alternative approach to this inequality.]

18. Let n and m be nonnegative integers. Show that there is an injective map from $\{1, \ldots, n\}$ to $\{1, \ldots, m\}$ if and only if $n \leq m$. Also show that there is a surjective map from $\{1, \ldots, n\}$ to $\{1, \ldots, m\}$ if and only if $n \geq m$. Deduce that there is a bijective map from $\{1, \ldots, n\}$ to $\{1, \ldots, m\}$ if and only if $n = m$. (Hint: Use induction.)

19. Given any function $f : \mathbb{R} \to \mathbb{R}$, prove the following:
   (i) If $f(xy) = f(x) + f(y)$ for all $x, y \in \mathbb{R}$, then $f(x) = 0$ for all $x \in \mathbb{R}$.
   [Note: It is, however, possible that there are nonzero functions defined on subsets of $\mathbb{R}$, such as $(0, \infty)$ that satisfy $f(xy) = f(x) + f(y)$ for all x, y in the domain of f. A prominent example of this is the logarithmic function, which will be discussed in Section 7.1.]

(ii) If $f(xy) = f(x)f(y)$ for all $x, y \in \mathbb{R}$, then either $f(x) = 0$ for all $x \in \mathbb{R}$, or $f(x) = 1$ for all $x \in \mathbb{R}$, or $f(0) = 0$ and $f(1) = 1$. Further, f is either an even function or an odd function.

Give examples of even as well as odd functions $f : \mathbb{R} \to \mathbb{R}$ satisfying $f(xy) = f(x)f(y)$ for all $x, y \in \mathbb{R}$.

20. Prove that the absolute value function, that is, $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = |x|$, is not a rational function.

21. Prove that the function $f : [0, \infty) \to \mathbb{R}$ defined by the following is an algebraic function:

$$\text{(i) } f(x) = 1 + \sqrt[3]{x}. \quad \text{(ii) } f(x) = \sqrt{x} + \sqrt{2x}, \quad \text{(iii) } f(x) = \sqrt{x} + \sqrt[3]{x}.$$

22. Let $f : \mathbb{R} \to \mathbb{R}$ be the function defined by the following:

$$\text{(i) } f(x) = x^3, \quad \text{(ii) } f(x) = x^4, \quad \text{(iii) } f(x) = |x|, \quad \text{(iv) } f(x) = \sqrt{|x|}.$$

Sketch the graph of f and determine the points at which f has local extrema as well as the points of inflection of f, if any, in each case.

23. Let $f : (0, \infty) \to \mathbb{R}$ be the function defined by the following:

$$\text{(i) } f(x) = \sqrt{x}, \quad \text{(ii) } f(x) = x^{3/2}, \quad \text{(iii) } f(x) = \frac{1}{|x|}, \quad \text{(iv) } f(x) = \frac{1}{x^2}.$$

Sketch the graph of f and determine the points at which f has local extrema as well as the points of inflection of f, if any, in each case.

24. Given any $f : D \to \mathbb{R}$ and $c \in \mathbb{R}$, define functions g_c, h_c, k_c, and ℓ_c from $\mathbb{R}$ to $\mathbb{R}$ as follows:

$$g_c(x) = f(x) + c, \quad h_c(x) = cf(x), \quad k_c(x) = f(x + c), \quad \ell_c(x) = f(cx).$$

If f is given by $f(x) = x^n$ for all $x \in \mathbb{R}$, then sketch the graph of g_c, h_c, k_c, and ℓ_c when $n = 1$, 2 or 3 and $c = 0$, 1, 2, -1, -2, $\frac{1}{2}$ or $-\frac{1}{2}$.

25. Consider $D \subseteq \mathbb{R}$ and $f : D \to \mathbb{R}$ defined by the following. Determine whether f is bounded above on D. If yes, find an upper bound for f on D. Also, determine whether f is bounded below on D. If yes, find a lower bound for f on D. Also, determine whether f attains its upper bound or lower bound.

$$\text{(i) } D = (-1, 1) \text{ and } f(x) = x^2 - 1, \quad \text{(ii) } D = (-1, 1) \text{ and } f(x) = x^3 - 1,$$

$$\text{(iii) } D = (-1, 1] \text{ and } f(x) = x^2 - 2x - 3, \quad \text{(iv) } D = \mathbb{R} \text{ and } f(x) = \frac{1}{1 + x^2}.$$

26. Let D be a bounded subset of $\mathbb{R}$ and $f : D \to \mathbb{R}$ be a polynomial function. Prove that f is bounded on D.

27. Let I be an interval and $f : I \to \mathbb{R}$ be a monotonically increasing function. Given any $r \in \mathbb{R}$, show that $rf : I \to \mathbb{R}$ is a monotonically increasing function if $r \geq 0$ and a monotonically decreasing function if $r < 0$.

28. Let I be an interval and $f : I \to \mathbb{R}$ be a convex function. Given any $r \in \mathbb{R}$, show that $rf : I \to \mathbb{R}$ is a convex function if $r \geq 0$ and a concave function if $r < 0$.

29. Let I be an interval and $f, g : I \to \mathbb{R}$ be convex functions. Show that $f + g : I \to \mathbb{R}$ is also convex.

30. Give an example of $f : (0, 1) \to \mathbb{R}$ such that f is
    (i) strictly increasing and convex,   (ii) strictly increasing and concave,
    (iii) strictly decreasing and convex,   (iv) strictly decreasing and concave.

31. Give an example of a nonconstant function $f : (-1, 1) \to \mathbb{R}$ such that f has a local extremum at 0, and 0 is a point of inflection for f.

32. Let I be an interval and $f : I \to \mathbb{R}$ be any function.
    (i) If f is monotonically increasing as well as monotonically decreasing on I, then show that f is constant on I.
    (ii) If f is convex as well as concave on I, then show that f is given by a linear polynomial (that is, there are $a, b \in \mathbb{R}$ such that $f(x) = ax + b$ for all $x \in I$).

33. Let I be an interval and $f : I \to \mathbb{R}$ be any function. Show that f is convex on I if and only if the slope of the chord joining $(x_1, f(x_1))$ and $(x, f(x))$ is less than or equal to the slope of the chord joining $(x, f(x))$ and $(x_2, f(x_2))$ for all $x_1 < x < x_2$ in I.

34. Let I be an interval and $f : I \to \mathbb{R}$ be any function. If f is convex on I, then show that for any $x_1, \ldots, x_n \in I$ and any nonnegative real numbers $t_1, \ldots, t_n$ with $t_1 + \cdots + t_n = 1$, we have

$$f(t_1 x_1 + \cdots + t_n x_n) \leq t_1 f(x_1) + \cdots + t_n f(x_n).$$

[Note: The above inequality is sometimes called **Jensen's inequality**.]

35. Use Jensen's inequality in Exercise 34 to show that if $n \in \mathbb{N}$ and $a_1, \ldots, a_n$ are nonnegative real numbers, then $(a_1 + \cdots + a_n)^2 \leq n(a_1^2 + \cdots + a_n^2)$.

36. For $f, g : \mathbb{R} \to \mathbb{R}$, which of the following statements are true? Why?
    (i) If f and g have a local maximum at $x = c$, then so does $f + g$.
    (ii) If f and g have a local maximum at $x = c$, then so does fg. What if $f(x) \geq 0$ and $g(x) \geq 0$ for all $x \in \mathbb{R}$?
    (iii) If c is a point of inflection for f as well as for g, then it is a point of inflection for $f + g$.
    (iv) If c is a point of inflection for f as well as for g, then it is a point of inflection for fg.

Part B

37. Given any $\ell, m \in \mathbb{Z}$ with $\ell \neq 0$, prove that there are unique integers q and r such that $m = \ell q + r$ and $0 \leq r < |\ell|$. (Hint: Consider the least element of the subset $\{m - \ell n : n \in \mathbb{Z} \text{ with } m - \ell n \geq 0\}$ of $\mathbb{Z}$.)

38. Given any integers m and n, not both zero, a positive integer d satisfying

$$(i) \; d \mid m \text{ and } d \mid n \quad \text{and} \quad (ii) \; e \in \mathbb{Z}, \; e \mid m \text{ and } e \mid n \Longrightarrow e \mid d$$

is called a **greatest common divisor**, or simply a **GCD**, of m and n. If $m = n = 0$, we set the GCD of m and n to be 0. Given any $m, n \in \mathbb{Z}$, show that a GCD of m and n exists and is unique; it is denoted by $\mathrm{GCD}(m, n)$. Also show that $\mathrm{GCD}(m, n) = um + vn$ for some $u, v \in \mathbb{Z}$. (Hint: Consider the least element of $\{um + vn : u, v \in \mathbb{Z} \text{ with } um + vn > 0\}$.)

39. Let $m, n \in \mathbb{Z}$. Show that m and n are relatively prime if and only if $\mathrm{GCD}(m, n) = 1$. Also show that m and n are relatively prime if and only if $um + vn = 1$ for some $u, v \in \mathbb{Z}$. Is it true that if a positive integer d satisfies $um + vn = d$ for some $u, v \in \mathbb{Z}$, then $d = \mathrm{GCD}(m, n)$?

40. Let $m, n \in \mathbb{Z}$ be relatively prime integers different from ± 1. Show by an example that the integers u, v satisfying $um + vn = 1$ need not be unique. Show, however, that there are unique $u, v \in \mathbb{Z}$ such that $um + vn = 1$ and $0 \le u < |n|$. In this case show that $|v| < |m|$. (Hint: Exercise 37.)

41. Given any integers m and n, both nonzero, a positive integer ℓ satisfying

$$\text{(i) } m \mid \ell \text{ and } n \mid \ell \qquad \text{and} \qquad \text{(ii) } k \in \mathbb{N}, \ m \mid k \text{ and } n \mid k \Longrightarrow \ell \mid k$$

is called a **least common multiple**, or simply an **LCM**, of m and n. If $m = 0$ or $n = 0$, we set the LCM of m and n to be 0. Given any $m, n \in \mathbb{Z}$, show that an LCM of m and n exists and is unique; it is denoted by $\mathrm{LCM}(m, n)$. Also show that if m and n are nonnegative integers and we let $d = \mathrm{GCD}(m, n)$ and $\ell = \mathrm{LCM}(m, n)$, then $d\ell = mn$.

42. If $m, n, n' \in \mathbb{Z}$ are such that m and n are relatively prime and $m \mid nn'$, then show that $m \mid n'$. Deduce that if p is a **prime** (which means that p is an integer > 1 and the only positive integers that divide p are 1 and p) and if p divides a product of two integers, then it divides one of them. (Hint: Exercise 38.)

43. Prove that every rational number r can be written as

$$r = \frac{p}{q}, \text{ where } p, q \in \mathbb{Z}, \ q > 0 \text{ and } p, q \text{ are relatively prime,}$$

and moreover the integers p and q are uniquely determined by r.

44. Show that if a rational number α satisfies a monic polynomial with integer coefficients, that is, if $\alpha \in \mathbb{Q}$ and $\alpha^n + c_{n-1}\alpha^{n-1} + \cdots + c_1\alpha + c_0 = 0$ for some $n \in \mathbb{N}$ and $c_0, c_1, \ldots, c_{n-1} \in \mathbb{Z}$, then α must be an integer. Use this to solve Exercise 10 above.

45. Consider the set $\{(x, y) : x \in \mathbb{R} \text{ and } y \in \mathbb{R}\}$ of ordered pairs of real numbers. Define the operations of addition and multiplication on this set as follows:
$$(x_1, y_1) + (x_2, y_2) = (x_1 + x_2, \ y_1 + y_2)$$

and

$$(x_1, y_1)(x_2, y_2) = (x_1 x_2 - y_1 y_2, \ x_1 y_2 + y_1 x_2).$$

Show that with respect to these operations, all the algebraic properties on page 3 are satisfied with $\mathbb{R}$ replaced by the above set of ordered pairs.

[Note: It is customary to write $i = (0, 1)$ and identify a real number x with the pair $(x, 0)$, so that a pair (x, y) in the above set may be written as $x+iy$. The elements $x+iy$ considered above are called **complex numbers**. The set of all complex numbers is denoted by $\mathbb{C}$. By identifying x with $x + i0 = (x, 0)$, we may regard $\mathbb{R}$ as a subset of $\mathbb{C}$.]

46. Show that $\mathbb{C}$ does not satisfy the order properties. In other words, it is impossible to find a subset $\mathbb{C}^+$ of $\mathbb{C}$ that satisfies properties similar to those stated for $\mathbb{R}^+$ on page 4. (Hint: -1 is a square in $\mathbb{C}$.)

47. A set D is said to be **countable** if it is finite or if there is a bijective map from $\mathbb{N}$ to D. A set that is not countable is said to be **uncountable**.
    (i) Show that the set $\{0, 1, 2, \ldots\}$ of all nonnegative integers is countable.
    (ii) Show that the set $\{1, 3, 5, \ldots\}$ of all odd positive integers is countable. Also, the set $\{2, 4, 6, \ldots\}$ of all even positive integers is countable.
    (iii) Show that the set $\mathbb{Z}$ of all integers is countable.

48. Prove that every subset of a countable set is countable.

49. (i) If A and B are any countable sets, then show that the set $A \times B := \{(a, b) : a \in A \text{ and } b \in B\}$ is also countable. (Hint: Write $A = \{a_m : m \in \mathbb{N}\}$ and $B = \{b_n : n \in \mathbb{N}\}$. List the elements (a_m, b_n) of $A \times B$ as a two-dimensional array and move diagonally. Alternatively, consider the map $2^m 3^n \mapsto (a_m, b_n)$ from an appropriate subset of $\mathbb{N}$ onto $A \times B$.)
    (ii) Let $\{A_n : n \in \mathbb{N}\}$ denote a family of sets indexed by $\mathbb{N}$. If A_n is countable for each $n \in \mathbb{N}$, then show that the union $\bigcup_{n \in \mathbb{N}} A_n$ is countable.
    (iii) Show that the set $\mathbb{Q}$ of all rational numbers is countable.

50. Let $\{0, 1\}^{\mathbb{N}}$ denote the set of all maps from $\mathbb{N}$ to the two–element set $\{0, 1\}$. Prove that $\{0, 1\}^{\mathbb{N}}$ is uncountable. (Hint: Write elements of $\{0, 1\}^{\mathbb{N}}$ as $(s_1, s_2, \ldots)$, where $s_n \in \{0, 1\}$ for $n \in \mathbb{N}$. Given any $f : \mathbb{N} \to \{0, 1\}^{\mathbb{N}}$, consider $(t_1, t_2, \ldots)$ defined by $t_n = 1$ if the nth entry of $f(n)$ is 0, and $t_n = 0$ if the nth entry of $f(n)$ is 1.)

51. Let $I_n = [a_n, b_n]$, $n \in \mathbb{N}$, be closed intervals in $\mathbb{R}$ such that $I_n \supseteq I_{n+1}$ for each $n \in \mathbb{N}$. If $x = \sup\{a_n : n \in \mathbb{N}\}$ and $y = \inf\{a_n : n \in \mathbb{N}\}$, then show that $x \in I_n$ and $y \in I_n$ for every $n \in \mathbb{N}$.

52. Let $m, n \in \mathbb{N}$ and let $x \in \mathbb{R}$ be such that $x \geq 0$ and $x \neq 1$. Show that

$$\frac{x^m - 1}{m} > \frac{x^n - 1}{n} \quad \text{if } m > n \quad \text{and} \quad \frac{x^m - 1}{m} < \frac{x^n - 1}{n} \quad \text{if } m < n.$$

(Hint: It suffices to assume that $m > n$. Write $n(x^m - 1) - m(x^m - 1)$ as $(x - 1)[n(x^n + x^{n+1} + \cdots + x^{m-1}) - (m - n)(1 + x + \cdots + x^{n-1})]$. Compare the $m - n$ elements $x^n, x^{n+1}, \ldots, x^{m-1}$ as well as the n elements $1, x, \ldots, x^{n-1}$ with x^n, when $x < 1$ and $x > 1$.)

53. Let $r \in \mathbb{Q}$ and $a, b \in \mathbb{R}$ be such that $a > 0$, $b > 0$, and $a \neq b$. Prove that

$$ra^{r-1}(a - b) < a^r - b^r < rb^{r-1}(a - b) \quad \text{if } r > 1, \text{ and}$$
$$rb^{r-1}(a - b) < a^r - b^r < ra^{r-1}(a - b) \quad \text{if } 0 < r < 1.$$

(Hint: Write $r = m/n$, where $m, n \in \mathbb{N}$ and use Exercise 52 with $x = (a/b)^{1/n}$ and with $x = (b/a)^{1/n}$.)

54. Use Exercise 53 to deduce the following:
    (i) [**Basic Inequality for Rational Powers**] Given any $a, b \in \mathbb{R}$ and $r \in \mathbb{Q}$ such that $a \geq 0$, $b \geq 0$ and $r \geq 1$, we have $|a^r - b^r| \leq rM^{r-1}|a - b|$, where $M = \max\{|a|, |b|\}$.
    (ii) [**Inequality for Rational Roots**] Given any $a, b \in \mathbb{R}$ and $r \in \mathbb{Q}$ such that $a \geq 0$, $b \geq 0$ and $0 < r < 1$, we have $|a^r - b^r| \leq 2|a - b|^r$. (Hint: It suffices to assume that $0 < b < a$. Now, if $b \leq a/2$, then $a^r - b^r \leq a^r \leq 2^r(a - b)^r$, whereas if $b > a/2$, then $a^r - b^r \leq rb^{r-1}(a - b) \leq r(a - b)^r$. Note that $\max\{r, 2^r\} \leq 2$.)
    (iii) [**Binomial Inequality for Rational Powers**] Given any $a \in \mathbb{R}$ and $r \in \mathbb{Q}$ such that $1 + a \geq 0$ and $r \geq 1$, we have $(1 + a)^r \geq 1 + ra$. Further, if $r > 1$, then $(1 + a)^r > 1 + ra$.
55. Give an alternative proof of the A.M.-G.M. inequality as follows.
    (i) First prove the inequality for n numbers $a_1, \ldots, a_n$ when $n = 2^m$ by using induction on m.
    (ii) In the general case, choose $m \in \mathbb{N}$ such that $2^m > n$, and apply (i) to the 2^m numbers $a_1, \ldots, a_n, g, \ldots, g$, with g repeated $2^m - n$ times, where $g = \sqrt[n]{a_1 a_2 \cdots a_n}$.
56. Use the Cauchy–Schwarz inequality to prove the **A.M.-H.M. inequality**, namely, if $n \in \mathbb{N}$ and $a_1, \ldots, a_n$ are positive real numbers, then prove that

$$\frac{a_1 + \cdots + a_n}{n} \geq \frac{n}{r}, \quad \text{where} \quad r := \frac{1}{a_1} + \cdots + \frac{1}{a_n}.$$

57. Given any $n \in \mathbb{N}$ and positive real numbers $a_1, \ldots, a_n$ let

$$M_p = \left(\frac{a_1^p + \cdots + a_n^p}{n}\right)^{1/p} \quad \text{for } p \in \mathbb{Q} \text{ with } p \neq 0.$$

Prove that if $p \in \mathbb{Q}$ is positive, then $M_p \leq M_{2p}$ and equality holds if and only if $a_1 = \cdots = a_n$. (Hint: Cauchy–Schwarz inequality.)
    [Note: M_p is called the pth **power mean** of $a_1, \ldots, a_n$. In fact, M_1 is the arithmetic mean and M_{-1} is the harmonic mean, while M_2 is called the **root mean square** of $a_1, \ldots, a_n$. A general *power mean inequality*, which includes as a special case the above inequality $M_p \leq M_{2p}$, the A.M.-G.M. inequality, and the G.M.-H.M. inequality, is described in Exercise 27 in the list of Revision Exercises at the end of Chapter 7.]
58. Given any $\ell(x), p(x) \in \mathbb{R}[x]$ with $\ell(x) \neq 0$, use induction on $\deg \ell(x)$ to prove that there are unique polynomials $q(x)$ and $r(x)$ in $\mathbb{R}[x]$ such that

$$p(x) = q(x)\ell(x) + r(x), \text{ and either } r(x) = 0 \text{ or } \deg r(x) < \deg \ell(x).$$

59. Given any $p(x) \in \mathbb{R}[x]$ and $\alpha \in \mathbb{R}$, show that there is a unique polynomial $q(x) \in \mathbb{R}[x]$ such that $p(x) = (x - \alpha)q(x) + p(\alpha)$. Deduce that α is a root of $p(x)$ if and only if the polynomial $(x - \alpha)$ divides $p(x)$.
60. Show that a nonzero polynomial in $\mathbb{R}[x]$ of degree n has at most n roots in $\mathbb{R}$. (Hint: Exercise 59.)

61. Given any $p(x), q(x) \in \mathbb{R}[x]$, not both zero, a polynomial $d(x)$ in $\mathbb{R}[x]$ satisfying
    (i) $d(x) \mid p(x)$ and $d(x) \mid q(x)$, and
    (ii) $e(x) \in \mathbb{R}[x]$, $e(x) \mid p(x)$ and $e(x) \mid q(x) \implies e(x) \mid d(x)$
    is called a **greatest common divisor**, or simply a **GCD**, of $p(x)$ and $q(x)$. In case $p(x) = q(x) = 0$, we set the GCD of $p(x)$ and $q(x)$ to be 0. Prove that for any $p(x), q(x) \in \mathbb{R}[x]$, a GCD of $p(x)$ and $q(x)$ exists, and is unique up to multiplication by a nonzero constant, that is, if $d_1(x)$ as well as $d_2(x)$ is a GCD of $p(x)$ and $q(x)$, then $d_2(x) = cd_1(x)$ for some $c \in \mathbb{R}$ with $c \neq 0$. Further, show that any GCD of $p(x)$ and $q(x)$ can be expressed as $u(x)p(x) + v(x)q(x)$ for some $u(x), v(x) \in \mathbb{R}[x]$. (Hint: Consider a polynomial of least degree in the set $\{u(x)p(x) + v(x)q(x) : u(x), v(x) \in \mathbb{R}[x] \text{ with } u(x)p(x) + v(x)q(x) \neq 0\}$.)

62. Let $p(x), q(x) \in \mathbb{R}[x]$. Show that $p(x)$ and $q(x)$ are relatively prime if and only if $u(x)p(x) + v(x)q(x) = 1$ for some $u(x), v(x) \in \mathbb{R}[x]$. Is it true that if a nonzero polynomial $d(x) \in \mathbb{R}[x]$ satisfies $u(x)p(x) + v(x)q(x) = d(x)$ for some $u(x), v(x) \in \mathbb{R}[x]$, then $d(x)$ is a GCD of $p(x)$ and $q(x)$?

63. Let $p(x), q(x) \in \mathbb{R}[x]$ be relatively prime polynomials of positive degree. Show by an example that the polynomials $u(x), v(x) \in \mathbb{R}[x]$ such that $u(x)p(x) + v(x)q(x) = 1$ need not be unique. Show, however, that there are unique $u(x), v(x) \in \mathbb{R}[x]$ such that $u(x)p(x) + v(x)q(x) = 1$ and either $u(x) = 0$ or $\deg u(x) < \deg q(x)$. In this case show that either $v(x) = 0$ or $\deg v(x) < \deg p(x)$. (Hint: Exercise 58.)

64. Let $p(x)$, $q_1(x)$ and $q_2(x)$ be nonzero polynomials in $\mathbb{R}[x]$ of degrees m, d_1, and d_2, respectively. Assume that $q_1(x)$ and $q_2(x)$ are relatively prime and that d_1 and d_2 are positive. If $m < d_1 + d_2$, then show that there are unique polynomials $u_1(x), u_2(x) \in \mathbb{R}[x]$ such that $p(x) = u_1(x)q_2(x) + u_2(x)q_1(x)$ and for $i = 1, 2$, either $u_i(x) = 0$ or $\deg u_i(x) < \deg q_i(x)$. Deduce that if $q(x) := q_1(x)q_2(x)$, then
$$\frac{p(x)}{q(x)} = \frac{u_1(x)}{q_1(x)} + \frac{u_2(x)}{q_2(x)}.$$

65. Let $q(x) \in \mathbb{R}[x]$ be a nonzero polynomial of degree n. Use the Real Fundamental Theorem of Algebra to write $q(x) = q_1(x)^{e_1} \cdots q_k(x)^{e_k}$, where $e_1, \ldots, e_k$ are positive integers and $q_1(x), \ldots, q_k(x)$ are distinct polynomials in $\mathbb{R}[x]$ that are either linear of the form $x - \alpha$ with $\alpha \in \mathbb{R}$ or quadratic of the form $x^2 + \beta x + \gamma$ with $\beta, \gamma \in \mathbb{R}$ such that $\beta^2 - 4\gamma < 0$. Show that the polynomials $q_i(x)^{e_i}$ and $q_j(x)^{e_j}$ are relatively prime for $i, j = 1, \ldots, k$ with $i \neq j$. Use Exercise 64 and induction to show that given any nonzero polynomial $p(x) \in \mathbb{R}[x]$ with $\deg p(x) < \deg q(x)$, there are unique polynomials $u_1(x), \ldots, u_k(x) \in \mathbb{R}[x]$ such that either $u_i(x) = 0$ or $\deg u_i(x) < e_i \deg q_i(x)$ for $i = 1, \ldots, k$ and
$$\frac{p(x)}{q(x)} = \frac{u_1(x)}{q_1(x)^{e_1}} + \cdots + \frac{u_k(x)}{q_k(x)^{e_k}}.$$

66. Let $p(x), q(x) \in \mathbb{R}[x]$ be nonzero polynomials of degrees m and n respectively. Let $e \in \mathbb{N}$ be such that $m < en$. Show that there are polynomials $A_1(x), \ldots, A_e(x)$ in $\mathbb{R}[x]$ such that either $A_i(x) = 0$ or $\deg A_i(x) < n$ for $i = 1, \ldots, e$ and $p(x) = A_1(x)q(x)^{e-1} + A_2(x)q(x)^{e-2} + \cdots + A_e(x)$. Deduce that
$$\frac{p(x)}{q(x)} = \frac{A_1(x)}{q(x)} + \frac{A_2(x)}{q(x)^2} + \cdots + \frac{A_e(x)}{q(x)^e}.$$
(Hint: Let $R_0(x) := p(x)$. Use Exercise 58 to find $A_1(x), \ldots, A_e(x)$ and $R_1(x), \ldots, R_e(x)$ successively in such a way that $R_{i-1}(x) = A_i(x)q(x)^{e-i} + R_i(x)$ for $i = 1, \ldots, e$.)

67. Use Exercises 58, 64, 65, and 66 to show that every rational function has a partial fraction decomposition as stated in this chapter.

68. Let $\mathbb{C}[x]$ denote the set of all polynomials in one variable x with coefficients in $\mathbb{C}$. Elements of $\mathbb{C}[x]$ look like $c_n x^n + c_{n-1} x^{n-1} + \cdots + c_1 x + c_0$, where n is a nonnegative integer and $c_0, c_1, \ldots, c_n \in \mathbb{C}$. In particular, the set $\mathbb{R}[x]$ of all polynomials in x with coefficients in $\mathbb{R}$ is a subset of $\mathbb{C}[x]$. For polynomials in $\mathbb{C}[x]$, the notions of equality, coefficients, degree, addition, multiplication, evaluation, and division are defined in a similar way as in the case of $\mathbb{R}[x]$. Given any $p(x) \in \mathbb{C}[x]$ (in particular, any $p(x) \in \mathbb{R}[x]$) and $\alpha \in \mathbb{C}$, we say that α is a (complex) **root** of $p(x)$ if $p(\alpha) = 0$. Show that the results in Exercises 58, 59, 61, and 63 are valid if $\mathbb{R}$ is replaced throughout by $\mathbb{C}$.

69. The **Fundamental Theorem of Algebra** states that if $p(x)$ is a polynomial in $\mathbb{C}[x]$ of positive degree, then $p(x)$ has at least one root in $\mathbb{C}$.
    (i) Assuming the Fundamental Theorem of Algebra, show that if $p(x)$ is a polynomial in $\mathbb{C}[x]$ of positive degree $n \in \mathbb{N}$, then we can write
    $$p(x) = c(x - \alpha_1) \cdots (x - \alpha_n),$$
    where c is the leading coefficient of $p(x)$ and $\alpha_1, \ldots, \alpha_n$ are (not necessarily distinct) complex numbers.
    (ii) Show that if $p(x) \in \mathbb{R}[x]$ and if a complex number $\alpha = a + ib$ is a root of $p(x)$, then its **conjugate** $\bar{\alpha} := a - ib$ is also a root of $p(x)$.
    (iii) Show that the Fundamental Theorem of Algebra, as stated above, and the Real Fundamental Theorem of Algebra, as stated in this chapter, are equivalent to each other, that is, assuming one of them, we can deduce the other.

70. Let $\mathbb{C}[x, y]$ denote the set of all polynomials in two variables x and y with coefficients in $\mathbb{C}$. Elements of $\mathbb{C}[x, y]$ look like $P(x, y) = \sum c_{i,j} x^i y^j$, where i and j vary over finite sets of nonnegative integers, and $c_{i,j} \in \mathbb{C}$. If $P(x, y)$ is not the zero polynomial, that is, if some $c_{i,j}$ is nonzero, then the **(total) degree** of $P(x, y)$ is defined be $\max\{i + j : c_{i,j} \neq 0\}$. We say that $P(x, y)$ is **homogeneous** of degree m if each term has degree m, that is, $i + j = m$ whenever $c_{i,j} \neq 0$. As in the case of polynomials in one variable, we can substitute real or complex numbers for the variables x and y. A pair (α, β), where $\alpha, \beta \in \mathbb{C}$, is called a **root** of $P(x, y)$ if $P(\alpha, \beta) = 0$.

(i) Show that there are nonzero polynomials in $\mathbb{C}[x, y]$ with infinitely many roots. Show, however, that there is no nonzero polynomial $P(x, y) \in \mathbb{C}[x, y]$ such that $P(\alpha, \beta) = 0$ for all $\alpha \in D$ and $\beta \in E$, where both D and E are infinite subsets of $\mathbb{C}$.

(ii) Show that if $P(x, y)$ is a homogeneous polynomial of positive degree m, then $P(x, y)$ factors as a product of homogeneous polynomials of degree 1, that is,

$$P(x, y) = \prod_{i=1}^{m} (\alpha_i x + \beta_i y) \quad \text{for some } \alpha_i, \beta_i \in \mathbb{C}, \ 1 \le i \le m.$$

Deduce that the pair $(\beta_i, -\alpha_i)$ is a root of $P(x, y)$ for $i = 1, \dots, m$, and up to proportionality, these are the only roots of $P(x, y)$, that is, if (α, β) is a root of $P(x, y)$, then $(\alpha, \beta) = (\lambda \beta_i, -\lambda \alpha_i)$ for some $\lambda \in \mathbb{C}$ and $i \in \{1, \dots, m\}$. (Hint: Consider $P(x, 1)$ or $P(y, 1)$, and use Exercise 69.)

71. Let $f : [a, b] \to \mathbb{R}$ be a function and c be any point of (a, b).
   (i) If f is monotonically (resp. strictly) increasing on $[a, c]$ and on $[c, b]$, then show that f is monotonically (resp. strictly) increasing on $[a, b]$.
   (ii) If f is convex (resp. strictly convex) on $[a, c]$ and on $[c, b]$, then is it true that f is convex (resp. strictly convex) on $[a, b]$?

72. Let I be an interval containing more than one point and $f : I \to \mathbb{R}$ be any function. Given any $x_1, x_2 \in I$ with $x_1 \ne x_2$, define

$$\phi(x_1, x_2) := \frac{f(x_1) - f(x_2)}{x_1 - x_2}.$$

Show that f is convex on I if and only if ϕ is a monotonically increasing function of x_1, that is, $\phi(x_1, x) \le \phi(x_2, x)$ for all $x_1, x_2 \in I$ with $x_1 < x_2$ and $x \in I \setminus \{x_1, x_2\}$.

73. Let I be an interval containing more than one point and $f : I \to \mathbb{R}$ be any function. Given any distinct points $x_1, x_2, x_3 \in I$, define

$$\Psi(x_1, x_2, x_3) = \frac{(x_3 - x_2)f(x_1) + (x_1 - x_3)f(x_2) + (x_2 - x_1)f(x_3)}{(x_1 - x_2)(x_2 - x_3)(x_3 - x_1)}.$$

(i) Show that Ψ is a symmetric function of x_1, x_2, x_3, that is, show that $\Psi(x_1, x_2, x_3) = \Psi(x_2, x_1, x_3) = \Psi(x_3, x_2, x_1) = \Psi(x_1, x_3, x_2) = \Psi(x_2, x_3, x_1) = \Psi(x_3, x_1, x_2)$.

(ii) If ϕ is as in Exercise 72 above, then show that

$$\Psi(x_1, x_2, x_3) = \frac{\phi(x_1, x_3) - \phi(x_2, x_3)}{x_1 - x_2} \quad \text{for distinct } x_1, x_2, x_3 \in I.$$

(iii) Show that f is convex on I if and only if $\Psi(x_1, x_2, x_3) \ge 0$ for all distinct points $x_1, x_2, x_3 \in I$.

2

Sequences

The word *sequence* is almost self-explanatory. It refers to a succession of certain objects. For us, these objects will be real numbers. A sequence of real numbers looks like an infinite succession such as

$$a_1, a_2, a_3, a_4, a_5, \ldots$$

where the a_n's are real numbers. For example,

$$1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \ldots$$

is a sequence of real numbers. If we see a sequence, it is natural to ask where it leads. For example, the above sequence $1, \frac{1}{2}, \frac{1}{3}, \ldots$ seems to approach 0. We shall make this idea precise by defining the notion of convergence of sequences.

In Section 2.1 below, we begin with the formal definition of a sequence and go on to discuss a number of basic concepts and results. Along the way, we will look at numerous examples of sequences and see whether they approach a fixed number. Next, in Section 2.2, we consider the notion of a subsequence of a given sequence, and also a special class of sequences known as Cauchy sequences. We show that any sequence in $\mathbb{R}$ that satisfies a mild condition of being 'bounded' has a convergent subsequence. Also, we show that a sequence in $\mathbb{R}$ is convergent if and only if it is a Cauchy sequence.

2.1 Convergence of Sequences

A **sequence** (in $\mathbb{R}$) is a real-valued function whose domain is the set $\mathbb{N}$ of all natural numbers. Usually, we shall denote sequences by (a_n), (b_n), and so on, or sometimes by (A_n), (B_n), and so on. The value of a sequence (a_n) at $n \in \mathbb{N}$ is given by a_n and this is called the nth **term** of that sequence. Here are some simple examples: For $n \in \mathbb{N}$, consider

(i) $a_n := 1$, (ii) $a_n := (-1)^n$, (iii) $a_n := \dfrac{1}{n}$, (iv) $a_n := n$, (v) $a_n := (-1)^n n$.

We shall use the terms 'bounded', 'unbounded', 'bounded above', 'bounded below' for a sequence just as we use them for any function. Thus the sequences defined in (i), (ii), and (iii) are bounded, but those defined in (iv) and (v) are not bounded. The sequence defined in (iv) is bounded below (by 1), but it is not bounded above, while the sequence defined in (v) is neither bounded above nor bounded below.

We say that a sequence (a_n) is **convergent** if there is $a \in \mathbb{R}$ that satisfies the following condition: For every $\epsilon > 0$, there is $n_0 \in \mathbb{N}$ such that

$$|a_n - a| < \epsilon \quad \text{for all} \ \ n \geq n_0.$$

In this case, we say that (a_n) **converges** to a or that a is a **limit** of (a_n), and write $a_n \to a$ (as $n \to \infty$). A sequence that is not convergent is said to be **divergent**.

Examples 2.1. (i) If $a_n := 1$ for $n \in \mathbb{N}$, then obviously $a_n \to 1$.

(ii) If $a_n := (-1)^n$ for $n \in \mathbb{N}$, then (a_n) is divergent. This can be seen as follows. Let $a \in \mathbb{R}$. If $|a| \neq 1$, let $\epsilon := \min\{|a - 1|, |a + 1|\}$. Then $\epsilon > 0$. Observe that $|a_n - a| \geq \epsilon$ for all n. If $|a| = 1$, then let $\epsilon := 2$, and observe that $|a_n - 1| \geq \epsilon$ for all odd n and $|a_n - (-1)| \geq \epsilon$ for all even n.

(iii) If $a_n := 1/n$ for $n \in \mathbb{N}$, then $a_n \to 0$. Indeed, given $\epsilon > 0$, we can take $n_0 := [1/\epsilon] + 1$, and then $|a_n - 0| < \epsilon$ for all $n \geq n_0$.

(iv) If $a_n := n$ for $n \in \mathbb{N}$, then (a_n) is divergent. Indeed, given $a \in \mathbb{R}$, we have $|a_n - a| \geq 1$ for all $n \geq [|a|] + 1$. Similarly, if $a_n := (-1)^n n$ for $n \in \mathbb{N}$, then (a_n) is divergent.

Before we begin our discussion of convergent sequences, we make two observations.

First, the convergence of a sequence (a_n) is not altered if a finite number of a_n's are replaced by some other b_n's. Thus if we replace $a_{n_1}, \ldots, a_{n_k}$ by $b_{n_1}, \ldots, b_{n_k}$ respectively, then the altered sequence converges if and only if the original sequence converges. With this in view, we may sometimes regard $(1/a_n)$ as a sequence if we know that all except finitely many a_n's are nonzero.

Next, if $a_n \to a$, then the inequality

$$\big| \, |a_n| - |a| \, \big| \leq |a_n - a|, \quad n \in \mathbb{N},$$

shows that $|a_n| \to |a|$. The converse is not true as can be seen by considering $a_n = (-1)^n$ for $n \in \mathbb{N}$. However, if $|a_n| \to 0$, then clearly $a_n \to 0$.

Proposition 2.2. (i) *A convergent sequence has a unique limit.*
(ii) *A convergent sequence is bounded.*

Proof. (i) Suppose $a_n \to a$ as well as $a_n \to b$. If $b \neq a$, let $\epsilon := |a - b|$. Since $a_n \to a$, there is $n_1 \in \mathbb{N}$ such that $|a_n - a| < \epsilon/2$ for all $n \geq n_1$, and since $a_n \to b$, there is $n_2 \in \mathbb{N}$ such that $|a_n - b| < \epsilon/2$ for all $n \geq n_2$. Let $n_0 := \max\{n_1, n_2\}$. Then

$$|a - b| \leq |a - a_{n_0}| + |a_{n_0} - b| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = |a - b|.$$

This contradiction shows that $b = a$.

(ii) Let $a_n \to a$. Then there is $n_0 \in \mathbb{N}$ such that $|a_n - a| < 1$ for all $n > n_0$. If $\alpha := \max\{|a_1|, \ldots, |a_{n_0}|, |a| + 1\}$, then $|a_n| \leq \alpha$ for all $n \in \mathbb{N}$. Hence (a_n) is bounded. $\qquad\square$

Let $a_n \to a$. In view of part (i) of Proposition 2.2, we say that a is *the* limit of (a_n) and write

$$\lim_{n \to \infty} a_n = a.$$

Example 2.1 (ii) shows that the converse of part (ii) of the above proposition does not hold.

We now prove some results that are useful in proving convergence or divergence of a variety of sequences.

First we consider how the algebraic operations on $\mathbb{R}$ are related to the concept of the convergence of a sequence of real numbers. The following result is known as the **Limit Theorem for Sequences**.

Proposition 2.3. *Let $a_n \to a$ and $b_n \to b$. Then*

(i) $a_n + b_n \to a + b$,
(ii) $ra_n \to ra$ *for any $r \in \mathbb{R}$,*
(iii) $a_n b_n \to ab$,
(iv) *if $a \neq 0$, then there is $m \in \mathbb{N}$ such that $a_n \neq 0$ for all $n \geq m$, and*

$$\frac{1}{a_n} \to \frac{1}{a}.$$

Proof. Let $\epsilon > 0$ be given. There are $n_1, n_2 \in \mathbb{N}$ such that

$$|a_n - a| < \epsilon \text{ for all } n \geq n_1 \quad \text{and} \quad |b_n - b| < \epsilon \text{ for all } n \geq n_2.$$

(i) Let $n_0 := \max\{n_1, n_2\}$. Then for all $n \geq n_0$,

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

(ii) Let $n_0 := n_1$. Then for all $n \geq n_0$,

$$|ra_n - ra| = |r|\,|a_n - a| < |r|\epsilon.$$

(iii) By part (ii) of Proposition 2.2, there is $\alpha \in \mathbb{R}$ such that $|a_n| \leq \alpha$ for all $n \in \mathbb{N}$. Let $n_0 := \max\{n_1, n_2\}$. Then for all $n \geq n_0$,

$$|a_n b_n - ab| = |a_n(b_n - b) + (a_n - a)b|$$
$$\leq |a_n|\,|b_n - b| + |a_n - a|\,|b|$$
$$\leq \alpha\epsilon + \epsilon|b| = (\alpha + |b|)\epsilon.$$

(iv) Since $|a| > 0$, there is $m \in \mathbb{N}$ such that $|a_n - a| < |a|/2$ for all $n \geq m$. But then $|a_n| \geq |a| - |a - a_n| > |a|/2$ for all $n \geq m$. Let $n_0 := \max\{n_1, m\}$. Then for all $n \geq n_0$, we have $a_n \neq 0$ and

$$\left| \frac{1}{a_n} - \frac{1}{a} \right| = \frac{|a - a_n|}{|a_n|\,|a|} < \frac{2\epsilon}{|a|^2}.$$

Since $\epsilon > 0$ is arbitrary, the desired conclusions follow.   □

With notation and hypotheses as in the above proposition, a combined application of parts (i) and (ii) of Proposition 2.3 shows that $a_n - b_n \to a - b$. Likewise, a combined application of parts (iii) and (iv) of Proposition 2.3 shows that if $b \neq 0$, then $a_n/b_n \to a/b$. Further, given any $m \in \mathbb{Z}$, successive applications of part (iii) or part (iv) of Proposition 2.3 show that $a_n^m \to a^m$, provided $a \neq 0$ in case $m < 0$.

Next, we show how the order relation on $\mathbb{R}$ and the operation of taking the kth root are preserved under convergence.

Proposition 2.4. *Let (a_n) and (b_n) be sequences and a, b be real numbers such that $a_n \to a$ and $b_n \to b$.*

(i) *If there is $n_0 \in \mathbb{N}$ such that $a_n \leq b_n$ for all $n \geq n_0$, then $a \leq b$. Conversely, if $a < b$, then there is $m_0 \in \mathbb{N}$ such that $a_n < b_n$ for all $n \geq m_0$.*
(ii) *If $a_n \geq 0$ for all $n \in \mathbb{N}$, then $a \geq 0$ and $a_n^{1/k} \to a^{1/k}$ for any $k \in \mathbb{N}$.*

Proof. (i) Suppose $b < a$. Let $\epsilon := (a-b)/2$. Since $b_n \to b$, there is $n_1 \in \mathbb{N}$ such that $n_1 \geq n_0$ and $b_n < b + \epsilon$ for all $n \geq n_1$. But since $b + \epsilon = (a+b)/2 = a - \epsilon$, we have $a_n \leq b_n < a - \epsilon$ for all $n \geq n_1$. This is a contradiction to $a_n \to a$. Hence $a \leq b$.

Conversely, suppose $a < b$. Let $\epsilon := (b - a)/2$. There is $m_1 \in \mathbb{N}$ such that $a_n < a + \epsilon$ for all $n \geq m_1$ and there is $m_2 \in \mathbb{N}$ such that $b_n > b - \epsilon$ for all $n \geq m_2$. Let $m_0 := \max\{m_1, m_2\}$. Since $b - \epsilon = (a + b)/2 = a + \epsilon$, we have $a_n < (a + b)/2 < b_n$ for all $n \geq m_0$.

(ii) Part (i) implies that $a \geq 0$. Let $k \in \mathbb{N}$ and $\epsilon > 0$ be given. Since $\epsilon^k > 0$, there is $n_2 \in \mathbb{N}$ such that $|a_n - a| < \epsilon^k$ for all $n \geq n_2$. Hence by the basic inequality for roots (part (ii) of Proposition 1.9), we obtain

$$|a_n^{1/k} - a^{1/k}| \leq |a_n - a|^{1/k} < \epsilon \quad \text{for all } n \geq n_2.$$

It follows that $a_n^{1/k} \to a^{1/k}$.   □

We note that it is possible to have $a_n \to a$, $b_n \to b$, $a_n < b_n$ for all $n \in \mathbb{N}$ and yet $a = b$. For example, let $a_n := 0$ and $b_n := 1/n$ for all $n \in \mathbb{N}$ and observe that $a = 0 = b$.

With notation and hypotheses as in the above proposition, a combined application of part (iii) of Proposition 2.3 and part (ii) of Proposition 2.4 shows that if $a_n \geq 0$ for all $n \in \mathbb{N}$, then $a_n^r \to a^r$, where r is any positive rational number, since we can write $r = m/k$, where m, $k \in \mathbb{N}$. This, together with part (iv) of Proposition 2.3, shows that if $a > 0$, then $a_n^r \to a^r$, where r is any negative rational number.

Proposition 2.5 (Sandwich Theorem). *Let (a_n), (b_n), (c_n) be sequences and $c \in \mathbb{R}$ be such that $a_n \leq c_n \leq b_n$ for all $n \in \mathbb{N}$ and $a_n \to c$ as well as $b_n \to c$. Then $c_n \to c$.*

Proof. Let $\epsilon > 0$ be given. Since $a_n \to c$, there is $n_1 \in \mathbb{N}$ such that $a_n - c > -\epsilon$ for all $n \geq n_1$, and since $b_n \to c$, there is $n_2 \in \mathbb{N}$ such that $b_n - c < \epsilon$ for all $n \geq n_2$. Let $n_0 := \max\{n_1, n_2\}$. Then

$$-\epsilon < a_n - c \leq c_n - c \leq b_n - c < \epsilon \quad \text{for all } n \geq n_0.$$

It follows that $c_n \to c$. $\qquad\square$

We now use the above result to show that the supremum and the infimum of a subset of $\mathbb{R}$ are limits of sequences in that subset.

Corollary 2.6. *Let E be a nonempty subset of $\mathbb{R}$.*

(i) *If E is bounded above and $a := \sup E$, then there is a sequence (a_n) such that $a_n \in E$ for all $n \in \mathbb{N}$ and $a_n \to a$.*

(ii) *If E is bounded below and $b := \inf E$, then there is a sequence (b_n) such that $b_n \in E$ for all $n \in \mathbb{N}$ and $b_n \to b$.*

Proof. To prove (i), suppose E is bounded above. Let $a := \sup E$. Then for every $n \in \mathbb{N}$, there is $a_n \in E$ such that $a_n > a - (1/n)$. Since $a_n \leq a$, we see that $0 \leq a - a_n < (1/n)$ for all $n \in \mathbb{N}$. Thus, by the Sandwich Theorem, $a_n \to a$.

The proof of (ii) is similar. $\qquad\square$

Examples 2.7. (i) Let $a \in \mathbb{R}$ with $|a| < 1$. Then

$$\lim_{n \to \infty} a^n = 0.$$

If $a = 0$, this is obvious. Suppose $a \neq 0$. Write $1/|a| = 1 + h$. Then $h > 0$ and so by the binomial inequality given in Proposition 1.10, we have

$$\frac{1}{|a|^n} = (1 + h)^n \geq 1 + nh > nh \quad \text{for all } n \in \mathbb{N}.$$

Hence $0 < |a|^n < 1/nh$. By part (ii) of Proposition 2.3 and Example 2.1 (iii), $1/nh \to 0$. Therefore, by the Sandwich Theorem, $|a|^n \to 0$, and thus $a^n \to 0$. As a consequence, we can find the limit of the sequence (A_n), where $A_n := 1 + a + \cdots + a^n$ for $n \in \mathbb{N}$. Since $A_n = (1 - a^{n+1})/(1 - a)$ for $n \in \mathbb{N}$, it follows that

$$\lim_{n\to\infty} (1 + a + \cdots + a^n) = \frac{1}{1 - a}.$$

(ii) Let $a \in \mathbb{R}$. Then

$$\lim_{n\to\infty} \frac{a^n}{n!} = 0.$$

Choose $m \in \mathbb{N}$ such that $|a| < m$. Then for $n > m$, we have

$$0 \le \left|\frac{a^n}{n!}\right| = \frac{|a|^m}{m!} \prod_{j=m+1}^{n} \frac{|a|}{j} < \frac{|a|^m}{m!} \left(\frac{|a|}{m}\right)^{n-m} = \frac{m^m}{m!} \left(\frac{|a|}{m}\right)^n.$$

Since m is a constant and $(|a|/m)^n \to 0$ by (i) above, the Sandwich Theorem shows that $|a^n/n!| \to 0$, that is, $a^n/n! \to 0$.

(iii) Let $a \in \mathbb{R}$ and $a > 0$. Then

$$\lim_{n\to\infty} a^{1/n} = 1.$$

This is obvious if $a = 1$. Suppose $a > 1$ and $\delta_n := a^{1/n} - 1$ for $n \in \mathbb{N}$. By the Binomial Theorem,

$$a = (1 + \delta_n)^n = 1 + n\delta_n + \cdots + \delta_n^n > n\delta_n,$$

so that $0 < \delta_n < a/n$ for all $n \in \mathbb{N}$. Since $a/n \to 0$, it follows from the Sandwich Theorem that $\delta_n \to 0$, that is, $a^{1/n} \to 1$. If $0 < a < 1$, let $b := 1/a$, so that $1/a^{1/n} = b^{1/n} \to 1$ because $b > 1$. Hence by part (iv) of Proposition 2.3, $a^{1/n} \to 1/1 = 1$.

(iv) For $n \in \mathbb{N}$, let

$$C_n := \frac{n}{n^2 + 1} + \frac{n}{n^2 + 2} + \cdots + \frac{n}{n^2 + n}.$$

Then $C_n \to 1$. To see this, let

$$A_n := \sum_{k=1}^{n} \frac{n}{n^2 + n} = \frac{n^2}{n^2 + n} \quad \text{and} \quad B_n := \sum_{k=1}^{n} \frac{n}{n^2 + 1} = \frac{n^2}{n^2 + 1} \quad \text{for } n \in \mathbb{N}.$$

Then $A_n \le C_n \le B_n$ for all $n \in \mathbb{N}$, and by Proposition 2.3,

$$A_n = \frac{1}{1 + (1/n)} \to 1 \quad \text{and} \quad B_n = \frac{1}{1 + (1/n^2)} \to 1.$$

Hence by the Sandwich Theorem, $C_n \to 1$. $\diamond$

We have seen in part (ii) of Proposition 2.2 that every convergent sequence is bounded. On the other hand, not every bounded sequence is convergent, as is shown by the example $a_n = (-1)^n$ for $n \in \mathbb{N}$. We shall now consider a class of sequences for which convergence is equivalent to boundedness.

A sequence (a_n) is called **(monotonically) increasing** if the corresponding function $n \mapsto a_n$ is (monotonically) increasing, that is, if $a_n \leq a_{n+1}$ for all $n \in \mathbb{N}$. Likewise, it is called **(monotonically) decreasing** if the corresponding function is (monotonically) decreasing, that is, if $a_n \geq a_{n+1}$ for all $n \in \mathbb{N}$. A sequence is said to be **monotonic** if it is either monotonically increasing or monotonically decreasing.

Proposition 2.8. (i) *A monotonically increasing sequence is convergent if and only if it is bounded above. Moreover, if a sequence (a_n) is monotonically increasing and bounded above, then*

$$\lim_{n \to \infty} a_n = \sup\{a_n : n \in \mathbb{N}\}.$$

(ii) *A monotonically decreasing sequence is convergent if and only if it is bounded below. Moreover, if a sequence (a_n) is monotonically increasing and bounded above, then*

$$\lim_{n \to \infty} a_n = \inf\{a_n : n \in \mathbb{N}\}.$$

Proof. (i) Let (a_n) be a monotonically increasing sequence. Suppose it is bounded above. Then the set $\{a_n : n \in \mathbb{N}\}$ has a supremum. Let $a := \sup\{a_n : n \in \mathbb{N}\}$. Given $\epsilon > 0$, there is $n_0 \in \mathbb{N}$ such that $a - \epsilon < a_{n_0}$. But since (a_n) is monotonically increasing, we have $a_{n_0} \leq a_n$ for all $n \geq n_0$. Hence

$$a - \epsilon < a_n \leq a < a + \epsilon \quad \text{for all } n \geq n_0.$$

Thus $a_n \to a$. Conversely, if (a_n) is convergent, then it is bounded above by part (ii) of Proposition 2.2.

(ii) A proof similar to the one above can be given. Alternatively, one may observe that if (a_n) is monotonically decreasing, and if we let $b_n := -a_n$, then (b_n) monotonically increasing. Also, (a_n) is bounded below if and only if (b_n) is bounded above, and in this case, $\inf\{a_n : n \in \mathbb{N}\} = -\sup\{b_n : n \in \mathbb{N}\}$. Also, by part (ii) of Proposition 2.3, $a_n \to a$ if and only if $-a_n \to -a$. Thus the desired results follow from (i) above. $\square$

Corollary 2.9. *A monotonic sequence is convergent if and only if it is bounded.*

Proof. Follows from parts (i) and (ii) of Proposition 2.8. $\square$

Examples 2.10. (i) Consider the sequence (A_n) defined by

$$A_n := \sum_{k=0}^{n} \frac{1}{k!} = 1 + \frac{1}{1!} + \frac{1}{2!} + \cdots + \frac{1}{n!} \quad \text{for } n \in \mathbb{N}.$$

Clearly, (A_n) is a monotonically increasing sequence. Also, for any $n \in \mathbb{N}$,

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

Hence (A_n) is bounded above. So by part (i) of Proposition 2.8, (A_n) is convergent.

(ii) Consider the sequence (B_n) defined by

$$B_n := \left(1 + \frac{1}{n}\right)^n \qquad \text{for } n \in \mathbb{N}.$$

We show that (B_n) is convergent and its limit is equal to the limit of the sequence (A_n) considered in (i) above. By the Binomial Theorem, we have

$$B_n = \sum_{k=0}^{n} \frac{n(n-1)\cdots(n-k+1)}{k!}\frac{1}{n^k} = \sum_{k=0}^{n} 1 \left(1 - \frac{1}{n}\right)\cdots\left(1 - \frac{k-1}{n}\right)\frac{1}{k!}.$$

This implies that $B_n \leq A_n$ for all $n \in \mathbb{N}$.

To find a lower bound for B_n in terms of A_n, we use the generalized binomial inequality given in Proposition 1.10, and obtain for $k = 1, \ldots, n$,

$$(1-0)\left(1 - \frac{1}{n}\right)\cdots\left(1 - \frac{k-1}{n}\right) \geq 1 - \left(0 + \frac{1}{n} + \cdots + \frac{k-1}{n}\right).$$

Now, since $0 + 1 + 2 + \cdots + (k-1) = (k-1)k/2$, we see that

$$1 - \frac{(k-1)k}{2n} \leq \frac{n(n-1)\cdots(n-k+1)}{n^k} \qquad \text{for } k = 0, 1, \ldots, n.$$

Dividing by $k!$ and summing from $k = 0$ to $k = n$, we have

$$\sum_{k=0}^{n} \frac{1}{k!} - \frac{1}{2n}\sum_{k=2}^{n} \frac{1}{(k-2)!} \leq B_n.$$

Moreover, from (i) above,

$$\sum_{k=2}^{n} \frac{1}{(k-2)!} \leq A_n < 3 \quad \text{and hence} \quad A_n - \frac{3}{2n} < B_n \leq A_n \quad \text{for all } n \in \mathbb{N}.$$

Therefore, by the Sandwich Theorem, (B_n) is convergent and its limit is equal to the limit of (A_n) .

Alternatively, we may argue as follows. Let $A_n \to A$ and $\epsilon > 0$ be given. Then there is $n_0 \in \mathbb{N}$ such that $A - (\epsilon/2) < a_n$ for all $n \geq n_0$. In particular,

$$A - \frac{\epsilon}{2} < A_{n_0}.$$

For $n \in \mathbb{N}$, define

$$C_n := \sum_{k=0}^{n_0} 1 \left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right) \frac{1}{k!}.$$

Then $C_n \leq B_n$ for all $n \geq n_0$. By parts (i) and (iii) of Proposition 2.3, $C_n \to \sum_{k=0}^{n_0}(1/k!) = A_{n_0}$ as $n \to \infty$. Hence there is $n_1 \in \mathbb{N}$ such that

$$A_{n_0} - \frac{\epsilon}{2} < C_n \quad \text{for all } n \geq n_1.$$

Now for all $n \geq \max\{n_0, n_1\}$, we have

$$A - \epsilon < A_{n_0} - \frac{\epsilon}{2} < C_n \leq B_n \leq A_n \leq A < A + \epsilon.$$

This proves that $B_n \to A$. We remark that yet another proof of the convergence of the sequence (B_n) is indicated in Exercise 8. The common limit of the sequences (A_n) and (B_n) is an important real number. This real number, denoted by e, will be introduced in Chapter 7. See Section 7.1 and, in particular, Corollary 7.6.

(iii) Consider the sequence (A_n) defined by

$$A_n := 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} \quad \text{for } n \in \mathbb{N}.$$

Clearly, (A_n) is a monotonically increasing sequence. Also, for $n \in \mathbb{N}$, we have

$$A_{2^n} = 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) + \cdots + \left(\frac{1}{2^{n-1}+1} + \cdots + \frac{1}{2^n}\right)$$
$$\geq 1 + \frac{1}{2} + \frac{2}{4} + \cdots + \frac{2^{n-1}}{2^n}$$
$$= 1 + \frac{n}{2}.$$

Hence there is no $\alpha \in \mathbb{R}$ such that $A_n \leq \alpha$ for all $n \in \mathbb{N}$, that is, (A_n) is not bounded above. Hence (A_n) is not convergent. Let us modify the sequence (A_n) by changing the signs of alternate summands of its terms and define

$$B_n := 1 - \frac{1}{2} + \frac{1}{3} - \cdots + (-1)^{n-1}\frac{1}{n} \quad \text{for } n \in \mathbb{N}.$$

We shall show that (B_n) is convergent. First note that $\frac{1}{2} \leq B_n \leq 1$. Let $C_n := B_{2n-1}$ and $D_n := B_{2n}$ for $n \in \mathbb{N}$. Then for every $n \in \mathbb{N}$, we have

$$C_{n+1} - C_n = \frac{1}{2n+1} - \frac{1}{2n} \leq 0 \quad \text{and} \quad D_{n+1} - D_n = \frac{1}{2n+1} - \frac{1}{2n+2} \geq 0.$$

Hence (C_n) is a monotonically decreasing sequence that is bounded below by $\frac{1}{2}$, and (D_n) is a monotonically increasing sequence that is bounded above by 1. By Proposition 2.8, both (C_n) and (D_n) are convergent. Let $C_n \to C$ and $D_n \to D$. Now since $D_n - C_n = \frac{1}{2n} \to 0$, we must have $D = C$. It follows that (B_n) is convergent and $B_n \to C$.

(iv) Consider the sequence (A_n) defined by

$$A_n := 1 + \frac{1}{2^2} + \frac{1}{3^2} + \cdots + \frac{1}{n^2} \quad \text{for } n \in \mathbb{N}.$$

Clearly, (A_n) is a monotonically increasing sequence. Also, for $n \in \mathbb{N}$,

$$A_n \le 1 + \frac{1}{1.2} + \frac{1}{2.3} + \cdots + \frac{1}{(n-1)n}$$

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

$$= 2 - \frac{1}{n} < 2.$$

Hence (A_n) is bounded above. So by part (i) of Proposition 2.8, (A_n) is convergent.

(v) Let p be a rational number and consider the sequence (A_n) defined by

$$A_n := 1 + \frac{1}{2^p} + \cdots + \frac{1}{n^p} \quad \text{for } n \in \mathbb{N}.$$

Clearly, (A_n) is a monotonically increasing sequence. We have seen in (iii) and (iv) above that if $p = 1$, then the sequence (A_n) is divergent, whereas if $p = 2$, then it is convergent. This implies that (A_n) is divergent if $p \le 1$, and it is convergent if $p \ge 2$, because for each $n \in \mathbb{N}$,

$$0 < \frac{1}{n} \le \frac{1}{n^p} \quad \text{if } p \le 1, \quad \text{while} \quad 0 < \frac{1}{n^p} \le \frac{1}{n^2} \quad \text{if } p \ge 2.$$

We now give an alternative argument that shows that (A_n) is convergent if $p > 1$. Suppose $p > 1$. For $n \in \mathbb{N}$, we have

$$A_{2n+1} = 1 + \left(\frac{1}{2^p} + \frac{1}{4^p} + \cdots + \frac{1}{(2n)^p}\right) + \left(\frac{1}{3^p} + \frac{1}{5^p} + \cdots + \frac{1}{(2n+1)^p}\right)$$

$$< 1 + 2\left(\frac{1}{2^p} + \frac{1}{4^p} + \cdots + \frac{1}{(2n)^p}\right)$$

$$= 1 + \frac{2}{2^p}\left(1 + \frac{1}{2^p} + \cdots + \frac{1}{n^p}\right)$$

$$= 1 + 2^{1-p}A_n$$

$$< 1 + 2^{1-p}A_{2n+1}.$$

Since $2^{1-p} < 1$, we see that $A_{2n+1} < 1/(1 - 2^{p-1})$ for $n \in \mathbb{N}$. Also, since $A_{2n} < A_{2n+1}$ for all $n \in \mathbb{N}$, it follows that (A_n) is bounded above. So by part (i) of Proposition 2.8, we see that (A_n) is convergent.

(vi) Consider a sequence defined by a linear recurrence relation, that is, let $\alpha, \beta, \gamma \in \mathbb{R}$ and (a_n) be defined by

$$a_1 := \alpha \quad \text{and} \quad a_{n+1} := \beta a_n + \gamma \quad \text{for } n \in \mathbb{N}.$$

Assume that α, β, γ are nonnegative and $\beta < 1$. Then each a_n is nonnegative, and

$$a_2 - a_1 = \beta a_1 + \gamma - a_1 = \gamma - (1 - \beta)\alpha,$$

whereas for $n > 1$,

$$a_{n+1} - a_n = \beta a_n + \gamma - (\beta a_{n-1} + \gamma) = \beta(a_n - a_{n-1}).$$

As a consequence, for $n > 1$,

$$a_{n+1} - a_n = \beta^{n-1}(a_2 - a_1) = \beta^{n-1}[\gamma - (1 - \beta)\alpha].$$

Thus, if $\gamma \leq (1 - \beta)\alpha$, it follows that (a_n) is a monotonically decreasing sequence which is bounded below by 0. Now, assume that $\gamma > (1 - \beta)\alpha$. Then (a_n) is a monotonically increasing sequence. Further, $a_1 = \alpha < \gamma/(1 - \beta)$ and for $n > 1$,

$$a_n \leq \frac{\gamma}{1 - \beta} \implies a_{n+1} = \beta a_n + \gamma \leq \frac{\beta\gamma}{1 - \beta} + \gamma = \frac{\gamma}{1 - \beta}.$$

Hence, in this case, (a_n) is bounded above by $\gamma/(1 - \beta)$. Thus, in any case, Proposition 2.8 shows that (a_n) is convergent. Let $a_n \to a$. Then $a_{n+1} \to a$ and since $a_{n+1} = \beta a_n + \gamma \to \beta a + \gamma$, we obtain that $a = \beta a + \gamma$, that is, $a = \gamma/(1 - \beta)$. $\diamond$

Remark 2.11. We introduce some notation for comparing the orders of magnitude of two sequences (a_n) and (b_n).

If there are $K > 0$ and $n_0 \in \mathbb{N}$ such that $|a_n| \leq K|b_n|$ for all $n \geq n_0$, then we write $a_n = O(b_n)$ [read as (a_n) is big-oh of (b_n)]. In particular, if $b_n = 1$ for all large n, then $a_n = O(1)$, and this means that the sequence (a_n) is bounded. Broadly speaking, $a_n = O(b_n)$ if the order of magnitude of (a_n) is at most the order of magnitude of (b_n). In case (a_n) and (b_n) are monotonically increasing sequences and $a_n = O(b_n)$, then we also say that the **growth rate** of (a_n) is at most the growth rate of (b_n). For example,

$$(-1)^n 10\, n + 100 = O(n) \quad \text{and} \quad (-1)^n \frac{10}{n} + \frac{100}{n\sqrt{n}} = O\left(\frac{1}{n}\right).$$

Given $\epsilon > 0$, if there is $n_0 \in \mathbb{N}$ such that $|a_n| \leq \epsilon|b_n|$ for all $n \geq n_0$, then we write $a_n = o(b_n)$ [read as (a_n) is little-oh of (b_n)]. If $b_n \neq 0$ for all large n, then $a_n = o(b_n)$ means that $\lim_{n\to\infty}(a_n/b_n)$ exists and is zero. In particular, if $b_n = 1$ for all large n, then $a_n = o(1)$, and this means that $a_n \to 0$. Broadly speaking, $a_n = o(b_n)$ if the order of magnitude of (a_n) is less

than the order of magnitude of (b_n). In case (a_n) and (b_n) are monotonically increasing sequences and $a_n = o(b_n)$, then we also say that the **growth rate** of (a_n) is less than the growth rate of (b_n). For example,

$$10\,n + 100 = o(n\sqrt{n}) \quad \text{and} \quad (-1)^n \frac{10}{n} + \frac{100}{n\sqrt{n}} = o\left(\frac{1}{\sqrt{n}}\right).$$

Suppose there is nonzero $\ell \in \mathbb{R}$ that satisfies the following condition: Given $\epsilon > 0$, there is $n_0 \in \mathbb{N}$ such that $|a_n - \ell b_n| < \epsilon$ for all $n \geq n_0$. In this case, we write $a_n \sim b_n$ [read as (a_n) is asymptotically equivalent to (b_n)]. Broadly speaking, $a_n \sim b_n$ if (a_n) is of the same order of magnitude as (b_n). It can be easily seen that $\sim$ is an equivalence relation on the set of all sequences of real numbers. If $b_n \neq 0$ for all large n, then $a_n \sim b_n$ means that $\lim_{n\to\infty}(a_n/b_n)$ exists and is nonzero. If (a_n) and (b_n) are monotonically increasing sequences and $a_n \sim b_n$, then we also say that (a_n) and (b_n) have the same **growth rate**. For example,

$$10\,n^2 + 100\,n + 1000 \sim n^2 \quad \text{and} \quad \frac{10}{n^2} + \frac{100}{n^3} + \frac{1000}{n^4} \sim \frac{1}{n^2}.$$

The notation introduced above is useful in understanding relative asymptotic behavior of two sequences. $\diamond$

Remark 2.12. Before concluding this section, we describe how in some cases ∞ or $-\infty$ can be regarded as a 'limit' of a sequence (a_n). We say that (a_n) **tends** to ∞ or **diverges** to ∞ if for every $\alpha \in \mathbb{R}$, there is $n_0 \in \mathbb{N}$ such that $a_n > \alpha$ for all $n \geq n_0$, and then we write $a_n \to \infty$. We write $a_n \not\to \infty$ if (a_n) does not tend to ∞. Similarly, we say that (a_n) **tends** to $-\infty$ or **diverges** to ∞ if for every $\beta \in \mathbb{R}$, there is $n_0 \in \mathbb{N}$ such that $a_n < \beta$ for all $n \geq n_0$, and then we write $a_n \to -\infty$. We write $a_n \not\to -\infty$ if (a_n) does not tend to $-\infty$.

If $a_n \to \infty$ and $b_n \to \ell$, where $\ell \in \mathbb{R}$, then it is easy to see that $a_n + b_n \to \infty$, $a_n b_n \to \infty$ provided $\ell > 0$, and $a_n b_n \to -\infty$ provided $\ell < 0$, whereas if $\ell = 0$, then nothing can be said about the convergence of $(a_n b_n)$, as the examples (i) $a_n := n$ and $b_n := 0$, (ii) $a_n := n$ and $b_n := 1/n$, (iii) $a_n := n$ and $b_n := 1/\sqrt{n}$, (iv) $a_n := n$ and $b_n := -1/\sqrt{n}$, (v) $a_n := n$ and $b_n := (-1)^n/n$ show. Further, if $a_n \to \infty$ and $b_n \to \infty$, then $a_n + b_n \to \infty$ and $a_n b_n \to \infty$. Similar conclusions hold if $a_n \to -\infty$ and $b_n \to -\infty$. On the other hand, if $a_n \to \infty$ and $b_n \to -\infty$, then nothing can be said about the convergence of $a_n + b_n$, as the examples (i) $a_n := n$, $b_n := -n$, (ii) $a_n := n$, $b_n := -n+1$, (iii) $a_n := n$, $b_n := -\sqrt{n}$, (iv) $a_n := n$, $b_n := -n^2$, (v) $a_n := n$, $b_n := -n + (-1)^n$ show. Some of these 'indeterminate' cases can be tested using the method indicated in Remark 4.44.

If $a_n > 0$ for all $n \in \mathbb{N}$, then it is clear that $a_n \to \infty$ if and only if $1/a_n \to 0$. Also, if (a_n) is monotonically increasing, then it is easy to see that $a_n \to \infty$ if and only if (a_n) is not bounded above. Similarly, if $a_n < 0$ for all $n \in \mathbb{N}$, then $a_n \to -\infty$ if and only if $1/a_n \to 0$. Also, if (a_n) is monotonically decreasing, then $a_n \to -\infty$ if and only if (a_n) is not bounded below. $\diamond$

Examples 2.13. (i) Let p be a positive rational number and $a_n := n^p$ for $n \in \mathbb{N}$. Then $a_n \to \infty$. Indeed, given any $\alpha \in \mathbb{R}$, we have $a_n > \alpha$ if $n > \lceil |\alpha|^{1/p} \rceil$.

(ii) Let $a_n := \sum_{k=1}^{n}(1/k)$. Then $a_n \to \infty$, since (a_n) is monotonically increasing and, as shown in Example 2.10 (iii), (a_n) is not bounded above.

(iii) Let $a \in \mathbb{R}$ be such that $|a| > 1$ and let $a_n := a^n$ for $n \in \mathbb{N}$. If $a > 1$, then $1/a_n = (1/a)^n \to 0$ (Example 2.7 (i)) and so $a_n \to \infty$. If $a < -1$, then consider $b_n := a_{2n-1}$, $c_n := a_{2n}$ for $n \in \mathbb{N}$, and note that $b_n = (a^2)^n/a \to -\infty$, $c_n = (a^2)^n \to \infty$, and so $a_n \not\to \infty$, $a_n \not\to -\infty$.

2.2 Subsequences and Cauchy Sequences

Let (a_n) be a sequence. If $n_1, n_2, \ldots$ are positive integers such that $n_k < n_{k+1}$ for each $k \in \mathbb{N}$, then the sequence (a_{n_k}), whose terms are

$$a_{n_1}, a_{n_2}, \ldots,$$

is called a **subsequence** of (a_n). Note that $n_1 < n_2 < \cdots$ implies that $n_k \to \infty$ as $k \to \infty$.

It is easy to see that a sequence (a_n) converges to a if and only if every subsequence of (a_n) converges to a. This follows by observing that (a_n) is itself a subsequence of (a_n) (if we take $n_k = k$ for $k \in \mathbb{N}$), and on the other hand, if (a_{n_k}) is a subsequence of (a_n), then for any $n_0 \in \mathbb{N}$, there is $k_0 \in \mathbb{N}$ such that $n_k \geq n_0$ for all $k \geq k_0$.

Similarly, it can be seen that a sequence (a_n) tends to ∞ if and only if every subsequence of (a_n) tends to ∞, and that (a_n) tends to $-\infty$ if and only if every subsequence of (a_n) tends to $-\infty$.

We now prove a remarkable fact about monotonic subsequences.

Proposition 2.14. *Every sequence in $\mathbb{R}$ has a monotonic subsequence.*

Proof. Let (a_n) be a sequence in $\mathbb{R}$. Consider the 'peaks' in (a_n), that is, those terms that are greater than all the succeeding terms. Let E be the set of all positive integers n for which a_n is a 'peak', that is, let

$$E = \{n \in \mathbb{N} : a_n > a_m \text{ for all } m > n\}.$$

First, assume that E is a finite set. Then there is $n_1 \in \mathbb{N}$ such that $n_1 > n$ for every $n \in E$. Since $n_1 \notin E$, there is $n_2 \in \mathbb{N}$ such that $n_2 > n_1$ and $a_{n_1} \leq a_{n_2}$. Again, since $n_2 \notin E$, there is $n_3 \in \mathbb{N}$ such that $n_3 > n_2$ and $a_{n_2} \leq a_{n_3}$. Having chosen n_k for $k \in \mathbb{N}$ in this manner, we note that $n_k \notin E$ and hence there is $n_{k+1} \in \mathbb{N}$ such that $n_{k+1} > n_k$ and $a_{n_k} \leq a_{n_{k+1}}$. Thus we obtain a monotonically increasing subsequence (a_{n_k}) of (a_n).

Next, assume that E is an infinite set. If we enumerate E as $n_1, n_2, \ldots$, where $n_1 < n_2 < \cdots$, then since $n_k \in E$ for each $k \in \mathbb{N}$, we have $a_{n_k} > a_m$ for all $m > n_k$. In particular, taking $m = n_{k+1}$, we get $a_{n_k} > a_{n_{k+1}}$ for each $k \in \mathbb{N}$. Thus we obtain a monotonically decreasing subsequence (a_{n_k}) of (a_n).

In any case, we have proved that (a_n) has a monotonic subsequence. $\square$

We shall use the above result to prove two important results in analysis, known as the Bolzano–Weierstrass Theorem and the Cauchy Criterion. To put the former result in perspective, note that every convergent sequence in $\mathbb{R}$ is bounded (part (ii) of Proposition 2.2), but a bounded sequence need not be convergent.

Proposition 2.15 (Bolzano–Weierstrass Theorem). *Every bounded sequence in $\mathbb{R}$ has a convergent subsequence.*

Proof. Let (a_n) be a bounded sequence in $\mathbb{R}$. By Proposition 2.14, (a_n) has a monotonic subsequence (a_{n_k}). Since (a_n) is bounded, so is its subsequence (a_{n_k}). Hence by Corollary 2.9, (a_{n_k}) is convergent. $\qquad\square$

The following result may be viewed as a more elaborate version of the Bolzano–Weierstrass Theorem.

Corollary 2.16. *Let (a_n) be a sequence in $\mathbb{R}$. If either (a_n) is bounded above and $a_n \nrightarrow -\infty$, or if (a_n) is bounded below and $a_n \nrightarrow \infty$, then (a_n) has a convergent subsequence.*

Proof. First assume that (a_n) is bounded above and $a_n \nrightarrow -\infty$. The statement $a_n \nrightarrow -\infty$ means there is $\beta \in \mathbb{R}$ such that for every $n_0 \in \mathbb{N}$, there is $n \in \mathbb{N}$ with $n > n_0$ and $a_n \geq \beta$. Hence there are $n_1 < n_2 < \cdots$ in $\mathbb{N}$ such that $a_{n_k} \geq \beta$ for each $k \in \mathbb{N}$. The subsequence (a_{n_k}) in $\mathbb{R}$ is thus bounded above as well as bounded below. So by the Bolzano–Weierstrass Theorem, (a_{n_k}) has a convergent subsequence. Finally, we note that a subsequence of (a_{n_k}) is a subsequence of (a_n) itself.

If (a_n) is bounded below and $a_n \nrightarrow \infty$, the proof is similar. $\qquad\square$

As a consequence of the Bolzano–Weierstrass Theorem, we obtain a useful characterization of convergent sequences as follows.

Proposition 2.17. *A sequence in $\mathbb{R}$ is convergent if and only if it is bounded and all of its convergent subsequences have the same limit.*

Proof. If a sequence (a_n) converges to a, then it is bounded by part (ii) of Proposition 2.2 and clearly, every subsequence (and not just every convergent subsequence) of (a_n) converges to a.

Conversely, assume that (a_n) is a bounded sequence and there is $a \in \mathbb{R}$ such that every convergent subsequence of (a_n) converges to a. We claim that $a_n \to a$. For otherwise, there are $\epsilon > 0$ and positive integers $n_1 < n_2 < \cdots$ such that $|a_{n_k} - a| \geq \epsilon$ for all $k \in \mathbb{N}$. By the Bolzano–Weierstrass Theorem, the bounded sequence (a_{n_k}) has a convergent subsequence, which cannot possibly converge to a. This is a contradiction. $\qquad\square$

In general, proving the convergence of a sequence (a_n) is difficult since we must correctly guess the limit of (a_n) beforehand. There is a way of avoiding this guesswork, which we now describe.

A sequence (a_n) in $\mathbb{R}$ is called a **Cauchy sequence** if for every $\epsilon > 0$, there is $n_0 \in \mathbb{N}$ such that

$$|a_n - a_m| < \epsilon \quad \text{for all } n,\, m \geq n_0.$$

It is clear that if (a_n) is a Cauchy sequence in $\mathbb{R}$, then $(a_{n+1} - a_n) \to 0$ as $n \to \infty$. The converse, however, does not hold. For example, if $a_n := \sqrt{n}$ for $n \in \mathbb{N}$, then

$$a_{n+1} - a_n = \sqrt{n+1} - \sqrt{n} = \frac{1}{\sqrt{n+1} + \sqrt{n}} \to 0,$$

but (a_n) is not a Cauchy sequence, because given any $n_0 \in \mathbb{N}$, we have

$$a_{4n_0} - a_{n_0} = \sqrt{4n_0} - \sqrt{n_0} = \sqrt{n_0} \geq 1.$$

The following result gives a useful sufficient condition for a sequence to be Cauchy. A more general sufficient condition is given in Exercise 25.

Proposition 2.18. *Let (a_n) be a sequence of real numbers and α be a real number such that $\alpha < 1$ and*

$$|a_{n+1} - a_n| \leq \alpha |a_n - a_{n-1}| \quad \text{for all } n \in \mathbb{N} \text{ with } n \geq 2.$$

Then (a_n) is a Cauchy sequence.

Proof. For $n \in \mathbb{N}$, we have

$$|a_{n+1} - a_n| \leq \alpha |a_n - a_{n-1}| \leq \alpha^2 |a_{n-1} - a_{n-2}| \leq \cdots \leq \alpha^{n-1} |a_2 - a_1|.$$

Hence for all $m,\, n \in \mathbb{N}$ with $m > n$, we have

$$
\begin{aligned}
|a_m - a_n| &\leq |a_m - a_{m-1}| + |a_{m-1} - a_{m-2}| + \cdots + |a_{n+1} - a_n| \\
&\leq |a_2 - a_1| \left(\alpha^{m-2} + \alpha^{m-3} + \cdots + \alpha^{n-1} \right) \\
&= |a_2 - a_1| \alpha^{n-1} \frac{(1 - \alpha^{m-n})}{(1 - \alpha)} \\
&\leq |a_2 - a_1| \alpha^{n-1} \frac{1}{1 - \alpha}.
\end{aligned}
$$

If $a_2 = a_1$, then it is clear that $a_n = a_1$ for all $n \in \mathbb{N}$, and (a_n) is a Cauchy sequence. Suppose $a_2 \neq a_1$ and let $\epsilon > 0$ be given. Since $\alpha < 1$, by Example 2.7 (i), we see that $\alpha^n \to 0$. Consequently, there is $n_0 \in \mathbb{N}$ such that

$$\alpha^{n-1} < \frac{\epsilon(1 - \alpha)}{|a_2 - a_1|} \quad \text{for all } n \in \mathbb{N} \text{ with } n \geq n_0.$$

It follows that $|a_m - a_n| < \epsilon$ for all $m,\, n \in \mathbb{N}$ with $m, n \geq n_0$. Thus (a_n) is a Cauchy sequence. $\qquad\square$

It may be noted that the condition

$$|a_{n+1} - a_n| \leq \alpha |a_n - a_{n-1}| \quad \text{for some } \alpha < 1$$

in the above proposition cannot be weakened to $|a_{n+1} - a_n| < |a_n - a_{n-1}|$. For example, if $a_n := \sqrt{n}$ for $n \in \mathbb{N}$, then

$$|a_{n+1} - a_n| = \sqrt{n+1} - \sqrt{n} < \sqrt{n} - \sqrt{n-1} = |a_n - a_{n-1}|,$$

but, as we have just seen, the sequence (a_n) is not Cauchy.

We are now ready to state and prove the Cauchy Criterion, which was alluded to earlier. Briefly, it says that in $\mathbb{R}$, the notions of convergent sequences and Cauchy sequences are equivalent.

Proposition 2.19 (Cauchy Criterion). *A sequence (a_n) in $\mathbb{R}$ is convergent if and only if it is a Cauchy sequence.*

Proof. Let (a_n) be a convergent sequence and $a_n \to a$. Given any $\epsilon > 0$, there is $n_0 \in \mathbb{N}$ such that $|a_n - a| < \epsilon/2$ for all $n \geq n_0$. Consequently,

$$|a_n - a_m| \leq |a_n - a| + |a - a_m| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \quad \text{for all } n, m \geq n_0.$$

Hence (a_n) is a Cauchy sequence.

Conversely, let (a_n) be a Cauchy sequence. First, we show that (a_n) is a bounded sequence. Since (a_n) is Cauchy, there is $n_1 \in \mathbb{N}$ such that

$$|a_n - a_m| < 1 \quad \text{for all } n, m \geq n_1.$$

Hence
$$|a_n| \leq |a_n - a_{n_1}| + |a_{n_1}| < 1 + |a_{n_1}| \quad \text{for all } n \geq n_1.$$

If we let $\alpha := \max\{|a_1|, \ldots, |a_{n_1-1}|, 1 + |a_{n_1}|\}$, then we have $|a_n| \leq \alpha$ for all $n \in \mathbb{N}$. Hence (a_n) is a bounded sequence. Next, by the Bolzano–Weierstrass Theorem, (a_n) has a convergent subsequence (a_{n_k}). Let $a_{n_k} \to a$. We show that in fact $a_n \to a$. Let $\epsilon > 0$ be given. Since (a_n) is Cauchy, there is $n_0 \in \mathbb{N}$ such that
$$|a_n - a_m| < \frac{\epsilon}{2} \quad \text{for all } n, m \geq n_0.$$

Also, since $a_{n_k} \to a$, there is $k_0 \in \mathbb{N}$ such that

$$|a_{n_k} - a| < \frac{\epsilon}{2} \quad \text{for all } k \geq k_0.$$

Further, since $n_1 < n_2 < \cdots$, there is $j \in \mathbb{N}$ such that $j \geq k_0$ and $n_j \geq n_0$. Now,

$$|a_n - a| \leq |a_n - a_{n_j}| + |a_{n_j} - a| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \quad \text{for all } n \geq n_0.$$

Hence the sequence (a_n) is convergent. $\square$

The following example shows how the Cauchy Criterion can be used to prove the convergence of a sequence.

Example 2.20. Consider the sequence (a_n) defined by

$$a_1 := 1 \quad \text{and} \quad a_{n+1} := 1 + \frac{1}{a_n} \quad \text{for } n \in \mathbb{N}.$$

First we show that (a_n) is a Cauchy sequence. It is clear that $a_n \geq 1$ for all $n \in \mathbb{N}$ and hence

$$a_n a_{n-1} = \left(1 + \frac{1}{a_{n-1}}\right) a_{n-1} = a_{n-1} + 1 \geq 2 \quad \text{for all } n \in \mathbb{N} \text{ with } n \geq 2.$$

Since

$$a_{n+1} - a_n = \left(1 + \frac{1}{a_n}\right) - \left(1 + \frac{1}{a_{n-1}}\right) = \frac{1}{a_n} - \frac{1}{a_{n-1}} = \frac{a_{n-1} - a_n}{a_n a_{n-1}},$$

we see that

$$|a_{n+1} - a_n| \leq \frac{1}{2}|a_n - a_{n-1}| \quad \text{for all } n \in \mathbb{N} \text{ with } n \geq 2.$$

Hence by Proposition 2.18, (a_n) is a Cauchy sequence and by Proposition 2.19, it is convergent. Let $a_n \to a$. Then $a_{n+1} \to a$, and since $a_{n+1} = 1 + (1/a_n)$, we have $a = 1 + (1/a)$. Also, $a_n \geq 1$ for all $n \in \mathbb{N}$ implies that $a \geq 1$. Hence $a = (1 + \sqrt{5})/2$.

It may be noted that (a_n) is not a monotonic sequence. In fact, for any $n \in \mathbb{N}$ with $n \geq 2$, we have $a_n \leq a_{n+1}$ if and only if $a_{n-1} \geq a_n$. So we cannot appeal to Proposition 2.19 to deduce the convergence of (a_n). $\diamond$

We remark that the Completeness Property of $\mathbb{R}$ is crucially used (via Corollary 2.9 and the Bolzano–Weierstrass Theorem) in the proof that every Cauchy sequence in $\mathbb{R}$ is convergent. Conversely, assuming that every Cauchy sequence in $\mathbb{R}$ is convergent, it is possible to establish the Completeness Property of $\mathbb{R}$. (See Exercise 42.) In view of this, the result in Proposition 2.19 is sometimes referred to as the **Cauchy completeness** of $\mathbb{R}$.

Notes and Comments

The concept of the convergence of a sequence is extremely crucial in calculus and analysis. It is the point of departure from 'discrete mathematics' to 'continuous mathematics'. It makes precise the idea of serially numbered real numbers coming arbitrarily close to a fixed real number. In the next chapter, this concept will be further extended to state what is meant by the 'limit' of a function defined on a subset of $\mathbb{R}$.

The arguments used to prove the convergence of sequences in some of our examples are not so standard. The convergence of the sequence (a_n), where $a_n = 1+(1/2^p)+\cdots+(1/n^p)$ for $n \in \mathbb{N}$ and $p > 1$, is proved following an article of Cohen and Knight [18]. Also, the proof (using the generalized binomial inequality) of the convergence of the sequence (b_n), where $b_n = (1 + (1/n))^n$ for $n \in \mathbb{N}$, is based on the article of Lyon and Ward [46]. Several examples of sequences that we have discussed in this chapter are, in fact, examples of 'infinite series' in disguise. A systematic study of infinite series will be taken up in Chapter 9.

The Bolzano–Weierstrass Theorem given in the second half of this chapter is the cornerstone of much of mathematical analysis. Many of the properties of a nice class of functions (the 'continuous' functions, which we shall introduce in the next chapter) are based on this result. Classically, the Bolzano–Weierstrass Theorem is proved by considering a bounded interval that contains infinitely many terms of the sequence, dividing it in equal halves, and picking a half that contains infinitely many terms of the sequence. This process can be continued and it leads to a nested sequence of intervals in which a limit of a subsequence is trapped. (See Exercises 33 and 34.) Another standard approach to prove the Bolzano–Weierstrass Theorem as well as the Cauchy Criterion is to use the notions of 'cluster point', $\limsup$ and $\liminf$. (See Exercise 16 and Exercises 35 through 41.) We have bypassed either of these methods and instead used a neat result that every sequence in $\mathbb{R}$ has a monotonic subsequence. This result is easy to prove and can be of interest in itself. It appears, for example, in the books of Spivak [57] (Lemma on page 378 of the first edition or page 451 of the third edition), Newman [49] (Problem 6 and its solution), and Ross [51] (Theorem 11.3).

Exercises

Part A

1. Which of the following sequences are bounded? Which of them are convergent? In case of convergence, find the limit.

    (i) $a_n := \dfrac{1}{n^2}$,   (ii) $a_n := \sqrt{n}$,   (iii) $a_n := (-1)^n$,   (iv) $a_n := \dfrac{n}{2n+1}$,

    (v) $a_n := \sqrt{n}(\sqrt{n+1} - \sqrt{n})$,   (vi) $a_n := n^{3/2}(\sqrt{n^3 + 1} - \sqrt{n^3})$.

2. Let (a_n) and (b_n) be sequences in $\mathbb{R}$. Under which of the following conditions is the sequence $(a_n b_n)$ convergent? Justify.

    (i) (a_n) is convergent.

    (ii) (a_n) is convergent and (b_n) is bounded.

    (iii) (a_n) converges to 0 and (b_n) is bounded.

    (iv) (a_n) and (b_n) are convergent.

3. Let $a, b \in \mathbb{R}$ and (a_n) be a sequence in $\mathbb{R}$ such that $a_n \to a$. and $a_n \geq b$ for all $n \in \mathbb{N}$. Show that $a \geq b$. Give an example in which $a_n > a$ for all $n \in \mathbb{N}$, but $a_n \to a$.

4. Let a and x be real numbers. If (b_n) and (c_n) are sequences in $\mathbb{R}$ such that

$$\lim_{n\to\infty} b_n = 0 = \lim_{n\to\infty} c_n \quad \text{and} \quad a - b_n \leq x \leq a + c_n \quad \text{for } n \in \mathbb{N},$$

then show that $x = a$.

5. If (a_n) is a sequence in $\mathbb{R}$ such that $a_n \neq 0$ for all n, $\lim_{n\to\infty} |a_{n+1}/a_n|$ exists and it is less than 1, then show that $a_n \to 0$.

6. If $k \in \mathbb{N}$ and $x \in \mathbb{R}$ with $|x| < 1$, then show that

$$\lim_{n\to\infty} n^k x^n = 0.$$

7. For $n \in \mathbb{N}$, let $a_n := n^{1/n}$. Show that $a_1 < a_2 < a_3$ and $a_n > a_{n+1}$ for all $n \geq 3$. Further, show that

$$1 < a_n < 1 + \frac{\sqrt{2}}{\sqrt{n-1}} \quad \text{for all } n \geq 2$$

and deduce that $a_n \to 1$ as $n \to \infty$.

8. Show that the sequence (B_n) defined by

$$B_n := \left(1 + \frac{1}{n}\right)^n \quad \text{for } n \in \mathbb{N}$$

is monotonically increasing. Deduce that the sequence (B_n) is convergent. (Hint: Given $n \in \mathbb{N}$, use the A.M.-G.M. inequality for $a_1 = \cdots = a_n := 1/(n+1)$ and $a_{n+1} := 1$. Also, note that $B_n \leq 3$ for all $n \in \mathbb{N}$.)

9. Show that the sequence (a_n) is convergent and find its limit if (a_n) is given by the following.
   (i) $a_1 := 1$ and $a_{n+1} := (3a_n + 2)/6$ for $n \in \mathbb{N}$.
   (ii) $a_1 := 1$ and $a_{n+1} := a_n/(2a_n + 1)$ for $n \in \mathbb{N}$.
   (iii) $a_1 := 1$ and $a_{n+1} := 2a_n/(4a_n + 1)$ for $n \in \mathbb{N}$.
   (iv) $a_1 := 2$ and $a_{n+1} := \sqrt{1 + a_n}$ for $n \in \mathbb{N}$.
   (v) $a_1 := 1$ and $a_{n+1} := \sqrt{2 + a_n}$ for $n \in \mathbb{N}$.
   (vi) $a_1 := 2$ and $a_{n+1} := (1/2) + \sqrt{a_n}$ for $n \in \mathbb{N}$.
   (vii) $a_1 := 1$ and $a_{n+1} := (1/2) + \sqrt{a_n}$ for $n \in \mathbb{N}$.

10. For $n \in \mathbb{N}$, let

$$a_n := \frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2n} \quad \text{and} \quad b_n := \frac{1}{1} + \frac{1}{3} + \cdots + \frac{1}{2n-1}.$$

Show that $a_n \to \infty$ and $b_n \to \infty$. (Hint: Example 2.10 (iii).)

11. Show that $(n!)^{1/n} \to \infty$.

12. Suppose α and β are real numbers such that $0 \leq \beta \leq \alpha$. Let

$$a_1 := \alpha, \ \ b_1 := \beta \ \ \text{and} \ \ a_{n+1} := \frac{a_n + b_n}{2}, \ \ b_{n+1} := \sqrt{a_n b_n} \ \ \text{for } n \in \mathbb{N}.$$

Show that (a_n) is a monotonically decreasing sequence that is bounded below by β, and (b_n) is a monotonically increasing sequence that is bounded above by α. Further, show that $0 \leq \alpha - \beta \leq (\alpha - \beta)/2^{n-1}$ for $n \in \mathbb{N}$. Deduce that (a_n) and (b_n) are convergent and have the same limit.
[Note: The common limit of the sequences (a_n) and (b_n) is called the **arithmetic-geometric mean** of the nonnegative real numbers α and β. It was introduced and studied by Gauss. For further details, see [20].]

13. If a monotonic sequence (a_n) has a subsequence (a_{n_k}) such that $a_{n_k} \to a$, where $a \in \mathbb{R}$ or $a = \infty$ or $a = -\infty$, then show that $a_n \to a$.

14. Prove that a sequence (a_n) in $\mathbb{R}$ has no convergent subsequence if and only if $|a_n| \to \infty$.

15. Let (a_n) be a sequence of real numbers and let $a \in \mathbb{R}$. Show that $a_n \to a$ if and only if every subsequence of (a_n) has a subsequence converging to a. (Hint: Proposition 2.17.)

16. A real number a is called a **cluster point** of a sequence (a_n) in $\mathbb{R}$ if there is a subsequence (a_{n_k}) of (a_n) such that $a_{n_k} \to a$.
   (i) Show that if $a_n \to a$, then a is the only cluster point of (a_n).
   (ii) Show that the converse of (i) is not true. In other words, show that there is a divergent sequence that has a unique cluster point. (Hint: $a_{2k} := \frac{1}{2k}$ and $a_{2k+1} := 2k + 1$ for $k \in \mathbb{N}$.)
   (iii) Show that if $a_n \to \infty$ or if $a_n \to -\infty$, then (a_n) has no cluster point.
   (iv) Show that the converse of (iii) is not true. In other words, show that there is a sequence without a cluster point that neither tends to ∞ nor tends to $-\infty$. (Hint: $a_n := (-1)^n n$ for $n \in \mathbb{N}$.)

17. Let $A_n := 1 + (1/2) + \cdots + (1/n)$ for $n \in \mathbb{N}$. Show that $(A_{n+1} - A_n) \to 0$ as $n \to \infty$, but (A_n) is not a Cauchy sequence.

18. Let $A_n := 1 + (1/2^2) + \cdots + (1/n^2)$ for $n \in \mathbb{N}$. Show that there is no real number $\alpha < 1$ such that $|A_{n+1} - A_n| \leq \alpha |A_n - A_{n-1}|$ for all $n \in \mathbb{N}$ with $n \geq 2$, but (A_n) is a Cauchy sequence.

Part B

19. Let $x \in \mathbb{R}$ and $x > 0$. Define

$$A_n := 1 + \frac{x}{1!} + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} \ \ \text{and} \ \ B_n := \left(1 + \frac{x}{n}\right)^n \ \ \text{for } n \in \mathbb{N}.$$

Show that (A_n) and (B_n) are convergent and have the same limit.

20. Show that the number $e := \lim_{n \to \infty} \sum_{k=0}^{n} (1/k!)$ is irrational. (Hint: For every $n \in \mathbb{N}$, $0 < e - \sum_{k=0}^{n} (1/k!) < (1/n!n)$. Multiply by $n!$.)

21. (i) If (a_n) is a sequence and $a_n \to a$, then show that $(a_1 + \cdots + a_n)/n \to a$. Here $a \in \mathbb{R}$ or $a = \infty$ or $a = -\infty$. Give an example to show that the converse is not true.

(ii) Find $\lim\limits_{n\to\infty} \dfrac{1}{n}\left(\dfrac{2}{5} + \dfrac{5}{11} + \cdots + \dfrac{n^2+1}{2n^2+3}\right)$.

22. Suppose α, β, and γ are positive real numbers. Let

$$a_1 := \alpha \quad \text{and} \quad a_{n+1} := \frac{a_n}{\beta a_n + \gamma} \quad \text{for } n \in \mathbb{N}.$$

Show that (a_n) is convergent. Further, if $a := \lim_{n\to\infty} a_n$, then show that $a = 0$ if $\gamma \geq 1$ and $a = (1-\gamma)/\beta$ otherwise. (Hint: Consider the cases $\alpha\beta + \gamma \geq 1$ and $\alpha\beta + \gamma < 1$.)

23. Suppose α and β are nonnegative real numbers. Let

$$a_1 := \alpha \quad \text{and} \quad a_{n+1} := \sqrt{\beta + a_n} \quad \text{for } n \in \mathbb{N}.$$

Show that (a_n) is convergent. Further, if $a := \lim_{n\to\infty} a_n$, then show that $a = 0$ if $\alpha = 0 = \beta$, and $a = (1 + \sqrt{1 + 4\beta})/2$ otherwise. (Hint: Consider the cases $\sqrt{\alpha + \beta} \leq \alpha$ and $\sqrt{\alpha + \beta} > \alpha$.)

24. Suppose α and β are nonnegative real numbers. Let

$$a_1 := \alpha \quad \text{and} \quad a_{n+1} := \beta + \sqrt{a_n} \quad \text{for } n \in \mathbb{N}.$$

Show that (a_n) is convergent. Further, if $a := \lim_{n\to\infty} a_n$, then show that $a = 0$ if $\alpha = 0 = \beta$, and $a = (1 + 2\beta + \sqrt{1 + 4\beta})/2$ otherwise. (Hint: Consider the cases $\sqrt{\alpha} + \beta \leq \alpha$ and $\sqrt{\alpha} + \beta > \alpha$.)

25. Let (a_n) and (b_n) be sequences such that $|a_{n+1} - a_n| \leq b_n$ for all $n \in \mathbb{N}$. Define

$$B_n := \sum_{k=1}^{n} b_k \quad \text{for } n \in \mathbb{N}.$$

If (B_n) is convergent, then show that (a_n) is a Cauchy sequence and hence it is convergent.

26. Let y be any real number with $0 \leq y < 1$. Define sequences (b_n) and (y_n) iteratively as follows. Let $y_1 := 10y$ and $b_1 := [y_1]$, and for each $n \in \mathbb{N}$,

$$y_{n+1} := 10(y_n - b_n) \quad \text{and} \quad b_{n+1} := [y_{n+1}].$$

Show that for each $n \in \mathbb{N}$ we have

$$0 \leq y_n < 10 \quad \text{and} \quad b_n \in \mathbb{Z} \text{ with } 0 \leq b_n \leq 9,$$

and moreover,

$$y = \frac{b_1}{10} + \frac{b_2}{10^2} + \cdots + \frac{b_n}{10^n} + \frac{y_{n+1}}{10^{n+1}}.$$

Deduce that

$$0 \leq \frac{y_{n+1}}{10^{n+1}} < \frac{1}{10^n} \quad \text{for each } n \in \mathbb{N}$$

and consequently,

$$y = \lim_{n \to \infty} \left(\frac{b_1}{10} + \frac{b_2}{10^2} + \cdots + \frac{b_n}{10^n} \right).$$

[Note: It is customary to call the nonnegative integers $b_1, b_2, \ldots$, the **digits** of y and write the above expression for y as $y = 0.b_1 b_2 \ldots$, and call it the **decimal expansion** of y.]

27. Given any $m \in \mathbb{N}$, show that there is a unique nonnegative integer k such that $10^k \leq m < 10^{k+1}$. Use Exercise 37 of Chapter 1 repeatedly to show that there are unique integers $a_0, a_1, \ldots, a_k$ between 0 and 9 such that

$$m = a_0 + a_1(10) + a_2(10^2) + \cdots + a_k \left(10^k \right).$$

28. Given any $x \in \mathbb{R}$, show that there is a nonnegative integer k and integers $a_k, a_{k-1}, \ldots, a_1, a_0, b_1, b_2, \ldots$ between 0 and 9 such that

$$x = \pm \lim_{n \to \infty} \left(a_k 10^k + a_{k-1} 10^{k-1} + \cdots + a_0 + \frac{b_1}{10} + \frac{b_2}{10^2} + \cdots + \frac{b_n}{10^n} \right).$$

(Hint: If $|x| < 1$, set $k = 0 = a_0$ and apply Exercise 26 to $y := |x|$, whereas if $|x| \geq 1$, apply Exercise 27 to $n := [|x|]$ and Exercise 26 to $y := |x| - n$.)
[Note: It is customary to call $a_k, a_{k-1}, \ldots, a_0, b_1, b_2, \ldots$ the **digits** of x and write the above expression for x as $x = \pm a_k a_{k-1} \ldots a_0.b_1 b_2 \ldots$, and call it the **decimal expansion** of x.]

29. Given any $y \in [0, 1)$, let (y_n) and (b_n) be the sequences associated to y as in Exercise 26. We say that the decimal expansion of y is **finite** if $y_n = 0$ for some $n \in \mathbb{N}$ and **recurring** if it not finite but $y_i = y_j$ for some $i, j \in \mathbb{N}$ with $i < j$. Show that if $y \in [0, 1)$ is a rational number, then its decimal expansion is either finite or recurring. (Hint: Write y in reduced form as $y = p/q$. Let $r_0 := p$. Use Exercise 37 of Chapter 1 successively to find integers $q_1, r_1, q_2, r_2, \ldots$ such that $10 r_{i-1} = q q_i + r_i$ and $0 \leq r_i < q$ for $i \geq 1$. Now $y_i = 10 r_{i-1}/q$ and the r_i's take only finitely many values.)
[Note: The converse also holds. see Remark 9.2.]

30. Show that the results of Exercises 26, 27, 28, and 29 are valid with the number 10 replaced by any integer $d > 1$ and the number 9 by $d - 1$.
[Note: The corresponding limiting expression of a real number x is called the d-**ary expansion** of x. When $d = 2$, it is called the **binary expansion** and when $d = 3$, it is called the **ternary expansion**.]

31. Define

$$a_1 := 1 \quad \text{and} \quad a_{n+1} := \left(1 + \frac{(-1)^n}{2^n} \right) a_n \quad \text{for } n \in \mathbb{N}.$$

(i) For every $n \in \mathbb{N}$, show that

$$|a_{n+1}| \leq \left(1 + \frac{1}{2^n} \right) \left(1 + \frac{1}{2^{n-1}} \right) \cdots \left(1 + \frac{1}{2} \right) \leq \left(\frac{n+1}{n} \right)^n < 3.$$

(Hint: Use the A.M.-G.M. inequality.)

(ii) Use (i) above to show that

$$|a_{n+1} - a_n| < \frac{3}{2^n} \quad \text{for all } n \in \mathbb{N}.$$

Deduce, using Exercise 25, that (a_n) is a Cauchy sequence.

(iii) Conclude that (a_n) is convergent. Is (a_n) monotonic?

32. Assuming only the algebraic and the order properties of $\mathbb{R}$, and assuming that every monotonically decreasing sequence that is bounded below is convergent in $\mathbb{R}$, establish the Completeness Property of $\mathbb{R}$. (Hint: Consider $S \subseteq \mathbb{R}$, $a_0 \in S$, and an upper bound α_0 of S. If $(a_0 + \alpha_0)/2$ is an upper bound of S, let $a_1 := a_0$ and $\alpha_1 := (a_0 + \alpha_0)/2$; otherwise, there is $a_1 \in S$ such that $(a_0 + \alpha_0)/2 < a_1$ and in this case, let $\alpha_1 := \alpha_0$. Continuing in this manner, obtain a monotonically decreasing sequence (α_n) that is bounded below.) [Compare part (ii) of Proposition 2.8.]

33. **(Nested Interval Theorem)** Let $I_n := [a_n, b_n]$, $n \in \mathbb{N}$, be closed intervals such that $I_n \supseteq I_{n+1}$ for all $n \in \mathbb{N}$ and $|b_n - a_n| \to 0$. Show that there is a unique $x \in \mathbb{R}$ such that $x \in I_n$ for all $n \in \mathbb{N}$. (Hint: Use Exercise 51 of Chapter 1.)

34. Use the Nested Interval Theorem in Exercise 33 to prove the Bolzano–Weierstrass Theorem.

35. Let (a_n) be a sequence in $\mathbb{R}$.

(i) Assume that (a_n) is bounded above and $a_n \not\to -\infty$. Define

$$M_n := \sup\{a_n, a_{n+1}, \ldots\} \quad \text{for} \quad n \in \mathbb{N} \quad \text{and} \quad M := \inf\{M_1, M_2, \ldots\}.$$

Show that the sequence (M_n) converges to M and M is the largest cluster point of (a_n).

(ii) Assume that (a_n) is bounded below and $a_n \not\to \infty$. Define

$$m_n := \inf\{a_n, a_{n+1}, \ldots\} \quad \text{for} \quad n \in \mathbb{N} \quad \text{and} \quad m := \sup\{m_1, m_2, \ldots\}.$$

Show that the sequence (m_n) converges to m and m is the smallest cluster point of (a_n).

[See Exercise 16 for the definition of a cluster point.]

36. Let (a_n) be a sequence in $\mathbb{R}$. Define the **limit superior** (or the **upper limit**) of (a_n) by

$$\limsup_{n \to \infty} a_n := \begin{cases} \lim_{n \to \infty} M_n & \text{if } (a_n) \text{ is bounded above and } a_n \not\to -\infty, \\ \infty & \text{if } (a_n) \text{ is not bounded above}, \\ -\infty & \text{if } a_n \to -\infty, \end{cases}$$

where the sequence (M_n) is as defined in Exercise 35. Similarly, define the **limit inferior** (or the **lower limit**) of (a_n) by

$$\liminf_{n \to \infty} a_n := \begin{cases} \lim_{n \to \infty} m_n & \text{if } (a_n) \text{ is bounded below and } a_n \not\to \infty, \\ -\infty & \text{if } (a_n) \text{ is not bounded below}, \\ \infty & \text{if } a_n \to \infty, \end{cases}$$

where the sequence (m_n) is as defined in Exercise 35. If (a_n) is bounded, then show that the set C of all cluster points of (a_n) is nonempty, and moreover,

$$\limsup_{n\to\infty} a_n = \lim_{n\to\infty} \sup\{a_n, a_{n+1}, \ldots\} = \max C$$

and

$$\liminf_{n\to\infty} a_n = \lim_{n\to\infty} \inf\{a_n, a_{n+1}, \ldots\} = \min C.$$

37. Determine $\limsup_{n\to\infty} a_n$ and $\liminf_{n\to\infty} a_n$ if (a_n) is as defined below.
   (i) $a_n := (-1)^n \left(1 + \frac{1}{n}\right)$ for $n \in \mathbb{N}$,
   (ii) $a_n := (-1)^n n$ for $n \in \mathbb{N}$,
   (iii) $a_1 := 0$ and for $k \in \mathbb{N}$, $a_{2k} := a_{2k-1}/2$ and $a_{2k+1} := (1/2) + a_{2k}$. (Hint: $a_{2k} = (1/2) - (1/2^k)$ for all $k \in \mathbb{N}$.)
38. Let (r_n) be a sequence such that $\mathbb{Q} = \{r_n : n \in \mathbb{N}\}$. [Note that by Exercise 49 (iii) of Chapter 1, such a sequence exists.] Determine the set of all cluster points of (r_n), and also $\liminf_{n\to\infty} r_n$ as well as $\limsup_{n\to\infty} r_n$.
39. Let (a_n) be a sequence in $\mathbb{R}$. Prove the following:
   (i) $\liminf_{n\to\infty} a_n \le \limsup_{n\to\infty} a_n$.
   (ii) (a_n) is bounded if and only if both $\liminf_{n\to\infty} a_n$ and $\limsup_{n\to\infty} a_n$ are real numbers.
   (iii) (a_n) is convergent if and only if both $\liminf_{n\to\infty} a_n$ and $\limsup_{n\to\infty} a_n$ are real numbers and are equal to each other. In this case,

$$\liminf_{n\to\infty} a_n = \lim_{n\to\infty} a_n = \limsup_{n\to\infty} a_n.$$

   (iv) $a_n \to \infty$ if and only if $\liminf_{n\to\infty} a_n = \infty = \limsup_{n\to\infty} a_n$.
   (v) $a_n \to -\infty$ if and only if $\liminf_{n\to\infty} a_n = -\infty = \limsup_{n\to\infty} a_n$.
40. Let (a_n) be a sequence in $\mathbb{R}$. Prove Corollary 2.16 (which is a more elaborate version of the Bolzano–Weierstrass Theorem) by showing that if (a_n) is bounded above and $a_n \not\to -\infty$, then (a_n) has a subsequence that converges to $\limsup_{n\to\infty} a_n$, while if (a_n) is bounded below and $a_n \not\to \infty$, then (a_n) has a subsequence that converges to $\liminf_{n\to\infty} a_n$.
41. Let (a_n) be a Cauchy sequence in $\mathbb{R}$. Prove that (a_n) is convergent by showing that it is bounded and $\limsup_{n\to\infty} a_n = \liminf_{n\to\infty} a_n$.
42. Assuming only the algebraic and the order properties of $\mathbb{R}$, and assuming that every Cauchy sequence in $\mathbb{R}$ is convergent, establish the Completeness Property of $\mathbb{R}$. (Hint: Consider $S \subseteq \mathbb{R}$ and a_n, α_n as in the Hint for Exercise 32. Then $\alpha_n - a_n \le (\alpha_0 - a_0)/2^n$ for all $n \in \mathbb{N}$.)

3

Continuity and Limits

In the previous chapter we studied real sequences, that is, real-valued functions defined on the subset $\mathbb{N}$ of $\mathbb{R}$. In this chapter we shall consider real-valued functions whose domains are arbitrary subsets of $\mathbb{R}$. The basic question we address is the following: Let $D \subseteq \mathbb{R}$, $c \in \mathbb{R}$, and let $f : D \to \mathbb{R}$ be a function. If a real number x in D is near c, then must there exist a real number l such that $f(x)$ is near l? In order to answer this and related questions, we develop the concepts of the continuity of a function and of the limit of a function.

In Section 3.1 below, we introduce the notion of continuity and derive a number of elementary results. Next, in Section 3.2, we examine this notion in relation to various properties of functions introduced in Section 1.3. Some important properties of real-valued continuous functions defined on a closed and bounded subset of $\mathbb{R}$ or on an interval will turn out to be of basic importance in our subsequent development of calculus and analysis. The fundamental notion of limit of a function is discussed in Section 3.3. All through these, our treatment will be based on the notion of convergence of sequences and the results proved in Chapter 2.

3.1 Continuity of Functions

Let $D \subseteq \mathbb{R}$. Consider a function $f : D \to \mathbb{R}$ and a point $c \in D$. We say that f is **continuous** at c if

$$(x_n) \text{ any sequence in } D \text{ and } x_n \to c \Longrightarrow f(x_n) \to f(c).$$

If f is not continuous at c, we say that f is **discontinuous** at c. In case f is continuous at every $c \in D$, we say that f is continuous on D.

Examples 3.1. (i) Let a and b be real numbers and $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = ax + b$ for $x \in \mathbb{R}$. Then f is continuous on $\mathbb{R}$. To see this, let $c \in \mathbb{R}$ and (x_n) be any sequence in $\mathbb{R}$ such that $x_n \to c$. By parts (i) and

(ii) of Proposition 2.3, $ax_n + b \to ac + b$, that is, $f(x_n) \to f(c)$. Thus f is continuous on $\mathbb{R}$.

(ii) Let $f(x) = |x|$ for $x \in \mathbb{R}$. [See Figure 1.4.] Again, f is continuous on $\mathbb{R}$. This follows by noting that whenever $c \in \mathbb{R}$ and $x_n \to c$, we have $|x_n| \to |c|$, because $|\,|x_n| - |c|\,| \le |x_n - c|$ for all $n \in \mathbb{N}$ by part (ii) of Proposition 1.8.

(iii) Let $f(x) = [x]$ for $x \in \mathbb{R}$. [See Figure 1.8.] If $c \in \mathbb{N}$, then f is not continuous at c, since $f(c) = c$, $c - (1/n) \to c$, and $f(c - (1/n)) = c - 1$ for all $n \in \mathbb{N}$, so $f(c - (1/n)) \nrightarrow f(c)$. On the other hand, if $c \in \mathbb{R} \setminus \mathbb{N}$, then f is continuous at c. To see this, let $\epsilon := \min\{c - [c], [c] + 1 - c\}$. Then $\epsilon > 0$ and since $x_n \to c$, there is $n_0 \in \mathbb{N}$ such that $|x_n - c| < \epsilon$, that is, $c - \epsilon < x_n < c + \epsilon$ and so $[c] < x_n < [c] + 1$ for all $n \ge n_0$. Thus $f(x_n) = [x_n] = [c]$ for all $n \ge n_0$. Therefore $f(x_n) \to f(c)$.

(iv) Let $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} 1 & \text{if } x \text{ is rational,} \\ 0 & \text{if } x \text{ is irrational.} \end{cases}$$

Then f is discontinuous at every $c \in \mathbb{R}$. To see this, we note that if c is rational and $x_n := c + (\sqrt{2}/n)$ for $n \in \mathbb{N}$, then each x_n is irrational and hence $f(x_n) = 0$ for all $n \in \mathbb{N}$, while $f(c) = 1$. On the other hand, if c is irrational and $x_n := ([nc] + 1)/n$ for $n \in \mathbb{N}$, then $c < x_n < c + (1/n)$, each x_n is rational, and hence $f(x_n) = 1$ for all $n \in \mathbb{N}$, while $f(c) = 0$. Thus in both cases, $x_n \to c$, but $f(x_n) \nrightarrow f(c)$. This function is known as the **Dirichlet function**. $\diamond$

We first prove a useful result regarding the sign of the values of a function that is continuous at a point.

Lemma 3.2. *Let $D \subseteq \mathbb{R}$, $c \in D$, and let $f : D \to \mathbb{R}$ be a function that is continuous at c. If $f(c) > 0$, then there is $\delta > 0$ such that $f(x) > 0$ whenever $x \in D$ and $|x - c| < \delta$. Likewise, if $f(c) < 0$, then there is $\delta > 0$ such that $f(x) < 0$ whenever $x \in D$ and $|x - c| < \delta$.*

Proof. Let $f(c) > 0$. Suppose that for every $\delta > 0$, there is $x \in D$ such that $|x - c| < \delta$ and $f(x) \le 0$. Taking the values $1, \frac{1}{2}, \frac{1}{3}, \dots$ for δ, we obtain $c_n \in D$ such that $|c_n - c| < 1/n$ and $f(c_n) \le 0$ for each $n \in \mathbb{N}$. Since $c_n \to c$ and f is continuous at c, we have $f(c_n) \to f(c)$. By part (i) of Proposition 2.4 we have $f(c) \le 0$. But this contradicts our assumption that $f(c) > 0$. Hence there is $\delta > 0$ such that $f(x) > 0$ whenever $x \in D$ and $|x - c| < \delta$.

The proof of the case $f(c) < 0$ is similar. $\square$

Proposition 3.3. *Let $D \subseteq \mathbb{R}$, $c \in D$, and let $f, g : D \to \mathbb{R}$ be functions that are continuous at c. Then*

(i) *$f + g$ is continuous at c,*

(ii) *rf is continuous at c for any $r \in \mathbb{R}$,*

(iii) fg is continuous at c,

(iv) if $f(c) \neq 0$, then there is $\delta > 0$ such that $f(x) \neq 0$ whenever $x \in D$ and $|x-c| < \delta$; moreover the function $1/f : D \cap (c-\delta, c+\delta) \to \mathbb{R}$ is continuous at c,

(v) if there is $\delta > 0$ such that $f(x) \geq 0$ whenever $x \in D$ and $|x-c| < \delta$, then for any $k \in \mathbb{N}$, the function $f^{1/k} : D \cap (c-\delta, c+\delta) \to \mathbb{R}$ is continuous at c.

Proof. Let (x_n) be any sequence in D such that $x_n \to c$. Then $f(x_n) \to f(c)$ and $g(x_n) \to g(c)$.

By parts (i), (ii), and (iii) of Proposition 2.3, it follows that

$$
\begin{aligned}
(f+g)(x_n) &= f(x_n) + g(x_n) \to f(c) + g(c) = (f+g)(c), \\
(rf)(x_n) &= rf(x_n) \to rf(c) = (rf)(c) \text{ for any } r \in \mathbb{R}, \\
(fg)(x_n) &= f(x_n)g(x_n) \to f(c)g(c) = fg(c).
\end{aligned}
$$

This proves (i), (ii) and (iii).

Next, assume that $f(c) \neq 0$. Then either $f(c) > 0$ or $f(c) < 0$. By Lemma 3.2, there is $\delta > 0$ such that $f(x) \neq 0$ whenever $x \in D$ and $|x-c| < \delta$. Now since $x_n \to c$, there is $n_0 \in \mathbb{N}$ such that $|x_n - c| < \delta$ for all $n \geq n_0$, and so $f(x_n) \neq 0$. By part (iv) of Proposition 2.3, it follows that

$$
\left(\frac{1}{f}\right)(x_n) = \frac{1}{f(x_n)} \to \frac{1}{f(c)} = \left(\frac{1}{f}\right)(c).
$$

This proves (iv).

Finally, if $f(x) \geq 0$ whenever $x \in D$ and $|x-c| < \delta$, then by part (ii) of Proposition 2.4, it follows that for any $k \in \mathbb{N}$, we have

$$
\left(f^{1/k}\right)(x_n) = (f(x_n))^{1/k} \to (f(c))^{1/k} = f^{1/k}(c).
$$

This proves (v). □

With notation and hypotheses as in the proposition above, a combined application of its parts (i) and (ii) shows that the difference $f - g$ is continuous at c. Likewise, a combined application of parts (iii) and (iv) shows that if $g(c) \neq 0$, then the quotient f/g is continuous at c. Further, since every positive rational number r is equal to n/k, where $n, k \in \mathbb{N}$, a combined application of parts (v) and (iii) shows that if there is $\delta > 0$ such that $f(x) \geq 0$ whenever $x \in D$ and $|x-c| < \delta$, then the function f^r is continuous at c for every positive rational number r. Similarly, a combined application of parts (v), (iv), and (iii) shows that if $f(c) > 0$, then the function f^r is continuous at c for every negative rational number r.

We now show that the composition of continuous functions is continuous.

Proposition 3.4. *Let $D, E \subseteq \mathbb{R}$, and let $f : D \to \mathbb{R}$ and $g : E \to \mathbb{R}$ be functions such that $f(D) \subseteq E$. Let $c \in D$ be given. Assume that f is continuous at c and g is continuous at $f(c)$. Then $g \circ f : D \to \mathbb{R}$ is continuous at c.*

Proof. Let (x_n) be any sequence in D such that $x_n \to c$. Then $f(x_n) \to f(c)$ since f is continuous at c. Now $(f(x_n))$ is a sequence in E and g is continuous at $f(c)$. Hence

$$(g \circ f)(x_n) = g(f(x_n)) \to g(f(c)) = (g \circ f)(c).$$

It follows that $g \circ f$ is continuous at c. □

One can piece together continuous functions to construct a continuous function, as the following result shows.

Proposition 3.5. *Let $D \subseteq \mathbb{R}$ and $c \in D$. Suppose*

$$D_1 := \{x \in D : x \le c\} \quad and \quad D_2 := \{x \in D : c \le x\}.$$

Let $f_1 : D_1 \to \mathbb{R}$ and $f_2 : D_2 \to \mathbb{R}$ be functions such that $f_1(c) = f_2(c)$. Then the function $f : D \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} f_1(x) & if\ x \in D_1, \\ f_2(x) & if\ x \in D_2, \end{cases}$$

is continuous on D if f_1 is continuous on D_1 and f_2 is continuous on D_2.

Proof. If $c_1 \in D$ and $c_1 < c$, then the continuity of f_1 at c_1 implies the continuity of f at c_1. Similarly if $c_2 \in D$ and $c < c_2$, then the continuity of f_2 at c_2 implies the continuity of f at c_2. Hence we only need to show that f is continuous at c. Let (x_n) be any sequence in D such that $x_n \to c$. If there is $n_1 \in \mathbb{N}$ such that $x_n \le c$ for all $n \ge n_1$, then $f(x_n) = f_1(x_n)$ for all $n \ge n_1$ and the continuity of f_1 at c implies that $f_1(x_n) \to f_1(c) = f(c)$. Similarly, if there is $n_2 \in \mathbb{N}$ such that $x \le x_n$ for all $n \ge n_2$, then $f(x_n) = f_2(x_n)$ for all $n \ge n_2$ and the continuity of f_2 at c implies that $f(x_n) \to f_2(c) = f(c)$. In the remaining case, there are positive integers $\ell_1 < \ell_2 < \cdots$ and $m_1 < m_2 < \cdots$ such that $x_{\ell_k} \le c < x_{m_k}$ for all $k \in \mathbb{N}$, and $\mathbb{N} = \{\ell_k : k = 1, 2, \ldots\} \cup \{m_k : k = 1, 2, \ldots\}$. Clearly $x_{\ell_k} \to c$ and $x_{m_k} \to c$ as $k \to \infty$. Moreover, $f(x_{\ell_k}) = f_1(x_{\ell_k})$ and $f(x_{m_k}) = f_2(x_{m_k})$ for all $k \in \mathbb{N}$. By the continuity of f_1 at c, we have $f(x_{\ell_k}) \to f_1(c) = f(c)$ and by the continuity of f_2 at c, we have $f(x_{m_k}) \to f_2(c) = f(c)$. It follows that $f(x_n) \to f(c)$. Thus f is continuous at c. We therefore conclude that f is continuous on D. □

Examples 3.6. (i) Consider a polynomial function $p : \mathbb{R} \to \mathbb{R}$ given by

$$p(x) = a_0 + a_1 x + \cdots + a_n x^n \qquad for \quad x \in \mathbb{R}.$$

Applying Proposition 3.3 repeatedly, we find that p is continuous on $\mathbb{R}$. Again, if $q : \mathbb{R} \to \mathbb{R}$ is another polynomial function, then the rational function p/q is continuous at a point $c \in \mathbb{R}$ if $q(c) \ne 0$. For example, if $D = \mathbb{R} \setminus \{1\}$ and

$$f(x) := \frac{x^4 + 3x + 2}{x - 1} \qquad for \quad x \in D,$$

then f is continuous on D.

(ii) Let $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) := |x^4 - 3x^3 + 2x - 1|^{1/4} \qquad \text{for} \quad x \in \mathbb{R},$$

By Propositions 3.3 and 3.4, we see that f is continuous on $\mathbb{R}$.

(iii) Consider a rational number r. Let $D := [0, \infty)$ if $r \geq 0$ and $D := (0, \infty)$ if $r < 0$. For $x \in D$, define $g(x) := x^r$. Since the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x$ is continuous on $\mathbb{R}$ and $g(x) = f^r(x)$ for $x \in D$, it follows from Proposition 3.3 and the remark following its proof that g is continuous on D.

(iv) Let $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) := \begin{cases} x^2 & \text{if } x \leq 0, \\ x & \text{if } x > 0. \end{cases}$$

Then by Proposition 3.5, f is continuous on $\mathbb{R}$.

(v) Let $f : [-1, 1] \to \mathbb{R}$ denote the zig-zag function given in Example 1.18. If $c \in [-1, 1]$ and $c \neq 0$, then Proposition 3.5 implies that f is continuous at c. To show that f is continuous at 0 as well, let (x_n) be a sequence in $[-1, 1]$ such that $x_n \to 0$. It can easily be seen that $|f(x_n)| \leq |x_n|$ for all $n \in \mathbb{N}$, and so $f(x_n) \to 0$. Thus f is continuous at 0. $\diamond$

We conclude this section by giving a criterion for the continuity of a function at a point that does not involve convergence of sequences.
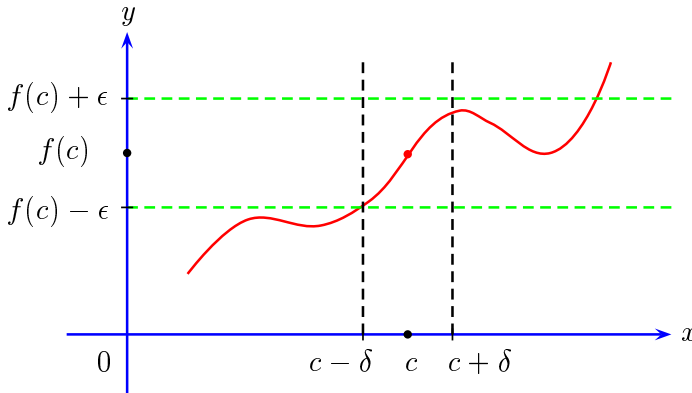
Fig. 3.1. Illustration of the ϵ-δ condition for continuity

Proposition 3.7. *Let $D \subseteq \mathbb{R}$, $c \in D$, and let $f : D \to \mathbb{R}$ be a function. Then f is continuous at c if and only if f satisfies the following ϵ-δ condition: For every $\epsilon > 0$, there is $\delta > 0$ such that*

$$x \in D \text{ and } |x - c| < \delta \implies |f(x) - f(c)| < \epsilon.$$

Proof. Let f be continuous at c. Suppose for a moment that the ϵ-δ condition does not hold. This means that there is $\epsilon > 0$ such that for every $\delta > 0$, there is $x \in D$ satisfying

$$|x - c| < \delta, \quad \text{but} \quad |f(x) - f(c)| \geq \epsilon.$$

Then there is a sequence (x_n) in D such that $|x_n - c| < 1/n$, but $|f(x_n) - f(c)| \geq \epsilon$ for all $n \in \mathbb{N}$. But then $x_n \to c$ and $f(x_n) \nrightarrow f(c)$. This contradicts the continuity of f at c.

Conversely, assume the ϵ-δ condition. Let (x_n) be any sequence in D such that $x_n \to c$. Let $\epsilon > 0$ be given. Then there is $\delta > 0$ such that

$$x \in D \text{ and } |x - c| < \delta \Longrightarrow |f(x) - f(c)| < \epsilon.$$

Since $x_n \to c$, there is $n_0 \in \mathbb{N}$ such that $|x_n - c| < \delta$ for all $n \geq n_0$. Hence $|f(x_n) - f(c)| < \epsilon$ for all $n \geq n_0$. Thus $f(x_n) \to f(c)$. This shows that f is continuous at c. $\qquad\square$

The ϵ-δ condition in the above result is illustrated in Figure 3.1.

3.2 Basic Properties of Continuous Functions

In this section we examine relations between the continuity of a function and various geometric properties of a function considered earlier in Section 1.3. Also, we shall introduce the notion of a uniformly continuous function and discuss its relation with continuity.

Continuity and Boundedness

A bounded function need not be continuous. For instance, we cite the Dirichlet function given in Example 3.1 (iv). Also, a continuous function need not be bounded. For example, let $D_1 := [0, \infty)$ and $f_1(x) := x$ for $x \in D_1$, or $D_2 := (0, 1]$ and $f_2(x) := 1/x$ for $x \in D_2$. An obvious reason why the continuous function f_1 is unbounded is that its domain D_1 is unbounded. To identify the reason why the continuous function f_2 is unbounded on its domain D_2, we introduce the following concept.

Let $D \subseteq \mathbb{R}$. We say that D is a **closed set** if

$$(x_n) \text{ any sequence in } D \text{ and } x_n \to x \Longrightarrow x \in D.$$

Notice that the interval $(0, 1]$ is not a closed set, since $(1/n) \in (0, 1]$ for each $n \in \mathbb{N}$ and $(1/n) \to 0$, but $0 \notin (0, 1]$. Similarly, it can be seen that the following intervals are not closed sets:

$$(a, b], \quad [a, b), \quad (a, b), \quad (a, \infty), \quad (-\infty, b), \quad \text{where } a, b \in \mathbb{R}.$$

On the other hand, the following intervals are closed sets:

$$[a, b], \quad [a, \infty), \quad (-\infty, b], \quad (-\infty, \infty), \quad \text{where} \quad a, b \in \mathbb{R}.$$

To show that the interval $[a, b]$ is a closed set, consider any sequence (x_n) in $[a, b]$ such that $x_n \to x$. Since $a \leq x_n \leq b$ and $x_n \to x$, part (i) of Proposition 2.4 shows that $a \leq x \leq b$, that is, $x \in [a, b]$. Similar proofs can be given to show that $[a, \infty)$ and $(-\infty, b]$ are closed sets. It is obvious that $\mathbb{R} = (-\infty, \infty)$ is a closed set.

We now show that if a function defined on a closed and bounded set is continuous, then it is necessarily bounded. In fact, we shall show that such a function attains its bounds on its domain.

Proposition 3.8. *Let D be a closed and bounded subset of $\mathbb{R}$, and $f : D \to \mathbb{R}$ be a continuous function. Then*

(i) *f is a bounded function, and*
(ii) *f attains its bounds on D, that is, there are r and s in D such that*

$$f(r) = \inf\{f(x) : x \in D\} \quad \text{and} \quad f(s) = \sup\{f(x) : x \in D\}.$$

Proof. (i) Suppose f is not bounded on D. Then for every $n \in \mathbb{N}$, there is $x_n \in D$ such that $|f(x_n)| > n$. Since D is a bounded set, the sequence (x_n) is bounded. By the Bolzano–Weierstrass Theorem, (x_n) has a convergent subsequence (x_{n_k}). If $x_{n_k} \to x$, then $x \in D$ since D is a closed set. Also, since f is continuous at x, we have $f(x_{n_k}) \to f(x)$. Being convergent, the sequence $(f(x_{n_k}))$ is bounded by part (ii) of Proposition 2.2. But $|f(x_{n_k})| > n_k$ for every $k \in \mathbb{N}$ and $n_k \to \infty$ as $k \to \infty$. This contradiction shows that f is a bounded function on D.

(ii) Since the function f is bounded on D, we have $m, M \in \mathbb{R}$ such that

$$m := \inf\{f(x) : x \in D\} \quad \text{and} \quad M := \sup\{f(x) : x \in D\}.$$

By Corollary 2.6, there are sequences (r_n) and (s_n) in D such that $f(r_n) \to m$ and $f(s_n) \to M$. Since D is a bounded set, the sequences (r_n) and (s_n) are bounded. By the Bolzano–Weierstrass Theorem, (r_n) has a convergent subsequence, say (r_{n_k}), and (s_n) has a convergent subsequence, say, (s_{m_j}). If $r_{n_k} \to r$ and $s_{m_j} \to s$, then r and s belong to D, since D is a closed set. Also, since f is continuous at r and s, we have $f(r_{n_k}) \to f(r)$ and $f(s_{m_j}) \to f(s)$. Hence

$$f(r) = \lim_{k \to \infty} f(r_{n_k}) = \lim_{n \to \infty} f(r_n) = m$$

and

$$f(s) = \lim_{j \to \infty} f(s_{m_j}) = \lim_{n \to \infty} f(s_n) = M.$$

Thus f attains its bounds on D. □

Examples 3.9. (i) Let a and b be real numbers such that $a < b$. Since the interval $[a, b]$ is a closed and bounded subset of $\mathbb{R}$, it follows from the preceding result that every continuous function defined on $[a, b]$ is bounded and attains its bounds on $[a, b]$. For example, if $f(x) := x$ for $x \in [-1, 2]$, then f attains its lower bound at -1 and it attains its upper bound at 2. Also, if $f(x) := x^2$ for $x \in [-1, 2]$, then f attains its lower bound at 0 and its upper bound at 2. In general, it is not easy to determine the lower and the upper bounds of a continuous function on $[a, b]$ and to locate the points in $[a, b]$ at which they are attained. We shall return to this question when we consider applications of 'differentiation' in Section 5.1.

(ii) If a subset D of $\mathbb{R}$ is not closed, then a continuous function on D may not be bounded on D, and even if it is bounded, it may not attain its bounds on D. For example, let $D := (a, b]$. If $f(x) := 1/(x - a)$ for $x \in D$, then f is continuous on D, but it is not bounded on D. Also, if $f(x) := x$, for $x \in D$, then f is continuous and bounded on D, but it does not attain its lower bound on D since $\inf\{f(x) : x \in D\} = a$ and $a \notin D$.

(iii) If a subset D of $\mathbb{R}$ is not bounded, then a continuous function on D may not be bounded on D and even if it is bounded on D, it may not attain its lower or upper bounds on D. For example, let $D = [a, \infty)$. If $f(x) := x$ for $x \in D$, then f is continuous on D, but it is not bounded on D. Also, if $f(x) := (x - a)/(x - a + 1)$ for $x \in D$, then f is continuous and bounded on D, but it does not attain its upper bound on D, because $\sup\{f(x) : x \in D\} = 1$ and $f(x) \neq 1$ for any $x \in D$. $\diamond$

Continuity and Monotonicity

It is easy to see that a function that is monotonic on an interval need not be continuous. For example, if $f(x) := [x]$ for $x \in \mathbb{R}$, then f is monotonic on $\mathbb{R}$, but it is discontinuous at every $c \in \mathbb{Z}$. (See Example 3.1 (iii).) Similarly, a continuous function defined on an interval need not be monotonic there, as the example $f(x) := |x|$ for $x \in \mathbb{R}$ shows. However, we now prove a peculiar result that says that if a function is strictly monotonic on an interval, then its inverse function (defined on the range of the given function) is continuous, irrespective of the continuity of the given function. Note also that the range of the given function, that is, the domain of the inverse function need not be an interval. This phenomenon is illustrated by Figure 3.2.

Proposition 3.10. *Let I be an interval and $f : I \to \mathbb{R}$ be a function that is strictly monotonic on I. Then $f^{-1} : f(I) \to \mathbb{R}$ is continuous.*

Proof. Since f is strictly monotonic on I, we see that f is one-one and its inverse $f^{-1} : f(I) \to \mathbb{R}$ is well defined. Consider $d \in f(I)$. Then there is unique $c \in I$ such that $f(c) = d$.

Assume first that f is strictly increasing on I. Let $\epsilon > 0$ be given. Suppose that c is neither the left (hand) endpoint nor the right (hand) endpoint of the interval I. Then there are $c_1, c_2 \in I$ such that

$$c - \epsilon < c_1 < c < c_2 < c + \epsilon.$$

Let $d_1 := f(c_1)$ and $d_2 := f(c_2)$. Since f is strictly increasing on I, we have $d_1 < d < d_2$, and since f^{-1} is also strictly increasing on $f(I)$, we obtain

$$y \in f(I), \ \ d_1 < y < d_2 \Longrightarrow c_1 = f^{-1}(d_1) < f^{-1}(y) < f^{-1}(d_2) = c_2,$$

so that $f^{-1}(d) - \epsilon < f^{-1}(y) < f^{-1}(d) + \epsilon$. Thus if we let $\delta := \min\{d - d_1, d_2 - d\}$, we see that $\delta > 0$ and

$$y \in f(I), \ \ |y - d| < \delta \Longrightarrow |f^{-1}(y) - f^{-1}(d)| < \epsilon.$$

Hence f^{-1} is continuous at d. [See Figure 3.2.] If $c = f^{-1}(d)$ is the left (hand) endpoint of the interval I, then since f and f^{-1} are strictly increasing on I and $f(I)$ respectively, we have

$$y \in f(I) \Longrightarrow d \le y \quad \text{and} \quad f^{-1}(d) \le f^{-1}(y),$$

and so the earlier argument works if we let $\delta := d_2 - d$. If $c = f^{-1}(d)$ is the right (hand) endpoint of the interval I, then similarly we have

$$y \in f(I) \Longrightarrow y \le d \quad \text{and} \quad f^{-1}(y) \le f^{-1}(d),$$

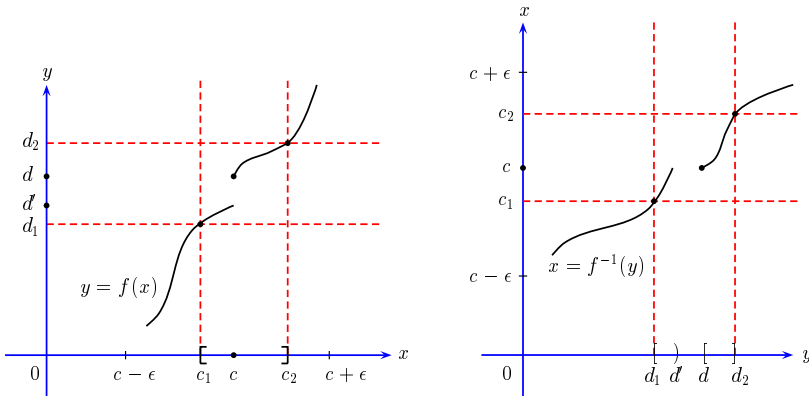and so the earlier argument works if we let $\delta := d - d_1$.

Fig. 3.2. A discontinuous strictly increasing function with continuous inverse

Thus in all the cases, Proposition 3.7 shows that f^{-1} is continuous at d. Since d is an arbitrary point of $f(I)$, we see that $f^{-1} : f(I) \to \mathbb{R}$ is a continuous function. It may be noted that $f(I)$ need not be an interval, as Figure 3.2 shows.

If f is strictly decreasing on I, then $-f$ is strictly increasing on I, and by what we have proved above, $(-f)^{-1} : (-f)(I) \to \mathbb{R}$ is continuous. Since $f^{-1}(y) = (-f)^{-1}(-y)$ for every $y \in f(I)$, it follows from Proposition 3.4 that f^{-1} is a continuous function. $\square$

Example 3.11. Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) := x + [x]$. If $x_1, x_2 \in \mathbb{R}$ and $x_1 < x_2$, then

$$f(x_1) = x_1 + [x_1] < x_2 + [x_1] \leq x_2 + [x_2] = f(x_2).$$

Hence f is strictly increasing on $\mathbb{R}$. If $m \in \mathbb{Z}$, then we have

$$f(x) = x + m \quad \text{for} \quad x \in [m, m+1).$$

Thus $f(\mathbb{R})$ is the union of the semiopen intervals $[2m, 2m+1)$, $m \in \mathbb{Z}$, that is, $f(\mathbb{R}) = \{y \in \mathbb{R} : [y] \text{ is even}\}$, and if $m \in \mathbb{Z}$, then we have

$$f^{-1}(y) = y - m \quad \text{for} \quad y \in [2m, 2m+1).$$

In other words,

$$f^{-1}(y) = y - \frac{[y]}{2} \quad \text{for} \quad y \in f(\mathbb{R}).$$

Observe that f^{-1} is continuous at each point of $f(\mathbb{R})$ even though f is not continuous at any $m \in \mathbb{Z}$. [See Figure 3.3.] $\diamond$
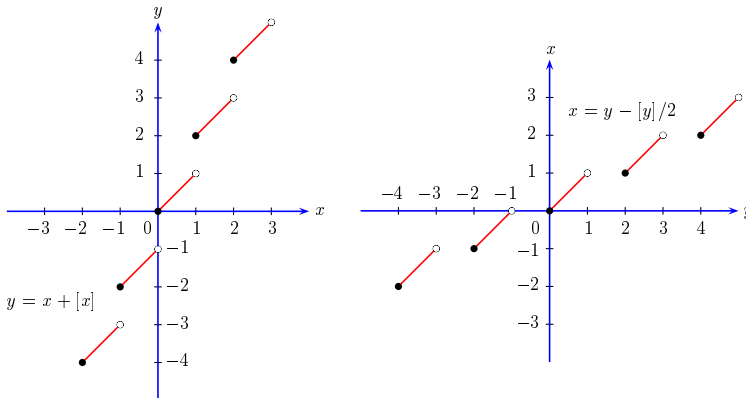

Fig. 3.3. Graphs of $f(x) = x + [x]$ and its inverse $f^{-1}(y) = y - [y]/2$

Corollary 3.12. *Let I be an interval and $f :\to \mathbb{R}$ be a strictly monotonic function such that $f(I)$ is an interval. Then f is one-one and continuous.*

Proof. Since f is strictly increasing on I, it is clear that it is one-one and its inverse $f^{-1} : f(I) \to \mathbb{R}$ is well defined. Let $J := f(I)$ and $g := f^{-1}$. Then g is strictly increasing on the interval J, and by Proposition 3.10, $g^{-1} : g(J) \to \mathbb{R}$ is continuous, that is, $f : I \to \mathbb{R}$ is continuous. $\square$

We shall prove in Proposition 3.14 that the converse of the above corollary holds. We shall later prove a stronger version of the above corollary (where strict monotonicity is replaced by monotonicity) in Proposition 3.36.

Continuity and Convexity

It is easy to see that a continuous function defined on an interval need not be either convex on that interval or concave on that interval. For example, if $f(x) := x^3$ for $x \in \mathbb{R}$, then f is continuous on $\mathbb{R}$, but it is neither convex nor concave on $\mathbb{R}$. (Example 1.15 (ii).) On the other hand, if a function is convex on an interval or if it is concave on an interval, then it is continuous at all points of that interval other than its endpoints. (See Exercise 47.) At an endpoint of an interval, a convex (or a concave) function may be discontinuous. For example, let $I := [-1, 1]$ and let $f : I \to \mathbb{R}$ be given by

$$f(x) := \begin{cases} |x| & \text{if } |x| < 1, \\ 2 & \text{if } x = -1 \text{ or } 1. \end{cases}$$

It can be easily seen that f is convex on I, but f is discontinuous at 1 as well as at -1.

Continuity and Intermediate Value Property

The following important result shows that a continuous function on an interval always has the Intermediate Value Property (IVP).

Proposition 3.13 (Intermediate Value Theorem). *Let I be an interval and $f : I \to \mathbb{R}$ be a continuous function. Then f has the IVP on I. In particular, $f(I)$ is an interval.*

Proof. Let a, b in I with $a < b$. Then $[a, b] \subseteq I$. Let r be an intermediate value between a and b, that is, $r \in (f(a), f(b))$ or $r \in (f(b), f(a))$.

Assume first that $f(a) < f(b)$, so that $r \in (f(a), f(b))$. Define

$$S := \{x \in [a, b] : f(x) < r\}.$$

Now $a \in S$, since $f(a) < r$. Hence $S \neq \emptyset$. Also, the set S is bounded above by b. If $c := \sup S$, then by part (i) of Corollary 2.6, there is a sequence (c_n) in S such that $c_n \to c$. Since $c \in [a, b] \subseteq I$, f is continuous at c. Hence $f(c_n) \to f(c)$. Also, $f(c_n) < r$, since $c_n \in S$. By part (i) of Proposition 2.4, we conclude that $f(c) \leq r$. We note that $c \neq b$, because $r < f(b)$. Let

$$b_n := c + \frac{b - c}{n} \in [a, b] \quad \text{for each } n \in \mathbb{N}.$$

Clearly, $b_n \to c$. The continuity of f at c implies that $f(b_n) \to f(c)$. But since $b_n > c$ and $c = \sup S$, we have $b_n \notin S$, that is, $f(b_n) \geq r$ for all $n \in \mathbb{N}$. As before, part (i) of Proposition 2.4 shows that $f(c) \geq r$. In particular, $c \neq a$. Thus $c \in (a, b)$ and $f(c) = r$.

The case $r \in (f(b), f(a))$ can be proved similarly. □

The above result (together with Propositions 1.23 and 3.10) has the following striking consequence.

Proposition 3.14 (Continuous Inverse Theorem). *Let I be an interval and $f : I \to \mathbb{R}$ be a one-one continuous function. Then the inverse function $f^{-1} : f(I) \to \mathbb{R}$ is continuous. In fact, f is strictly monotonic on I and $f(I)$ is an interval.*

Proof. By the Intermediate Value Theorem (Proposition 3.13), the one-one function f has the IVP. Hence by Proposition 1.23, f is strictly monotonic and $f(I)$ is an interval. Thus, by Proposition 3.10, f^{-1} is continuous. $\square$

The above result shows that the converse of Corollary 3.12 holds. However, the converse of the Intermediate Value Theorem does not hold in general, that is, a discontinuous function may have the IVP on an interval I. We illustrate this by the following example.

Example 3.15. Let $D := [0, 1]$ and consider a 'criss-cross' function defined on D whose graph is obtained by the line segments joining $(1, 1)$ to $(\frac{1}{2}, -1)$, $(\frac{1}{2}, -1)$ to $(\frac{1}{3}, 1)$, $(\frac{1}{3}, 1)$ to $(\frac{1}{4}, -1)$, and so on. [See Figure 3.4.] More precisely, let $f : D \to \mathbb{R}$ be defined as follows: $f(0) := 0$ and for $x \in (0, 1]$,

$$f(x) := \begin{cases} 2k(k+1)x - 2k - 1 & \text{if } \dfrac{1}{k+1} \le x \le \dfrac{1}{k}, \ k \in \mathbb{N}, \ k \text{ odd}, \\[2ex] -2k(k+1)x + 2k + 1 & \text{if } \dfrac{1}{k+1} \le x \le \dfrac{1}{k}, \ k \in \mathbb{N}, \ k \text{ even}. \end{cases}$$

We note that $|f(x)| \le 1$ for all $x \in [0, 1]$. Also, for every $k \in \mathbb{N}$, the function f assumes every value between -1 and 1 on the interval $[1/(k+1), 1/k]$. Let $I_k := [1/(k+1), 1/k]$ and for $x \in I_k$ for $k = 1, 2, \ldots$, define $f_k(x) := f(x)$. Then f is continuous on I_k and $f_k(1/(k+1)) = f_{k+1}(1/(k+1))$ for each $k \in \mathbb{N}$. Since $(0, 1] = \bigcup_{k=1}^{\infty} I_k$, by Proposition 3.5 we see that f is continuous on $(0, 1]$.

On the other hand, f is not continuous at 0. To see this, let $x_n := 1/(2n-1)$ and $y_n := 1/(2n)$ for $n \in \mathbb{N}$. Then $x_n \to 0$ and $y_n \to 0$, but $f(x_n) = 1$, while $f(y_n) = -1$ for all $n \in \mathbb{N}$, so that $f(x_n) \not\to f(0)$ and $f(y_n) \not\to f(0)$. This argument in fact shows that f cannot be made continuous at 0 by redefining its value at 0.

Next, we show that f has the IVP on the interval $[0, 1]$. Let $[a, b]$ be any subinterval of $[0, 1]$. If $a > 0$, then f is continuous on $[a, b]$ and hence f assumes every value between $f(a)$ and $f(b)$ by the Intermediate Value Theorem (Proposition 3.13). Also, if $a = 0$ and $b > 0$, then there is a positive integer k such that $(1/k) < b$. Now $a < 1/(k+1) < 1/k < b$ and f assumes every value between -1 and 1 on the interval $[1/(k+1), 1/k]$. This shows that f has the IVP on $[0, 1]$ although it is not continuous on $[0, 1]$. $\diamondsuit$
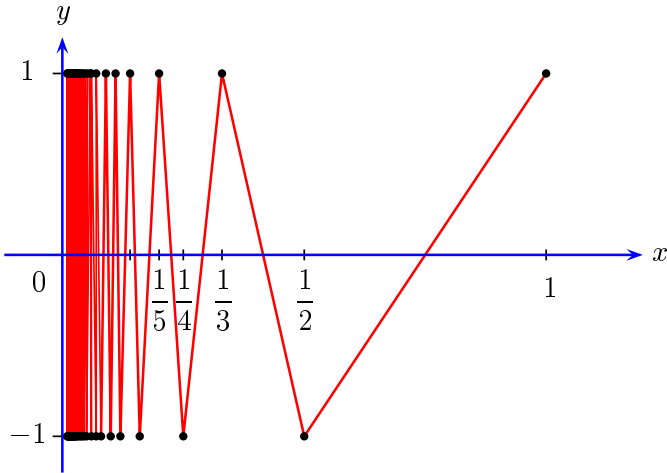
Fig. 3.4. Graph of the criss-cross function in Example 3.15

Remark 3.16. The following partial converse of the Intermediate Value Theorem holds. If a one-one function defined on an interval has the IVP, then it is continuous. This can be seen by noting that such a function is strictly monotonic and its range is an interval (Proposition 1.23), and then appealing to Corollary 3.12. Further, Proposition 3.36 will show that any monotonic function having the IVP on an interval is continuous.   ◇

Uniform Continuity

We introduce a concept that in general is stronger than the concept of continuity of a function. It will be useful in Chapter 6 when we consider 'integrable' functions.

Let $D \subseteq \mathbb{R}$ and let $f : D \to \mathbb{R}$ be a function. We say that f is **uniformly continuous** on D if

$$(x_n), (y_n) \text{ any sequences in } D \text{ and } x_n - y_n \to 0 \Longrightarrow f(x_n) - f(y_n) \to 0.$$

The following result establishes a relation between the continuity and the uniform continuity of a function.

Proposition 3.17. *Let $D \subseteq \mathbb{R}$. Every uniformly continuous function on D is continuous on D. Moreover, if D is a closed and bounded set, then every continuous function on D is uniformly continuous on D.*

Proof. Let $f : D \to \mathbb{R}$ be given. First assume that f is uniformly continuous on D. If $c \in D$ and (x_n) is any sequence in D such that $x_n \to c$, then let $y_n := c$ for all $n \in \mathbb{N}$. Since $x_n - y_n \to 0$, we have $f(x_n) - f(c) = f(x_n) - f(y_n) \to 0$,

that is, $f(x_n) \to f(c)$. Thus f is continuous at c. Since this holds for every $c \in D$, f is continuous on D.

Now assume that D is a closed and bounded set, and f is continuous on D. Suppose f is not uniformly continuous on D. Then there are sequences (x_n) and (y_n) in D such that $x_n - y_n \to 0$, but $f(x_n) - f(y_n) \nrightarrow 0$. Consequently, there exist $\epsilon > 0$ and positive integers $n_1 < n_2 < \cdots$ such that $|f(x_{n_k}) - f(y_{n_k})| \geq \epsilon$ for all $k \in \mathbb{N}$. Since D is a bounded set, the sequence (x_{n_k}) is bounded. By the Bolzano–Weierstrass Theorem, it has a convergent subsequence, say, $(x_{n_{k_j}})$. Let us denote the sequences $(x_{n_{k_j}})$ and $(y_{n_{k_j}})$ by $(\tilde{x}_j)$ and $(\tilde{y}_j)$ for simplicity. Let $\tilde{x}_j \to c$. Then $c \in D$ since D is a closed set. Because $x_n - y_n \to 0$, we have $\tilde{x}_j - \tilde{y}_j \to 0$ and hence $\tilde{y}_j \to c$ as well. Since f is continuous at c, we obtain $f(\tilde{x}_j) \to f(c)$ and $f(\tilde{y}_j) \to f(c)$. Thus

$$f(\tilde{x}_j) - f(\tilde{y}_j) \to f(c) - f(c) = 0.$$

But this is a contradiction, since $|f(\tilde{x}_j) - f(\tilde{y}_j)| \geq \epsilon$ for all $j \in \mathbb{N}$. Hence f is uniformly continuous on D. $\square$

We remark that the continuity of a function on a set D is a local concept, that is, a function is defined to be continuous on D if it is continuous at every $c \in D$. On the other hand, the uniform continuity of a function on a set D takes into account the behavior of the function on the entire set D. In this sense, uniform continuity is a global concept.

Examples 3.18. (i) Since the interval $[a, b]$ is a closed and bounded set, it follows from the preceding proposition that every continuous function on $[a, b]$ is uniformly continuous on $[a, b]$. This result will be of crucial importance in our discussion of Riemann integration in Chapter 6.

(ii) If a subset D of $\mathbb{R}$ is not closed, then a continuous function on D may not be uniformly continuous on D. For example, consider $D := (a, b]$ and $f : D \to \mathbb{R}$ defined by $f(x) := 1/(x - a)$. Clearly f is continuous on D. But f is not uniformly continuous on D. To see this, let

$$x_n := a + \frac{b - a}{n} \quad \text{and} \quad y_n := a + \frac{b - a}{n + 1}, \quad \text{for } n \in \mathbb{N}.$$

Then $x_n - y_n = (b - a)/n(n + 1) \to 0$, but

$$f(x_n) - f(y_n) = \frac{n - (n + 1)}{b - a} = \frac{1}{a - b} \quad \text{for all } n \in \mathbb{N},$$

and hence $f(x_n) - f(y_n) \nrightarrow 0$.

(iii) If a subset D of $\mathbb{R}$ is not bounded, then a continuous function on D may not be uniformly continuous on D. For example, let $D = [a, \infty)$ and $f(x) = x^2$ for $x \in D$. Then f is continuous on D. But f is not uniformly continuous on D. To see this, let

$$x_n := a + n \quad \text{and} \quad y_n := a + n - \frac{1}{n}, \quad n \in \mathbb{N}.$$

Then $x_n - y_n = 1/n \to 0$, but

$$f(x_n) - f(y_n) = (a+n)^2 - \left(a + n - \frac{1}{n}\right)^2 = 2 + \frac{2a}{n} - \frac{1}{n^2}$$

for all $n \in \mathbb{N}$, and so $f(x_n) - f(y_n) \not\to 0$. ◇

Finally, we give a criterion for the uniform continuity of a function that does not involve convergence of sequences. The following result may be compared with the ϵ-δ condition for continuity given in Proposition 3.7.

Proposition 3.19. *Let $D \subseteq \mathbb{R}$ and $f : D \to \mathbb{R}$ be a function. Then f is uniformly continuous on D if and only if f satisfies the following ϵ-δ condition: For every $\epsilon > 0$, there is $\delta > 0$ such that*

$$x, y \in D \text{ and } |x - y| < \delta \Longrightarrow |f(x) - f(y)| < \epsilon.$$

Proof. Let f be uniformly continuous on D. Suppose that there is $\epsilon > 0$ such that for any given $\delta > 0$, there are x and y in D such that $|x - y| < \delta$, but $|f(x) - f(y)| \geq \epsilon$. Considering $\delta := 1/n$ for $n \in \mathbb{N}$, we obtain sequences (x_n) and (y_n) in D such that $|x_n - y_n| < 1/n$ but $|f(x_n) - f(y_n)| \geq \epsilon$ for all $n \in \mathbb{N}$. Then $x_n - y_n \to 0$, but $f(x_n) - f(y_n) \not\to 0$. This contradicts the assumption that f is uniformly continuous on D.

Conversely, assume that the ϵ-δ condition holds. Let (x_n) and (y_n) be any sequences in D such that $x_n - y_n \to 0$. Let $\epsilon > 0$ be given. Then there is $\delta > 0$ such that $|f(x) - f(y)| < \epsilon$, whenever $x, y \in D$ and $|x - y| < \delta$. Since $x_n - y_n \to 0$, we can find $n_0 \in \mathbb{N}$ such that $|x_n - y_n| < \delta$ for all $n \geq n_0$. But then $|f(x_n) - f(y_n)| < \epsilon$ for all $n \geq n_0$. Thus $f(x_n) - f(y_n) \to 0$. Hence f is uniformly continuous on D. □

3.3 Limits of Functions of a Real Variable

In Chapter 2 we have seen what is meant by the limit of a sequence. As we know, a sequence is a function whose domain is the set $\mathbb{N}$ of all natural numbers. We shall now define the concept of a limit of a function at a point in $\mathbb{R}$ provided the domain of the function satisfies certain conditions. For defining this concept as well as for proving several basic properties, we shall utilize the notion of sequences and their properties.

Let $D \subseteq \mathbb{R}$ and $c \in \mathbb{R}$ be such that D contains $(c - r, c)$ and $(c, c + r)$ for some $r > 0$, that is, D contains an open interval about c except possibly the point c itself. Consider a function $f : D \to \mathbb{R}$. We say that a **limit** of f as x tends to c exists if there is a real number ℓ such that

$$(x_n) \text{ any sequence in } D \setminus \{c\} \text{ and } x_n \to c \Longrightarrow f(x_n) \to \ell.$$

We then write

$$f(x) \to \ell \ \text{ as } \ x \to c \quad \text{or} \quad \lim_{x \to c} f(x) = \ell.$$

Note that there does exist a sequence in $D \setminus \{c\}$ that converges to c. For example,

$$x_n := c - \frac{r}{n+1}$$

belongs to $D \setminus \{c\}$ for all $n \in \mathbb{N}$ and $x_n \to c$. In particular, it follows from part (i) of Proposition 2.2 that $\lim_{x \to c} f(x)$ is unique whenever it exists.

Examples 3.20. (i) Consider the function whose graph is as in Figure 3.5. More precisely, let $D := \mathbb{R}$, $c := 0$, and let $f : D \to \mathbb{R}$ be defined by

$$f(x) := \begin{cases} 1 & \text{if } x < 0, \\ 2 & \text{if } x = 0, \\ x+1 & \text{if } x > 0. \end{cases}$$

Then $\lim_{x \to 0} f(x) = 1$. To see this, let (x_n) be a sequence in $D \setminus \{0\}$ such that $x_n \to 0$. If $x_n < 0$ for any $n \in \mathbb{N}$, then $f(x_n) - 1 = 1 - 1 = 0$, and if $x_n > 0$ for any $n \in \mathbb{N}$, then $f(x_n) - 1 = (x_n + 1) - 1 = x_n$. It follows that $f(x_n) \to 1$. Thus $\ell = 1$ is the limit of f as x tends to 0. Note that $f(0) = 2$.
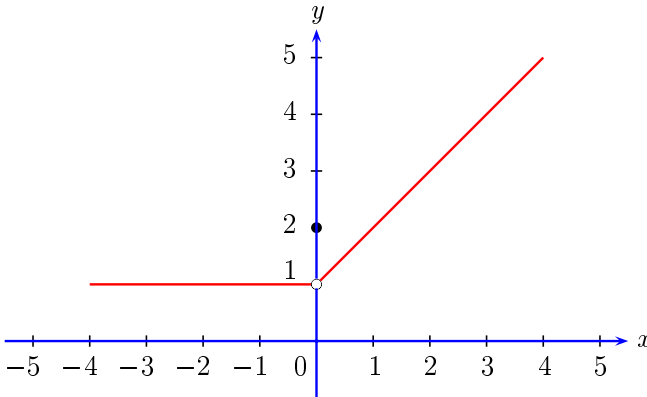

Fig. 3.5. Graph of $f(x) := \begin{cases} 1 & \text{if } x < 0, \\ 2 & \text{if } x = 0, \\ x+1 & \text{if } x > 0 \end{cases}$

(ii) Let $D := \mathbb{R} \setminus \{0\}$, $c := 0$ and let $f : D \to \mathbb{R}$ be defined by

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

If $x_n := (-1)^n/n$ for $n \in \mathbb{N}$, then (x_n) is a sequence in D and $x_n \to 0$, but since $f(x_n) = (-1)^n$ for $n \in \mathbb{N}$, the sequence $(f(x_n))$ is divergent, as we have noted in Example 2.1(ii). Hence a limit of f as x tends to 0 does not exist. This can also be seen by letting $y_n := 1/n$, $z_n := -1/n$ for $n \in \mathbb{N}$ and observing that $y_n \to 0$, $z_n \to 0$, whereas $f(y_n) \to 1$, $f(z_n) \to -1$.

(iii) Let $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) := \begin{cases} 1 & \text{if } x \text{ is rational,} \\ 0 & \text{if } x \text{ is irrational.} \end{cases}$$

Then for any $c \in \mathbb{R}$, a limit of $f(x)$ as x tends to c does not exist. To see this, let $c \in \mathbb{R}$ and consider

$$x_n := \frac{[nc] + 1}{n} \quad \text{and} \quad y_n := \frac{[nc] + \sqrt{2}}{n}.$$

Then for each $n \in \mathbb{N}$, $c < x_n \le c + (1/n)$ and $c < y_n \le c + (\sqrt{2}/n)$. So $x_n \to c$ and $y_n \to c$. Since each x_n is rational, $f(x_n) = 1$ and since each y_n is irrational, $f(y_n) = 0$. Hence $f(x_n) \to 1$, whereas $f(y_n) \to 0$. Thus a limit of $f(x)$ as x tends to c does not exist. $\diamond$

We now relate the concepts of continuity and limit.

Proposition 3.21. *Let $D \subseteq \mathbb{R}$ and $c \in \mathbb{R}$ be such that $(c - r, c + r) \subseteq D$ for some $r > 0$. Consider a function $f : D \to \mathbb{R}$. Then f is continuous at c if and only if $\lim_{x \to c} f(x)$ exists and is equal to $f(c)$.*

Proof. Assume that f is continuous at c. Let (x_n) be any sequence in D such that $x_n \to c$. By the continuity of f at c, we have $f(x_n) \to f(c)$. Thus $\lim_{x \to c} f(x)$ exists and equals $f(c)$.

Conversely, assume that $\lim_{x \to c} f(x)$ exists and is equal to $f(c)$. Let (x_n) be any sequence in D such that $x_n \to c$. If there is $n_0 \in \mathbb{N}$ such that $x_n = c$ for all $n \ge n_0$, then clearly $f(x_n) \to f(c)$. Otherwise, there are positive integers $n_1, n_2, \ldots$ such that $n_1 < n_2 < \cdots$ and $\{n \in \mathbb{N} : x_{n_k} \ne c\} = \{n_k : k \in \mathbb{N}\}$. Now, (x_{n_k}) is a sequence in $D \setminus \{c\}$ that converges to c, and therefore, $f(x_{n_k}) \to f(c)$. Since $f(x_n) = f(c)$ for all $\mathbb{N} \setminus \{n_k : k \in \mathbb{N}\}$, it follows that $f(x_n) \to f(c)$. Hence f is continuous at c. $\square$

Example 3.22. The function $f : \mathbb{R} \to \mathbb{R}$ given in Example 3.20(i) is not continuous at 0 since $\lim_{x \to 0} f(x) = 1$ and $f(0) = 2$. On the other hand, if we define $g : \mathbb{R} \to \mathbb{R}$ by

$$g(x) := \begin{cases} 1 & \text{if } x \le 0, \\ x + 1 & \text{if } x > 0, \end{cases}$$

then, as before, we have $\lim_{x \to 0} g(x) = 1$ and also $g(0) = 1$. Hence g is continuous at 0. $\diamond$

We now prove some results that are useful in calculating the limits of several functions. First we consider how the algebraic operations on $\mathbb{R}$ are related to limits of functions of a real variable. The following result is known as the **Limit Theorem for Functions**.

Proposition 3.23. *Let $D \subseteq \mathbb{R}$ and $c \in \mathbb{R}$ be such that D contains $(c - r, c)$ and $(c, c + r)$ for some $r > 0$, and $f, g : D \to \mathbb{R}$ be functions such that*

$$\lim_{x \to c} f(x) = \ell \quad and \quad \lim_{x \to c} g(x) = m.$$

Then

(i) $\lim_{x \to c} (f + g)(x) = \ell + m,$

(ii) $\lim_{x \to c} (rf)(x) = r\ell$ *for any $x \in \mathbb{R}$,*

(iii) $\lim_{x \to c} (fg)(x) = \ell m,$

(iv) *if $\ell \neq 0$, then there is $\delta > 0$ such that $\delta \leq r$ and $f(x) \neq 0$ for all x satisfying $0 < |x - c| < \delta$; moreover, for the function $1/f : \{x \in \mathbb{R} : 0 < |x - c| < \delta\} \to \mathbb{R}$, we have*

$$\lim_{x \to c} \left(\frac{1}{f}\right)(x) = \frac{1}{\ell}.$$

Proof. Consider any sequence (x_n) in $D \setminus \{c\}$ such that $x_n \to c$. Then

$$f(x_n) \to \ell \quad and \quad g(x_n) \to m.$$

By parts (i), (ii), and (iii) of Proposition 2.3, we have

$$\begin{aligned}
(f + g)(x_n) &= f(x_n) + g(x_n) \to \ell + m, \\
(rf)(x_n) &= rf(x_n) \to r\ell \text{ for any } r \in \mathbb{R}, \\
(fg)(x_n) &= f(x_n)g(x_n) \to \ell m.
\end{aligned}$$

This proves (i), (ii), and (iii).

To prove (iv), suppose $\ell \neq 0$. If there is no $\delta > 0$ such that $\delta \leq r$ and $f(x) \neq 0$ for all x satisfying $0 < |x - c| < \delta$, then we can find a sequence (c_n) in $D \setminus \{c\}$ such that

$$|c_n - c| < \frac{r}{n} \quad and \quad f(c_n) = 0 \quad \text{for every } n \in \mathbb{N}.$$

Since $c_n \to c$, we have $f(c_n) \to \ell$. But this is not possible since $f(c_n) = 0$ for every $n \in \mathbb{N}$, whereas $\ell \neq 0$. Hence there is $\delta > 0$ such that $f(x) \neq 0$ for all x satisfying $0 < |x - c| < \delta$, and the function $1/f$ is defined at all such x. Moreover, if (x_n) is any sequence in $\{x \in \mathbb{R} : 0 < |x - c| < \delta\}$ such that $x_n \to c$, then $f(x_n) \to \ell$ and hence by part (iv) of Proposition 2.3, $1/f(x_n) \to 1/\ell$. $\square$

With notation and hypotheses as in the proposition above, a combined application of its parts (i) and (ii) shows that $\lim_{x \to c}(f - g)(x) = \ell - m$. Likewise, a combined application of parts (iii) and (iv) shows that if $m \neq 0$, then $\lim_{x \to c}(f/g)(x) = \ell/m$.

Next, we show how the order relation on $\mathbb{R}$ and the operation of taking the kth root are preserved under limits.

Proposition 3.24. *Let D, c, f, g, ℓ, and m be as in Proposition 3.23.*

(i) *If there is $\delta > 0$ such that*

$$f(x) \leq g(x) \text{ for all } x \in \mathbb{R} \text{ satisfying } 0 < |x - c| < \delta,$$

then $\ell \leq m$. Conversely, if $\ell < m$, then there is $\delta > 0$ such that

$$f(x) < g(x) \text{ for all } x \in \mathbb{R} \text{ satisfying } 0 < |x - c| < \delta.$$

In particular, if there is $\delta > 0$ such that $g(x) \geq 0$ for all $x \in \mathbb{R}$ satisfying $0 < |x - c| < \delta$, then $\lim_{x \to c} g(x) \geq 0$, and conversely, if $\lim_{x \to c} g(x) > 0$, then there is $\delta > 0$ such that $g(x) > 0$ for all $x \in \mathbb{R}$ satisfying $0 < |x - c| < \delta$.

(ii) *If $f(x) \geq 0$ for all $x \in D$, then $\ell \geq 0$ and for any $k \in \mathbb{N}$, we have*

$$\lim_{x \to c} f^{1/k}(x) = \ell^{1/k}.$$

Proof. Consider a sequence (x_n) in $D \setminus \{c\}$ such that $x_n \to c$.

(i) Suppose there is $\delta > 0$ such that $f(x) \leq g(x)$ for all $x \in \mathbb{R}$ satisfying $0 < |x - c| < \delta$. Then there is $n_0 \in \mathbb{N}$ such that $|x_n - c| < \delta$ for all $n \geq n_0$, and hence $f(x_n) \leq g(x_n)$ for all $n \geq n_0$. Since $f(x_n) \to \ell$ and $g(x_n) \to m$, part (i) of Proposition 2.4 shows that $\ell \leq m$.

Conversely, suppose $\ell < m$. If there is no $\delta > 0$ such that $f(x) < g(x)$ for all $x \in \mathbb{R}$ satisfying $0 < |x - c| < \delta$, then we can find a sequence (c_n) in $D \setminus \{c\}$ such that

$$|c_n - c| < \frac{1}{n} \quad \text{and} \quad f(c_n) \geq g(c_n) \quad \text{for all } n \in \mathbb{N}.$$

Then $c_n \to c$ and so

$$\ell = \lim_{n \to \infty} f(c_n) \geq \lim_{n \to \infty} g(c_n) = m,$$

again by part (i) of Proposition 2.4. This contradicts $\ell < m$. Hence there is $\delta > 0$ such that $f(x) < g(x)$ for all $x \in \mathbb{R}$ satisfying $0 < |x - c| < \delta$.

The particular case follows by letting $f = 0$.

(ii) Part (i) implies that $\ell \geq 0$. Let $k \in \mathbb{N}$. By part (ii) of Proposition 2.4, it follows that $(f(x_n))^{1/k} \to \ell^{1/k}$. Since (x_n) is an arbitrary sequence in $D \setminus \{c\}$ such that $x_n \to c$, we have $\lim_{x \to c} f^{1/k}(x) = \ell^{1/k}$. $\qquad \square$

With notation and hypotheses as in the above proposition, a combined application of part (iii) above and part (iii) of Proposition 3.23 shows that if r is any positive rational number and $f(x) \geq 0$ for all $x \in D$, then $\lim_{x \to c} f^r(x) = \ell^r$, since $r = m/k$, where $m,\, k \in \mathbb{N}$. This, together with part (iv) of Proposition 3.23, shows that if $\ell > 0$, then $\lim_{x \to c} f^r(x) = \ell^r$ for any negative rational number r.

Proposition 3.25 (Sandwich Theorem). *Let $D \subseteq \mathbb{R}$ and $c \in \mathbb{R}$ be such that $(c - r, c)$ and $(c, c + r)$ are contained in D for some $r > 0$. Assume that $f, g, h : D \to \mathbb{R}$ are such that*

$$f(x) \leq h(x) \leq g(x) \quad \text{for all} \ \ x \in D \quad \text{and} \quad \lim_{x \to c} f(x) = \ell = \lim_{x \to c} g(x).$$

Then

$$\lim_{x \to c} h(x) = \ell.$$

Proof. Let (x_n) be any sequence in $D \setminus \{c\}$ such that $x_n \to c$. Then $f(x_n) \to \ell$, $g(x_n) \to \ell$, and $f(x_n) \leq h(x_n) \leq g(x_n)$ for all $n \in \mathbb{N}$. Hence $h(x_n) \to \ell$ by the Sandwich Theorem for sequences. This proves that $\lim_{x \to c} h(x) = \ell$. □

Example 3.26. Consider the 'criss-cross' function $f : [0, 1] \to \mathbb{R}$ given in Example 3.15. Let $D := [-1, 1]$ and define

$$\tilde{f}(x) = \begin{cases} f(x) & \text{if } x \in [0, 1], \\ f(-x) & \text{if } x \in [-1, 0). \end{cases}$$

Then $-1 \leq \tilde{f}(x) \leq 1$ for all $x \in D$. Let $h : D \to \mathbb{R}$ be defined by $h(x) = x\tilde{f}(x)$ and $g : D \to \mathbb{R}$ be defined by $g(x) = |x|$. Since $-g(x) \leq h(x) \leq g(x)$ for all $x \in D$ and $\lim_{x \to 0} g(x) = 0 = \lim_{x \to 0}(-g)(x)$, the Sandwich Theorem shows that $\lim_{x \to 0} h(x) = 0$. ◇

Let us now give a criterion for the existence of a limit of a function of a real variable that does not involve convergence of sequences. The following result may be compared with Proposition 3.7.

Proposition 3.27. *Let $D \subseteq \mathbb{R}$ and $c \in \mathbb{R}$ be such that $(c - r, c)$ and $(c, c + r)$ are contained in D for some $r > 0$, and let consider $f : D \to \mathbb{R}$. Then $\lim_{x \to c} f(x)$ exists if and only if there is $\ell \in \mathbb{R}$ satisfying the following ϵ-δ condition: For every $\epsilon > 0$, there is $\delta > 0$ such that*

$$x \in D \text{ and } 0 < |x - c| < \delta \Longrightarrow |f(x) - \ell| < \epsilon.$$

Proof. Assume that $\lim_{x \to c} f(x)$ exists and is equal to ℓ. Suppose the ϵ-δ condition does not hold. This means that there is $\epsilon > 0$ such that for every $\delta > 0$, there is $x \in D$ satisfying

$$0 < |x - c| < \delta, \text{ but } |f(x) - \ell| \geq \epsilon.$$

Taking $\delta = 1/n$ for $n \in \mathbb{N}$, we see that there is a sequence (x_n) in $D \setminus \{c\}$ such that $|x_n - c| < 1/n$, but $|f(x_n) - \ell| \geq \epsilon$ for all $n \in \mathbb{N}$. Now $x_n \to c$ and $f(x_n) \not\to \ell$. This contradicts the assumption that $\lim_{x \to c} f(x) = \ell$.

Conversely, assume the ϵ-δ condition. Let (x_n) be any sequence in $D \setminus \{c\}$ such that $x_n \to c$. Let $\epsilon > 0$ be given. Then there is $\delta > 0$ such that

$$x \in D \text{ and } 0 < |x - c| < \delta \Longrightarrow |f(x) - \ell| < \epsilon.$$

Since $x_n \to c$, there is $n_0 \in \mathbb{N}$ such that $|x_n - c| < \delta$ for all $n \geq n_0$. Hence $|f(x_n) - \ell| < \epsilon$ for all $n \geq n_0$. Thus $f(x_n) \to \ell$. It follows that $\lim_{x \to c} f(x)$ exists and is equal to ℓ. $\qquad \square$

Next, we consider an analogue of the Cauchy Criterion for the convergence of a sequence (Proposition 2.19).

Proposition 3.28 (Cauchy Criterion for Limits of Functions). *Let $D \subseteq \mathbb{R}$ and $c \in \mathbb{R}$ be such that $(c - r, c)$ and $(c, c + r)$ are contained in D for some $r > 0$, and consider $f : D \to \mathbb{R}$. Then $\lim_{x \to c} f(x)$ exists if and only if for every $\epsilon > 0$, there is $\delta > 0$ such that*

$$x, y \in D, \ 0 < |x - c| < \delta \text{ and } 0 < |y - c| < \delta \Longrightarrow |f(x) - f(y)| < \epsilon.$$

Proof. Assume that $\lim_{x \to c} f(x)$ exists and is equal to ℓ. Let $\epsilon > 0$ be given. By Proposition 3.27, there is $\delta > 0$ such that

$$x \in D \text{ and } 0 < |x - c| < \delta \Longrightarrow |f(x) - \ell| < \frac{\epsilon}{2}.$$

Hence for $x, y \in D$ satisfying $0 < |x - c| < \delta$ and $0 < |y - c| < \delta$, we have

$$|f(x) - f(y)| \leq |f(x) - \ell| + |\ell - f(y)| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Conversely, assume that the condition given in the statement of the proposition holds. Let $\epsilon > 0$ be given. Then there is $\delta > 0$ such that

$$x, y \in D, \ 0 < |x - c| < \delta \text{ and } 0 < |y - c| < \delta \Longrightarrow |f(x) - f(y)| < \epsilon.$$

Consider a sequence (x_n) in $D \setminus \{c\}$ such that $x_n \to c$. Then there is $n_0 \in \mathbb{N}$ such that $|x_n - c| < \delta$ for all $n \geq n_0$. Consequently,

$$|f(x_n) - f(x_m)| < \epsilon \quad \text{for all } n, m \geq n_0.$$

Thus $(f(x_n))$ is a Cauchy sequence in $\mathbb{R}$. By the Cauchy Criterion for sequences (Proposition 2.19), there is $\ell \in \mathbb{R}$ such that $f(x_n) \to \ell$. Hence there is $n_1 \in \mathbb{N}$ such that $n_1 \geq n_0$ and $|f(x_{n_1}) - \ell| < \epsilon$. Since $0 < |x_{n_1} - c| < \delta$, it follows that

$$x \in D \text{ and } 0 < |x - c| < \delta \Longrightarrow |f(x) - \ell| \leq |f(x) - f(x_{n_1})| + |f(x_{n_1}) - \ell| < 2\epsilon.$$

Since $\epsilon > 0$ is arbitrary, Proposition 3.27 shows that $\lim_{x \to c} f(x)$ exists and is equal to ℓ. $\qquad \square$

We now consider one-sided limits. Let D be a subset of $\mathbb{R}$ and $f : D \to \mathbb{R}$ be a function. Suppose that $c \in \mathbb{R}$ is such that $(c - r, c) \subseteq D$ for some $r > 0$. We say that a **left (hand) limit** of f as x tends to c (from the left) exists if there is a real number ℓ such that

$$(x_n) \text{ any sequence in } D, \ x_n < c, \text{ and } x_n \to c \Longrightarrow f(x_n) \to \ell.$$

We then write

$$f(x) \to \ell \text{ as } x \to c^- \quad \text{or} \quad \lim_{x \to c^-} f(x) = \ell.$$

It is easy to see that if $\lim_{x \to c^-} f(x)$ exists, then it is unique.

Similarly, if $c \in \mathbb{R}$ is such that $(c, c+r) \subseteq D$ for some $r > 0$, then we define the **right (hand) limit** of $f(x)$ as x tends c (from the right) upon replacing the requirement '$x_n < c$' by the requirement '$x_n > c$' in the definition above. We then write

$$f(x) \to \ell \text{ as } x \to c^+ \quad \text{or} \quad \lim_{x \to c^+} f(x) = \ell.$$

Results similar to Propositions 3.23, 3.24, 3.25, 3.27, and 3.28 hold for left limits and for right limits. We also have the following result.

Proposition 3.29. *Let $D \subseteq \mathbb{R}$ and $c \in \mathbb{R}$ be such that $(c - r, c)$ and $(c, c+r)$ are contained in D for some $r > 0$, and consider $f : D \to \mathbb{R}$. Then*

$$\lim_{x \to c} f(x) = \ell \iff \lim_{x \to c^-} f(x) = \ell = \lim_{x \to c^+} f(x).$$

If in addition $c \in D$, then

$$f \text{ is continuous at } c \iff \lim_{x \to c^-} f(x) = f(c) = \lim_{x \to c^+} f(x).$$

Proof. If $\lim_{x \to c} f(x) = \ell$, then clearly $\lim_{x \to c^-} f(x) = \ell = \lim_{x \to c^+} f(x)$.

Conversely, assume that $\lim_{x \to c^-} f(x) = \ell = \lim_{x \to c^+} f(x)$. Let (x_n) be any sequence in $D \setminus \{c\}$. Then $x_n < c$ or $x_n > c$ for any $n \in \mathbb{N}$. If there is $n_1 \in \mathbb{N}$ such that $x_n < c$ for all $n \geq n_1$, then since $\lim_{x \to c^-} f(x) = \ell$, we see that $f(x_n) \to \ell$. Also, if there is $n_2 \in \mathbb{N}$ such that $x_n > c$ for all $n \geq n_2$, then since $\lim_{x \to c^+} f(x) = \ell$, we see that again $f(x_n) \to \ell$. In the remaining case, there are positive integers $j_1 < j_2 < \cdots$ and $m_1 < m_2 < \cdots$ such that $x_{j_k} < c < x_{m_k}$ for all $k \in \mathbb{N}$ and $\mathbb{N} = \{j_k : k = 1, 2, \ldots\} \cup \{m_k : k = 1, 2, \ldots\}$. Since $\lim_{x \to c^-} f(x) = \ell$, we have $f(x_{j_k}) \to \ell$. Also, since $\lim_{x \to c^+} f(x) = \ell$, we have $f(x_{m_k}) \to \ell$. Since every $n \in \mathbb{N}$ is equal to j_k or to m_k for some $k \in \mathbb{N}$, it follows that $f(x_n) \to \ell$. We therefore conclude that $\lim_{x \to c} f(x) = \ell$.

The last statement of the proposition follows from the equivalence proved above and Proposition 3.21. $\square$

Next, suppose that $D \subseteq \mathbb{R}$ and D contains a semi-infinite interval (a, ∞), where $a \in \mathbb{R}$. Consider a function $f : D \to \mathbb{R}$. We say that a **limit** of f as x tends to infinity exists if there is a real number ℓ such that

$$(x_n) \text{ any sequence in } D \text{ and } x_n \to \infty \implies f(x_n) \to \ell.$$

We then write

$$f(x) \to \ell \text{ as } x \to \infty \quad \text{or} \quad \lim_{x \to \infty} f(x) = \ell.$$

Since there does exist a sequence (x_n) in D such that $x_n \to \infty$, we see that $\lim_{x \to \infty} f(x)$ is unique. If $D \subseteq \mathbb{R}$ contains a semi-infinite interval $(-\infty, a)$ where $a \in \mathbb{R}$, then we define $\lim_{x \to -\infty} f(x)$ analogously, and write

$$f(x) \to \ell \text{ as } x \to -\infty \quad \text{or} \quad \lim_{x \to -\infty} f(x) = \ell.$$

Results similar to 3.23, 3.24, and 3.25 hold for limits as $x \to \infty$ or as $x \to -\infty$. We now give an analogue of Proposition 3.27 for such limits.

Proposition 3.30. *Let $D \subseteq \mathbb{R}$ be such that (a, ∞) is contained in D for some $a \in \mathbb{R}$ and let $f : D \to \mathbb{R}$ be a function. Then $\lim_{x \to \infty} f(x)$ exists if and only if there is $\ell \in \mathbb{R}$ satisfying the following ϵ-α condition: For every $\epsilon > 0$, there is $\alpha \in \mathbb{R}$ such that*

$$x \in D \text{ and } x \geq \alpha \implies |f(x) - \ell| < \epsilon.$$

Proof. Let $\lim_{x \to \infty} f(x)$ exist and equal ℓ. Suppose for a moment that the ϵ-α condition does not hold. This means that there is $\epsilon > 0$ such that for every $\alpha \in \mathbb{R}$, there is $x \in D$ satisfying

$$x \geq \alpha, \text{ but } |f(x) - \ell| \geq \epsilon.$$

By choosing $\alpha = n$ for each $n \in \mathbb{N}$, we may find a sequence (x_n) in D such that $x_n \geq n$, but $|f(x_n) - \ell| \geq \epsilon$ for all $n \in \mathbb{N}$. Now $x_n \to \infty$ and $f(x_n) \nrightarrow \ell$. This contradicts $\lim_{x \to \infty} f(x) = \ell$.

Conversely, assume the ϵ-α condition. Let (x_n) be any sequence in D such that $x_n \to \infty$. Let $\epsilon > 0$ be given. Then there is $\alpha \in \mathbb{R}$ such that

$$x \in D \text{ and } x \geq \alpha \implies |f(x) - \ell| < \epsilon.$$

Since $x_n \to \infty$, there is $n_0 \in \mathbb{N}$ such that $x_n \geq \alpha$ for all $n \geq n_0$. Hence $|f(x_n) - \ell| < \epsilon$ for all $n \geq n_0$. Thus $f(x_n) \to \ell$. So $\lim_{x \to \infty} f(x)$ exists and equals ℓ. $\qquad \square$

Remark 3.31. Let $D \subseteq \mathbb{R}$ be such that (a, ∞) is contained in D for some $a \in \mathbb{R}$, and let $f, g : D \to \mathbb{R}$ be functions. We may compare the orders of magnitude of f and g as $x \to \infty$ just as we compared the orders of magnitude of sequences (a_n) and (b_n) in Remark 2.11.

If there are $K > 0$ and $\alpha \in \mathbb{R}$ such that $|f(x)| \leq K|g(x)|$ for all $x \geq \alpha$, then we write $f(x) = O(g(x))$ as $x \to \infty$ [read $f(x)$ is big-oh of $g(x)$ as x tends to infinity]. In particular, if $g(x) = 1$ for all large x, then $f(x) = O(1)$ as $x \to \infty$, and this means that the function f is bounded. Broadly speaking, $f(x) = O(g(x))$ as $x \to \infty$ if the order of magnitude of f is at most the order of magnitude of g as $x \to \infty$. In case f and g are monotonically increasing functions and $f(x) = O(g(x))$ as $x \to \infty$, then we also say that the **growth rate** of f is at most the growth rate of g as $x \to \infty$. For example,

$$10\,[x] + 100 = O(x) \quad \text{and} \quad \frac{10}{[x]} + \frac{100}{x\sqrt{x}} = O\left(\frac{1}{x}\right).$$

Given $\epsilon > 0$, if there is $\alpha \in \mathbb{R}$ such that $|f(x)| \leq \epsilon|g(x)|$ for all $x \geq \alpha$, then we write $f(x) = o(g(x))$ as $x \to \infty$ [read $f(x)$ is little-oh of $g(x)$ as x tends to infinity]. If $g(x) \neq 0$ for all large $x \in \mathbb{R}$, then $f(x) = o(g(x))$ as $x \to \infty$ means that $\lim_{x\to\infty}(f(x)/g(x))$ exists and is zero. In particular, if $g(x) = 1$ for all large x, then $f(x) = o(1)$ as $x \to \infty$, and this means that $f(x) \to 0$ as $x \to \infty$. Broadly speaking, $f(x) = o(g(x))$ as $x \to \infty$ if the order of magnitude of f is less than the order of magnitude of g as $x \to \infty$. In case f and g are monotonically increasing functions and $f(x) = o(g(x))$ as $x \to \infty$, then we also say that the **growth rate** of f is less than the growth rate of g as $x \to \infty$. For example,

$$10\,x + 100 = o(x\sqrt{x}) \quad \text{and} \quad \frac{10}{[x]} + \frac{100}{x\sqrt{x}} = o\left(\frac{1}{\sqrt{x}}\right).$$

Suppose there is nonzero $\ell \in \mathbb{R}$ that satisfies the following condition: Given $\epsilon > 0$, there is $\alpha \in \mathbb{R}$ such that $|f(x) - \ell g(x)| < \epsilon$ for all $x \geq \alpha$. In this case, we write $f(x) \sim g(x)$ as $x \to \infty$ [read $f(x)$ is asymptotically equivalent to $g(x)$ as x tends to ∞]. Broadly speaking, $f(x) \sim g(x)$ as $x \to \infty$ if $f(x)$ is of the same order of magnitude as $g(x)$ as $x \to \infty$. It can be easily seen that $\sim$ is an equivalence relation on the set of all real-valued functions defined on D. If $g(x) \neq 0$ for all large x, then $f(x) \sim g(x)$ as $x \to \infty$ means that $\lim_{x\to\infty}(f(x)/g(x))$ exists and is nonzero. If f and g are monotonically increasing functions and $f(x) \sim g(x)$ as $x \to \infty$, then we also say that f and g have the same **growth rate** as $x \to \infty$. For example,

$$10\,x^2 + 100\,x + 1000 \sim x^2 \quad \text{and} \quad \frac{10}{x^2} + \frac{100}{x^3} + \frac{1000}{x^4} \sim \frac{1}{x^2}.$$

More interesting examples will be given in Section 7.1.

If $(-\infty, a)$ is contained in D for some $a \in \mathbb{R}$, then we may compare the orders of magnitude of f and g as $x \to -\infty$, or if $c \in \mathbb{R}$ is such that $(c - r, c + r) \subseteq D$ for some $r > 0$, then we may compare the orders of magnitude of f and g as $x \to c$ in a similar manner. $\diamond$

As we have described for sequences, we now describe how in some cases ∞ or $-\infty$ can be regarded as a 'limit' of a function of a real variable. Let $D \subseteq \mathbb{R}$

and $c \in \mathbb{R}$ be such that $(c - r, c)$ and $(c, c + r)$ are contained in D for some $r > 0$. We say that $f(x)$ tends to ∞ as x tends to c if

$$(x_n) \text{ any sequence in } D \text{ and } x_n \to c \Longrightarrow f(x_n) \to \infty.$$

We then write

$$f(x) \to \infty \text{ as } x \to c.$$

We give an analogue of Proposition 3.27 for a function that tends to infinity.

Proposition 3.32. *Let $D \subseteq \mathbb{R}$, $c \in \mathbb{R}$ be such that $(c - r, c)$ and $(c, c+r)$ are contained in D for some $r > 0$, and let $f : D \to \mathbb{R}$ be a function. Then $f(x)$ tends to ∞ as x tends to c if and only if the following α-δ condition holds: For every $\alpha \in \mathbb{R}$, there is $\delta > 0$ such that*

$$x \in D \text{ and } 0 < |x - c| < \delta \Longrightarrow f(x) > \alpha.$$

Proof. Let $f(x)$ tend to ∞ as x tends to c. Suppose for a moment that the α-δ condition does not hold. This means that there is $\alpha \in \mathbb{R}$ such that for every $\delta > 0$, there is $x \in D$ satisfying

$$0 < |x - c| < \delta, \text{ but } f(x) \leq \alpha.$$

Then there is a sequence (x_n) in $D \backslash \{c\}$ such that $|x_n - c| < 1/n$, but $f(x_n) \leq \alpha$ for all $n \in \mathbb{N}$. Now $x_n \to c$ and $f(x_n) \nrightarrow \infty$. This contradicts the assumption that $f(x) \to \infty$ as $x \to c$.

Conversely, assume the α-δ condition. Let (x_n) be any sequence in $D \backslash \{c\}$ such that $x_n \to c$. Let $\alpha \in \mathbb{R}$ be given. Then there is $\delta > 0$ such that

$$x \in D \text{ and } 0 < |x - c| < \delta \Longrightarrow f(x) > \alpha.$$

Since $x_n \to c$, there is $n_0 \in \mathbb{N}$ such that $|x_n - c| < \delta$ for all $n \geq n_0$. Hence $f(x_n) > \alpha$ for all $n \geq n_0$. Thus $f(x_n) \to \infty$. So $f(x) \to \infty$ as $x \to c$.   □

In a similar manner, we may define '$f(x) \to -\infty$ as $x \to c$' and an equivalent 'β-δ condition' can be formulated. Also, one-sided limits (as $x \to c^-$ or as $x \to c^+$) are similarly treated. Further, we may define

$$f(x) \to \infty \text{ as } x \to \infty \quad \text{and} \quad f(x) \to -\infty \text{ as } x \to \infty$$

as well as

$$f(x) \to \infty \text{ as } x \to -\infty \quad \text{and} \quad f(x) \to -\infty \text{ as } x \to -\infty$$

analogously.

If $f(x) \to \infty$ and $g(x) \to \ell$, where $\ell \in \mathbb{R}$ or $\ell = \infty$ or $\ell = -\infty$, then results regarding the existence of the 'limits' of $f(x) + g(x)$ and $f(x)g(x)$ can be stated on the lines of the results stated in Remark 2.12.

Examples 3.33. (i) We have

$$\lim_{x \to \infty} \frac{1}{x} = 0 = \lim_{x \to -\infty} \frac{1}{x},$$

while

$$\frac{1}{x} \to \infty \ \text{ as } \ x \to 0^+ \quad \text{and} \quad \frac{1}{x} \to -\infty \ \text{ as } \ x \to 0^-.$$

(ii) We have $\displaystyle \lim_{x \to \infty} \frac{x^2 + 2x + 3}{4x^2 + 5x + 6} = \frac{1}{4}$, since

$$\frac{x^2 + 2x + 3}{4x^2 + 5x + 6} = \frac{1 + \dfrac{2}{x} + \dfrac{3}{x^2}}{4 + \dfrac{5}{x} + \dfrac{6}{x^2}} \qquad \text{for all } x \in \mathbb{R}, \ x \ne 0.$$

(iii) We have $x^3 \to \infty$ as $x \to \infty$ and $x^3 \to -\infty$ as $x \to -\infty$.

The concept of a limit involving ∞ or $-\infty$ is useful in considering 'asymptotes of curves'. Roughly speaking, a straight line is considered to be an **asymptote** of a curve if it comes arbitrarily close to that curve. A classification of the asymptotes, depending on their slopes, is given below.

Let $D \subseteq \mathbb{R}$ and $f : D \to \mathbb{R}$ be a function. We tacitly assume that D satisfies an appropriate condition needed for defining the relevant limit.

1. A straight line given by $y = b$, where $b \in \mathbb{R}$, is called a **horizontal asymptote** of the curve $y = f(x)$ if

$$\lim_{x \to \infty} (f(x) - b) = 0 \quad \text{or} \quad \lim_{x \to -\infty} (f(x) - b) = 0.$$

2. A straight line given by $y = ax + b$, where $a, b \in \mathbb{R}$ and $a \ne 0$, is called an **oblique asymptote** of the curve $y = f(x)$ if

$$\lim_{x \to \infty} (f(x) - ax - b) = 0 \quad \text{or} \quad \lim_{x \to -\infty} (f(x) - ax - b) = 0.$$

3. A straight line given by $x = c$, where $c \in \mathbb{R}$, is called a **vertical asymptote** of the curve $y = f(x)$ if one or more of the following holds:

$$f(x) \to \infty \text{ as } x \to c^-, \quad f(x) \to -\infty \text{ as } x \to c^-,$$
$$f(x) \to \infty \text{ as } x \to c^+, \quad f(x) \to -\infty \text{ as } x \to c^+.$$

Examples 3.34. (i) Let $D := (-\infty, 0) \cup (1, \infty)$ and define $f : D \to \mathbb{R}$ as follows:

$$f(x) = \begin{cases} \dfrac{2x - 1}{x - 1} & \text{if } x > 1, \\[3mm] \dfrac{3x^2 + 4x + 1}{x} & \text{if } x < 0. \end{cases}$$

For $x > 1$, we have $f(x) = 2 + [1/(x-1)]$, and so $\lim_{x \to \infty}(f(x) - 2) = 0$. Hence the straight line given by $y = 2$ is a horizontal asymptote of the curve $y = f(x)$. Also, $f(x) \to \infty$ as $x \to 1^+$. Hence the straight line given by $x = 1$ is a vertical asymptote of the curve $y = f(x)$.

For $x < 0$, we have $f(x) = 3x + 4 + (1/x)$, and so $\lim_{x \to -\infty}(f(x) - 3x - 4) = 0$. Hence the straight line given by $y = 3x + 4$ is an oblique asymptote of the curve $y = f(x)$. Also, $f(x) \to -\infty$ as $x \to 0^-$. Hence the straight line given by $x = 0$ is a vertical asymptote of the curve $y = f(x)$.

We can use this information to draw a graph of f as in Figure 3.6.
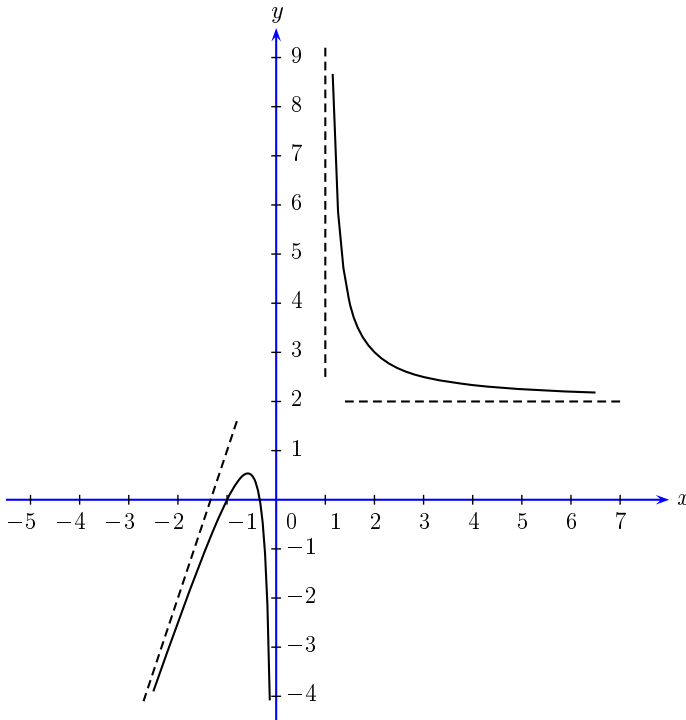
Fig. 3.6. Graph of $f(x) = \begin{cases} (2x-1)/(x-1) & \text{if } x > 1, \\ (3x^2 + 4x + 1)/x & \text{if } x < 0, \end{cases}$ with its horizontal, oblique, and vertical asymptotes

(ii) Let P and Q be nonzero polynomial functions that do not have a common real root. Consider $D := \mathbb{R} \setminus E$, where E is the set of all real roots of the polynomial function Q. Define $f(x) = P(x)/Q(x)$ for $x \in D$.

If the degree of P is equal to the degree of Q, then

$$f(x) = b + \frac{R(x)}{Q(x)} \quad \text{for } x \in D,$$

where $b \in \mathbb{R}$ and R is a polynomial function whose degree is less than the degree of Q. Since $R(x)/Q(x) \to 0$ as $x \to \infty$ and also as $x \to -\infty$, we see that the straight line given by $y = b$ is a horizontal asymptote of the curve $y = f(x)$.

If the degree of P is greater than the degree of Q by 1, then

$$f(x) = ax + b + \frac{R(x)}{Q(x)} \quad \text{for } x \in D,$$

where $a, b \in \mathbb{R}$, $a \neq 0$ and R is a polynomial function whose degree is less than the degree of Q. Again, since $R(x)/Q(x) \to 0$ as $x \to \infty$ and also as $x \to -\infty$, we see that the straight line given by $y = ax + b$ is an oblique asymptote of the curve $y = f(x)$.

Let now $c \in E$, that is, $c \in \mathbb{R}$ and $Q(c) = 0$. Then it is easy to see that $f(x) \to \infty$ or $f(x) \to -\infty$ (depending on the signs of the leading coefficients of the polynomial functions P and Q) as $x \to c^-$ and also $x \to c^+$. Hence the straight line given by $x = c$ is a vertical asymptote of the curve $y = f(x)$.

We now consider limits of monotonic functions. The results given below may be compared with the corresponding results for limits of monotonic sequences (Proposition 2.8 and Remark 2.12).

Proposition 3.35. *Let $f : (a, b) \to \mathbb{R}$ be a monotonically increasing function. Then*

(i) $\lim_{x \to b^-} f(x)$ *exists if and only if f is bounded above; in this case, we have*

$$\lim_{x \to b^-} f(x) = \sup\{f(x) : x \in (a, b)\}.$$

If f is not bounded above, then $f(x) \to \infty$ as $x \to b^-$.

(ii) $\lim_{x \to a^+} f(x)$ *exists if and only if f is bounded below; in this case, we have*

$$\lim_{x \to a^+} f(x) = \inf\{f(x) : x \in (a, b)\}.$$

If f is not bounded below, then $f(x) \to -\infty$ as $x \to a^+$.

Here $a \in \mathbb{R}$ or $a = -\infty$, and $b \in \mathbb{R}$ or $b = \infty$.

Proof. Consider a sequence (b_n) in (a, b) such that $b_n \to b$.

(i) Assume that f is bounded above and let $M := \sup\{f(x) : x \in (a, b)\}$. Given $\epsilon > 0$, there is $c \in (a, b)$ such that $M - \epsilon < f(c)$. Now since f is monotonically increasing, we have $M - \epsilon < f(x)$ for all $x \in (c, b)$. Also, since $b_n \to b$ and $c < b$, there is $n_0 \in \mathbb{N}$ such that $c < b_n$ for all $n \geq n_0$. Hence $M - \epsilon < f(b_n)$ for all $n \geq n_0$. On the other hand, $f(b_n) \leq M$ for all $n \in \mathbb{N}$. Hence $f(b_n) \to M$ as $b_n \to b$. Since (b_n) is an arbitrary sequence such that $b_n \to b$, we obtain $f(x) \to M$ as $x \to b$.

Assume now that f is not bounded above. Let $\alpha \in \mathbb{R}$. Then there is $c \in (a, b)$ such that $\alpha < f(c)$. Again, since f is monotonically increasing,

$\alpha < f(x)$ for all $x \in (c, b)$. Since $b_n \to b$ and $c < b$, there is $n_0 \in \mathbb{N}$ such that $c < b_n$ for all $n \geq n_0$. Hence $\alpha < f(b_n)$ for all $n \geq n_0$. This shows that $f(b_n) \to \infty$ as $b_n \to b$. Since (b_n) is an arbitrary sequence such that $b_n \to b$, we obtain $f(x) \to \infty$ as $x \to b$.

(ii) The proof of this part is similar to the proof of part (i) above. $\square$

A result similar to the one above holds for a monotonically decreasing function. (See Exercise 32.)

We shall now use Proposition 3.35 to prove the converse of the Intermediate Value Theorem (Proposition 3.13) for monotonic functions.

Proposition 3.36. *Let I be an interval and $f : I \to \mathbb{R}$ be a function that is monotonic on I. If f has the IVP on I, then f is continuous on I.*

Proof. Assume first that f is monotonically increasing on I and f has the IVP on I. Consider $c \in I$.

Suppose that c is neither the left (hand) endpoint nor the right (hand) endpoint of the interval I. Then there are $c_1, c_2 \in I$ such that $c_1 < c < c_2$. Now the function f is bounded above on the interval (c_1, c) and it is bounded below on the interval (c, c_2) by $f(c)$. Hence by Proposition 3.35, $f(x) \to \ell_1$ as $x \to c^-$ and $f(x) \to \ell_2$ as $x- \to c^+$, where

$$\ell_1 := \sup\{f(x) : c_1 < x < c\} \quad \text{and} \quad \ell_2 := \inf\{f(x) : c < x < c_2\}.$$

Clearly, $\ell_1 \leq f(c) \leq \ell_2$. In fact, since f has the IVP on I, we have $\ell_1 = f(c) = \ell_2$. Thus by Proposition 3.29, f is continuous at c.

If c is the left (hand) endpoint of the interval I, then the above argument shows that $f(x) \to \ell_2$ as $x \to c^+$ and $\ell_2 = f(c)$, while if c is the right (hand) endpoint of the interval I, then we have $f(x) \to \ell_1$ as $x \to c^-$ and $\ell_1 = f(c)$.

Thus in all cases, f is continuous at c. Since c is an arbitrary point of I, we see that f is continuous on I.

If f is monotonically decreasing on I and f has the IVP on I, then $-f$ is monotonically increasing on I and $-f$ has the IVP on I. Hence by what we have proved above, $-f$ is continuous on I, that is, f is continuous on I. $\square$

We conclude this section by stating that Proposition 3.29 about left limits and right limits can be used to prove that if a function is convex or concave on an interval, then it is continuous at every point of that interval other than its endpoints. (See Exercise 47.)

Notes and Comments

Most books on calculus and analysis treat limits of functions of a real variable first and then discuss the continuity of such a function. We follow, however, the reverse order. Our definition of continuity of a function relies on the

concept of limit of a sequence, which is introduced in Chapter 2; the domain of such a function can be an arbitrary subset of $\mathbb{R}$*. Such an approach is unusual but not new. See, for example, the book by Goffman [27]. Our definition of the limit of a function at a point* $c \in \mathbb{R}$ *also uses the concept of a limit of a sequence; the assumption that an open interval about* c*, except possibly* c *itself, is contained in the domain of the function ensures the uniqueness of the limit, whenever it exists. Our discussion of the relationship between continuity and various geometric properties of a function is based on a remarkable result, which states that the inverse of a strictly monotonic function defined on an interval is always continuous.*

Our approach of utilizing the limits of sequences to introduce continuity and limits of functions of a real variable seems to be simple-minded and easier to understand than the standard approach, which uses the ϵ*-*δ *condition. We have shown the equivalence of these two approaches toward the end of our discussion of continuity and of limits.*

It is possible to define the limit of a function at a point c *under a less-restrictive assumption on the domain* D *of the function than what we have imposed. In fact, it is sufficient to assume that for every* $r > 0$*, there is* $x \in D$ *such that* $0 < |x - c| < r$*. See Exercises 43–45. The assumption imposed by us in the text is merely for the sake of simplicity.*

Exercises

Part A

1. State whether there is a function $f : [0, 3] \to \mathbb{R}$ that is continuous at 2 and satisfies
   (i) $f(x) = (x^3 - 3x - 2)/(x - 2)$ for $x \neq 2$,
   (ii) $f(x) = \begin{cases} x & \text{if } x \in [1, 2), \\ x/2 & \text{if } x \in (2, 3]. \end{cases}$
2. State whether there is a continuous function $f : [0, 1] \to \mathbb{R}$ such that for every $n \in \mathbb{N}$,
   (i) $f((2n - 1)/n) = (-1)^n$, (ii) $f((2n + 1)/n) = 2^{1/n}$.
3. Let k be an odd positive integer and $f(x) = \sqrt[k]{x}$ for $x \in \mathbb{R}$. Show that f is continuous on $\mathbb{R}$.
4. Let $f : \mathbb{R} \to \mathbb{R}$ satisfy $f(x + y) = f(x) + f(y)$ for all x, $y \in \mathbb{R}$. If f is continuous at 0, then show that (i) f is continuous at every $c \in \mathbb{R}$ and (ii) $f(sx) = sf(x)$ for all s, $x \in \mathbb{R}$. Deduce that there exists $r \in \mathbb{R}$ such that $f(x) = rx$ for all $x \in \mathbb{R}$.
5. Let $f : \mathbb{R} \to \mathbb{R}$ satisfy $f(x + y) = f(x)f(y)$ for all x, $y \in \mathbb{R}$. If f is continuous at 0, then show that f is continuous at every $c \in \mathbb{R}$.
   [Note: An important example of such a function, known as the exponential function, will be given in Section 7.1.]

6. Let $f : \mathbb{R} \to \mathbb{R}$ satisfy $f(xy) = f(x)f(y)$ for all $x, y \in \mathbb{R}$. If f is continuous at 1, then show that f is continuous at every $c \in \mathbb{R}$, except possibly at $c = 0$. Give an example of such a function that is continuous at 1 as well as at 0. Also, give an example of such a function that is continuous at 1, but not at 0. (Compare Exercise 19 of Chapter 1.)

7. Let $f : (0, \infty) \to \mathbb{R}$ satisfy $f(xy) = f(x) + f(y)$ for all $x, y \in (0, \infty)$. If f is continuous at 1, then show that f is continuous at every $c \in (0, \infty)$. [Note: An important example of such a function, known as the logarithmic function, will be given in Section 7.1.]

8. Let $f : \mathbb{R} \to \mathbb{R}$ be given by

$$f(x) = \begin{cases} ax & \text{if } x \leq 0, \\ \sqrt{x} & \text{if } x > 0, \end{cases}$$

where $a \in \mathbb{R}$. Show that f is continuous on $\mathbb{R}$.

9. Let $f : \mathbb{R} \to \mathbb{R}$ be given by

$$f(x) = \begin{cases} x & \text{if } x \text{ is rational,} \\ 1 - x & \text{if } x \text{ is irrational.} \end{cases}$$

Show that f is continuous only at $\frac{1}{2}$.

10. Let $D := \{1/n : n \in \mathbb{N}\} \cup \{0\}$ and $f : D \to \mathbb{R}$ be any function. Show that f is continuous at $1/n$ for every $n \in \mathbb{N}$, and f is continuous at 0 if and only if $f(1/n) \to f(0)$.

11. Let $f : [0, 2] \to \mathbb{R}$ be given by

$$f(x) = \begin{cases} x & \text{if } 0 \leq x < 1, \\ 3 - x & \text{if } 1 \leq x \leq 2. \end{cases}$$

Show that f assumes every value between 0 and 2 exactly once on $[0, 2]$, but f is not continuous on $[0, 2]$.

12. Let $f : [0, 1] \to \mathbb{R}$ be given by

$$f(x) = \begin{cases} 3x/2 & \text{if } 0 \leq x < \frac{1}{2}, \\ (3x - 1)/2 & \text{if } \frac{1}{2} \leq x \leq 1. \end{cases}$$

Show that $f([0, 1]) = [0, 1]$. Is f continuous on $[0, 1]$? Does f have the IVP on $[0, 1]$?

13. Show that the cubic $x^3 - 6x + 3$ has exactly three real roots. (Hint: Find $f(-3)$, $f(0)$, $f(1)$, and $f(2)$, and use the IVP.)

14. Let $D \subseteq \mathbb{R}$, $c \in D$, and $f : D \to \mathbb{R}$ be such that f is continuous at c. Show that $|f| : D \to \mathbb{R}$ is continuous at c. Is the converse true?

15. Let $D \subseteq \mathbb{R}$, $c \in D$, and $f, g : D \to \mathbb{R}$ be such that f and g are continuous at c. Show that the functions $\max(f, g), \min(f, g) : [a, b] \to \mathbb{R}$ given by

$$\max(f, g)(x) = \max\{f(x), g(x)\} \quad \text{and} \quad \min(f, g)(x) = \min\{f(x), g(x)\}$$

for $x \in [a, b]$ are continuous at c. (Hint: $\max(f, g) = (f + g + |f - g|)/2$ and $\min(f, g) = (f + g - |f - g|)/2$.)

16. Let $f : [a, b] \to \mathbb{R}$ be continuous. Use the IVP to show that for any $c_1, \ldots, c_n \in [a, b]$, there is $c \in [a, b]$ such that

$$f(c) = \frac{f(c_1) + \cdots + f(c_n)}{n}.$$

17. If a function f satisfies one of the following conditions, then can it be continuous? Why?
    (i) $f : [1, 10] \to \mathbb{R}$, $f(1) = 0$, $f(10) = 11$, range of $f \subseteq [-1, 0] \cup [1, 11]$.
    (ii) $f : [0, 1] \to \mathbb{R}$ and range of $f = (-1, 1)$.
    (iii) $f : [-1, 1] \to \mathbb{R}$ and range of $f = [0, \infty)$.
18. Let $f : \mathbb{R} \to \mathbb{R}$ be given by

$$f(x) = \begin{cases} x/(1 + x) & \text{if } x \geq 0, \\ x/(1 - x) & \text{if } x < 0. \end{cases}$$

    Show that f is continuous and bounded on $\mathbb{R}$. Also, prove that

$$\inf\{f(x) : x \in \mathbb{R}\} = -1 \quad \text{and} \quad \sup\{f(x) : x \in \mathbb{R}\} = 1,$$

    but there are no r, s in $\mathbb{R}$ such that $f(r) = -1$ and $f(s) = 1$.
19. Let D and E be subsets of $\mathbb{R}$ such that D is closed and bounded. If $f : D \to E$ is bijective and continuous, then show that $f^{-1} : E \to D$ is continuous. (Hint: Proposition 2.17.) In particular, this result holds if $D = [a, b]$. (Compare Proposition 3.14.)
20. Analyze the following functions for uniform continuity:
    (i) $f(x) = x$, $x \in \mathbb{R}$,   (ii) $f(x) = 1/x$, $x \in (0, 1]$,
    (iii) $f(x) = x^2$, $x \in (0, 1)$,   (iv) $f(x) = \sqrt{1 - x^2}$, $x \in [-1, 1]$.
21. Let D and E be subsets of $\mathbb{R}$, and let $f : D \to \mathbb{R}$ and $g : E \to \mathbb{R}$ be functions such that the range of f is contained in E. If f is uniformly continuous on D and g is uniformly continuous on E, then show that $g \circ f$ is uniformly continuous on D.
22. Let $f : \mathbb{R} \to \mathbb{R}$ be given by

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

    Then $\lim_{x \to 0} f(x) = 0$, but there is a sequence (x_n) such that $x_n \to 0$ and $f(x_n) \not\to 0$. Explain.
23. Show that $\lim_{x \to c} f(x)$ does not exist if $f : \mathbb{R} \to \mathbb{R}$ is given by
    (i) $f(x) = [x] - x$, $c = 1$,   (ii) $f(x) = \dfrac{|x + 1|}{x + 1}$, $c = -1$.
24. Consider f, $g : \mathbb{R} \to \mathbb{R}$ and $c \in \mathbb{R}$. Under which of the following conditions does $\lim_{x \to c} f(x)g(x)$ exist? Justify.
    (i) $\lim_{x \to c} f(x)$ exists.
    (ii) $\lim_{x \to c} f(x)$ exists and g is bounded on $\{x \in \mathbb{R} : 0 < |x - c| < \delta\}$ for some $\delta > 0$.

(iii) $\lim_{x \to c} f(x) = 0$ and g is bounded on $\{x \in \mathbb{R} : 0 < |x - c| < \delta\}$ for some $\delta > 0$.

(iv) $\lim_{x \to c} f(x)$ and $\lim_{x \to c} g(x)$ exist.

25. Prove that the following limits exist.

(i) $\lim_{x \to 0} x[x]$, (ii) $\lim_{x \to 0} \dfrac{\sqrt{1 + x} - 1}{x}$, (iii) $\lim_{x \to \infty} \dfrac{7x - 1}{x^2}$, (iv) $\lim_{x \to \infty} \dfrac{x^4 + x}{x^4 + 1}$,

(v) $\lim_{x \to 0^+} \dfrac{\sqrt{x}}{\sqrt{7 + \sqrt{x + 5}}}$, (vi) $\lim_{x \to 1} \dfrac{|x - 1| + 1}{x + |x + 1|}$, (iv) $\lim_{x \to 3} ([x] - [2x - 1])$.

26. Show that $f(x) \to \infty$ as $x \to \infty$, if $f : [0, \infty) \to \mathbb{R}$ is given by

(i) $f(x) = \dfrac{3x^2 + 1}{2x + 1}$, (ii) $f(x) = [x]$.

27. Let f and g be polynomial functions given by

$$f(x) = a_n x^n + \cdots + a_1 x + a_0 \ \ \text{and} \ \ g(x) = b_m x^m + \cdots + b_1 x + b_0,$$

where $a_n, \ldots, a_0, b_m, \ldots, b_0$ are in $\mathbb{R}$, $a_n \neq 0$ and $b_m \neq 0$. Show that

$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = \begin{cases} 0 & \text{if } m > n, \\ a_m/b_m & \text{if } m = n. \end{cases}$$

In case $m < n$, show that

$$\frac{f(x)}{g(x)} \to \infty \text{ as } x \to \infty \text{ if } \frac{a_n}{b_m} > 0 \,, \ \text{ and } \ \frac{f(x)}{g(x)} \to -\infty \text{ as } x \to \infty \text{ if } \frac{a_n}{b_m} < 0.$$

28. Let $f : \mathbb{R} \to \mathbb{R}$ and $c \in \mathbb{R}$. If $\lim_{x \to c} f(x)$ exists, then show that

$$\lim_{h \to 0^+} [f(c + h) - f(c - h)] = 0.$$

Is the converse true? Justify your answer.

29. **(Limit of Composition)** Let $D \subseteq \mathbb{R}$, $s_0 \in \mathbb{R}$ be such that $(s_0 - r, s_0)$ and $(s_0, s_0 + r)$ are contained in D for some $r > 0$, and let $u : D \to \mathbb{R}$ be a function such that $\lim_{s \to s_0} u(s)$ exists. Let $t_0 := \lim_{s \to s_0} u(s)$. Suppose $E \subseteq \mathbb{R}$ is such that $u(D \setminus \{s_0\}) \subseteq E$ and consider a function $v : E \to \mathbb{R}$. Assume either that $(t_0 - \delta, t_0) \cup (t_0, t_0 + \delta)$ is contained in E for some $\delta > 0$, $\lim_{t \to t_0} v(t)$ exists, and $u(s) \neq t_0$ for every $s \in D \setminus \{s_0\}$, or that $t_0 \in E$ and v is continuous at t_0. Then prove that $\lim_{s \to s_0} v \circ u(s) = v(t_0)$. Show also that the condition '$u(s) \neq t_0$ for every $s \in D \setminus \{s_0\}$' or the requirement of continuity of the function v at t_0 cannot be dropped from this result.

30. Given $\epsilon > 0$, find $\delta > 0$ such that $|f(x) - \ell| < \epsilon$ whenever $0 < |x - c| < \delta$, if

(i) $f(x) = x^2 + 1$, $c = 1$, $\ell = 2$, (ii) $f(x) = \dfrac{1}{x}$, $c \neq 0$, $\ell = \dfrac{1}{c}$,

(iii) $f(x) = \dfrac{3x^2 + 7x + 2}{2x + 4}$, $c = -2$, $\ell = -5/2$.

31. Find the asymptotes of the following curves:

(i) $y = \dfrac{x}{x+1}$, $x \neq -1$, (ii) $y = \dfrac{x}{x-1}$, $x \neq 1$, (iii) $y = \dfrac{x^2}{x^2-1}$, $x \neq \pm 1$,

(vi) $y = \dfrac{x^2}{x^2+1}$, (v) $y = \dfrac{x^2+1}{x}$, $x \neq 0$, (vi) $y = \dfrac{x^2+x-2}{x-2}$, $x \neq 2$.

32. Let $f : (a,b) \to \mathbb{R}$ be a monotonically decreasing function. Prove the following results. (Compare Proposition 3.35.)

   (i) $\lim_{x \to b^-} f(x)$ exists if and only if f is bounded below; in this case, we have
$$\lim_{x \to b^-} f(x) = \inf\{f(x) : x \in (a,b)\}.$$

   If f is not bounded below, then $f(x) \to -\infty$ as $x \to b^-$.

   (ii) $\lim_{x \to a^+} f(x)$ exists if and only if f is bounded above; in this case, we have
$$\lim_{x \to a^+} f(x) = \sup\{f(x) : x \in (a,b)\}.$$

   If f is not bounded above, then $f(x) \to \infty$ as $x \to a^+$.

Part B

33. Let $k \in \mathbb{N}$ and $f(x) = x^{1/k}$ for $x \in [0,\infty)$. If $\epsilon \in \mathbb{R}$ is such that $0 < \epsilon \leq 1$, define $\delta := \min\{(1+\epsilon)^n - 1, 1 - (1-\epsilon)^n\}$. Show that $\delta > 0$ and

$$x \in [0,\infty) \text{ and } |x - 1| < \delta \implies |f(x) - 1| < \epsilon.$$

   Also, show that δ is the greatest real number for which this holds.

34. Let $f : [0,\infty) \to \mathbb{R}$ be given by

$$f(x) = \begin{cases} 1 & \text{if } x = 0, \\ 1/q & \text{if } x = p/q \text{ where } p,\, q \in \mathbb{N} \text{ and } p, q \text{ have no common factor,} \\ 0 & \text{if } x \text{ is irrational.} \end{cases}$$

   Show that f is discontinuous at each rational in $[0,\infty)$ and it is continuous at each irrational in $[0,\infty)$.
   [Note: This function is known as **Thomae's function**.]

35. Let $f : [a,b] \to \mathbb{R}$ be a continuous function satisfying $f(a) = f(b)$. Show that there are $c, d \in [a,b]$ such that $d - c = (b-a)/2$ and $f(c) = f(d)$. Deduce that for every $\epsilon > 0$, there are $x, y \in [a,b]$ such that $0 < y - x < \epsilon$ and $f(x) = f(y)$.

36. Prove that a function $f : (a,b) \to \mathbb{R}$ is convex if and only if it is continuous on (a,b) and satisfies

$$f\left(\frac{x_1 + x_2}{2}\right) \leq \frac{f(x_1) + f(x_2)}{2} \quad \text{for all } x_1, x_2 \in (a,b).$$

(Hint: To prove convexity, first show that

$$f\left(\frac{x_1 + \cdots + x_n}{n}\right) \le \frac{f(x_1) + \cdots + f(x_n)}{n} \quad \text{for all } x_1, \ldots, x_n \in (a, b)$$

by observing that it holds if $n = 2^k$, where $k \in \mathbb{N}$, and it holds for any $n \in \mathbb{N}$ whenever it holds for $n + 1$; then show that $f(\lambda x + (1 - \lambda)y) \le \lambda f(x) + (1 - \lambda)f(y)$ for all $x, y \in (a, b)$ and $\lambda \in \mathbb{Q}$ with $0 < \lambda < 1$, and finally use the continuity of f to complete the argument.)

37. Let $D \subseteq \mathbb{R}$ and $f : D \to \mathbb{R}$. Prove the following.
    (i) If D is bounded and f is uniformly continuous on D, then f is bounded on D. Is this true if f is merely continuous on D?
    (ii) Let (x_n) be a Cauchy sequence in D. If f is uniformly continuous on D, then $(f(x_n))$ is also a Cauchy sequence. Is this true if f is merely continuous on D?

38. Let $f : (a, b) \to \mathbb{R}$ be a continuous function. Show that f can be extended to a continuous function on $[a, b]$ if and only if f is uniformly continuous on (a, b). (Hint: Exercise 37 (ii) and Proposition 3.17.)

39. Suppose $f : D \to \mathbb{R}$ satisfies $|f(x) - f(y)| \le \alpha |x - y|^r$ for all x, y in D and some constants $\alpha \in \mathbb{R}$, $r \in \mathbb{Q}$, $r > 0$. Show that f is uniformly continuous on D.

40. Let $r \in \mathbb{Q}$ and $r \ge 0$. If $f : [0, \infty) \to \mathbb{R}$ is defined by $f(x) = x^r$, show that f is uniformly continuous if and only if $r \le 1$. (Hint: If $r \le 1$, then $|x^r - y^r| \le 2|x - y|^r$ for all $x, y \in [0, \infty)$ by Exercise 54 (ii) of Chapter 1. If $r > 1$, then consider $x_n := n$, $y_n := n + (1/n^{r-1})$ for $n \in \mathbb{N}$.)

41. Let $f, g : D \to \mathbb{R}$ be uniformly continuous on D. Are the functions $f + g$, fg, $1/f$ (provided $f(x) \ne 0$ for all $x \in D$) uniformly continuous on D? What if D is a bounded subset of $\mathbb{R}$? What if D is a closed subset of $\mathbb{R}$? What if $D = [a, b]$? Justify your answers.

42. Let $D \subseteq \mathbb{R}$ and $c \in \mathbb{R}$ be such that D contains $(c - r, c)$ and $(c, c + r)$ for some $r > 0$. Given any $f : D \to \mathbb{R}$, show that $\lim_{x \to c} f(x)$ exists if and only if the following conditions hold:
    (i) For any sequence (x_n) in $D \setminus \{c\}$ such that $x_n \to c$, the sequence $(f(x_n))$ is bounded.
    (ii) For any sequences (x_n) and (y_n) in $D \setminus \{c\}$ such that $x_n \to c$, $y_n \to c$, and moreover, both $(f(x_n))$ and $(f(y_n))$ are convergent, we have $\lim_{n \to \infty} f(x_n) = \lim_{n \to \infty} f(y_n)$.
    (Hint: Proposition 2.17.)

43. Let $D \subseteq \mathbb{R}$ and $c \in \mathbb{R}$. If for every $r > 0$, there is $x \in D$ such that $0 < |x - c| < r$, then c is called a **limit point** (or an **accumulation point**) of D.
    (i) Show that c is a limit point of D if and only if there is a sequence (x_n) in $D \setminus \{c\}$ such that $x_n \to c$.
    (ii) If c is a limit point of D, then show that for every $r > 0$, the set $\{x \in D : 0 < |x - c| < r\}$ is infinite.
    (iii) If D is a finite subset of $\mathbb{R}$, show that D has no limit point.

(iv) Find all limit points of D if $D := \mathbb{N}$, or $D := \{1/n : n \in \mathbb{N}\}$, or $D := \mathbb{Q}$, or $D := (a, b)$.

(v) Let (a_n) be a sequence in $\mathbb{R}$ and let $D := \{a_n : n \in \mathbb{N}\}$. If c is limit point of D, then show that c is a cluster point of (a_n). On the other hand, if $a_n := (-1)^n$ for all $n \in \mathbb{N}$, then 1 and -1 are cluster points of (a_n), but the set $D := \{a_n : n \in \mathbb{N}\} = \{1, -1\}$ has no limit point. (See Exercise 16 of Chapter 2 for the definition of a cluster point of a sequence.)

44. Let $D \subseteq \mathbb{R}$, $c \in \mathbb{R}$, and let $f : D \to \mathbb{R}$ be a function.

(i) If c is a limit point of D, then we say that a **limit** of f as x tends to c exists if there is a real number ℓ such that

$$(x_n) \text{ any sequence in } D \setminus \{c\} \text{ and } x_n \to c \Longrightarrow f(x_n) \to \ell.$$

Show that if a limit of f as x tends to c exists, then it is unique.

(ii) If c is not a limit point of D, then show that for any $\ell \in \mathbb{R}$, the condition

$$(x_n) \text{ any sequence in } D \setminus \{c\} \text{ and } x_n \to c \Longrightarrow f(x_n) \to \ell$$

holds vacuously.

45. Let $D \subseteq \mathbb{R}$ and $c \in \mathbb{R}$ be a limit point of D. Prove analogues of Propositions 3.21, 3.23, 3.24, 3.25, and 3.27.

46. Let $f : (a, b) \to \mathbb{R}$ be a monotonically increasing function. Show that for every $c \in (a, b)$, both $\lim_{x \to c+} f(x)$ and $\lim_{x \to c-} f(x)$ exist, and

$$\lim_{x \to c-} f(x) = \sup_{a < x < c} f(x) \leq f(c) \leq \inf_{c < x < b} f(x) = \lim_{x \to c+} f(x).$$

Also, if $d \in (a, b)$ and $c < d$, then show that

$$\lim_{x \to c+} f(x) = \lim_{x \to d-} f(x).$$

Further, show that similar results hold for a monotonically decreasing function.

47. Let I be an interval and $f : I \to \mathbb{R}$ be a function that is convex on I, or concave on I. Show that f is continuous at every point of I except possibly the endpoints of I. (Hint: Let c be an interior point of I and $c_1, c_2 \in I$ be such that $c_1 < c < c_2$. If f is convex on I, then

$$f(c_1) + \frac{f(c) - f(c_1)}{c - c_1}(x - c_1) \leq f(x) \leq f(c) + \frac{f(c_2) - f(c)}{c_2 - c}(x - c)$$

for all $x \in (c, c_2)$ and

$$f(c) + \frac{f(c) - f(c_2)}{c_2 - c}(c - x) \leq f(x) \leq f(c_1) + \frac{f(c) - f(c_1)}{c - c_1}(x - c_1)$$

for all $x \in (c_1, c)$.)

4

Differentiation

Differentiation is a process that associates to a real-valued function f another function f', called the derivative of f. This process is *local* in the sense that the value of f' at a point c depends only on the values of f in a small interval around c. The concept of differentiation originated from two classical problems:

1. The geometric problem of determining a tangent at a point to a curve in the plane.
2. The physical problem of determining the speed or the velocity of an object, such as a particle or a vehicle or a planet.

The notion of a derivative, which we shall study in this chapter, and the fact that it can often be computed effectively, turns out to be a key to solving the above two problems. Furthermore, the notion of a derivative has an enormous number of applications both within and outside mathematics. Some of these applications will be considered in Chapter 5.

In the first section below, we begin by describing in greater detail the second problem above. This leads to the definition of differentiability. The concept of differentiability of a function is intimately related to the continuity of an associated function, and this connection is made explicit by a lemma of Carathéodory. We first prove Carathéodory's Lemma and then use it to the fullest extent possible to derive a number of basic properties of differentiation. Next, in Section 4.2, we present results known as the Mean Value Theorem and Taylor's Theorem, which are extremely useful in calculus and analysis. In Section 4.3, we show that for differentiable functions, geometric properties of functions such as monotonicity, convexity, and concavity can be effectively determined by looking at their derivatives. Finally, in Section 4.4, we describe L'Hôpital's Rules, which show how differentiation can be used to compute certain limits.

4.1 The Derivative and Its Basic Properties

Suppose we are traveling in a car from one place to another. If at an instant t_1 we have covered a distance $s_1 = s(t_1)$ from the starting point and at another instant t_2 we have covered a distance $s_2 = s(t_2)$, then it is clear that the average speed for the journey between these two instants, is

$$\frac{\text{distance}}{\text{time}} = \frac{s(t_2) - s(t_1)}{t_2 - t_1}.$$

Now, what if we want to know the precise speed at a particular instant t_0? If the procedure above is followed blindly, then the answer would come out as zero, and that does not makes sense. So, a natural thing to do is to consider the average speed

$$\frac{s(t_0 + h) - s(t_0)}{(t_0 + h) - t_0} = \frac{s(t_0 + h) - s(t_0)}{h},$$

where h is rather small (but can be positive or negative), so that $t_0 + h$ varies over points close to t_0. It is conceivable that as h approaches 0, the quotient above approaches what the speed at t_0 should be. Also it is clear that the notion of limit, which was discussed in the previous chapter, would readily make the last statement precise. Thus, we simply set[1]

$$\text{the instantaneous speed at } t_0 = \lim_{h \to 0} \frac{s(t_0 + h) - s(t_0)}{h}.$$

The idea here can easily be extended from a 'distance function' s to an arbitrary real-valued function f. It is, however, desirable that to form quotients such as those above, near a point $x = c$, the function should at least be defined at points around c. We thus make the following definition.

Let D be a subset of $\mathbb{R}$. An element $c \in D$ is said to be an **interior point** of D if there is $r > 0$ such that $(c - r, c + r) \subseteq D$. A function $f : D \to \mathbb{R}$ is said to be **differentiable** at an interior point c of D if the limit

$$\lim_{h \to 0} \frac{f(c + h) - f(c)}{h}$$

exists. In this case, the value of the limit is denoted by $f'(c)$ and is called the **derivative** of f at c.

If $D \subseteq \mathbb{R}$ is such that every point of D is an interior point of D, then a function $f : D \to \mathbb{R}$ is said to be **differentiable on** D if f is differentiable at

---

[1] In this example, the distance function s is evidently increasing and thus the limit here would be nonnegative. For an arbitrary linear motion, the function s may not be increasing and thus the limit could also be negative. It is then customary to call it the **instantaneous velocity** rather than the instantaneous speed. In general, speed is given by the absolute value of the velocity.

every point of D. In case f is differentiable on D, we obtain a new function from D to $\mathbb{R}$ whose value at $c \in D$ is $f'(c)$. Quite naturally, this function is denoted by f' and is called the **derivative (function)** of f.

Some alternative notations for the derivative f' are

$$\frac{df}{dx}, \text{ or also } \frac{dy}{dx} \text{ when one writes } y = f(x).$$

Likewise, $f'(c)$ is sometimes denoted by

$$\left.\frac{df}{dx}\right|_{x=c}, \text{ or } \left.\frac{dy}{dx}\right|_{x=c}.$$

At times, physicists use the notation $\dot{f}$ instead of f'.

Examples 4.1. (i) If $f : \mathbb{R} \to \mathbb{R}$ is a constant function, then clearly $f'(c) = 0$ for every $c \in \mathbb{R}$. Thus, the derivative of a constant function is the zero function.

(ii) If $f : \mathbb{R} \to \mathbb{R}$ is the identity function given by $f(x) = x$, then

$$\frac{f(c+h) - f(c)}{h} = \frac{(c+h) - c}{h} = 1 \quad \text{for any } c \in \mathbb{R}.$$

It follows that f is differentiable on $\mathbb{R}$ and $f'(x) = 1$.

(iii) If $f : \mathbb{R} \to \mathbb{R}$ is the absolute value function given by $f(x) = |x|$, then

$$\frac{f(0+h) - f(0)}{h} = \frac{|h|}{h},$$

and from part (ii) of Example 3.20, we see that the limit of this quotient as $h \to 0$ does not exist. So f is not differentiable at $c = 0$. On the other hand, f is differentiable at each $c \in \mathbb{R}$, $c \neq 0$, and $f'(c)$ is 1 if $c > 0$ and -1 if $c < 0$.

(iv) If $f : (-1, 1) \to \mathbb{R}$ is defined by $f(x) = \sqrt{x^2 + x^3} = |x|\sqrt{x+1}$, then

$$\frac{f(0+h) - f(0)}{h} = \frac{|h|\sqrt{h+1}}{h}$$

and thus, as in the previous example, the limit of this quotient as $h \to 0$ does not exist. So f is not differentiable at $c = 0$.

(v) If $f : \mathbb{R} \to \mathbb{R}$ is defined by $f(x) = \sqrt[3]{x^2} = x^{2/3}$, then

$$\frac{f(0+h) - f(0)}{h} = \frac{1}{\sqrt[3]{h}}$$

and the limit of this quotient as $h \to 0$ clearly does not exist. So f is not differentiable at $c = 0$. $\diamond$
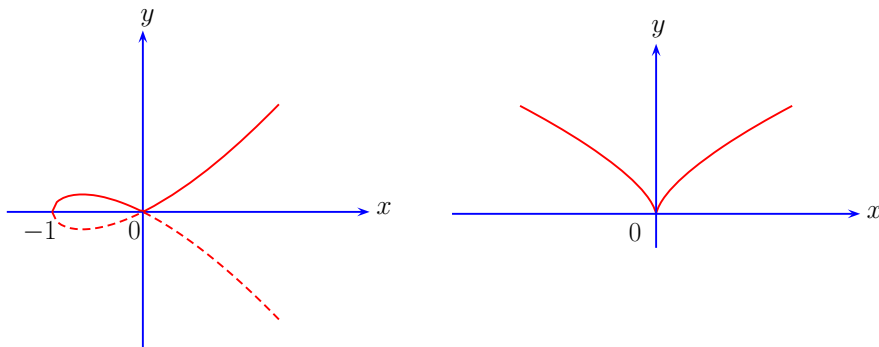
Fig. 4.1. Graphs of (iv) $y = \sqrt{x^2 + x^3}$ and (v) $y = \sqrt[3]{x^2}$

Now let us turn to a geometric interpretation of the notion of derivative and in particular, a 'solution' to the first problem stated at the beginning of this chapter. So let $D \subseteq \mathbb{R}$ be such that every point of D is an interior point of D and let $f : D \to \mathbb{R}$ be a function. Given any $c \in D$ and $h \neq 0$ such that $c + h \in D$, the quotient

$$\frac{f(c + h) - f(c)}{h}$$

gives the slope of the chord joining the points $(c, f(c))$ and $(c + h, f(c + h))$ on the curve $y = f(x)$, $x \in D$. As $h \to 0$, these chords seem to approach a 'tangent' to the curve $y = f(x)$ at the point $(c, f(c))$. It is therefore, reasonable to *define* the tangent to the curve $y = f(x)$ at the point $(c, f(c))$ to be the line given by the equation

$$y - f(c) = m(x - c), \quad \text{where} \quad m = f'(c) = \lim_{h \to 0} \frac{f(c + h) - f(c)}{h},$$

provided the limit above exists, that is, provided f is differentiable at c. Notice that the form of the equation for the tangent is such that a vertical line (such as the one given by $x = $ constant) can never be a tangent to a curve of the form $y = f(x)$. Further, the (geometric) condition that there is a unique nonvertical tangent to the curve $y = f(x)$ at a point $(c, f(c))$ is equivalent to the (analytic) condition that f is differentiable at c. Thus, intuitively speaking, to say that f is differentiable at c means that the graph of f is 'smooth' at $(c, f(c))$, that is to say, the graph has a unique nonvertical tangent at c. In this case the graph of f has no breaks or sharp edges or cusps at c. This is similar to the intuitive meaning of the continuity of f at c, namely, that the graph of f is unbroken at c. For an illustration of these remarks, take another look at Examples 4.1 (iii), (iv), and (v) as well as Figure 1.4 of Chapter 1 and Figure 4.1 above.

We shall now describe a number of basic properties of derivatives and to prove these, the following characterization of differentiability will be very useful.

Proposition 4.2 (Carathéodory's Lemma). *Let $D \subseteq \mathbb{R}$ and let c be an interior point of D. Then a function $f : D \to \mathbb{R}$ is differentiable at c if and only if there exists a function $f_1 : D \to \mathbb{R}$ such that $f(x) - f(c) = (x - c)f_1(x)$ for all $x \in D$, and f_1 is continuous at c. Moreover, if these conditions hold, then $f'(c) = f_1(c)$.*

Proof. If f is differentiable at c, then we can define $f_1 : D \to \mathbb{R}$ by

$$
f_1(x) := \begin{cases} \dfrac{f(x) - f(c)}{x - c} & \text{if } x \in D, \ x \neq c, \\[2mm] f'(c) & \text{if } x = c. \end{cases}
$$

It is clear that f_1 satisfies the required properties. Conversely, if there exists a function $f_1 : D \to \mathbb{R}$ with the given properties, then

$$
\lim_{h \to 0} \frac{f(c + h) - f(c)}{h} = \lim_{h \to 0} f_1(c + h) = f_1(c),
$$

where the first equality follows by putting $x = c + h$ in the relation $f(x) - f(c) = (x - c)f_1(c)$ and the second equality follows from Proposition 3.21 since f_1 is continuous at c. This proves that f is differentiable at c and $f'(c) = f_1(c)$. $\square$

Let $D \subseteq \mathbb{R}$ and c be an interior point of D. Given a function $f : D \to \mathbb{R}$, a function $f_1 : D \to \mathbb{R}$ satisfying

(i) $f(x) - f(c) = (x - c)f_1(x)$ for all $x \in D$, and (ii) f_1 is continuous at c

is called an **increment function** associated with f and c. It is clear that such an increment function, if it exists, is uniquely determined by f and c. Carathéodory's Lemma can be paraphrased by saying that differentiability of a function f at a point c is equivalent to the existence of an increment function associated with f and c, and in this case the derivative of f at c is the value of the increment function at c.

An immediate corollary of Carathéodory's Lemma is the following.

Proposition 4.3. *Let $D \subseteq \mathbb{R}$ and c be an interior point of D. If a function $f : D \to \mathbb{R}$ is differentiable at c, then f is continuous at c.*

Proof. Let f_1 be the increment function associated with f and c. Continuity of f_1 at c implies the continuity of f at c since $f(x) = f(c) + (x - c)f_1(c)$ for all $x \in D$. $\square$

Notice that the converse of the above Proposition is not true. In other words, continuity need not imply differentiability. For example, the absolute value function is continuous at 0, but it is not differentiable at 0.

Remark 4.4. Since differentiability implies continuity, all the properties of continuous functions such as those discussed in Section 3.2 are inherited by differentiable functions. On the other hand, it may be worthwhile to examine whether the conditions that imply continuity also imply differentiability. Often, this is not the case. For example, we have shown in Corollary 3.12 that if I is an interval and $f : I \to \mathbb{R}$ is a strictly monotonic function such that $f(I)$ is an interval, then f is continuous. However, such a function need not be differentiable. To see this, consider $I := [0, 2]$ and $f : I \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} x & \text{if } 0 \le x \le 1, \\ 3x - 2 & \text{if } 1 < x \le 2. \end{cases}$$

Then f is strictly increasing and $f(I) = [0, 4]$ is an interval but f is not differentiable at 1 [since $f'_-(1) = 1 \ne 3 = f'_+(1)$]. The same example shows that the hypothesis of Proposition 3.36, namely that I is an interval and $f : I \to \mathbb{R}$ is monotonic and has the IVP on I, does not imply the differentiability of f on I. As another example, we have indicated in Exercise 47 of Chapter 3 that if I is an interval and $f : I \to \mathbb{R}$ is convex on I, or concave on I, then f is continuous at every interior point of I. However, convexity or concavity does not imply differentiability. To see this, consider the absolute value function $f : [-1, 1] \to \mathbb{R}$ defined by $f(x) = |x|$. As seen earlier, f is convex on $[-1, 1]$ but fails to be differentiable at an interior point of $[-1, 1]$, namely, at 0. Similarly, $-f$ is concave on $[-1, 1]$ but fails to be differentiable at 0. $\diamond$

Proposition 4.5. *Let $D \subseteq \mathbb{R}$, c be an interior point of D, and $f, g : D \to \mathbb{R}$ be functions that are differentiable at c. Then*

  (i) *$f + g$ is differentiable at c and $(f + g)'(c) = f'(c) + g'(c)$,*
 (ii) *rf is differentiable at c and $(rf)'(c) = rf'(c)$ for any $r \in \mathbb{R}$,*
(iii) *fg is differentiable at c and $(fg)'(c) = f'(c)g(c) + f(c)g'(c)$,*
(iv) *if $f(c) \ne 0$, then there is $\delta > 0$ such that $(c - \delta, c + \delta) \subseteq D$ and $f(x) \ne 0$ for all $x \in (c - \delta, c + \delta)$; moreover, the function $1/f : (c - \delta, c + \delta) \to \mathbb{R}$ is differentiable at c, and*

$$\left(\frac{1}{f}\right)'(c) = -\frac{f'(c)}{f(c)^2},$$

 (v) *if $f(c) > 0$, then there is $\delta > 0$ such that $(c - \delta, c + \delta) \subseteq D$ and $f(x) > 0$ for all $x \in (c - \delta, c + \delta)$; moreover, for any $k \in \mathbb{N}$, the function $f^{1/k} : (c - \delta, c + \delta) \to \mathbb{R}$ is differentiable at c and*

$$\left(f^{1/k}\right)'(c) = \frac{1}{k}f(c)^{(1/k)-1}f'(c).$$

Proof. Let f_1 and g_1 denote, respectively, the increment functions associated with f and g and the point c. Using part (i) of Proposition 3.3, we easily see that $f_1 + g_1$ is the increment function associated with $f + g$ and c. Likewise,

using part (ii) of Proposition 3.3, we see that rf_1 is the increment function associated with rf and c for any $r \in \mathbb{R}$. This proves (i) and (ii).

Next, for any $x \in D$, the difference $f(x)g(x) - f(c)g(c)$ can be written as

$$[f(x) - f(c)]g(x) + f(c)[g(x) - g(c)] = (x - c)[f_1(x)g(x) + f(c)g_1(x)].$$

Moreover, by Proposition 4.3, g is continuous at c and thus by parts (i), (ii), and (iii) of Proposition 3.3, the function $f_1 g + f(c)g_1$ is continuous at c. This implies (iii).

Since c is an interior point of D, by Proposition 4.3 and part (iv) of Proposition 3.3, we see that if $f(c) \neq 0$, then there is $\delta > 0$ such that $(c - \delta, c + \delta) \subseteq D$ and $f(x) \neq 0$ for all $x \in (c - \delta, c + \delta)$, and moreover, the function $1/f : (c - \delta, c + \delta) \to \mathbb{R}$ is continuous at c. Thus, if $f(c) \neq 0$, we have

$$\frac{1}{f(x)} - \frac{1}{f(c)} = \frac{-[f(x) - f(c)]}{f(x)f(c)} = (x - c)\left[\frac{-f_1(x)}{f(x)f(c)}\right] \quad \text{for } x \in (c - \delta, c + \delta).$$

This yields (iv) since the function $-f_1/f(c)f$ is continuous at c.

Finally, suppose $k \in \mathbb{N}$ and $f(c) > 0$. Since c is an interior point of D, it follows from Lemma 3.2 that there is $\delta > 0$ such that $(c - \delta, c + \delta) \subseteq D$ and $f(x) > 0$ for all $x \in (c - \delta, c + \delta)$. For simplicity, write $F(x) := f(x)^{1/k}$ for $x \in (c - \delta, c + \delta)$. Then part (v) of Proposition 3.3 shows that $F : (c - \delta, c + \delta) \to \mathbb{R}$ is continuous at c, and for any $x \in (c - \delta, c + \delta)$, we have

$$f(x) - f(c) = [F(x) - F(c)]\left[F(x)^{k-1} + F(c)F(x)^{k-2} + \cdots + F(c)^{k-1}\right].$$

Now since $F(x) > 0$ for any $x \in (c - \delta, c + \delta)$, we obtain

$$F(x) - F(c) = (x - c)\left[\frac{f_1(x)}{F(x)^{k-1} + F(c)F(x)^{k-2} + \cdots + F(c)^{k-1}}\right].$$

This implies (v) since the function $f_1/\left(F^{k-1} + F(c)F^{k-2} + \cdots + F(c)^{k-1}\right)$ is continuous at c. $\square$

Remark 4.6. With notation and hypothesis as in the above proposition, a combined application of its parts (i) and (ii) shows that the difference $f - g$ is differentiable at c and $(f - g)'(c) = f'(c) - g'(c)$. Likewise, a combined application of parts (iii) and (iv) shows that if $g(c) \neq 0$, then the quotient f/g is differentiable at c and its derivative is given by the following **quotient rule**:

$$\left(\frac{f}{g}\right)'(c) = \frac{f'(c)g(c) - f(c)g'(c)}{g(c)^2}.$$

Further, given any $n \in \mathbb{N}$, successive applications of part (iii) of above proposition [or, if you prefer, induction on n] shows that the nth power f^n is differentiable at c and $(f^n)'(c) = nf(c)^{n-1}f'(c)$. Moreover, if $f(c) \neq 0$, then using the previous formula and part (iv), we see that

$$\left(\frac{1}{f^n}\right)'(c) = -\frac{nf(c)^{n-1}f'(c)}{f(c)^{2n}} = -nf(c)^{-n-1}f'(c).$$

Since the derivative of a constant function is zero, it follows that for any $m \in \mathbb{Z}$, f^m is differentiable at c and

$$(f^m)'(c) = mf(c)^{m-1}f'(c), \quad \text{provided } f(c) \neq 0 \text{ in case } m < 0.$$

Furthermore, given any $r \in \mathbb{Q}$, we can write $r = m/k$, where $m \in \mathbb{Z}$ and $k \in \mathbb{N}$, and then the last formula together with part (v) of the above proposition shows that if $f(c) > 0$, then the rth power $f^r = (f^m)^{1/k}$ is differentiable at c and

$$\left((f^m)^{1/k}\right)'(c) = \frac{1}{k}f^{m[(1/k)-1]}(c)\,(f^m)'(c) = \frac{m}{k}f(c)^{m[(1/k)-1]+m-1}f'(c).$$

In other words, for any $r \in \mathbb{Q}$, we have

$$(f^r)'(c) = rf(c)^{r-1}f'(c),$$

provided $f(c) \neq 0$ if r is a negative integer and $f(c) > 0$ if r is not an integer.
$\diamond$

Example 4.7. As a particular case of the results in Remark 4.6 and Examples 4.1 (i), (ii), we see that the nth power function is differentiable on $\mathbb{R}$ for each nonnegative integer n, and

$$\frac{d}{dx}(x^n) = nx^{n-1}.$$

Moreover, the above result is valid for negative integral powers, provided $x \neq 0$, and for rational nonintegral powers, provided $x > 0$. In particular,

$$\frac{d}{dx}(x^r) = rx^{r-1} \quad \text{for every } r \in \mathbb{Q} \text{ and } x \in (0, \infty).$$

$\diamond$

Example 4.8. Using Proposition 4.5 and Example 4.7, it follows that every polynomial function is differentiable on $\mathbb{R}$ and every rational function is differentiable at each point of $\mathbb{R}$ where it is defined. Moreover, the derivatives of such functions can also be readily computed. For instance, if $f : \mathbb{R} \setminus \{1\} \to \mathbb{R}$ is given by $f(x) = (x^4 + 3x + 2)/(x - 1)$, then we have

$$f'(x) = \frac{(4x^3 + 3)(x - 1) - (x^4 + 3x + 2)}{(x - 1)^2} = \frac{3x^4 - 4x^3 - 5}{(x - 1)^2}$$

for $x \neq 1$.
$\diamond$

To compute the derivative of a composite function $u = g(y)$ where y, in turn, is a function of x, say $y = f(x)$, the Chain Rule or Substitution Rule described below is quite useful. Roughly speaking, the Chain Rule can be stated as follows:

$$\frac{du}{dx} = \frac{du}{dy} \cdot \frac{dy}{dx}.$$

It may be tempting to prove this by just canceling out dy. But that wouldn't be correct because we haven't defined the quantities dy and dx by themselves even though we have defined $\frac{dy}{dx}$. We give below a precise statement of the Chain Rule and a proof using Carathéodory's Lemma.

Proposition 4.9 (Chain Rule). *Let $D, E \subseteq \mathbb{R}$ and $f : D \to \mathbb{R}$, $g : E \to \mathbb{R}$ be functions such that $f(D) \subseteq E$. Suppose c is an interior point of D such that $f(c)$ is an interior point of E. If f is differentiable at c and g differentiable at $f(c)$, then $g \circ f$ is differentiable at c and*

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

Proof. Let $f_1 : D \to \mathbb{R}$ be the increment functions associated with f and c. Then

$$f(x) - f(c) = (x - c)f_1(x) \quad \text{for all } x \in D.$$

Also, let $g_1 : E \to \mathbb{R}$ be the increment functions associated with g and $f(c)$. Then

$$g(y) - g(f(c)) = (y - f(c))g_1(y) \quad \text{for all } y \in E.$$

Since $f(D) \subseteq E$, we can use the above two equations to obtain

$$g(f(x)) - g(f(c)) = [f(x) - f(c)]g_1(f(x)) = (x - c)g_1(f(x))f_1(x) \quad \text{for } x \in D.$$

Now, using Propositions 3.3 and 3.4, we see that the function $(g_1 \circ f) \cdot f_1 : D \to \mathbb{R}$ is continuous at c. Hence by Carathéodory's Lemma, $g \circ f$ is differentiable at c and $(g \circ f)'(c) = g_1(f(c))f_1(c) = g'(f(c))f'(c)$. ☐

Example 4.10. Consider $F : \mathbb{R} \to \mathbb{R}$ defined by $F(x) = (4x^3 + 3)^7 + 2$. We can of course compute $F'(x)$ by expanding the seventh power and using Proposition 4.5 together with the formula for the derivative of the nth power function. But it is simpler to view F as the composite $g \circ f$, where $u = g(y) = y^7 + 2$ and $y = f(x) = 4x^3 + 2$, and apply the Chain Rule. This gives

$$F'(x) = g'(f(x))f'(x) = \left[7(4x^3 + 2)^6\right]\left(12x^2\right) = 84x^2 \left(4x^3 + 2\right)^6$$

for $x \in \mathbb{R}$. ◇

We shall now prove a general result about the derivatives of inverse functions. Roughly speaking, this result can be stated as follows.

$$\frac{dx}{dy} = \text{ is the reciprocal of } \frac{dy}{dx} \quad \text{when } \frac{dy}{dx} \text{ is nonzero.}$$

A precise statement and a proof using Carathéodory's Lemma appears below.

Proposition 4.11 (Differentiable Inverse Theorem). *Let I be an interval and c be an interior point of I. Suppose $f : I \to \mathbb{R}$ is a one-one and continuous function. Let $f^{-1} : f(I) \to I$ be the inverse function. Then $f(c)$ is an interior point of $f(I)$. Moreover, if f is differentiable at c and $f'(c) \neq 0$, then f^{-1} is differentiable at $f(c)$ and*

$$\left(f^{-1}\right)'(f(c)) = \frac{1}{f'(c)}.$$

Proof. Let $J := f(I)$. By Proposition 3.14, f is strictly monotonic and J is an interval. Hence $f(c)$ is an interior point of J. Suppose f is differentiable at c and $f'(c) \neq 0$. Let $f_1 : I \to \mathbb{R}$ be the increment function associated with f and c. Then f_1 is continuous at c with $f_1(c) = f'(c)$ and we have

$$f(x) - f(c) = (x - c)f_1(x) \quad \text{for all } x \in I.$$

Since f is one-one, for any $x \in I$ with $x \neq c$, we have $f(x) \neq f(c)$, and therefore, $f_1(x) \neq 0$. Also, $f_1(c) = f'(c) \neq 0$. Thus f_1 is never zero on I, and so $1/f_1$ is defined on I. Further, since f_1 is continuous at c, so is $1/f_1$. Hence the equation displayed above can be written as $(x - c) = [f(x) - f(c)]/f_1(x)$ for all $x \in I$. This implies that

$$f^{-1}(y) - f^{-1}(f(c)) = (y - f(c))\frac{1}{f_1\left(f^{-1}(y)\right)} \quad \text{for all } y \in J.$$

Moreover, by Propositions 3.5 and 3.14, we see that $f_1 \circ f^{-1}$ is continuous at $f(c)$. Hence by part (iv) of Proposition 3.3, the reciprocal of $f_1 \circ f^{-1}$ is also continuous at $f(c)$. Thus, by Carathéodory's Lemma, it follows that f^{-1} is differentiable at $f(c)$ and $\left(f^{-1}\right)'(f(c)) = 1/f_1\left(f^{-1}(f(c))\right) = 1/f'(c)$. $\square$

Differentiation applied successively leads to the notion of higher derivatives. More formally, suppose $D \subseteq \mathbb{R}$ and $f : D \to \mathbb{R}$ is a function that is differentiable at every point of an interval $(c-\delta, c+\delta) \subseteq D$. Then we have the derivative function f' defined on $(c-\delta, c+\delta)$. In case f' is differentiable at c, then we say that f is **twice differentiable** at c and denote the derivative of f' at c by $f''(c)$. The quantity $f''(c)$ is called the **second derivative** (or the **second-order derivative**) of f at c. Further, if f' is differentiable at every point of an interval about c, then the second derivative function f'' is defined on this interval. In case f'' is also differentiable at c, then we say that f is **thrice differentiable** at c and denote the derivative of f'' at c by $f'''(c)$. Similarly, one defines n-times differentiability of f and the nth derivative $f^{(n)}(c)$ for any $n \in \mathbb{N}$. The notations

$$\left.\frac{d^2 f}{dx^2}\right|_{x=c}, \quad \left.\frac{d^3 f}{dx^3}\right|_{x=c}, \quad \text{and} \quad \left.\frac{d^n f}{dx^n}\right|_{x=c}$$

are sometimes used instead of $f''(c)$, $f'''(c)$, and $f^{(n)}(c)$, respectively. In case f is n times differentiable at c for every $n \in \mathbb{N}$, then f is said to be **infinitely differentiable** at c.

To compute higher derivatives, the following formula, known as **Leibniz's Rule for derivatives**, can be quite useful:

$$(fg)^{(n)} = \sum_{k=0}^{n} \binom{n}{k} f^{(k)} g^{(n-k)} = f^{(n)}g + nf^{(n-1)}g' + \cdots + nf'g^{(n-1)} + fg^{(n)},$$

Here f, g are real-valued functions, both of which are assumed to be n times differentiable at c and $\binom{n}{k}$ denotes the binomial coefficient as defined in Exercise 2 of Chapter 1. Also, $f^{(0)}$ and $g^{(0)}$ denote, by convention, f and g respectively. Leibniz's Rule for derivatives can be easily proved by induction on n, using Proposition 4.5 and the Pascal triangle identity:

$$\binom{n}{k} + \binom{n}{k-1} = \binom{n+1}{k}.$$

Example 4.12. If r is any rational number and $f : (0, \infty) \to (0, \infty)$ is given by $f(x) = x^r$, then f is infinitely differentiable at every $c \in (0, \infty)$, and $f^{(n)}(c) = r(r-1) \cdots (r - n + 1)c^{r-n}$. $\diamond$

Let us now consider the case of a real-valued function defined on a closed interval. While we can talk about the differentiability of such a function at any interior point, the definition we have given of differentiability does not apply to the endpoints. To take care of this omission, we introduce the notions of left derivative and right derivative as follows.

Let $D \subseteq \mathbb{R}$ and $c \in D$ be such that $(c - r, c] \subseteq D$ for some $r > 0$. The left (hand) limit

$$\lim_{x \to c^-} \frac{f(x) - f(c)}{x - c},$$

if it exists, is called the **left (hand) derivative** of f at c and is denoted by $f'_-(c)$. In the case $[c, c+r) \subseteq D$ for some $r > 0$, the **right (hand) derivative** of f at c is defined similarly and is denoted by $f'_+(c)$.

If c happens to be an interior point, then it follows from Proposition 3.29 that f is differentiable at c if and only if both $f'_-(c)$ and $f'_+(c)$ exist and are equal.

For example, if $f : \mathbb{R} \to \mathbb{R}$ is the absolute value function, then $f'_-(0) = -1$, whereas $f'_+(0) = 1$. On the other hand, if $f : \mathbb{R} \to \mathbb{R}$ is defined by $f(x) = x^{2/3}$, then neither $f'_-(0)$ nor $f'_+(0)$ exists. In each of these examples, we find that the function f is not differentiable at 0.

We say that a real-valued function f defined on a closed interval $[a, b]$ is **differentiable** if f is differentiable at every point of (a, b), and moreover if $f'_+(a)$ and $f'_-(b)$ exist. In this case, the function $f' : [a, b] \to \mathbb{R}$ defined by

$$f'(x) = \begin{cases} f'_+(a) & \text{if } x = a, \\ f'(c) & \text{if } x \in (a, b), \\ f'_-(b) & \text{if } x = b, \end{cases}$$

is called the **derivative** of f on $[a, b]$. If f' is differentiable on $[a, b]$, then f is said to be **twice differentiable** on $[a, b]$, and we let f'' be the derivative of f' on $[a, b]$. More generally, the nth derivative $f^{(n)}$ of f on $[a, b]$ is defined for any $n \in \mathbb{N}$ in a similar way.

It should be noted that Carathéodory's Lemma continues to be valid for derivatives at the endpoints of a function $f : [a, b] \to \mathbb{R}$. The proof is identical to that of Proposition 4.2, provided we take limits as $h \to 0^+$ in case $c = a$ and as $h \to 0^-$ in case $c = b$. As a consequence, results similar to Propositions 4.3 and 4.5 as well as those in Remark 4.6 are valid for functions $f : D \to \mathbb{R}$ when $D = [a, b]$ and $c = a$ or $c = b$. Moreover, the Chain Rule (Proposition 4.9) is also valid if D is an interval and c is an endpoint of D such that $f(c)$ is an endpoint or an interior point of an interval contained in E. Likewise, the Differentiable Inverse Theorem (Proposition 4.11) is valid if c is an endpoint of I, provided in the conclusion we write "$f(c)$ is an endpoint of J" instead of "$f(c)$ is an interior point of J". The proof is essentially the same as before.

Tangents and Normals to Curves

We have discussed earlier the notion of tangent to plane curves of the form $y = f(x)$. We shall now see how it can be extended to plane curves of more general type. Also, we will discuss the related notion of normal in the context of more general curves.

Plane curves of the form $y = f(x)$ admit generalizations to two distinct, yet overlapping, classes of plane curves. These are as follows.

1. **Parametrically Defined Curves**: These are the plane curves C given by $(x(t), y(t))$, where x, y are real-valued functions[2] defined on some subset D of $\mathbb{R}$, and the parameter t varies over the points of D. Usually, we express this by simply saying that C is the (parametrically defined) curve $(x(t), y(t))$, $t \in D$. For example, the rectangular hyperbola is the curve $(t, 1/t)$, $t \in \mathbb{R} \setminus \{0\}$.

2. **Implicitly Defined Curves**: These are the plane curves C given by an equation of the form $F(x, y) = 0$, where F is a real-valued function defined on some subset E of the plane $\mathbb{R}^2$, and (x, y) vary over the points of E. Usually, we express this by simply saying that C is the (implicitly defined) curve $F(x, y) = 0$, $(x, y) \in E$. The reference to the domain E of F is skipped if $E = \mathbb{R}^2$. For example, the circle centered at the origin with unit radius is the curve $x^2 + y^2 - 1 = 0$.

Notice that if $D \subseteq \mathbb{R}$ and $f : D \to \mathbb{R}$ is any function, then the curve $y = f(x)$ can be viewed as a parametrically defined curve $(x(t), y(t))$, $t \in D$, where $x(t) := t$ and $y(t) := f(t)$. Also, it can be viewed as an implicitly defined

---

[2] Generally, one requires that the set D be an interval and the two functions $x, y : D \to \mathbb{R}$ be continuous. In most applications, this will be so but we do not make it a part of the definition.

curve $F(x, y) = 0$, $(x, y) \in E$, where $E = D \times \mathbb{R}$ and $F : E \to \mathbb{R}$ is given by $F(x, y) := y - f(x)$.

If C is a parametrically defined curve $(x(t), y(t))$, $t \in D$, and t_0 is an interior point of D such that both x and y are differentiable at t_0 and $x'(t_0), y'(t_0)$ are not both zero, then we define the **tangent** to C at the point $(x(t_0), y(t_0))$ to be the line

$$[y - y(t_0)]x'(t_0) - [x - x(t_0)]y'(t_0) = 0.$$

The line passing through $(x(t_0), y(t_0))$ and perpendicular to the tangent at this point, namely, the line given by

$$[x - x(t_0)]x'(t_0) + [y - y(t_0)]y'(t_0) = 0,$$

is called the **normal** to the curve C at the point $(x(t_0), y(t_0))$. In case $x'(t_0)$ or $y'(t_0)$ does not exist or $(x'(t_0), y'(t_0)) = (0, 0)$, we say that the tangent to C (as well as the normal to C) at $(x(t_0), y(t_0))$ is *not defined*. It may be noted that the definition of tangent to parametrically defined curves is consistent with the previous definition for tangent to curves of the form $y = f(x)$. More generally, if $x'(t_0) \neq 0$ and y can be considered a function of x in an open interval about $x_0 := x(t_0)$, then by the Chain Rule,

$$\left.\frac{dy}{dx}\right|_{x=x_0} = \left(\left.\frac{dy}{dt}\right|_{t=t_0}\right)\left(\left.\frac{dx}{dt}\right|_{t=t_0}\right)^{-1} = \frac{y'(t_0)}{x'(t_0)}.$$

The Chain Rule also helps us to formulate the notion of tangents to implicitly defined curves. For example, if $F(x, y) = x^2 + y^2 - 25$, then $F(x, y) = 0$ defines the circle of radius 5 centered at the origin. To find the tangent at a point, say $(3, 4)$, we differentiate $F(x, y)$ with respect to x, treating y as a function of x. Thus, using Chain Rule, we obtain

$$2x + 2y\frac{dy}{dx} = 0 \quad \text{and hence} \quad \left.\frac{dy}{dx}\right|_{(3,4)} = -\left.\frac{x}{y}\right|_{(3,4)} = -\frac{3}{4}.$$

This suggests that the tangent to this circle at the point $(3, 4)$ is given by the line $y - 4 = -\frac{3}{4}(x - 3)$, that is, $3x + 4y - 25 = 0$. In general, given any equation $F(x, y) = 0$, we can try to differentiate with respect to x, treating y as a function of x. This process is known as **implicit differentiation**, and it leads to an equation of the type

$$P(x, y) + Q(x, y)\frac{dy}{dx} = 0.$$

At a point $(x_0, y_0) \in \mathbb{R}^2$ on the curve $F(x, y) = 0$ (that is, $(x_0, y_0) \in \mathbb{R}^2$ satisfying $F(x_0, y_0) = 0$) with the additional property that $P(x_0, y_0)$ and $Q(x_0, y_0)$ are defined and $Q(x_0, y_0) \neq 0$, we *define* the tangent to $F(x, y) = 0$ at (x_0, y_0) to be the line

$$y - y_0 = m(x - x_0), \quad \text{where} \quad m = \left.\frac{dy}{dx}\right|_{(x_0, y_0)} = -\frac{P(x_0, y_0)}{Q(x_0, y_0)}.$$

It may be checked that this definition is consistent with our previous definition in the case $F(x, y) = y - f(x)$ for some function f which is differentiable at x_0; note that in this case

$$\left.\frac{dy}{dx}\right|_{(x_0, y_0)} = \left.\frac{dy}{dx}\right|_{x=x_0}.$$

Sometimes, when dealing with curves defined by $F(x, y) = 0$, it is useful to reverse the roles of x and y. Thus, we may also try to differentiate with respect to y, treating x as a function of y. This leads to an equation of the type

$$R(x, y) + S(x, y)\frac{dx}{dy} = 0.$$

Now suppose $(x_0, y_0) \in \mathbb{R}^2$ is a point on the curve $F(x, y) = 0$ such that $\frac{dy}{dx}$ is not defined at (x_0, y_0). [Roughly speaking, this corresponds to the case $Q(x_0, y_0) = 0$.] If, however, $R(x_0, y_0)$ and $S(x_0, y_0)$ are defined and $S(x_0, y_0) \neq 0$, then we *define* the tangent to $F(x, y) = 0$ at (x_0, y_0) to be the line

$$x - x_0 = \widetilde{m}(y - y_0) \quad \text{where} \quad \widetilde{m} = \left.\frac{dx}{dy}\right|_{(x_0, y_0)} = -\frac{R(x_0, y_0)}{S(x_0, y_0)}.$$

It may be noted that if both $\frac{dy}{dx}$ and $\frac{dx}{dy}$ are defined and are nonzero at a point (x_0, y_0) on the curve $F(x, y) = 0$, then it follows from the Differentiable Inverse Theorem (Proposition 4.11) that $\widetilde{m} = 1/m$, and hence the lines obtained by either of the two approaches are identical. If, however, both $\frac{dy}{dx}$ and $\frac{dx}{dy}$ are not defined, or if both of them are zero at (x_0, y_0), then we say that the tangent to the curve $F(x, y) = 0$ at the point (x_0, y_0) is *not defined*.[3]

As before, at a point (x_0, y_0) on a plane curve $F(x, y) = 0$ where the tangent is defined, we define the **normal** to be the unique line passing through (x_0, y_0) and perpendicular to the tangent to this curve at (x_0, y_0).

For example, for the circle $x^2 + y^2 - 25 = 0$, the tangent at the point $(5, 0)$ can be determined by the latter method of differentiating with respect to y, treating x as a function of y. Indeed, we see that the tangent is given by the vertical line $x - 5 = 0$. On the other hand, for the curve $y^2 - x^2 - x^3 = 0$, neither $\frac{dy}{dx}$ nor $\frac{dx}{dy}$ are defined at the origin, and hence the tangent at the origin is not defined.

---

[3] For algebraic plane curves $F(x, y) = 0$, where $F(x, y)$ is a polynomial in two variables, there is an alternative, purely algebraic, method to determine tangents at a point. In fact, this algebraic approach could be used to *define* tangents at every point on the curve including those points at which the tangent is not defined as far as calculus is concerned. We stick to the method of calculus in this book but briefly outline the algebraic approach in Exercise 43, and refer to the book of Abhyankar [1] for more details.

4.2 The Mean Value and Taylor Theorems

The Mean Value Theorem or, for short, the MVT, is a result that in geometric terms may be described as follows. For any nice curve of the form $y = f(x)$, $x \in [a, b]$, there exists a point $(c, f(c))$ on the curve, where $c \in (a, b)$, at which the tangent is parallel to the line joining the endpoints $(a, f(a))$ and $(b, f(b))$ of the curve. [See Figure 4.2.] Here, by 'nice' we mean that the curve is unbroken, that is, f is continuous, and that tangents can be drawn everywhere except perhaps at the endpoints, that is, f is differentiable on (a, b).
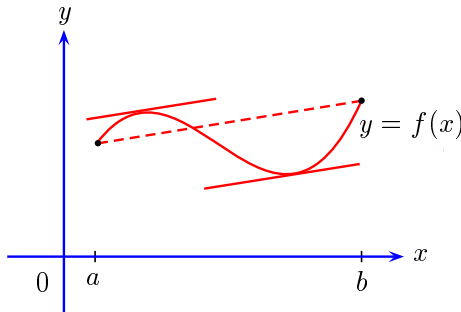

Fig. 4.2. Illustration of the MVT: Tangents at some intermediate points are parallel to the line joining the endpoints $(a, f(a))$ and $(b, f(b))$

As we shall see in the sequel, the MVT is a very useful result in calculus. In particular, an extension of the MVT, known as Taylor's Theorem, will allow us to approximate a large class of functions by polynomial functions.

A special case of the MVT is that in which the end points of the curve $y = f(x)$, $x \in [a, b]$, lie on a horizontal line, that is, $f(a) = f(b)$. In this case, the MVT amounts to asserting the existence of a point $(c, f(c))$ on the curve, where $c \in (a, b)$, at which the tangent is parallel to the x-axis, that is, $f'(c) = 0$. We will, in fact, prove this special case first, and deduce the MVT as a consequence. This special case, known as Rolle's Theorem, will, in turn, be deduced from the following simple but useful fact about points of local extrema.[4]

Lemma 4.13. *Let $D \subseteq \mathbb{R}$ and c be an interior point of D. If $f : D \to \mathbb{R}$ is differentiable at c and has a local extremum at c, then $f'(c) = 0$.*

Proof. Suppose f is differentiable at c. By Carathéodory's Lemma, there is $f_1 : D \to \mathbb{R}$ such that f_1 is continuous at c and

---

[4] It may be a good idea to recall the definitions of points of local extrema, that is, points of local maximum or local minimum, of a function. See Section 1.3.

$$f(x) - f(c) = (x - c)f_1(x) \quad \text{for all } x \in D.$$

Now, if f has a local maximum at c, then there is $\delta > 0$ such that $(c-\delta, c+\delta) \subseteq D$ and $f(x) - f(c) \leq 0$ for all $x \in (c - \delta, c + \delta)$. So, in this case,

$$f_1(x) \geq 0 \quad \text{for all } x \in (c - \delta, c) \quad \text{and} \quad f_1(x) \leq 0 \quad \text{for all } x \in (c, c + \delta).$$

Similarly, if f has a local minimum at c, then there is $\delta > 0$ such that $(c - \delta, c + \delta) \subseteq D$ and $f(x) - f(c) \geq 0$ for all $x \in (c - \delta, c + \delta)$. So, in this case,

$$f_1(x) \leq 0 \quad \text{for all } x \in (c - \delta, c) \quad \text{and} \quad f_1(x) \geq 0 \quad \text{for all } x \in (c, c + \delta).$$

In any case, by the continuity of f_1 at c, it follows that $f_1(c) = 0$, that is, $f'(c) = 0$.   □

We pause to give an interesting application of the above lemma before moving on to Rolle's Theorem and the MVT. This is a result that is sometimes ascribed to Darboux and called the **IVP for derivatives**. To put this result in perspective, let us first note that if I is an interval and a function $f : I \to \mathbb{R}$ is **continuously differentiable**, that is, if f' exists and is continuous, then by Proposition 3.13, f' has the IVP on I. But in general, f' need not be continuous.[5] Yet, as the following result shows, f' has the IVP.

Proposition 4.14. *Let I be an interval and $f : I \to \mathbb{R}$ be a differentiable function. Then the derivative function f' has the IVP on I.*

Proof. Let $a, b \in I$ with $a < b$ and $r \in \mathbb{R}$ be such that $f'(a) < r < f'(b)$. Consider the function $g : [a, b] \to \mathbb{R}$ defined by

$$g(x) = f(x) - rx \quad \text{for } x \in [a, b].$$

Then g is differentiable and

$$\lim_{x \to a^+} \frac{g(x) - g(a)}{x - a} = g'(a) = f'(a) - r < 0.$$

Therefore, in view of part (i) of Proposition 3.24, it follows that there is $\delta > 0$ such that for all $x \in (a, a + \delta)$, we have

$$\frac{g(x) - g(a)}{x - a} < 0 \quad \text{and hence} \quad g(x) < g(a).$$

Thus, g cannot attain its minimum at a. In a similar way, since $g'(b) = f'(b) - r > 0$, we see that there is $\delta > 0$ such that $g(x) < g(b)$ for all $x \in (b - \delta, b)$. Thus, g cannot attain its minimum at b as well. But g is a continuous function on the closed and bounded set $[a, b]$, and hence by Proposition 3.8, g attains its minimum at some $c \in [a, b]$. Moreover, c is an interior point of $[a, b]$, since c cannot equal a or b. Hence it follows from Lemma 4.13 that $g'(c) = 0$, that is, $f'(c) = r$. Thus, f' has the IVP on I.   □

---

[5] Simple examples to show that the derivative of a differentiable function may not be continuous can be constructed using trigonometric functions. See, for instance, Example 7.19 of Chapter 7.

The above result is sometimes useful for showing that a given function cannot be the derivative of any other function. For example, if $f : [-1, 1] \to \mathbb{R}$ is the integral part function given by $f(x) = [x]$, then it is clear that f attains the values -1, 0, and 1 but none in between. Hence, $f \neq F'$ for any differentiable function $F : [-1, 1] \to \mathbb{R}$.

Proposition 4.15 (Rolle's Theorem). *If $f : [a, b] \to \mathbb{R}$ is continuous on $[a, b]$ and differentiable on (a, b) and if $f(a) = f(b)$, then there is $c \in (a, b)$ such that $f'(c) = 0$.*

Proof. Since f is a continuous function on the closed and bounded set $[a, b]$, it follows from Proposition 3.8 that f is bounded and attains its bounds on $[a, b]$. Thus, there are $c_1, c_2 \in [a, b]$ such that

$$f(c_1) = \max\{f(x) : x \in [a, b]\} \quad \text{and} \quad f(c_2) = \min\{f(x) : x \in [a, b]\}.$$

Now if c_1 or c_2 is an interior point of $[a, b]$, then by Lemma 4.13, we have $f'(c_1) = 0$ or $f'(c_2) = 0$, and the result is proved. Otherwise, both c_1 and c_2 are endpoints of $[a, b]$, and since $f(a) = f(b)$, we have $f(c_1) = f(c_2)$. Thus, the maximum and the minimum values of f on $[a, b]$ coincide. Hence f is constant on $[a, b]$, and therefore, $f'(c) = 0$ for every $c \in (a, b)$. $\square$

Rolle's Theorem can be used together with the IVP of continuous functions to check the uniqueness and the existence of roots in certain intervals, especially for polynomials with real coefficients. This is illustrated by the following examples.

Examples 4.16. (i) If $f(x) = x^3 + px + q$ for $x \in \mathbb{R}$, where $p, q \in \mathbb{R}$ and $p > 0$, then f has a unique real root. To see this, note that if f had more than one real root, then there would be $a, b \in \mathbb{R}$ with $a < b$ and $f(a) = f(b) = 0$. Hence by Rolle's Theorem, there would be $c \in (a, b)$ such that $f'(c) = 0$. But $f'(x) = 3x^2 + p$ is not zero for any $x \in \mathbb{R}$ since $p > 0$. On the other hand, $f(x) \to -\infty$ as $x \to -\infty$ and $f(x) \to \infty$ as $x \to \infty$, and thus f takes negative as well as positive values. Hence, $f(c) = 0$ for some $c \in \mathbb{R}$, since f has the IVP on $\mathbb{R}$. Thus f has a unique real root.

(ii) If $f(x) = x^4 + 2x^3 - 2$ for $x \in \mathbb{R}$, then f has a unique root in $[0, \infty)$. Indeed, $f'(x) = 4x^3 + 6x^2$ is positive for all $x \in (0, \infty)$, while $f(0) = -2 < 0$ and $f(1) = 1 > 0$. Thus Rolle's Theorem implies that f has at most one root in $[0, \infty)$, while the IVP implies that f has at least one root in $[0, 1]$. $\diamond$

Note that in the above examples, the functions were clearly continuous and differentiable. The following negative examples show that the conclusion of Rolle's Theorem may not be true if any one of the three conditions on f is dropped.

Examples 4.17. (i) Consider $f : [0, 1] \to \mathbb{R}$ defined by $f(x) = x$ for $x \in [0, 1)$ and $f(1) = 0$. Then f is differentiable on $(0, 1)$ and $f(0) = f(1) = 0$ but $f'(c) = 1 \neq 0$ for every $c \in (0, 1)$. Rolle's Theorem does not apply here since f is not continuous on $[0, 1]$. [In fact, continuity fails only at $x = 1$.]

(ii) Consider $f : [-1, 1] \to \mathbb{R}$ defined by $f(x) = |x|$ for $x \in [-1, 1]$. Then f is continuous on $[-1, 1]$ and $f(-1) = f(1) = 0$. But $f'(c) = 1$ or -1 for $c \neq 0$. Rolle's Theorem does not apply here since f is not differentiable on $(-1, 1)$. [In fact, differentiability fails only at $x = 0$.]

(iii) Consider $f : [0, 1] \to \mathbb{R}$ defined by $f(x) = x$ for $x \in [0, 1]$. Then f is continuous on $[0, 1]$ and differentiable on $(0, 1)$ but $f'(c) = 1 \neq 0$ for every $c \in (0, 1)$. Rolle's Theorem does not apply here since $f(0) \neq f(1)$. $\diamond$

We are now ready to state and prove the Mean Value Theorem. It may be remarked that there are, in fact, several versions of the Mean Value Theorem. The one we state below is among the most commonly used. It is usually ascribed to Lagrange and sometimes referred to as Lagrange's Mean Value Theorem.

Proposition 4.18 (Mean Value Theorem). *If a function $f : [a, b] \to \mathbb{R}$ is continuous on $[a, b]$ and differentiable on (a, b), then there is $c \in (a, b)$ such that*

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

Proof. Consider $F : [a, b] \to \mathbb{R}$ defined by

$$F(x) = f(x) - f(a) - s(x - a), \quad \text{where} \quad s = \frac{f(b) - f(a)}{b - a}.$$

Then $F(a) = 0$ and our choice of the constant s is such that $F(b) = 0$. So Rolle's Theorem applies to F, and as a result, there is $c \in (a, b)$ such that $F'(c) = 0$. This implies that $f'(c) = s$, as desired. $\square$

Remark 4.19. If we write $b = a + h$, then the conclusion of the MVT may be stated as follows:

$$f(a + h) = f(a) + h f'(a + \theta h) \quad \text{for some } \theta \in (0, 1).$$

The equivalence with the MVT is easily verified. $\diamond$

Corollary 4.20 (Mean Value Inequality). *If a function $f : [a, b] \to \mathbb{R}$ is continuous on $[a, b]$ and differentiable on (a, b), and if $m, M \in \mathbb{R}$ are such that $m \leq f'(x) \leq M$ for all $x \in (a, b)$, then*

$$m(b - a) \leq f(b) - f(a) \leq M(b - a).$$

Proof. The desired inequality is an immediate consequence of the MVT. $\square$

The corollary below is perhaps the most important consequence of the MVT in calculus.

Corollary 4.21. *Let I be an interval containing more than one point, and $f : I \to \mathbb{R}$ be any function. Then f is a constant function on I if and only if f' exists and is identically zero on I.*

Proof. If f is a constant function on I, then it is obvious that f' exists on I and $f'(x) = 0$ for all $x \in I$. Conversely, if f' exists and vanishes identically on I, then for any $x_1, x_2 \in I$ with $x_1 < x_2$, we have $[x_1, x_2] \subseteq I$ and applying the MVT to the restriction of f to $[x_1, x_2]$, we obtain

$$f(x_2) - f(x_1) = f'(c)(x_2 - x_1) \quad \text{for some } c \in (x_1, x_2).$$

Since $f'(c) = 0$, we have $f(x_1) = f(x_2)$. This proves that f is a constant function on I. $\square$

Remark 4.22. If $D \subseteq \mathbb{R}$ is not an interval, then there can be nonconstant differentiable functions on D whose derivative is identically zero. For example, if $D = (0, 1) \cup (1, 2)$ is a disjoint union of two open intervals and $f : D \to \mathbb{R}$ is defined by $f(x) = 1$ if $x \in (0, 1)$ and $f(x) = 2$ if $x \in (1, 2)$, then f is differentiable and f' is identically zero on D but f is not a constant function. Thus, the hypothesis that I is an interval is essential in Corollary 4.21. $\diamond$

The MVT or the mean value inequality may also be used to approximate a differentiable function around a point. For example, if $m \in \mathbb{N}$ and $f(x) = \sqrt{x}$ for $x \in [m, m + 1]$, then

$$\sqrt{m+1} - \sqrt{m} = f(m+1) - f(m) = f'(c) = \frac{1}{2\sqrt{c}}$$

for some $c \in \mathbb{R}$ such that $m < c < m + 1$. Hence

$$\frac{1}{2\sqrt{m+1}} < \sqrt{m+1} - \sqrt{m} < \frac{1}{2\sqrt{m}}.$$

For example, by putting $m = 1$, we obtain

$$1 + \frac{1}{2\sqrt{2}} < \sqrt{2} < 1 + \frac{1}{2} \quad \text{and hence} \quad \frac{4}{3} < \sqrt{2} < \frac{3}{2}.$$

Similarly, putting $m = 2, 3$, and 4, we can obtain estimates for $\sqrt{3}$ and $\sqrt{5}$. (See Exercise 28 (i).)

If we want a better approximation, a natural candidate for an approximating function is a polynomial function. Suppose we want to approximate f by a polynomial P around a. Naturally, we require $f(a) = P(a)$. Then the simplest approximation is the constant polynomial given by $P(x) = f(a)$, and the MVT can be used to estimate the error $f(x) - P(x) = f(x) - f(a)$. Next, if we require $f(a) = P(a)$ and further, $f'(a) = P'(a)$, then we may consider a linear polynomial instead of the constant polynomial $f(a)$. To be able to evaluate easily a linear polynomial P at a, let us write $P(x) = c_0 + c_1(x - a)$. Then the conditions $f(a) = P(a)$ and $f'(a) = P'(a)$ are equivalent to $c_0 = f(a)$ and $c_1 = f'(a)$. In general, if f has derivatives up to the nth order, then an nth-degree polynomial given by

$$P(x) = c_0 + c_1(x - a) + c_2(x - a)^2 + \cdots + c_n(x - a)^n$$

will satisfy the conditions

$$f(a) = P(a), \ f'(a) = P'(a), \ f''(a) = P''(a), \ \ldots, \ f^{(n)}(a) = P^{(n)}(a)$$

if we take

$$c_0 = f(a), \ c_1 = f'(a), \ c_2 = \frac{f''(a)}{2!}, \ \ldots, \ c_n = \frac{f^{(n)}(a)}{n!}.$$

Note that these values are simply obtained by successively differentiating $P(x)$, substituting $x = a$, and then comparing with the corresponding derivative of f at a. This time, the error in the nth-degree approximation $P(x)$ can be estimated by the following generalization of the MVT.

Proposition 4.23 (Taylor's Theorem). *Let $n \in \mathbb{Z}$, $n \geq 0$, and $f : [a, b] \to \mathbb{R}$ be such that $f', f'', \ldots, f^{(n)}$ exist on $[a, b]$ and further, $f^{(n)}$ is continuous on $[a, b]$ and differentiable on (a, b). Then there is $c \in (a, b)$ such that*

$$f(b) = f(a) + f'(a)(b - a) + \cdots + \frac{f^{(n)}(a)}{n!}(b - a)^n + \frac{f^{(n+1)}(c)}{(n+1)!}(b - a)^{n+1}.$$

Proof. For $x \in [a, b]$, let

$$P(x) = f(a) + f'(a)(x - a) + \frac{f''(a)}{2!}(x - a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x - a)^n.$$

Consider $F : [a, b] \to \mathbb{R}$ defined by

$$F(x) = f(x) - P(x) - s(x - a)^{n+1}, \quad \text{where} \quad s = \frac{f(b) - P(b)}{(b - a)^{n+1}}.$$

Then $F(a) = 0$ and our choice of s is such that $F(b) = 0$. So Rolle's Theorem applies to F, and as a result, there is $c_1 \in (a, b)$ such that $F'(c_1) = 0$. Next, $f'(a) = P'(a)$ and so $F'(a) = 0$ as well. Now, Rolle's Theorem applies to the restriction of F' to $[a, c_1]$, and so there is $c_2 \in (a, c_1)$ such that $F''(c_2) = 0$. Further, if $n > 1$, then $F''(a) = 0$, and so there is $c_3 \in (a, c_2)$ such that $F'''(c_3) = 0$. Continuing in this way, we see that there is $c := c_{n+1} \in (a, c_n)$ such that $F^{(n+1)}(c) = 0$. Now, $P^{(n+1)}$ is identically zero, since P is a polynomial of degree n. In particular, $P^{(n+1)}(c) = 0$. Hence $f^{(n+1)}(c) = s(n+1)!$, which, in turn, yields the desired result. $\qquad \square$

Remarks 4.24. (i) Note that the MVT corresponds to the case $n = 0$ of Taylor's Theorem. The case $n = 1$ is sometimes called the **Extended Mean Value Theorem**.

(ii) In our statement of Taylor's Theorem, the point a was the left endpoint of the interval on which the function f was defined. There is an analogous version for the right (hand) endpoint. Namely, if $f : [a, b] \to \mathbb{R}$ is as in the statement of Taylor's Theorem, then there is $c \in (a, b)$ such that

$$f(a) = f(b) + f'(b)(a - b) + \cdots + \frac{f^{(n)}(b)}{n!}(a - b)^n + \frac{f^{(n+1)}(c)}{(n + 1)!}(a - b)^{n+1}.$$

This can be proved in a similar manner or alternatively deduced from our version of Taylor's Theorem by applying it to the function $g : [a, b] \to \mathbb{R}$ defined by $g(x) = f(a + b - x)$ for $x \in [a, b]$, and noting that $g^{(k)}(x) = (-1)^k f^{(k)}(a + b - x)$ for $k \in \mathbb{N}$. It follows that if I is any interval, a is any point of I, and $f : I \to \mathbb{R}$ is such that $f', f'', \ldots, f^{(n)}$ exist on I and $f^{(n+1)}$ exists at every interior point of I, then for any $x \in I$, $x \neq a$, there is c between a and x such that

$$f(x) = f(a) + f'(a)(x - a) + \cdots + \frac{f^{(n)}(a)}{n!}(x - a)^n + \frac{f^{(n+1)}(c)}{(n + 1)!}(x - a)^{n+1}.$$

The last expression is sometimes referred to as the **Taylor formula** for f around a. The polynomial given by

$$P_n(x) = f(a) + f'(a)(x - a) + \cdots + \frac{f^{(n)}(a)}{n!}(x - a)^n$$

is called the nth **Taylor polynomial** of f around a. The difference $R_n = f - P_n$ is called the **remainder** of order n. Note that the Taylor formula for f around a shows that the remainder R_n is given by

$$R_n(x) = \frac{f^{(n+1)}(c)}{(n + 1)!}(x - a)^{n+1} \quad \text{for some } c \text{ between } a \text{ and } x.$$

The above expression for $R_n(x)$ is sometimes called the **Lagrange form of remainder** in Taylor formula. This is to distinguish it from some alternative expressions for the remainder in the Taylor formula that appear in Exercise 49 of this chapter and Exercise 46 of Chapter 6. $\diamond$

The following corollary of the Taylor formula generalizes Corollary 4.21 and gives a characterization of polynomial functions on intervals.

Corollary 4.25. *Let I be an interval containing more than one point, and $f : I \to \mathbb{R}$ be any function. Let n be a nonnegative integer. Then f is a polynomial function of degree $\leq n$ on I if and only if $f^{(n+1)}$ exists and is identically zero on I.*

Proof. If f is a polynomial function on I of degree $\leq n$, that is, if there are $a_0, a_1, \ldots, a_n \in \mathbb{R}$ such that

$$f(x) = a_n x^n + \cdots + a_1 x + a_0 \quad \text{for all } x \in I,$$

then it is obvious that $f^{(n+1)}(x) = 0$ for all $x \in I$. To prove the converse, it suffices to fix some $a \in I$ and apply Taylor's formula for f around a. $\square$

Example 4.26. Let $r \in \mathbb{Q}$ and $f : [-1, 1] \to \mathbb{R}$ be defined by $f(x) := (1+x)^r$. Given any nonnegative integer k and $c \in (-1, 1)$, we clearly have

$$f^{(k)}(c) = r(r-1)\cdots(r-k+1)(1+x)^{r-k}.$$

Hence, the nth Taylor polynomial of f around 0 is given by

$$P_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(0)}{k!} x^k = \sum_{k=0}^{n} \frac{r(r-1)\cdots(r-k+1)}{k!} x^k = \sum_{k=0}^{n} \binom{r}{k} x^k.$$

Notice that if r equals a nonnegative integer n, then $f^{(n+1)}$ is identically zero, and so the remainder of order n is zero. Thus in this case, by Taylor Theorem, we recover the binomial expansion for $(1+x)^n$. $\diamond$

Usually, the nth Taylor polynomial of f around a provides a progressively better approximation to f around a as n increases. We will study this aspect in greater detail in Section 5.3. For the moment, let us revisit the estimates for $\sqrt{2}$ that were obtained from the MVT and see what happens when we use Taylor's Theorem. Thus, let $m \in \mathbb{N}$ and $f : [m, m+1] \to \mathbb{R}$ be given by $f(x) := \sqrt{x}$. Applying the Taylor formula for f around m, with $n = 1$, we have

$$f(x) = f(m) + f'(m)(x-m) + \frac{f''(c)}{2!}(x-m)^2 \quad \text{for some } c \text{ between } m \text{ and } x.$$

In particular, for $x = m+1$, we get

$$\sqrt{m+1} = \sqrt{m} + \frac{1}{2\sqrt{m}} - \frac{1}{8c\sqrt{c}} \quad \text{for some } c \in (m, m+1).$$

For example, by putting $m = 1$, we obtain

$$1 + \frac{1}{2} - \frac{1}{8} < \sqrt{2} < 1 + \frac{1}{2} - \frac{1}{16\sqrt{2}} \quad \text{and hence} \quad \frac{11}{8} < \sqrt{2} < 1 + \frac{1}{2} - \frac{1}{16(3/2)} = \frac{35}{24},$$

where in the last inequality we have used the estimate $\sqrt{2} < \frac{3}{2}$, which is obvious from

$$\sqrt{2} < 1 + \frac{1}{2} - \frac{1}{16\sqrt{2}}.$$

The resulting bounds $\frac{11}{8} = 1.375$ and $\frac{35}{24} \approx 1.4583$ are, in fact, better than the bounds $\frac{4}{3} \approx 1.33$ and $\frac{3}{2} = 1.5$ obtained using the MVT. Needless to say, the higher order Taylor polynomials would give even better bounds. In this way, if you are stranded on an island without your calculator and a demon demands to know a reasonably correct value of $\sqrt{2}$, then Taylor's Theorem can save the day for you!

4.3 Monotonicity, Convexity, and Concavity

Let I be an interval in $\mathbb{R}$ and $f : I \to \mathbb{R}$ be any function. Recall from Chapter 1 that f is said to be **(monotonically) increasing** on I if $x_1, x_2 \in I$, $x_1 < x_2$ implies $f(x_1) \le f(x_2)$. Also, f is said to be **(monotonically) decreasing** on I if $x_1, x_2 \in I$, $x_1 < x_2$ implies $f(x_1) \ge f(x_2)$. One says that f is **monotonic** on I if it is monotonically increasing on I or monotonically decreasing on I. The function f is said to be **strictly increasing** [resp. **strictly decreasing**] on I if $x_1, x_2 \in I$, $x_1 < x_2$, implies $f(x_1) < f(x_2)$ [resp. $f(x_1) > f(x_2)$]. Also, one says that f is **strictly monotonic** on I if it is strictly increasing on I or strictly decreasing on I.

As has been pointed out and illustrated in Chapter 1, the notions of monotonicity and strict monotonicity are purely geometric, and a priori they have no relation with derivatives. However, in the case of differentiable functions, there is an intimate relationship between derivatives and the notions of monotonicity and strict monotonicity. The key idea can be easily grasped by looking at the graph of a function. The tangents to the graph of an increasing function have positive slopes, whereas the tangents to the graph of a decreasing function have negative slopes. [See, for example, the graph of $y = x^2$ in Figure 1.2 of Chapter 1.] A more precise analytic formulation of this is given in the proposition below. In practice, this greatly simplifies checking whether a differentiable function is increasing or decreasing.

Proposition 4.27. *Let I be an interval containing more than one point, and $f : I \to \mathbb{R}$ be a differentiable function. Then we have the following:*

(i) *f' is nonnegative throughout I $\iff$ f is monotonically increasing on I.*
(ii) *f' is nonpositive throughout I $\iff$ f is monotonically decreasing on I.*
(iii) *f' is positive throughout I $\implies$ f is strictly increasing on I.*
(iv) *f' is negative throughout I $\implies$ f is strictly decreasing on I.*

Proof. Suppose $x_1, x_2 \in I$ with $x_1 < x_2$. Then $[x_1, x_2] \subseteq I$ and we can apply the MVT to the restriction of f to $[x_1, x_2]$ to obtain

$$f(x_2) - f(x_1) = f'(c)(x_2 - x_1) \quad \text{for some } c \in (x_1, x_2).$$

Thus, if f' is nonnegative throughout I, then $f(x_1) \le f(x_2)$, whereas if f' is nonpositive throughout I, then $f(x_1) \ge f(x_2)$. This proves the implication "$\implies$" in (i) and (ii). Moreover, we also see from the MVT that if f' is positive throughout I, then $f(x_1) < f(x_2)$, whereas if f' is negative throughout I, then $f(x_1) > f(x_2)$. This proves (iii) and (iv).

Now, given any $x_0 \in I$ and $0 \ne h \in \mathbb{R}$ such that $x_0 + h \in I$, the quotient

$$\frac{f(x_0 + h) - f(x_0)}{h}$$

is always nonnegative if f is monotonically increasing and always nonpositive if f is monotonically decreasing. Therefore, $f'(x_0) \ge 0$ if f is monotonically

increasing on I, whereas $f'(x_0) \leq 0$ if f is monotonically decreasing on I. This proves the implication "$\Longleftarrow$" in (i) and (ii). $\square$

The following corollary is essentially obtained by combining the first two and the last two parts of the above proposition.

Corollary 4.28. *Let I be an interval containing more than one point, and $f : I \to \mathbb{R}$ be a differentiable function. Then we have the following.*

(i) *f' does not change sign throughout $I \Longleftrightarrow f$ is monotonic on I.*
(ii) *f' is nonzero throughout $I \Longrightarrow f$ is strictly monotonic on I.*

Proof. Using parts (i) and (ii) of Proposition 4.27, we obtain (i), while using the IVP for f' (Proposition 4.14) together with parts (iii) and (iv) of Proposition 4.27, we obtain (ii). $\square$

Examples 4.29. (i) Consider the polynomial function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = x^4 - 8x^3 + 22x^2 - 24x + 7.$$

Then f is differentiable and one can easily check that

$$f'(x) = 4x^3 - 24x^2 + 44x - 24 = 4(x-1)(x-2)(x-3).$$

Therefore, $f'(x) \geq 0$ if $x \geq 3$ or $1 \leq x \leq 2$, whereas $f'(x) \leq 0$ if $x \leq 1$ or $2 \leq x \leq 3$. Thus, f is monotonically increasing on $[1, 2]$ and on $[3, \infty)$, whereas f is monotonically decreasing on $[2, 3]$ and on $(-\infty, 1]$. In fact, since f' vanishes only at $x = 1, 2$, and 3, we see that f is strictly increasing on $(1, 2)$ and on $(3, \infty)$, whereas f is strictly decreasing on $(2, 3)$ and on $(-\infty, 1)$. Notice that in an example such as this, it would be extremely difficult to arrive at the above conclusions directly from the definition.
(ii) Let $n \in \mathbb{N}$ and consider the nth power function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^n$. Then $f'(x) = nx^{n-1}$ for $x \in \mathbb{R}$. First, assume that n is odd. Then $f'(x) \geq 0$ for all $x \in \mathbb{R}$. Thus, f is monotonically increasing on $\mathbb{R}$. In fact, since f' vanishes only at $x = 0$, we see that f is strictly increasing on $(-\infty, 0)$ as well as on $(0, \infty)$. Next, assume that n is even. Then $f'(x) \geq 0$ for $x \geq 0$ and $f'(x) \leq 0$ for $x \leq 0$. Thus, f is monotonically increasing on $[0, \infty)$ and monotonically decreasing on $(-\infty, 0]$. In fact, since f' vanishes only at $x = 0$, we see that f is strictly increasing on $(0, \infty)$ and strictly decreasing on $(-\infty, 0)$. Notice that in this example, we can reach these conclusions directly from the definition. In fact, we can do a little better. Namely, we can easily see that if n is odd, then f is strictly increasing on $\mathbb{R}$, whereas if n is even, then f is strictly increasing on $[0, \infty)$ and strictly decreasing on $(-\infty, 0]$. $\diamond$

As the last example shows, the converse of the implication in part (iii) of Proposition 4.27 is not true. In fact, it suffices to note that $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^3$ is strictly increasing but $f'(0) = 0$. Similarly, the function

$g : \mathbb{R} \to \mathbb{R}$ defined by $g(x) = -x^3$ is strictly decreasing but $g'(0) = 0$, which shows that the converse of the implication in part (iv) of Proposition 4.27 is not true as well. As a consequence, the converse of part (ii) of Corollary 4.28 is not true. In other words, these parts give only sufficient conditions for a differentiable function to be strictly increasing or strictly decreasing or strictly monotonic. However, with a little more effort, it is possible to give a necessary and sufficient condition, as shown by the proposition below.

Proposition 4.30. *Let I be an interval containing more than one point, and $f : I \to \mathbb{R}$ be a differentiable function. Then*

(i) *f is strictly increasing on I if and only if f' is nonnegative throughout I and f' does not vanish identically on any subinterval of I containing more than one point.*

(ii) *f is strictly decreasing on I if and only if f' is nonpositive throughout I and f' does not vanish identically on any subinterval of I containing more than one point.*

Proof. Let f be strictly increasing on I. Then by part (i) of Proposition 4.27, f' is nonnegative throughout I. Moreover, if f' were to vanish identically on any subinterval J of I containing more than one point, then by Corollary 4.21, f would be constant on J, and this is a contradiction because f is strictly increasing. For the converse, first note that by part (i) of Proposition 4.27, f is monotonically increasing on I. Further, if for some $x_1, x_2 \in I$ with $x_1 < x_2$ we have $f(x_1) = f(x_2)$, then f is constant throughout $[x_1, x_2]$ and hence f' vanishes identically on $[x_1, x_2]$, which is a contradiction. This proves (i).

The assertion (ii) is proved similarly. $\square$

We now turn to the notions of convexity and concavity. Let us recall that if I is an interval in $\mathbb{R}$, then $f : I \to \mathbb{R}$ is said to be convex on I if the graph of f lies below the line joining any two points on it, whereas $f : I \to \mathbb{R}$ is said to be concave on I if the graph of f lies above the line joining any two points on it. In other words, f is **convex** on I if for any $x_1, x_2, x \in I$ with $x_1 < x < x_2$, we have $f(x) \leq L(x)$, whereas f is **concave** on I if for any $x_1, x_2, x \in I$ with $x_1 < x < x_2$, we have $f(x) \geq L(x)$, where

$$L(x) := f(x_1) + \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_1) \quad \text{for } x \in I.$$

Recall also that f is **strictly convex** (resp. **strictly concave**) on I if for any $x_1, x_2, x \in I$ with $x_1 < x < x_2$, we have $f(x) < L(x)$ (resp. $f(x) > L(x)$). Equivalently, f is convex or strictly convex or concave or strictly concave on I according as $f(tx_1 + (1-t)x_2)$ is $\leq$ or $<$ or $\geq$ or $>$ than $tf(x_1) + (1-t)f(x_2)$ for all $x_1, x_2 \in I$ and $t \in (0, 1)$.

As noted before, convexity and concavity are purely geometric notions and a priori they have no relation with derivatives. However, in the case of differentiable functions, there is an intimate relation between derivatives and

the notions of convexity and concavity. The key idea can once again be gleaned by looking at the graphs. Namely, if we draw tangents at each point, then we see that as we move from left to right, the slopes increase if the function is convex, whereas the slopes decrease if the function is concave. [See, for example, the graph of $y = x^3$ in Figure 1.3 of Chapter 1.] A more precise analytic formulation of this is given in the proposition below. In practice, this greatly simplifies checking whether a differentiable function is convex or concave.

Proposition 4.31. *Let I be an interval containing more than one point, and $f : I \to \mathbb{R}$ be a differentiable function. Then we have the following:*

(i) *f' is monotonically increasing on $I \iff f$ is convex on I.*
(ii) *f' is monotonically decreasing on $I \iff f$ is concave on I.*
(iii) *f' is strictly increasing on $I \iff f$ is strictly convex on I.*
(iv) *f' is strictly decreasing on $I \iff f$ is strictly concave on I.*

Proof. First, assume that f' is monotonically increasing on I. Let $x_1, x_2, x \in I$ be such that $x_1 < x < x_2$. By the MVT, there are $c_1 \in (x_1, x)$ and $c_2 \in (x, x_2)$ satisfying

$$f(x) - f(x_1) = f'(c_1)(x - x_1) \quad \text{and} \quad f(x_2) - f(x) = f'(c_2)(x_2 - x).$$

Now, $c_1 < c_2$ and f' is monotonically increasing on I and so

$$\frac{f(x) - f(x_1)}{x - x_1} = f'(c_1) \le f'(c_2) = \frac{f(x_2) - f(x)}{x_2 - x}.$$

Collecting only the terms involving $f(x)$ on the left (hand) side, we obtain

$$f(x)\left(\frac{1}{x - x_1} + \frac{1}{x_2 - x}\right) \le \frac{f(x_1)}{x - x_1} + \frac{f(x_2)}{x_2 - x}.$$

Multiplying throughout by $(x - x_1)(x_2 - x)/(x_2 - x_1)$, we see that

$$f(x) \le \frac{1}{x_2 - x_1}[f(x_1)(x_2 - x) + f(x_2)(x - x_1)] = f(x_1) + \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_1).$$

Thus, f is convex on I.

Conversely, assume that f is convex on I. Let $x_1, x_2, x \in I$ be such that $x_1 < x < x_2$. Then

$$f(x) \le f(x_1) + \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_1) = f(x_2) - \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x_2 - x),$$

where the last equality follows by writing $x - x_1 = (x_2 - x_1) - (x_2 - x)$. As a consequence,

$$\frac{f(x) - f(x_1)}{x - x_1} \le \frac{f(x_2) - f(x_1)}{x_2 - x_1} \le \frac{f(x_2) - f(x)}{x_2 - x}.$$

Taking limits as $x \to x_1^+$ and $x \to x_2^-$, we obtain

$$f'(x_1) \leq \frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq f'(x_2).$$

Thus, f' is monotonically increasing on I. This proves (i). Moreover, the arguments in the preceding paragraph, with $\leq$ replaced by $<$, also prove that if f' is strictly increasing on I, then f is strictly convex on I. Conversely, assume that f is strictly convex on I. Then by part (i) above, f' is monotonically increasing on I. Further, if for some $x_1, x_2 \in I$ with $x_1 < x_2$, we have $f'(x_1) = f'(x_2)$, then f' is constant throughout $[x_1, x_2]$, and so f'' is identically zero on $[x_1, x_2]$. Hence by Corollary 4.25, there are constants $a_0, a_1 \in \mathbb{R}$ such that $f(x) = a_1 x + a_0$ for all $x \in [x_1, x_2]$. But this contradicts the strict convexity of f. Thus, (iii) is proved.

The corresponding results (ii) and (iv) about concave and strictly concave functions are proved similarly. Alternatively, (ii) and (iv) follow from applying (i) and (iii) to $-f$.   □

For twice differentiable functions, testing convexity or concavity can sometimes be simpler using the following result.

Proposition 4.32. *Let I be an interval containing more than one point, and $f : I \to \mathbb{R}$ be a twice differentiable function. Then we have the following:*

 (i) *f'' is nonnegative throughout I $\iff$ f is convex on I.*
 (ii) *f'' is nonpositive throughout I $\iff$ f is concave on I.*
(iii) *f'' is positive throughout I $\implies$ f is strictly convex on I.*
(iv) *f'' is negative throughout I $\implies$ f is strictly concave on I.*

Proof. Apply Proposition 4.31 to f and Proposition 4.27 to f'.   □

The following corollary is obtained by combining the first two and the last two parts of the above proposition.

Corollary 4.33. *Let I be an interval containing more than one point, and $f : I \to \mathbb{R}$ be a twice differentiable function. Then we have the following:*

 (i) *f'' does not change sign throughout I $\iff$ f is convex on I or f is concave on I.*
 (ii) *f'' is nonzero throughout I $\implies$ f is strictly convex on I or f is strictly concave on I.*

Proof. Apply Proposition 4.31 to f and Corollary 4.28 to f'.   □

Examples 4.34. (i) Consider the polynomial function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) := x^4 + 2x^3 - 36x^2 + 62x + 5.$$

Then f is twice differentiable with $f'(x) = 4x^3 + 6x^2 - 72x + 62$ and

$$f''(x) = 12x^2 + 12x - 72 = 12(x+3)(x-2).$$

Therefore, $f''(x) \geq 0$ if $x \geq 2$ or $x \leq -3$, whereas $f''(x) \leq 0$ if $-3 \leq x \leq 2$. Thus, f is convex on $[2, \infty)$ and on $(-\infty, -3]$, whereas f is concave on $[-3, 2]$. In fact, since f'' vanishes only at $x = -3$ and 2, we see that f is strictly convex on $(2, \infty)$ and on $(-\infty, -3)$, whereas f is strictly concave on $(-3, 2)$. Notice that in an example such as this, it would be extremely difficult to arrive at the above conclusions directly from the definition.

(ii) Let $n \in \mathbb{N}$ and consider the nth-power function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) := x^n$. Then $f''(x) = n(n-1)x^{n-2}$ for $x \in \mathbb{R}$. Thus if n is even, then f is convex on $\mathbb{R}$, whereas if n is odd and $n > 1$, then f is convex on $[0, \infty)$ and concave on $(-\infty, 0]$. In case $n = 1$, f is convex as well as concave on $\mathbb{R}$. In case $n > 1$, f'' vanishes only at $x = 0$ and hence if n is even, then f is strictly convex on $(0, \infty)$ as well as on $(-\infty, 0)$, whereas if n is odd and $n > 1$, then f is strictly convex on $(0, \infty)$ and strictly concave on $(-\infty, 0)$. Notice that in this example as well, it is not very easy to arrive at the above conclusions directly from the definition when n is large. [Compare Examples 1.15 (i), (ii) and Exercise 32 of Chapter 1.] However, we can directly appeal to Proposition 4.31 instead of Proposition 4.32 to get a stronger conclusion. Namely, if n is even, then $f'(x) = nx^{n-1}$ is strictly increasing on $\mathbb{R}$ and hence f is strictly convex on $\mathbb{R}$, whereas if n is odd and $n > 1$, then $f'(x) = nx^{n-1}$ is strictly increasing on $[0, \infty)$ and on $(-\infty, 0]$, and hence f is strictly convex on $[0, \infty)$ and strictly concave on $(-\infty, 0]$. $\diamond$

The converse of the implication in part (ii) of Corollary 4.33 is not true. For example, $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^4$ is strictly convex on $\mathbb{R}$ but $f''(0) = 0$. Similarly, $g : \mathbb{R} \to \mathbb{R}$ defined by $g(x) = -x^4$ is strictly concave on $\mathbb{R}$ but $g''(0) = 0$. Thus, part (ii) of Corollary 4.33 gives only a sufficient condition for a twice differentiable function to be strictly convex or strictly concave. However, with a little more effort, it is possible to give a necessary and sufficient condition, as shown by the proposition below.

Proposition 4.35. *Let I be an interval containing more than one point, and $f : I \to \mathbb{R}$ be a twice differentiable function. Then*

(i) *f is strictly convex on I if and only if f'' is nonnegative throughout I and f'' does not vanish identically on any subinterval of I containing more than one point.*

(ii) *f is strictly concave on I if and only if f'' is nonpositive throughout I and f'' does not vanish identically on any subinterval of I containing more than one point.*

Proof. Applying Proposition 4.30 to f' and using parts (iii) and (iv) of Proposition 4.31, we get the desired results. $\square$

4.4 L'Hôpital's Rule

In this section, we shall describe a useful method for finding limits that is known as L'Hôpital's Rule.[6] Actually, there are several versions of L'Hôpital's Rule and the formal statements of these will appear in the form of propositions or as a part of some remarks.

In its simplest form, L'Hôpital's Rule says the following. Suppose f, g are real-valued differentiable functions in an interval $(c - r, c + r)$ about a point c and suppose $f(c) = g(c) = 0$. If it so happens that the quotient $f'(x)/g'(x)$ of the derivatives is defined in an open interval around c (so that $g'(x) \neq 0$ in this interval), and moreover,

$$\lim_{x \to c} \frac{f'(x)}{g'(x)} = \frac{f'(c)}{g'(c)},$$

then the quotient $f(x)/g(x)$ has a limit as $x \to c$ and it is, in fact, the same limit as that of the quotient of the derivatives, that is,

$$\lim_{x \to c} \frac{f(x)}{g(x)} = \lim_{x \to c} \frac{f'(x)}{g'(x)}.$$

Since $g(c) = 0$, it follows from the MVT that the quotient $f(x)/g(x)$ is defined (that is, $g(x) \neq 0$) for all $x \neq c$ in the interval about c where $g'(x) \neq 0$.

For example, let us take $c = 0$ and $f, g : (-1, 1) \to \mathbb{R}$ given by

$$f(x) := \sqrt{1 + x^2} - \sqrt{1 - x^2} \quad \text{and} \quad g(x) := x \quad \text{for } x \in (-1, 1).$$

Then f, g are differentiable on $(-1, 1)$ and $f(0) = g(0) = 0$, while $g'(x) = 1$ is nonzero for every $x \in (-1, 1)$. Moreover,

$$\lim_{x \to c} \frac{f'(x)}{g'(x)} = \lim_{x \to 0} \frac{x \left[(1 + x^2)^{-1/2} + (1 - x^2)^{-1/2} \right]}{1} = \frac{0}{1} = \frac{f'(0)}{g'(0)}.$$

Hence from L'Hôpital's Rule, we can conclude that

$$\lim_{x \to 0} \frac{\sqrt{1 + x^2} - \sqrt{1 - x^2}}{x} = 0.$$

In this example, we could have avoided L'Hôpital's Rule and instead rationalized the quotient $f(x)/g(x)$ (that is, multiplied the numerator and denominator by $\sqrt{1 + x^2} + \sqrt{1 - x^2}$) to compute the limit. However, algebraic tricks such as rationalization become increasingly unwieldy if instead of $\sqrt{1 + x^2} - \sqrt{1 - x^2}$, $f(x)$ were given by $(1 + x^2)^{3/2} - (1 - x^2)^{3/2}$ or $(1 + x^2)^{5/2} - (1 - x^2)^{7/2}$. But L'Hôpital's Rule can still be applied to compute the limit just as easily.

---

[6] L'Hôpital, sometimes written L'Hospital, is pronounced *Lowpeetal*.

The reason why L'Hôpital's Rule works is quite simple. One just has to observe that

$$\lim_{x \to c} \frac{f'(x)}{g'(x)} = \frac{f'(c)}{g'(c)} = \frac{\displaystyle\lim_{x \to c} \frac{f(x) - f(c)}{x - c}}{\displaystyle\lim_{x \to c} \frac{g(x) - g(c)}{x - c}} = \lim_{x \to c} \frac{f(x) - f(c)}{g(x) - g(c)} = \lim_{x \to c} \frac{f(x)}{g(x)},$$

where the last step follows since $f(c) = g(c) = 0$.

It turns out that we can get rid of some of the assumptions in the simple formulation of L'Hôpital's Rule given above. Indeed, in the true spirit of dealing with limits as $x \to c$, we need not require that the concerned functions be defined at the point c. Thus the condition $f(c) = g(c) = 0$ may be replaced by the conditions

$$\lim_{x \to c} f(x) = 0 \quad \text{and} \quad \lim_{x \to c} g(x) = 0,$$

while the condition about the quotient of derivatives may be replaced by the condition

$$\frac{f'(x)}{g'(x)} \to \ell \quad \text{as } x \to c,$$

assuming, of course, that the above quotient is defined in an open interval about c except possibly at c. Now for the proof we will have to contend with some problems. First, a minor problem is that $f(c)$ and $g(c)$ are no longer defined. This is easily handled by simply *defining* $f(c) = g(c) = 0$. A more serious problem is that $f'(c)$ and $g'(c)$ don't make sense anymore. To handle this, one has to deal directly with the quotients such as $[f(x) - f(c)]/[g(x) - g(c)]$. What we need, in fact, is the following generalization of the MVT.

Proposition 4.36 (Cauchy's Mean Value Theorem). *If $f, g : [a, b] \to \mathbb{R}$ are continuous on $[a, b]$ and differentiable on (a, b), then there is $c \in (a, b)$ such that*

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

Proof. If $g(b) = g(a)$, the result follows by applying Rolle's Theorem to g. Otherwise, we consider $F : [a, b] \to \mathbb{R}$ defined by

$$F(x) = f(x) - f(a) - s[g(x) - g(a)], \quad \text{where} \quad s = \frac{f(b) - f(a)}{g(b) - g(a)}.$$

Then $F(a) = 0$ and our choice of the constant s is such that $F(b) = 0$. So Rolle's Theorem applies to F, and as a result, there is $c \in (a, b)$ such that $F'(c) = 0$. This implies that $sg'(c) = f'(c)$, as desired. $\square$

We are now ready to prove the first version of L'Hôpital's Rule. This one is called L'Hôpital's Rule for $\frac{0}{0}$ indeterminate forms since it applies to limits of quotients when both the numerator and the denominator tend to 0. For convenience, we state and prove below the version for left (hand) limits and remark later how other versions of L'Hôpital's Rule for $\frac{0}{0}$ indeterminate forms can be derived.

Proposition 4.37 (L'Hôpital's Rule for $\frac{0}{0}$ Indeterminate Forms). *Let $c \in \mathbb{R}$ and $f, g : (c - r, c) \to \mathbb{R}$ be differentiable functions such that*

$$\lim_{x \to c^-} f(x) = 0 \quad \text{and} \quad \lim_{x \to c^-} g(x) = 0.$$

Suppose $g'(x) \neq 0$ for all $x \in (c - r, c)$, and

$$\frac{f'(x)}{g'(x)} \to \ell \text{ as } x \to c^-.$$

Then

$$\frac{f(x)}{g(x)} \to \ell \text{ as } x \to c^-.$$

Here ℓ can be a real number or ∞ or $-\infty$.

Proof. Extend f, g to $(c - r, c]$ by putting $f(c) = g(c) = 0$. Let (x_n) be a sequence in $(c - r, c)$ such that $x_n \to c$. Since $g'(x) \neq 0$ for all $x \in (c - r, c)$, by Cauchy's Mean Value Theorem we see that for each $n \in \mathbb{N}$,

$$\frac{f(x_n)}{g(x_n)} = \frac{f(x_n) - f(c)}{g(x_n) - g(c)} = \frac{f'(c_n)}{g'(c_n)} \quad \text{for some } c_n \text{ between } x_n \text{ and } c.$$

Now $x_n \to c$ implies that $c_n \to c$, and hence the quotient above tends to ℓ as $n \to \infty$. Thus $f(x_n)/g(x_n) \to \ell$. $\square$

Remarks 4.38. (i) L'Hôpital's Rule for $\frac{0}{0}$ indeterminate forms is also valid for right (hand) limits. The statement and the proof is identical to the above, except we replace the interval $(c - r, c)$ by $(c, c + r)$ and the symbols $x \to c^-$ by $x \to c^+$. Combining the versions for left limits and right limits, we obtain L'Hôpital's Rule for (two-sided) limits of $\frac{0}{0}$ indeterminate forms, which may be stated as follows.

Let $c \in \mathbb{R}$ and $D = (c - r, c) \cup (c, c + r)$ for some $r > 0$. Let $f, g : D \to \mathbb{R}$ be differentiable functions such that $\lim_{x \to c} f(x) = 0$ and $\lim_{x \to c} g(x) = 0$. Suppose $g'(x) \neq 0$ for all $x \in D$, and $f'(x)/g'(x) \to \ell$ as $x \to c$. Then $f(x)/g(x) \to \ell$ as $x \to c$. Here ℓ can be a real number or ∞ or $-\infty$.

(ii) Analogues of L'Hôpital's Rule for $\frac{0}{0}$ indeterminate forms are also valid if instead of considering limits as $x \to c$, where c is a real number, we consider limits as $x \to \infty$ or as $x \to -\infty$. For example, a statement for limits as $x \to -\infty$ would be as follows.

Let $a \in \mathbb{R}$ and $f, g : (-\infty, a) \to \mathbb{R}$ be differentiable functions such that $f(x) \to 0$ and $g(x) \to 0$ as $x \to -\infty$. Suppose $g'(x) \neq 0$ for all $x \in (-\infty, a)$, and $f'(x)/g'(x) \to \ell$ as $x \to -\infty$. Then $f(x)/g(x) \to \ell$ as $x \to -\infty$. Here ℓ can be a real number or ∞ or $-\infty$.

As for the proof, it suffices to assume that $a < 0$ and apply L'Hôpital's Rule for left (hand) limits to the functions $F, G : (1/a, 0) \to \mathbb{R}$ defined by $F(x) = f(1/x)$ and $G(x) = g(1/x)$, considering limits as $x \to 0^-$. $\diamond$

The following examples illustrate how limits of certain functions that are in $\frac{0}{0}$ indeterminate form, or that can be converted to $\frac{0}{0}$ indeterminate form, can be computed by making one or more applications of L'Hôpital's Rule. The verification that L'Hôpital's Rule is indeed applicable (that is, the hypotheses of Proposition 4.37 are satisfied) in each case is left as an exercise.

Examples 4.39. (i) $\displaystyle\lim_{x\to 2}\frac{\sqrt{x^2+5}-3}{x^2-4}=\lim_{x\to 2}\frac{x/\sqrt{x^2+5}}{2x}=\frac{2/3}{4}=\frac{1}{6}.$

(ii) $\displaystyle\lim_{x\to 1}\frac{x^3-3x^2+3x-1}{x^3+x^2-5x+3}=\lim_{x\to 1}\frac{3x^2-6x+3}{3x^2+2x-5}=\lim_{x\to 1}\frac{6x-6}{6x+2}=0.$

(iii) We have

$$\lim_{x\to\infty}\left(x^3+4x^2+13x+1\right)^{1/3}-x=\lim_{y\to 0+}\frac{\left(1+4y+13y^2+y^3\right)^{1/3}-1}{y}.$$

Using L'Hôpital's Rule we see that the above limit exists and is equal to

$$\lim_{y\to 0+}\frac{1}{3}\left(1+4y+13y^2+y^3\right)^{-2/3}\left(4+26y+3y^2\right)=\frac{4}{3}. \qquad \diamond$$

Now we describe another version of L'Hôpital's Rule, which is useful in computing limits of $\frac{\infty}{\infty}$ indeterminate forms, that is, of quotients of functions where both the numerator and the denominator tend to infinity. It turns out here that the formulation as well as the proof of this rule is valid even when the numerator does not tend to infinity. But we may still refer to it as L'Hôpital's Rule for $\frac{\infty}{\infty}$ indeterminate forms. As before, we state and prove below the version for left (hand) limits. This time the statement and the proof are such that they are applicable to left (hand) limits as x approaches a real number and also (left hand!) limits as $x\to\infty$.

Proposition 4.40 (L'Hôpital's Rule for $\frac{\infty}{\infty}$ Indeterminate Forms). *Let I be the interval $[a,c)$ where $a\in\mathbb{R}$, and either $c\in\mathbb{R}$ with $a<c$ or $c=\infty$. Let $f,g:I\to\mathbb{R}$ be differentiable functions such that $|g(x)|\to\infty$ as $x\to c^-$. Suppose $g'(x)\neq 0$ for all $x\in I$ and*

$$\frac{f'(x)}{g'(x)}\to\ell \text{ as } x\to c^-.$$

Then

$$\frac{f(x)}{g(x)}\to\ell \text{ as } x\to c^-.$$

Here ℓ can be a real number or ∞ or $-\infty$.

Proof. Since $g'(x)\neq 0$ for all $x\in I$, by part (ii) of Corollary 4.27, either g is strictly increasing on I or g is strictly decreasing on I. Replacing g and f by $-g$ and $-f$ if necessary, we assume that g is strictly increasing on I. Now,

since g is strictly increasing on I and g, being continuous, has IVP on I, it follows that there is $a_1 \in I$ such that $g(x) > 0$ for all $x \in (a_1, c)$.

To begin with, let us consider the case that ℓ is a real number. Let $\epsilon > 0$ be given. Since $f'(x)/g'(x) \to \ell$ as $x \to c^-$, there is $a_2 \in (a_1, c)$ such that

$$\ell - \epsilon < \frac{f'(x)}{g'(x)} < \ell + \epsilon \quad \text{for all } x \in (a_2, c).$$

Let $h_\epsilon, h_{2\epsilon} : I \to \mathbb{R}$ be defined by

$$h_\epsilon := f - (\ell - \epsilon)g \quad \text{and} \quad h_{2\epsilon} := f - (\ell - 2\epsilon)g.$$

Then $h'_{2\epsilon}(x) > h'_\epsilon(x) > 0$ for all $x \in (a_2, c)$. Therefore, by part (iii) of Proposition 4.27, the functions h_ϵ and $h_{2\epsilon}$ are strictly increasing on (a_2, c). We claim that there is some $a_3 \in (a_2, c)$ such that $h_{2\epsilon}(x) > 0$ for all $x \in (a_3, c)$. To see this, assume the contrary. Then we can find an increasing sequence (x_n) in (a_2, c) such that $x_n \to c$ and $h_{2\epsilon}(x_n) \le 0$ for all $n \in \mathbb{N}$. Now, since $g(x) \to \infty$ as $x \to c^-$, we have $g(x_n) \to \infty$. On the other hand, since h_ϵ is (strictly) increasing on (a_2, c), we have $h_\epsilon(x_1) \le h_\epsilon(x_n)$ for all $n \in \mathbb{N}$, and hence

$$\epsilon g(x_n) = h_{2\epsilon}(x_n) - h_\epsilon(x_n) \le 0 - h_\epsilon(x_1), \text{ that is, } g(x_n) \le \frac{-h_\epsilon(x_1)}{\epsilon}$$

for all $n \in \mathbb{N}$. This contradicts the condition that $g(x_n) \to \infty$. So, our claim is proved. Thus, there is $a_3 \in (a_2, c)$ such that

$$h_{2\epsilon}(x) = f(x) - (\ell - 2\epsilon)g(x) > 0, \text{ that is, } \ell - 2\epsilon < \frac{f(x)}{g(x)} \quad \text{for all } x \in (a_3, c).$$

In a similar way, we see that there is $a_4 \in (a_2, c)$ such that

$$\frac{f(x)}{g(x)} < \ell + 2\epsilon \quad \text{for all } x \in (a_4, c).$$

Thus, if we let $a_5 := \max\{a_3, a_4\}$, then we have

$$\ell - 2\epsilon < \frac{f(x)}{g(x)} < \ell + 2\epsilon \quad \text{for all } x \in (a_5, c).$$

Since $\epsilon > 0$ is arbitrary, this proves that $f(x)/g(x) \to \ell$ as $x \to c^-$.

Next, suppose $\ell = \infty$. In this case we can proceed as above and the arguments are, in fact, simpler. Let $\alpha \in \mathbb{R}$ be given. Then there is $a_2 \in (a_1, c)$ such that $f'(x)/g'(x) > \alpha$ for all $x \in (a_2, c)$. Let $h_\alpha, h_{\alpha-1} : I \to \mathbb{R}$ be defined by $h_\alpha := f - \alpha g$ and $h_{\alpha-1} := f - (\alpha - 1)g$. Then $h'_{\alpha-1}(x) > h'_\alpha(x) > 0$ for all $x \in (a_2, c)$. Therefore, by part (iii) of Proposition 4.27, the functions h_α and $h_{\alpha-1}$ are strictly increasing on (a_2, c). We claim that there is some $a_3 \in (a_2, c)$ such that $h_{\alpha-1}(x) > 0$ for all $x \in (a_3, c)$. To see this, assume the contrary. Then we can find an increasing sequence (x_n) in (a_2, c) such that $x_n \to c$

and $h_{\alpha-1}(x_n) \leq 0$ for all $n \in \mathbb{N}$. Now, since $g(x) \to \infty$ as $x \to c^-$, we have $g(x_n) \to \infty$. On the other hand, since h_α is (strictly) increasing on (a_2, c), we have $h_\alpha(x_1) \leq h_\alpha(x_n)$ for all $n \in \mathbb{N}$, and hence

$$g(x_n) = h_{\alpha-1}(x_n) - h_\alpha(x_n) \leq 0 - h_\alpha(x_1) \quad \text{for all } n \in \mathbb{N}.$$

This contradicts the condition that $g(x_n) \to \infty$. So our claim is proved. Thus, there is $a_3 \in (a_2, c)$ such that

$$h_{\alpha-1}(x) = f(x) - (\alpha - 1)g(x) > 0, \text{ that is, } \frac{f(x)}{g(x)} > \alpha - 1 \quad \text{for all } x \in (a_3, c).$$

Since $\alpha \in \mathbb{R}$ is arbitrary, this proves that $f(x)/g(x) \to \infty$ as $x \to c^-$.

The case $\ell = -\infty$ is proved similarly. □

Remark 4.41. L'Hôpital's Rule for $\frac{\infty}{\infty}$ indeterminate forms is valid for right (hand) limits as x approaches a real number, and also (right hand!) limits as $x \to -\infty$. The statement is analogous to Proposition 4.40, and is also proved similarly. Combining the versions for left (hand) limits and right (hand) limits, we obtain L'Hôpital's Rule for (two-sided) limits of $\frac{\infty}{\infty}$ indeterminate forms, which may be stated as follows:

Let $c \in \mathbb{R}$ and $D = (c - r, c) \cup (c, c + r)$ for some $r > 0$. Let $f, g : D \to \mathbb{R}$ be differentiable functions such that $|g(x)| \to \infty$ as $x \to c$. Suppose $g'(x) \neq 0$ for all $x \in D$, and $f'(x)/g'(x) \to \ell$ as $x \to c$. Then $f(x)/g(x) \to \ell$ as $x \to c$. Here ℓ can be a real number or ∞ or $-\infty$. ◇

Examples 4.42. (i) $\displaystyle \lim_{x \to \infty} \frac{x^2 + 2x + 3}{3x^2 + 2x + 1} = \lim_{x \to \infty} \frac{2x + 2}{6x + 2} = \lim_{x \to \infty} \frac{2}{6} = \frac{1}{3}$.

(ii) $\displaystyle \lim_{x \to \infty} \frac{x^3}{x^2 - 1} - \frac{x^3}{x^2 + 1} = \lim_{x \to \infty} \frac{2x^3}{x^4 - 1} = \lim_{x \to \infty} \frac{6x^2}{4x^3} = \lim_{x \to \infty} \frac{3}{2x} = 0$. ◇

While L'Hôpital's Rule is extremely useful in computing limits, it is not a panacea! There are situations in which it is not applicable. Quite often, this happens for the simple reason that L'Hôpital's Rule is applied even when one of the conditions for it to hold fails. This is illustrated by the following examples.

Examples 4.43. (i) If $f, g : \mathbb{R} \to \mathbb{R}$ are defined by $f(x) := x$ and $g(x) := \sqrt{1 + x^2}$, then $f(x) \to \infty$ and $g(x) \to \infty$ as $x \to \infty$. If we try to apply L'Hôpital's Rule, we get a loop:

$$\lim_{x \to \infty} \frac{x}{\sqrt{1 + x^2}} = \lim_{x \to \infty} \frac{1}{2x/2\sqrt{1 + x^2}}$$

$$= \lim_{x \to \infty} \frac{\sqrt{1 + x^2}}{x}$$

$$= \lim_{x \to \infty} \frac{2x/2\sqrt{1 + x^2}}{1}$$

$$= \lim_{x \to \infty} \frac{x}{\sqrt{1 + x^2}}.$$

However, the desired limit exists and can be found directly as follows:

$$\lim_{x \to \infty} \frac{x}{\sqrt{1 + x^2}} = \lim_{x \to \infty} \frac{1}{\sqrt{(1/x^2) + 1}} = \frac{1}{\sqrt{0 + 1}} = 1.$$

(ii) If we were to apply L'Hôpital's Rule indiscriminately to calculate limits of quotients such as $(x + 1)/x$ as $x \to 0$, we obtain

$$\lim_{x \to 0} \frac{x + 1}{x} = \lim_{x \to 0} \frac{1}{1} = 1.$$

But in fact, the limit does not exist; indeed,

$$\frac{x + 1}{x} = 1 + \frac{1}{x} \to \infty \text{ as } x \to 0^+ \quad \text{and} \quad \frac{x + 1}{x} \to -\infty \text{ as } x \to 0^-.$$

In this case L'Hôpital's Rule was not applicable since the given quotient is neither in $\frac{0}{0}$ form nor in $\frac{\infty}{\infty}$ form as $x \to 0$.   ◇

Remarks 4.44. (i) Evaluating limits of seemingly different indeterminate forms such as $0 \cdot \infty$ and $\infty - \infty$ is also possible using L'Hôpital's Rule, since such forms can be reduced to $\frac{0}{0}$ indeterminate forms. More precisely, if $c \in \mathbb{R}$ and $f, g : (c - r, c) \to \mathbb{R}$ are differentiable functions such that

$$f(x) \to 0 \quad \text{and} \quad g(x) \to \infty \text{ or } -\infty \quad \text{as } x \to c^-,$$

then there is $\delta > 0$ such that $\delta < r$ and $g(x) \neq 0$ for all $x \in (c - \delta, c)$. Now, $1/g(x) \to 0$ as $x \to c^-$ and

$$f(x)g(x) = \frac{f(x)}{1/g(x)} \quad \text{for } x \in (c - \delta, c),$$

and thus a $0 \cdot \infty$ indeterminate form is converted to a $\frac{0}{0}$ indeterminate form to which L'Hôpital's Rule can be applied. Likewise, if $f(x) \to \infty$ and $g(x) \to \infty$, then there is $\delta > 0$ such that $\delta < r$ and $f(x) > 0$ as well as $g(x) > 0$ for all $x \in (c - \delta, c)$. Now we can write

$$f(x) - g(x) = \frac{(1/g(x)) - (1/f(x))}{(1/f(x)g(x))} \quad \text{for } x \in (c - \delta, c),$$

and thus a $\infty - \infty$ indeterminate form is converted to a $\frac{0}{0}$ indeterminate form to which L'Hôpital's Rule can be applied.

(ii) The power of L'Hôpital's Rule will be especially evident when we add to our repertoire of functions the logarithmic, exponential, and trigonometric functions and try to compute limits involving these functions. This will also enable us to deal with other indeterminate forms such as 0^0, ∞^0, and 1^∞. These variants of L'Hôpital's Rule are explained in Remark 7.12. Examples of limits involving the logarithmic, exponential, and trigonometric functions appear in Example 7.4 (ii), and Exercises 18 and 19 in the list of Revision Exercises at the end of Chapter 7, and in all these, L'Hôpital's Rule is particularly useful. In Examples 7.18 and 7.19, it will be shown that the converse of L'Hôpital's Rule, for $\frac{\infty}{\infty}$ and for $\frac{0}{0}$ indeterminate forms, does not hold in general.   ◇

Notes and Comments

In this chapter, we have derived all the basic properties of differentiation by appealing to a characterization of differentiability in terms of continuity and the relevant properties of continuous functions. With this approach, the proofs seem to become simpler. Another advantage is that we obtain formulas for the sum, product, quotient, composite, and the inverse of functions in the course of proving their differentiability and it is not necessary to know them beforehand. The said characterization of differentiability appears, for example, in the book of Bartle and Sherbert [8], where it is ascribed to Carathéodory, and used to derive the Chain Rule and the Differentiable Inverse Theorem. Here, we have used it more extensively.

The Mean Value Theorem (MVT) and, more generally, Taylor's Theorem are among the most useful results in calculus. The importance of the MVT in calculus mainly stems from the fact that it is crucial in characterizing constant functions, monotonic functions, and convex/concave functions. Such characterizations can be proved using only the mean value inequality, which is obtained here as a corollary of the MVT. On the other hand, it is possible to give an alternative proof of the mean value inequality using properties of Riemann integration and without recourse to the MVT. This has prompted several articles with rather colorful titles. See, for example, the papers by Bers [10], Boas [14], Cohen [17], and Smith [56].

The study of convex functions, which was initiated in Chapter 1 and further continued in this chapter, is now a subject in itself. A quick and elegant introduction can be found in the first chapter of the little classic on gamma functions by Artin [2]. For more on the subject of convex analysis, see the introductory text of van Tiel [63].

Most books on calculus discuss L'Hôpital's Rule for $\frac{0}{0}$ and $\frac{\infty}{\infty}$ indeterminate forms but prove only the former. On the other hand, some relatively advanced books such as Rudin [53] give a sleek proof applicable to both versions at once. A unified proof such as that in Rudin [53] appears to have been inspired by the article of Taylor [61], where it is given as an improved version of a proof by Wazewski. For pedagogical reasons, we have chosen to avoid the sleek unified proof and given instead separate proofs for the two versions. The proof in the $\frac{0}{0}$ case is quite standard and follows quickly from Cauchy's MVT, thanks to our sequential approach to limits. The proof in the $\frac{\infty}{\infty}$ case uses the Intermediate Value Property of derivatives and is essentially based on an argument of Lettenmeyer, which is also outlined in the article of Taylor [61].

Exercises

Part A

1. Use the definition of a derivative to find $f'(x)$ if

(i) $f(x) = x^2$, $x \in \mathbb{R}$,   (ii) $f(x) = 1/x$, $0 \neq x \in \mathbb{R}$,
(iii) $f(x) = \sqrt{x^2 + 1}$, $x \in \mathbb{R}$,   (iv) $f(x) = 1/\sqrt{2x + 3}$, $x \in (-3/2, \infty)$.

2. Let $f : (a, b) \to \mathbb{R}$ be a function such that

$$|f(x + h) - f(x)| \leq C|h|^r \quad \text{for all } x, x + h \in (a, b),$$

where C is a constant and $r \in \mathbb{Q}$ with $r > 1$. Show that f is differentiable on (a, b) and compute $f'(x)$ for $x \in (a, b)$.

3. If $f : (a, b) \to \mathbb{R}$ is differentiable at $c \in (a, b)$, then show that

$$\lim_{h \to 0+} \frac{f(c + h) - f(c - h)}{2h}$$

exists and equals $f'(c)$. Is the converse true?

4. Let $f : (0, \infty) \to \mathbb{R}$ satisfy $f(xy) = f(x) + f(y)$ for all $x, y \in (0, \infty)$. If f is differentiable at 1, show that f is differentiable at every $c \in (0, \infty)$ and $f'(c) = f'(1)/c$. In fact, show that f is infinitely differentiable. If $f'(1) = 2$, find $f^{(n)}(3)$.

5. Let $f : \mathbb{R} \to \mathbb{R}$ satisfy $f(x + y) = f(x)f(y)$ for all $x, y \in \mathbb{R}$. If f is differentiable at 0, then show that f is differentiable at every $c \in \mathbb{R}$ and $f'(c) = f'(0)f(c)$. In fact, show that f is infinitely differentiable. If $f'(0) = 2$, find $f^{(n)}(1)$ for $n \in \mathbb{N}$, in terms of $f(1)$.

6. Let $f, g : \mathbb{R} \to \mathbb{R}$ satisfy $f(x + y) = f(x)g(y) + g(x)f(y)$ and $g(x + y) = g(x)g(y) - f(x)f(y)$ for all $x, y \in \mathbb{R}$. If f and g are differentiable at 0, then show that f and g are differentiable at every $c \in \mathbb{R}$, and we have $f'(c) = g'(0)f(c) + f'(0)g(c)$ and $g'(c) = g'(0)g(c) - f'(0)f(c)$. In fact, show that f and g are infinitely differentiable.

7. Suppose $f, g : \mathbb{R} \to \mathbb{R}$ satisfy $f(x - y) = f(x)g(y) - g(x)f(y)$ and $g(x - y) = g(x)g(y) + f(x)f(y)$ for all $x, y \in \mathbb{R}$. If $f'_+(0)$ exists, then show that f and g are differentiable at every $c \in \mathbb{R}$, and $f'(c) = f'(0)g(c)$ and $g'(c) = -f'(0)f(c)$. In fact, show that f and g are infinitely differentiable. If $f'_+(0) = 2$, find $f^{(n)}(1)$ and $g^{(n)}(1)$ in terms of $f(1)$ and $g(1)$. (Hint: Prove that f is an odd function, g is an even function, f and g are differentiable at 0 and $g'(0) = 0$. Use Exercise 6.)

8. Find the points on the curve $x^2 + xy + y^2 = 7$ at which (i) the tangent is parallel to the x-axis, (ii) the tangent is parallel to the y-axis.

9. Find the equation of the tangent at $(\frac{1}{4}, 4)$ to the parametrically defined curve $x(t) = t^{-2}$, $y(t) = \sqrt{t^2 + 12}$ for $t \in (0.1, 1)$.

10. Find values of the constants a, b, and c for which the graphs of the two functions $f(x) = x^2 + ax + b$ and $g(x) = x^3 - c$, $x \in \mathbb{R}$, intersect at the point $(1, 2)$ and have the same tangent there.

11. Find the tangents to the implicitly defined curve $x^2 y + xy^2 = 6$ at points for which $x = 1$. Also, compute $\frac{d^2 y}{dx^2}$ at these points.

12. Given $n \in \mathbb{N}$, let $f_n : \mathbb{R} \to \mathbb{R}$ be defined by $f_n(x) := x^n$ if $x \geq 0$ and $f_n(x) := -x^n$ if $x < 0$. Show that f_n is $(n - 1)$-times differentiable on $\mathbb{R}$, $f_n^{(n-1)}$ is continuous on $\mathbb{R}$, but $f_n^{(n)}(0)$ does not exist.

13. Let $D \subseteq \mathbb{R}$ be symmetric about the origin, that is, $-x \in D$ whenever $x \in D$. If $c \in D$ and $f : D \to \mathbb{R}$ is either an even or an odd function, then show that the left (hand) derivative $f'_-(c)$ at c exists if and only if the right (hand) derivative $f'_+(-c)$ at $-c$ exists. Further, if either (and hence both) of these derivatives exists, then show that $f'_-(c) = -f'_+(-c)$ if f is even, and $f'_-(c) = f'_+(-c)$ if f is odd. Deduce that if f is differentiable, then f' is an odd (resp. even) function according as f is an even (resp. odd) function.

14. Let I be an interval, $c \in I$, and $f : I \to \mathbb{R}$ be any function. Let, as usual, $|f| : I \to \mathbb{R}$ be the function defined by $|f|(x) = |f(x)|$ for $x \in I$.
    (i) Suppose $(c, c + r) \subseteq I$ for some $r > 0$ and $f'_+(c)$ exists. Then show that $|f|'_+(c)$ exists.
    (ii) Suppose If $(c - r, c) \subseteq I$ for some $r > 0$ and $f'_-(c)$ exists. Then show that $|f|'_-(c)$ exists.
    (iii) Suppose $(c - r, c + r) \subseteq I$ for some $r > 0$ and $f'(c)$ exists. Then show that $|f|'(c)$ exists if and only if either there is $\delta > 0$ such that $\delta \le r$ and $f(x)$ has the same sign for all $x \in (c - \delta, c + \delta)$, or $f(c) = f'(c) = 0$.

15. Let $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$ be two points on the curve $y = ax^2 + bx + c$. If $P_3 = (x_3, y_3)$ lies on the arc $P_1 P_2$ and the tangent to the curve at P_3 is parallel to the chord $P_1 P_2$, show that $x_3 = (x_1 + x_2)/2$.

16. Show that the x-axis is a normal to the curve $y^2 = x$ at $(0, 0)$. If three normals can be drawn to this curve from a point $(a, 0)$, show that a must be greater than $\frac{1}{2}$. Find the value of a such that the two normals, other than the x-axis, are perpendicular to each other.

17. Let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). If $f(a)$ and $f(b)$ are of different signs and $f'(x) \ne 0$ for all $x \in (a, b)$, then show that there is a unique $x_0 \in (a, b)$ such that $f(x_0) = 0$.

18. Show that the cubic $2x^3 + 3x^2 + 6x + 10$ has exactly one real root.

19. Let $n \in \mathbb{N}$ and $f : [a, b] \to \mathbb{R}$ be such that $f^{(n-1)}$ is continuous on $[a, b]$ and $f^{(n)}$ exists in (a, b). If f vanishes at $n + 1$ distinct points in $[a, b]$, then show that $f^{(n)}$ vanishes at least once in (a, b).

20. Let $f : [-\frac{1}{2}, \frac{1}{2}] \to \mathbb{R}$ be given by

$$f(x) = \begin{cases} \sqrt{2x - x^2} & \text{if } 0 \le x \le \frac{1}{2}, \\ \sqrt{-2x - x^2} & \text{if } -\frac{1}{2} \le x \le 0. \end{cases}$$

Show that $f(\frac{1}{2}) = f(-\frac{1}{2})$ but $f'(x) \ne 0$ for all x with $0 < |x| < \frac{1}{2}$. Does this contradict Rolle's Theorem? Justify your answer.

21. Let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). If $f(a) < f(b)$, then show that $f'(c) > 0$ for some $c \in (a, b)$.

22. Let $a > 0$ and $f : [-a, a] \to \mathbb{R}$ be continuous. Suppose $f'(x)$ exists and $f'(x) \le 1$ for all $x \in (-a, a)$. If $f(a) = a$ and $f(-a) = -a$, then show that $f(x) = x$ for every $x \in (-a, a)$.

23. In each of the following cases, find a function f that satisfies all the given conditions, or else show that no such function exists.

(i) $f''(x) > 0$ for all $x \in \mathbb{R}$, $f'(0) = 1$, $f'(1) = 1$,

(ii) $f''(x) > 0$ for all $x \in \mathbb{R}$, $f'(0) = 1$, $f'(1) = 2$,

(iii) $f''(x) \geq 0$ for all $x \in \mathbb{R}$, $f'(0) = 1$, $f(x) \leq 100$ for all $x > 0$,

(iv) $f''(x) > 0$ for all $x \in \mathbb{R}$, $f'(0) = 1$, $f(x) \leq 1$ for all $x < 0$.

24. Let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). Suppose $f(a) = a$ and $f(b) = b$. Show that there is $c \in (a, b)$ such that $f'(c) = 1$. Further, show that there are distinct $c_1, c_2 \in (a, b)$ such that $f'(c_1) + f'(c_2) = 2$. More generally, show that for every $n \in \mathbb{N}$, there are n distinct points $c_1, \ldots, c_n \in (a, b)$ such that $f'(c_1) + \cdots + f'(c_n) = n$.

25. Let a function $f : [a, b] \to \mathbb{R}$ be continuous and its second derivative f'' exist everywhere on the open interval (a, b). Suppose the line segment joining $(a, f(a))$ and $(b, f(b))$ intersects the graph of f at a third point $(c, f(c))$, where $a < c < b$. Prove that $f''(t) = 0$ for some $t \in (a, b)$.

26. Use the MVT to prove that for all $n \in \mathbb{N}$ and $a, b \in \mathbb{R}$ such that $0 < a \leq b$, we have $na^{n-1}(b - a) \leq b^n - a^n \leq nb^{n-1}(b - a)$.

27. Use the MVT to prove that

$$\frac{1}{3(m+1)^{2/3}} < (m+1)^{1/3} - m^{1/3} < \frac{1}{3m^{2/3}} \quad \text{for all } m \in \mathbb{N}.$$

28. Use the MVT to prove the following inequalities.

(i) $\dfrac{27}{16} < \sqrt{3} < \dfrac{7}{4}$ and $\dfrac{20}{9} < \sqrt{5} < \dfrac{9}{4}$.

(ii) $\dfrac{19}{16} < 2^{1/3} < \dfrac{4}{3}$, $\dfrac{17}{9} < 7^{1/3} < \dfrac{23}{12}$ and $\dfrac{1298}{625} < 9^{1/3} < \dfrac{25}{12}$.

29. Use the MVT to show that $10.049 < \sqrt{101} < 10.05$ and $10.24 < \sqrt{105} < 10.25$. Also, find better estimates using Taylor's Theorem with $n = 1$, that is, using the Extended MVT.

30. Let $f : (a, b) \to \mathbb{R}$ and $c \in (a, b)$ be such that f is continuous at c and $f'(x)$ exist for every $x \in (a, c) \cup (c, b)$. If $\lim_{x \to c} f'(x)$ exists, then show that $f'(c)$ exists and is equal to this limit.

31. (i) Let $f, g : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). If $f(a) \leq g(a)$ and $f'(x) \leq g'(x)$ for all $x \in (a, b)$, then show that $f(b) \leq g(b)$.

(ii) Use (i) to show that $15x^2 \leq 8x^3 + 12 \leq 18x^2$ for all $x \in [1.25, 1.5]$. Deduce that the range of the function $h : [1.25, 1.5] \to \mathbb{R}$ given by $h(x) = (2x^3 + 3)/3x^2$ is contained in $[1.25, 1.5]$.

32. Find the nth Taylor polynomial of f around a, that is,

$$P_n(x) = f(a) + f'(a)(x - a) + \cdots + \frac{f^{(n)}(a)}{n!}(x - a)^n \quad \text{for } x \in \mathbb{R},$$

when $a = 0$ and $f(x)$ equals:

(i) $\dfrac{1}{1 - x}$,   (ii) $\dfrac{1}{1 + x}$,   (iii) $\dfrac{x}{1 + x^2}$.

33. Let I be an interval containing more than one point and $f : I \to \mathbb{R}$ be any function.

(i) Assume that f is differentiable. If f' is nonnegative on I and f' vanishes at only a finite number of points on any bounded subinterval of I, then show that f is strictly increasing on I.

(ii) Assume that f is twice differentiable. If f'' is nonnegative on I and f'' vanishes at only a finite number of points on any bounded subinterval of I, then show that f is strictly convex on I.

(iii) Consider $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = (x - 2n)^3 + 2n$, where $n \in \mathbb{Z}$ is such that $x \in [2n-1, 2n+1)$. Show that f is differentiable on $\mathbb{R}$ and f'' exists on $(2n-1, 2n+1)$, but $f''_+(2n+1) = 6$, whereas $f''_-(2n+1) = -6$ for each $n \in \mathbb{N}$. Also show that f is strictly increasing on $\mathbb{R}$ although $f'(2n) = 0$ for each $n \in \mathbb{N}$. (Compare (i) above and Exercise 12 in the list of Revision Exercises at the end of Chapter 7.)

(iv) Consider $g : \mathbb{R} \to \mathbb{R}$ given by $g(x) = (x - 2n)^4 + 8nx$, where $n \in \mathbb{Z}$ is such that $x \in [2n - 1, 2n + 1)$. Show that g is twice differentiable on $\mathbb{R}$ and g''' exists on $(2n - 1, 2n + 1)$, but $g'''_+(2n + 1) = 24$, whereas $g'''_-(2n + 1) = -24$ for each $n \in \mathbb{N}$. Also show that g is strictly convex on $\mathbb{R}$ although $g''(2n) = 0$ for each $n \in \mathbb{N}$. (Compare (ii) above and Exercise 13 in the list of Revision Exercises at the end of Chapter 7.)

34. Let I be an interval in $\mathbb{R}$ and $c \in I$ be an interior point. If $f : I \to \mathbb{R}$ is monotonically increasing and if the left and right derivatives of f at c, namely $f'_-(c)$ and $f'_+(c)$, exist, then show that $f'_-(c) \geq 0$ and $f'_+(c) \geq 0$. Further, give examples of monotonically increasing functions $f : I \to \mathbb{R}$ for which $f'_-(c) < f'_+(c)$ or for which $f'_-(c) > f'_+(c)$.

35. Let $f : [a, b] \to \mathbb{R}$ be such that f' is continuous on $[a, b]$ and f'' exists on (a, b). Show that there is $c \in (a, b)$ such that

$$f''(c)[f(b) - f(a)] = f'(c)[f'(b) - f'(a)].$$

36. Let $f, g, h : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). Show that there is $c \in (a, b)$ such that the 3×3 determinant

$$\begin{vmatrix} f(a) & f(b) & f'(c) \\ g(a) & g(b) & g'(c) \\ h(a) & h(b) & h'(c) \end{vmatrix}$$

is zero, that is, $f(a)[g(b)h'(c) - h(b)g'(c)] - f(b)[g(a)h'(c) - h(a)g'(c)] + f'(c)[g(a)h(b) - h(a)g(b)] = 0$. Deduce that if $h(x) = 1$ for all $x \in [a, b]$, we obtain the conclusion of Cauchy's Mean Value Theorem (Proposition 4.36). What does the result say if $g(x) = x$ and $h(x) = 1$ for all $x \in [a, b]$?

37. Let $f, g : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). If there is $\alpha \in \mathbb{R}$ such that $|f'(x)| \leq \alpha |g'(x)|$ for all $x \in (a, b)$ and if $g'(x) \neq 0$ for all $x \in (a, b)$, then show that $|f(b) - f(a)| \leq \alpha |g(b) - g(a)|$. Is the conclusion valid if the condition "$g'(x) \neq 0$ for all $x \in (a, b)$" is omitted?

38. Evaluate the following limits:

(i) $\displaystyle\lim_{x \to 1} \frac{(2x - x^4)^{1/2} - x^{1/3}}{1 - x^{3/4}},$ (ii) $\displaystyle\lim_{x \to \infty} \frac{5x^2 - 3x}{7x^2 + 1},$

(iii) $\lim_{x \to \infty} \left(x - \sqrt{x + x^2} \right)$,   (iv) $\lim_{x \to \infty} \dfrac{\sqrt{x + 2}}{\sqrt{x + 1}}$.

39. Show that if $f, g : \mathbb{R} \to \mathbb{R}$ are functions defined by

$$f(x) = \begin{cases} x + 2 & \text{if } x \neq 0, \\ 0 & \text{if } x = 0, \end{cases} \quad \text{and} \quad g(x) = \begin{cases} x + 1 & \text{if } x \neq 0, \\ 0 & \text{if } x = 0, \end{cases}$$

then

$$\lim_{x \to 0} \frac{f'(x)}{g'(x)} = 1 \quad \text{but} \quad \lim_{x \to 0} \frac{f(x)}{g(x)} = 2.$$

Does this contradict L'Hôpital's Rule?

40. Consider the following application of L'Hôpital's Rule:

$$\lim_{x \to 1} \frac{3x^2 - 2x - 1}{x^2 - x} = \lim_{x \to 1} \frac{6x - 2}{2x - 1} = \lim_{x \to 1} \frac{6}{2} = 3.$$

Is it correct? Justify.

41. Consider $f : \mathbb{R} \setminus \{1\} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) := \frac{1}{x - 1} \quad \text{for } x \neq 1 \quad \text{and} \quad g(x) := x \quad \text{for } x \in \mathbb{R}.$$

Show that

$$\frac{f'(x)}{g'(x)} \to -\infty \text{ as } x \to 1^+, \quad \text{but} \quad \frac{f(x)}{g(x)} \to \infty \text{ as } x \to 1^+.$$

Does this contradict L'Hôpital's Rule? Justify.

Part B

42. Let $f : (a, b) \to \mathbb{R}$ and $c \in (a, b)$. Show that the following are equivalent:
    (i) f is differentiable at c.
    (ii) There exist $\alpha \in \mathbb{R}$, $\delta > 0$ and a function $\epsilon_1 : (-\delta, \delta) \to \mathbb{R}$ such that

$$f(c + h) = f(c) + \alpha h + h \epsilon_1(h) \text{ for all } h \in (-\delta, \delta) \text{ and } \lim_{h \to 0} \epsilon_1(h) = 0.$$

    (iii) There exists $\alpha \in \mathbb{R}$ such that

$$\lim_{h \to 0} \frac{|f(c + h) - f(c) - \alpha h|}{|h|} = 0.$$

    If the above conditions hold, then show that $f'(c) = \alpha$.

43. Let C be an algebraic plane curve, that is, let C be implicitly defined by $F(x, y) = 0$, where $F(x, y)$ is a nonzero polynomial in two variables x and y with coefficients in $\mathbb{R}$. Let the (total) degree of $F(x, y)$ be n. Let $P = (x_0, y_0)$ be a point on C, so that $F(x_0, y_0) = 0$.

(i) If we let $X := x - c$ and $Y := y - d$ and define $g(X,Y) := f(x,y)$, then show that $g(X,Y)$ is a polynomial in X and Y with $g(0,0) = 0$. Deduce that there is a unique $m \in \mathbb{N}$ such that $m \leq n$ and

$$g(X,Y) = g_m(X,Y) + g_{m+1}(X,Y) + \cdots + g_n(X,Y),$$

where $g_i(X,Y)$ is either the zero polynomial or a nonzero homogeneous polynomial of degree i, for $m \leq i \leq n$, and $g_m(X,Y) \neq 0$. We denote the integer m by $\mathrm{mult}_P(C)$, and call it the **multiplicity** of C at the point P.

(ii) Show that a tangent to the curve C at the point P is defined (as far as calculus is concerned) if and only if $\mathrm{mult}_P(C) = 1$. Moreover, if $\mathrm{mult}_P(C) = 1$, then there are $\alpha_1, \beta_1 \in \mathbb{R}$ such that $g_1(X,Y) = \alpha_1 X + \beta_1 Y$, and then the line $\alpha_1(x-c) + \beta_1(y-d) = 0$ is the tangent to C at P.

(iii) Show that if $F(x,y) = y - f(x)$ for some polynomial $f(x)$ in one variable x, then for the corresponding curve C given by $F(x,y) = 0$, we have $\mathrm{mult}_P(C) = 1$ for every P on C.

(iv) Determine the integer $m = \mathrm{mult}_P(C)$ and a factorization of $g_m(X,Y)$ when $P = (0,0)$ and C is the curve implicitly defined by $F(x,y) := y^2 - x^2 - x^3 = 0$, or by $F(x,y) := y^2 - x^3 = 0$.

[Note: In view of Exercise 70 of Chapter 1, the initial form $g_m(X,Y)$ factors as a product of homogeneous linear polynomials, that is,

$$g_m(X,Y) = \prod_{i=1}^{m}(\alpha_i X + \beta_i Y) \quad \text{for some } \alpha_i, \beta_i \in \mathbb{C}, \ 1 \leq i \leq m.$$

In the algebraic approach to tangents, the m (complex) lines given by $\alpha_i(x - c) + \beta_i(y - d) = 0$ for $i = 1, \ldots, m$, are called the tangent lines to the curve C at the point P.]

44. Let I be an interval and $f : I \to \mathbb{R}$ be continuous on I and differentiable at every interior point of I. If there is a constant α such that $|f'(x)| \leq \alpha$ for all interior points x of I, then show that f is uniformly continuous on I. Is the converse true? In other words, is it true that if $f : I \to \mathbb{R}$ is uniformly continuous on I and differentiable at every interior point of I, then there is a constant α such that $|f'(x)| \leq \alpha$ for all interior points x of I?

45. Let $f : [a,b] \to \mathbb{R}$ be such that f' is continuous on $[a,b]$ and f'' exists on (a,b). Given any $\xi \in [a,b]$, show that there is $c \in (a,b)$ such that

$$f(\xi) - f(a) = \frac{f(b) - f(a)}{b - a}(\xi - a) + \frac{f''(c)}{2}(\xi - a)(\xi - b).$$

46. Let $f(x)$ be a polynomial. A real number c is called a **root** of $f(x)$ of **multiplicity** m if $f(x) = (x - c)^m g(x)$ for some polynomial $g(x)$ such that $g(c) \neq 0$.

(i) Let $f(x)$ have r roots (counting multiplicities) in an open interval (a, b). Show that the polynomial $f'(x)$ has at least $r-1$ roots in (a, b). Also, give an example where $f'(x)$ has more than $r-1$ roots in (a, b). More generally, for $k \in \mathbb{N}$, show that the polynomial $f^{(k)}(x)$ has at least $r-k$ roots in (a, b).

(ii) If $f^{(k)}(x)$ has s roots in (a, b), what can you conclude about the number of roots of $f(x)$ in (a, b)?

47. Let $f(x)$ be a polynomial of degree n. Given any $a \in \mathbb{R}$, show that

$$f(x) = f(a) + f'(a)(x - a) + \cdots + \frac{f^{(n)}(a)}{n!}(x - a)^n, \quad \text{for } x \in \mathbb{R}.$$

Deduce that a is a root of $f(x)$ of multiplicity m if and only if $f(a) = f'(a) = \cdots = f^{(m-1)}(a) = 0$ and $f^{(m)}(a) \neq 0$. Further, show that if a is a root of f of multiplicity m, then

$$\lim_{h \to 0} \frac{f(a + h) - f(a)}{h^m} = \frac{f^{(m)}(a)}{m!}.$$

48. Give an alternative proof of Taylor's Theorem with a single application of Rolle's Theorem by proceeding as follows. Let the notation and hypothesis be as in the statement of Taylor Theorem (Proposition 4.23). Also, as in the proof of Taylor's Theorem, for $x \in [a, b]$, let

$$P(x) = f(a) + f'(a)(x - a) + \frac{f''(a)}{2!}(x - a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x - a)^n.$$

Define $g : [a, b] \to \mathbb{R}$ by

$$g(x) = f(x) + f'(x)(b-x) + \frac{f''(x)}{2!}(b-x)^2 + \cdots + \frac{f^{(n)}(x)}{n!}(b-x)^n + s(b-x)^{n+1},$$

where $s = [f(b) - P(b)]/(b - a)^{n+1}$. Show that $g(a) = g(b) = f(b)$. Apply Rolle's Theorem to g to deduce Taylor's Theorem.

49. Let the notation and hypothesis be as in the statement of Taylor's Theorem (Proposition 4.23). Given any $p \in \mathbb{N}$ with $p \leq n+1$, show that there is $c \in (a, b)$ such that

$$f(b) = f(a) + f'(a)(b-a) + \cdots + \frac{f^{(n)}(a)}{n!}(b-a)^n + \frac{f^{(n+1)}(c)}{n!p}(b-a)^p(b-c)^{n-p+1}.$$

[Hint: Proceed as in the previous exercise except to change the $(n+1)$th power to the pth power in the definitions of $g(x)$ and s.] Show that Taylor's Theorem is a special case of this result with $p = n+1$. Further, show that if I is any interval containing more than one point, a is any point of I, and $f : I \to \mathbb{R}$ is such that $f', f'', \ldots, f^{(n)}$ exist on I and $f^{(n+1)}$ exists at every interior point of I, then for any $x \in I$, there is c between a and x such that

$$f(x) = P_n(x) + R_{n,p}(x), \quad \text{where} \quad R_{n,p}(x) = \frac{f^{(n+1)}(c)}{n!p}(x-a)^p(x-c)^{n-p+1},$$

and where $P_n(x)$ is the nth Taylor polynomial of f around a.

[Note: The remainder term in the above result, namely $R_{n,p}(x)$, is called the **Schlömilch form of remainder**. It reduces to the Lagrange form of remainder when $p = n + 1$, whereas it is called the **Cauchy form of remainder** when $p = 1$.]

50. Let I be an interval containing more than one point and $f : I \to \mathbb{R}$ be a convex function.

(i) Show that for every interior point c of I, both $f'_-(c)$ and $f'_+(c)$ exist and $f'_-(c) \le f'_+(c)$. (Hint: Use Exercise 72 of Chapter 1.)

(ii) Show that for any $x_1, x_2 \in I$ with $x_1 < x_2$, we have $f'_+(x_1) \le f'_-(x_2)$.

51. Let $m \in \mathbb{N}$ and $f, g : [a, b] \to \mathbb{R}$ be such that $f, f', \ldots, f^{(m-1)}$ as well as $g, g', \ldots, g^{(m-1)}$ are continuous on $[a, b]$ and $f^{(m)}, g^{(m)}$ exist on (a, b). Suppose $f'(a) = f''(a) = \cdots = f^{(m-1)}(a) = 0$ and $g'(a) = g''(a) = \cdots = g^{(m-1)}(a) = 0$, but $g^{(m)}(x) \ne 0$ for all $x \in (a, b)$. Prove that there exist $c_1, \ldots, c_m \in (a, b)$ such that

$$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(c_1)}{g'(c_1)} = \frac{f''(c_2)}{g''(c_2)} = \cdots = \frac{f^{(m)}(c_m)}{g^{(m)}(c_m)}.$$

[Note: This generalizes Cauchy's Mean Value Theorem.]

52. Let $c \in \mathbb{R}$, $r > 0$, and $f : (c - r, c + r) \to \mathbb{R}$ be such that $f''(c)$ exists. Show that

$$\lim_{h \to 0+} \frac{f(c + h) + f(c - h) - 2f(c)}{h^2}$$

exists and is equal to $f''(c)$. Give an example of a function that is differentiable on $(c - r, c + r)$, for which this limit exists, but $f''(c)$ does not exist. (Hint: L'Hôpital's Rule.)

53. Let $c \in \mathbb{R}$, $r > 0$, $f : (c - r, c + r) \to \mathbb{R}$, and $n \in \mathbb{N}$ be such that $f^{(n)}(c)$ exists. Show that

$$\lim_{h \to 0} \frac{f(c + h) - f(c) - hf'(c) - \cdots - h^{n-1}\left[f^{(n-1)}(c)/(n-1)!\right]}{h^n} = \frac{f^{(n)}(c)}{n!}.$$

(Hint: L'Hôpital's Rule.)

5

Applications of Differentiation

The notion of differentiation is remarkably effective in studying the geometric properties of functions. We have seen already how derivatives are useful in determining monotonicity, convexity, or concavity for differentiable functions defined on an interval. We shall study similar applications in this chapter.

First, in Section 5.1 we will see how one can determine the absolute (or global) minimum or maximum of a large class of functions. Next, in Section 5.2 we shall describe a number of useful tests to determine the local minima or maxima of a function and to detect the points of inflection. In this way, we shall be able to locate the ups and downs, the peaks and dips, and the twists and turns in the graph of a real-valued function. This information is extremely useful in curve sketching, that is, in drawing graphs of real-valued functions and identifying their key features. In Section 5.3, we revisit the idea of approximating functions by simpler functions, which we discussed in Chapter 4 in connection with the MVT and Taylor's Theorem. We shall discuss here in greater detail the most widely used methods of approximation, namely, linear and quadratic approximations. Finally, in Section 5.4, we discuss a method of Picard for finding fixed points of functions, and a method of Newton for finding zeros of functions.

5.1 Absolute Minimum and Maximum

We have seen in Proposition 3.8 that a continuous real-valued function defined on a closed and bounded subset of $\mathbb{R}$ is bounded and attains its bounds. In other words, if $D \subseteq \mathbb{R}$ is closed and bounded, and $f : D \to \mathbb{R}$ is continuous, then the **absolute minimum** and the **absolute maximum** of f on D, namely,

$$m := \min\{f(x) : x \in D\} \quad \text{and} \quad M := \max\{f(x) : x \in D\},$$

exist, and moreover, there are $r, s \in D$ such that $m = f(r)$ and $M = f(s)$. A question that arises naturally is the following: Knowing the function f, how

does one find the **absolute extrema** m and M, and the points r and s where they are attained? It turns out that we can considerably narrow down the search for the points where the absolute extrema are attained if we look at the derivative of f. To make this precise, let us first formulate a couple of definitions.

Given $D \subseteq \mathbb{R}$ and $f : D \to \mathbb{R}$, a point $c \in D$ is called a **critical point** of f if c is an interior point of D such that either f is not differentiable at c, or f is differentiable at c and $f'(c) = 0$.

Given $D \subseteq \mathbb{R}$, by a **boundary point** of D we shall mean a real number c such that for every $r > 0$, the interval $(c - r, c + r)$ contains a point of D as well as a point not belonging to D. For example, if $D = [a, b]$, then the endpoints a and b are the boundary points of D, whereas the points of (a, b) are the interior points of D.

Proposition 5.1. *Let D be a closed and bounded subset of $\mathbb{R}$, and $f : D \to \mathbb{R}$ be a continuous function. Then the absolute minimum as well as the absolute maximum of f is attained either at a critical point of f or at a boundary point of D.*

Proof. By Proposition 3.8, f attains its absolute minimum as well as its absolute maximum on D. Let $c \in D$ be a point at which the absolute minimum of f is attained. Suppose c is an interior point of D. Then clearly, f has a local minimum at c. Hence if f is differentiable at c, then by Lemma 4.13, $f'(c) = 0$. It follows that c must be a critical point of D. Thus, c is either a critical point of f or a boundary point of D.

A similar argument applies to a point at which the absolute maximum of f is attained.   □

In practice, the critical points of a function and the boundary points of its domain are few in number. Thus, in view of the above proposition, we have a simple recipe to determine the absolute extrema and the points where they are attained. Namely, determine the critical points of a function and the boundary points of its domain; then calculate the values at these points, and compare these values. The greatest value among them is the absolute maximum, while the least value is the absolute minimum. This recipe is illustrated by the following examples.

Examples 5.2. (i) Consider $f : [-1, 2] \to \mathbb{R}$ defined by

$$f(x) := \begin{cases} -x & \text{if } -1 \leq x \leq 0, \\ 2x^3 - 4x^2 + 2x & \text{if } 0 < x \leq 2. \end{cases}$$

Let us try to find the absolute extrema of f. First, note that by Proposition 3.5, f is continuous on $[0, 2]$. Next, f is not differentiable at 0 since $f'_-(0) = -1$ and $f'_+(0) = 2$. On the other hand,

$$f'(x) = \begin{cases} -1 & \text{if } -1 \leq x < 0, \\ 6x^2 - 8x + 2 = 2(3x - 1)(x - 1) & \text{if } 0 < x < 2. \end{cases}$$

So, $f'(x) = 0$ only at $x = \frac{1}{3}$ and $x = 1$. It follows that $x = 0$, $x = \frac{1}{3}$ and $x = 1$ are the only critical points of f. The boundary points of our domain $[-1, 2]$ are -1 and 2. Thus we make the following table.

x	-1	0	$\dfrac{1}{3}$	1	2
$f(x)$	1	0	$\dfrac{8}{27}$	0	4

From this we conclude that the absolute minimum of f is 0, which is attained at $x = 0$ as well as at $x = 1$, whereas the absolute maximum of f is 4, which is attained at $x = 2$. Note that although f has a local maximum at $x = \frac{1}{3}$, it is not the absolute maximum of f.
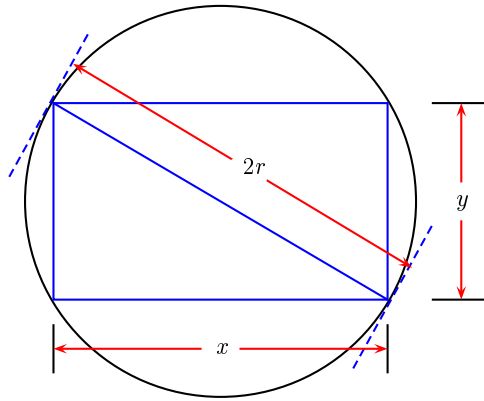
Fig. 5.1. Rectangle inscribed in a circle of radius r

(ii) Let us show that among all rectangles that can be inscribed in a given circle, the square has the greatest area. Let r be the radius of the given circle. If x is the length and y the breadth of an inscribed rectangle [see Figure 5.1], then $0 \le x, y \le 2r$ and $x^2 + y^2 = (2r)^2 = 4r^2$. Now, the area of the rectangle could be viewed as a function $A : [0, 2r] \to \mathbb{R}$ given by $A(x) := xy = x\sqrt{4r^2 - x^2}$. To compute $A'(x)$ we may use implicit differentiation. For example, the equation $x^2 + y^2 = 4r^2$ implies that

$$2x + 2y\frac{dy}{dx} = 0,$$

and hence at points where $y \ne 0$, that is, $x \ne 2r$, we obtain

$$\frac{dA}{dx} = y + x\frac{dy}{dx} = y - \frac{x^2}{y} = \frac{y^2 - x^2}{y} = \frac{4r^2 - 2x^2}{y}.$$

For $0 \leq x < 2r$, we have

$$\frac{dA}{dx} = 0 \iff x = \sqrt{2}\,r.$$

Thus $x = \sqrt{2}\,r$ is the only critical point of A, and so we make the following table:

x	0	$\sqrt{2}\,r$	$2r$
$A(x)$	0	$2r^2$	0

From this we conclude that the area $A(x)$ of the rectangle is maximal when $x = \sqrt{2}\,r = y$, that is, when the rectangle is, in fact, a square. $\diamond$

5.2 Local Extrema and Points of Inflection

Heuristically speaking, a local maximum of a function corresponds to a peak or a pinnacle in its graph, while a local minimum is something like a dip or a depression. Let us also recall the formal definition from Chapter 1. Namely, if $D \subseteq \mathbb{R}$ and c is an interior point in D, then $f : D \to \mathbb{R}$ is said to have a **local minimum** at c if there is $\delta > 0$ such that

$$(c - \delta, c + \delta) \subseteq D \quad \text{and} \quad f(x) \geq f(c) \text{ for all } x \in (c - \delta, c + \delta).$$

On the other hand, f is said to have a **local maximum** at c if there is $\delta > 0$ such that

$$(c - \delta, c + \delta) \subseteq D \quad \text{and} \quad f(x) \leq f(c) \text{ for all } x \in (c - \delta, c + \delta).$$

Also, recall that f is said to have a **strict local minimum** [resp. **strict local maximum**] at c if there is $\delta > 0$ such that $(c - \delta, c + \delta) \subseteq D$ and $f(x) > f(c)$ [resp. $f(x) < f(c)$] for all $x \in (c - \delta, c + \delta)$, $x \neq c$.

For example, consider a function whose graph looks as in Figure 5.2. In fact, this is the graph of the function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) := \begin{cases} 8 & \text{if } x \leq -2, \\ x^4 - 2x^2 & \text{if } x \in (-2, 2), \\ 10 - x & \text{if } x \geq 2. \end{cases}$$

We see that at $x = -1$ and $x = 1$, the function has a strict local minimum, whereas at $x = 0$ and $x = 2$, it has a strict local maximum. At $x = -2$, it has a local maximum that is not strict. In fact, there is a (nonstrict) local minimum as well as a local maximum at every point in $(-\infty, -2)$ where the function is constant, that is, its graph has a plateau.
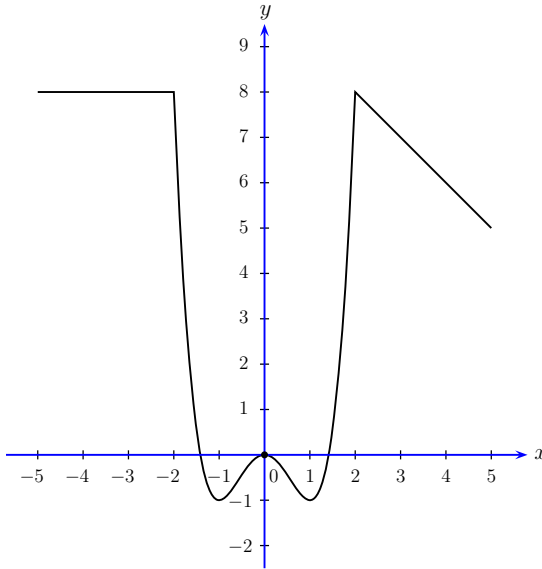
Fig. 5.2. Illustration of local extrema as peaks and dips

To get an idea of the relation between derivatives and the notions of local minimum/maximum, we may look at the behavior of the above graph around its peaks or dips (or even a plateau). We see that as we approach a dip (local minimum) from the left, the graph is decreasing and the tangents have negative slopes, whereas as we approach it from the right, the graph is increasing and the tangents have positive slopes. Similarly, as we approach a peak (local maximum) from the left, the graph is increasing and the tangents have positive slopes, whereas as we approach it from the right, the graph is decreasing and the tangents have negative slopes. In case the tangent is defined at a peak or a dip, then it is necessarily horizontal, that is, it has slope zero. We have already seen the analytic formulation of the latter property in the form of Lemma 4.13. This lemma says that if $f : D \to \mathbb{R}$ is differentiable at an interior point c of $D \subseteq \mathbb{R}$, then the vanishing of $f'(c)$ is a **necessary condition** for f to have a local extremum at c. We have seen examples $[f(x) := x^3$ for $x \in \mathbb{R}; c = 0]$ that show that this condition is not sufficient to guarantee a local extremum. However, the above observations about the behavior of the graph lead to some **sufficient conditions for a local extremum**. We first state the result for a local minimum.

Proposition 5.3. *Let $D \subseteq \mathbb{R}$, c be an interior point of D, and $f : D \to \mathbb{R}$ be any function. Then we have the following:*

(i) [**First Derivative Test for Local Minimum**] *If f is continuous at c, and also,*

(a) f is differentiable on $(c-r, c) \cup (c, c+r)$ for some $r > 0$, and
(b) there is $\delta > 0$ with $\delta \leq r$ such that $f'(x) \leq 0$ for all $x \in (c-\delta, c)$, and $f'(x) \geq 0$ for all $x \in (c, c+\delta)$,
then f has a local minimum at c.

(ii) [**Second Derivative Test for Local Minimum**] If f is twice differentiable at c and satisfies $f'(c) = 0$ and $f''(c) > 0$, then f has a local minimum at c.

Proof. (i) If the conditions in (i) are satisfied and $\delta > 0$ is as in subpart (b) of (i), then by parts (i) and (ii) of Proposition 4.27, we see that f is decreasing on $(c-\delta, c)$ and increasing on $(c, c+\delta)$. Now, the continuity of f at c implies that $f(x) \to f(c)$ as $x \to c$ (Proposition 3.21), and hence it follows that $f(x) \geq f(c)$ for all $x \in (c-\delta, c+\delta)$. Thus, f has a local minimum at c.

(ii) If $f''(c)$ exists, then it is tacitly assumed that f' exists on $(c-r, c+r)$ for some $r > 0$ with $(c-r, c+r) \subseteq D$. Now, if $f'(c) = 0$ and $f''(c) > 0$, then

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

Thus, by part (i) of Proposition 3.24, there is $\delta > 0$ such that $\delta < r$ and

$$\frac{f'(x)}{x-c} > 0 \quad \text{for all } x \in (c-\delta, c) \cup (c, c+\delta).$$

In view of this, we see that f satisfies the conditions in (i) above. Hence f has a local minimum at c.   □

The corresponding result for a local maximum is similar, and is stated below for ease of reference.

Proposition 5.4. Let $D \subseteq \mathbb{R}$, c be an interior point of D, and $f : D \to \mathbb{R}$ be any function. Then we have the following:

(i) [**First Derivative Test for Local Maximum**] If f is continuous at c, and also,
    (a) f is differentiable on $(c-r, c) \cup (c, c+r)$ for some $r > 0$, and
    (b) there is $\delta > 0$ with $\delta \leq r$ such that $f'(x) \geq 0$ for all $x \in (c-\delta, c)$, and $f'(x) \leq 0$ for all $x \in (c, c+\delta)$,
    then f has a local maximum at c.
(ii) [**Second Derivative Test for Local Maximum**] If f is twice differentiable at c and satisfies $f'(c) = 0$ and $f''(c) < 0$, then f has a local maximum at c.

Proof. Similar to the proof of Proposition 5.3.   □

Remarks 5.5. (i) An informal, but easy, way to remember the First Derivative Test (for local minimum as well as local maximum) is as follows:

f' changes from $-$ to $+$ at $c \Longrightarrow f$ has a local minimum at c;
f' changes from $+$ to $-$ at $c \Longrightarrow f$ has a local maximum at c;

Note, however, that apart from differentiability in an open interval about c, except possibly at c, the continuity at the point c is essential. For example, we can use this test to ascertain that the absolute value function has a local minimum at 0. Likewise, it can be used to ascertain that $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) := -|x|$ has a local maximum at 0. However, it can not be applied to check whether the integer part function has a local extremum at, say, $x = 0$. [See Figure 1.5.]

(ii) The Second Derivative Test (for local minimum as well as local maximum) is valid under a restrictive hypothesis, namely, twice differentiability, and usually needs more checking (values of both the derivatives). But it has the advantage of being short and easy to remember.

(iii) While the First Derivative Test and the Second Derivative Test provide sufficient conditions for a local extremum, neither of them is necessary, that is, a function can have a local extremum at a point but may not satisfy the hypothesis of either of these tests. We have, in fact, seen in Chapter 1 an example of a function [the piecewise linear zigzag function in Example 1.18] that has a local minimum at 0 but is not decreasing on $(-\delta, 0]$ and increasing on $[0, \delta)$ for any $\delta > 0$. It can easily be seen that this function does not satisfy the hypothesis of the First Derivative Test as well as of the Second Derivative Test, even though it has a local minimum at 0. Easier counterexamples appear in Examples 5.6 (ii) and (iii) below. Notice that the negative of these functions provide examples of functions that have a local maximum but do not satisfy the hypothesis of any of the tests.

(iv) If in subpart (b) of the First Derivative Test for local minimum, we change the inequalities '$f'(x) \leq 0$' and '$f'(x) \geq 0$' to the corresponding strict inequalities '$f'(x) < 0$' and '$f'(x) > 0$', then we can conclude that f has a strict local minimum at the corresponding point. Similarly in the case of a local maximum. More generally, we can reach the conclusion about f having a strict local extremum if in addition to the hypothesis of the First Derivative Test, we require that f' not vanish identically on any subinterval of $(c - \delta, c)$ or of $(c, c + \delta)$ containing more than one point. Thus, f' is allowed to vanish at a few stray points but not on a continuous segment with c as one of its endpoints. On the other hand, if the hypothesis of the Second Derivative Test is satisfied, then we have, in fact, a strict local extremum. These assertions are easily proved by gleaning through the proofs of Propositions 5.3 and 5.4 and using Propositions 4.27 and 4.30. $\diamond$

Examples 5.6. (i) Consider $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) := \frac{1}{x^4 - 2x^2 + 7}.$$

Note that since $x^4 - 2x^2 + 7 = (x^2 - 1)^2 + 6 > 0$ for all $x \in \mathbb{R}$, the function f is well defined and differentiable on $\mathbb{R}$. Moreover,

$$f'(x) = \frac{-(4x^3 - 4x)}{(x^4 - 2x^2 + 7)^2} = \frac{-4x(x-1)(x+1)}{(x^4 - 2x^2 + 7)^2} \quad \text{for } x \in \mathbb{R}.$$

Thus, f' vanishes only at $x = -1, 0, 1$. Now we can make a table as follows.

Interval	$(-\infty, -1)$	$(-1, 0)$	$(0, 1)$	$(1, \infty)$
Sign of f'	$+$	$-$	$+$	$-$

In view of this, from the First Derivative Test, we can conclude that f has a local minimum at $x = 0$ and local maxima at $x = -1$ and $x = 1$. Notice that in this example, it would be quite complicated to compute f'' and use the Second Derivative Test.

(ii) Consider $f : (-1, 1) \to \mathbb{R}$ defined by

$$f(x) := \begin{cases} x^2 & \text{if } 0 < |x| < 1, \\ -1 & \text{if } x = 0. \end{cases}$$

Then it is clear that $f(0) < f(x)$ for all nonzero $x \in (-1, 1)$, and thus f has a strict local minimum at $x = 0$. However, the conditions of the First Derivative Test are not satisfied. Indeed, f is differentiable on $(-1, 0)$ as well as on $(0, 1)$ and f' changes sign from $-$ to $+$ at $x = 0$ but f is not continuous at 0.

(iii) Consider $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) := x^4$. Then $f(0) = 0 < f(x)$ for all nonzero $x \in \mathbb{R}$, and thus f has a strict local minimum at $x = 0$. However, the conditions of the Second Derivative Test are not satisfied. Indeed, f is twice differentiable and $f'(0) = 0$, but $f''(0)$ is not positive. $\diamond$

Points of Inflection

We shall now move on to a more subtle attribute of (the graph of) a real-valued function, namely, the geometric notion of a point of inflection, which was defined in Chapter 1. Briefly, this is a point at which convexity changes to concavity or vice versa. More precisely, given any $D \subseteq \mathbb{R}$ and $f : D \to \mathbb{R}$, an interior point $c \in D$ is said to be a **point of inflection** for f if there is $\delta > 0$ with $(c - \delta, c + \delta) \subseteq D$ such that f is convex on $(c - \delta, c)$ and concave on $(c, c + \delta)$, or vice versa. Also, recall that c is said to be a **strict point of inflection** of f if there is $\delta > 0$ with $(c - \delta, c + \delta) \subseteq \mathbb{R}$ such that f is strictly convex on $(c - \delta, c)$ and strictly concave on $(c, c + \delta)$, or vice versa.

   A typical example is $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) := x^3$ [see Figure 1.3 (iv)], for which 0 is a point of inflection; in fact, 0 is a strict point of inflection for this function.

   Characterizations of convexity and concavity in terms of derivatives discussed in Chapter 4 lead to the following result about points of inflection.

Proposition 5.7 (Necessary and Sufficient Conditions for a Point of Inflection). *Let $D \subseteq \mathbb{R}$, c be an interior point of D. Let $f : D \to \mathbb{R}$ be any function. Then we have the following:*

(i) *Suppose f is differentiable on $(c-r, c) \cup (c, c+r)$ for some $r > 0$. Then c is a point of inflection for f if and only if there is $\delta > 0$ with $\delta \leq r$ such that f' is monotonically increasing on $(c - \delta, c)$, whereas f' is monotonically decreasing on $(c, c + \delta)$, or vice versa.*

(ii) *Suppose f is twice differentiable on $(c - r, c) \cup (c, c + r)$ for some $r > 0$. Then c is a point of inflection for f if and only if there is $\delta > 0$ with $\delta \leq r$ such that f'' is nonnegative throughout $(c-\delta, c)$, whereas f'' is nonpositive throughout $(c, c + \delta)$, or vice versa.*

Proof. Part (i) follows from parts (i) and (ii) of Proposition 4.31, while part (ii) follows from parts (i) and (ii) of Proposition 4.32.   □

The above results can be used to obtain weaker but concise conditions that are necessary or sufficient for an interior point in the domain of a function to be a point of inflection.

Proposition 5.8. *Let $D \subseteq \mathbb{R}$, c be an interior point of D, and $f : D \to \mathbb{R}$ be any function. Then we have the following:*

(i) [**Necessary Condition for a Point of Inflection**] *Let f be twice differentiable at c. If c is a point of inflection for f, then $f''(c) = 0$.*

(ii) [**Sufficient Conditions for a Point of Inflection**] *Let f be thrice differentiable at c. If $f''(c) = 0$ and $f'''(c) \neq 0$, then c is a point of inflection for f.*

Proof. (i) If $f''(c)$ exists, then it is tacitly assumed that f' exists on $(c-r, c+r)$ for some $r > 0$ with $(c - r, c + r) \subseteq D$. If c is a point of inflection for f, then by part (i) of Proposition 5.7, there is $\delta > 0$ with $\delta \leq r$ such that

f' is increasing on $(c - \delta, c)$ and decreasing on $(c, c + \delta)$, or vice versa.

Now f', being differentiable at c, is continuous at c, and therefore we have

$$f'(x) \leq f'(c) \text{ for all } x \in (c - \delta, c) \quad \text{and} \quad f'(c) \geq f'(x) \text{ for all } x \in (c, c + \delta),$$

or vice versa (that is, the inequalities $\leq$ and $\geq$ above are interchanged). Hence

$$0 \leq \lim_{x \to c^-} \frac{f'(x) - f'(c)}{x - c} = f''(c) = \lim_{x \to c^+} \frac{f'(x) - f'(c)}{x - c} \leq 0,$$

or vice versa (that is, the inequalities $\leq$ above change to $\geq$). In any case, we readily see that $f''(c) = 0$.

(ii) If $f'''(c)$ exists, then it is tacitly assumed that f'' exists on $(c-r, c+r)$ for some $r > 0$ with $(c - r, c + r) \subseteq D$. Suppose now that $f''(c) = 0$ and $f'''(c) \neq 0$. We may first assume that $f'''(c) < 0$. Then

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

Hence by part (i) of Proposition 3.24, there is $\delta > 0$ with $\delta \leq r$ such that $f''(x)/(x - c) < 0$ for all $x \in (c - \delta, c) \cup (c, c + \delta)$. Consequently,

$$f''(x) > 0 \text{ for all } x \in (c - \delta, c) \quad \text{and} \quad f''(x) < 0 \text{ for all } x \in (c, c + \delta).$$

Thus, by part (ii) of Proposition 5.7, c is a point of inflection for f. A similar argument holds if $f''(c) = 0$ and $f'''(c) > 0$.   □

Remarks 5.9. (i) The condition in part (i) of the above proposition is not sufficient. Consider, for example, $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) := x^4$. Then 0 is not a point of inflection for f, but $f''(0) = 0$.

(ii) The condition in part (ii) is not necessary. Consider, for example, $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) := x^5$. Then 0 is a point of inflection for f, but $f'''(0) = 0$.

(iii) If in part (i) of Proposition 5.7 we change the words 'monotonically increasing' and 'monotonically decreasing' to 'strictly increasing' and 'strictly decreasing', respectively, then we obtain a necessary and sufficient condition that c is a strict point of inflection for f. Likewise, in part (ii) of Proposition 5.7, if in addition to the condition about the sign of f'', we require that f'' not vanish identically on any subinterval of $(c - \delta, c)$ or of $(c, c + \delta)$ containing more than one point, then we obtain a necessary and sufficient condition for c to be a strict point of inflection for f. On the other hand, if the sufficient condition in part (ii) of Proposition 5.8 is satisfied, then c is, in fact, a strict point of inflection for f. These assertions are easily proved by gleaning through the proofs of Propositions 5.7 and 5.8 and appealing to parts (iii) and (iv) of Proposition 4.31 as well as parts (i) and (ii) of Proposition 4.35.   ◇

As an illustration of the various tests obtained in this section and also in Section 4.3, let us work out an example in which we first use these tests to identify several features of the function or its graph. We shall also see how, equipped with this knowledge, one can make a rough sketch of the graph.

Example 5.10. Consider $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) := x^3 - 6x^2 + 9x + 1$. Then

$$f'(x) = 3x^2 - 12x + 9 = 3(x - 1)(x - 3) \quad \text{and} \quad f''(x) = 6x - 12 = 6(x - 2).$$

Thus, $f'(x) = 0$ only at $x = 1$ and 3, while $f''(x) = 0$ only at $x = 2$. Moreover, we can make tables as follows:

Interval	$(-\infty, 1)$	$(1, 3)$	$(3, \infty)$		Interval	$(-\infty, 2)$	$(2, \infty)$
Sign of f'	$+$	$-$	$+$		Sign of f''	$-$	$+$

In view of this [together with Propositions 4.27, 5.3, 5.4, 4.31, 5.7, and 5.8], we obtain the following:

- f is (strictly) increasing on $(-\infty, 1)$ as well as on $(3, \infty)$, and f is (strictly) decreasing on $(1, 3)$.
- f has a (strict) local maximum at $x = 1$ and a (strict) local minimum at $x = 3$.
- f is (strictly) concave on $(-\infty, 2)$ and (strictly) convex on $(2, \infty)$.
- 2 is a (strict) point of inflection for f.

Now we can make a rough sketch of the curve $y = f(x)$ by plotting a few points [for example, $f(0) = 1$, $f(1) = 5$, $f(2) = 3$, $f(3) = 1$, and $f(4) = 5$] and using the above facts. It will look like the graph in Figure 5.3. $\diamond$
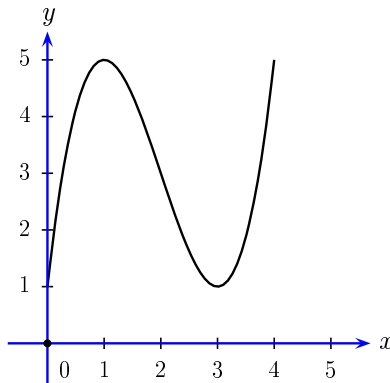
Fig. 5.3. Graph of $y = x^3 - 6x^2 + 9x + 1$

5.3 Linear and Quadratic Approximations

By way of motivating Taylor's Theorem, we have discussed in Chapter 4 how the MVT and its generalizations are helpful in evaluating functions approximately. In this section, we shall formalize these aspects and give some basic features of the simplest of such approximations that are used in practice, namely, the linear and quadratic approximations.

Let $D \subseteq \mathbb{R}$ and c be an interior point of D. If $f : D \to \mathbb{R}$ is differentiable at c, then the function $L : \mathbb{R} \to \mathbb{R}$ defined by

$$L(x) := f(c) + f'(c)(x - c) \quad \text{for } x \in \mathbb{R}$$

is called the **linear approximation** to f around c. Note that $L(x)$ is the first Taylor polynomial of f around c. Geometrically speaking, $y = L(x)$ represents a line, which is precisely the tangent to the curve $y = f(x)$ at the point
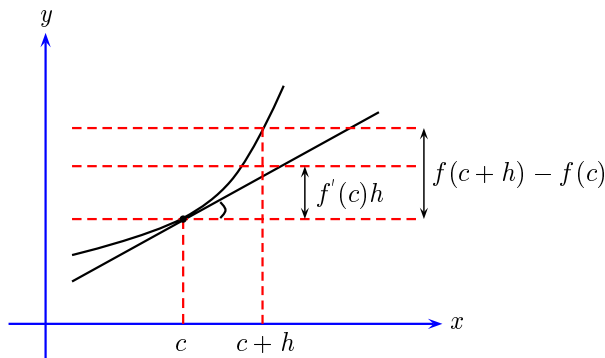
Fig. 5.4. Linear approximation or tangent line approximation around a point

$(c, f(c))$. For this reason, L is also called the **tangent line approximation** to f around c.

The difference
$$e_1(x) := f(x) - L(x) \quad \text{for } x \in D$$
is called the **error** at x in the linear approximation to f around c.

Proposition 5.11. *Let $D \subseteq \mathbb{R}$ and c be an interior point of D. If $f : D \to \mathbb{R}$ is differentiable at c, then the linear approximation L to f around c is indeed an approximation to f around c, that is,*

$$\lim_{x \to c} L(x) = f(c) \quad \text{or equivalently,} \quad \lim_{x \to c} e_1(x) = 0.$$

In fact, $e_1(x)$ rapidly approaches zero as $x \to c$ in the sense that

$$\lim_{x \to c} \frac{e_1(x)}{x - c} = 0.$$

Moreover, given any $b \in D$ with $b \neq c$, if I_b denotes the open interval with c and b as its endpoints and if f' exists and is continuous on I_b as well as its endpoints c and b, f'' exists on I_b, and $|f''(x)| \leq M_2(b)$ for all $x \in I_b$, then we have the following error bound:

$$|e_1(b)| \leq \frac{M_2(b)}{2} |b - c|^2.$$

Proof. It is obvious from the definition of L that

$$\lim_{x \to c} L(x) = f(c) \quad \text{or equivalently,} \quad \lim_{x \to c} e_1(x) = 0.$$

The assertion about rapid vanishing of $e_1(x)$ follows by noting that

$$\lim_{x \to c} \frac{e_1(x)}{x - c} = \lim_{x \to c} \frac{f(x) - f(c) - f'(c)(x - c)}{x - c} = f'(c) - f'(c) = 0.$$

Finally, if $b \in D$ and $b \neq c$, then applying Taylor's Formula (Remark 4.24), with $I = I_b$ and $n = 1$, we see that there is ξ between c and b such that

$$e_1(b) = f(b) - f(c) - f'(c)(b - c) = \frac{f''(\xi)}{2}(b - c)^2.$$

This implies the desired error bound for $|e_1(b)|$.   □

Example 5.12. Let $D := \{x \in \mathbb{R} : x \neq 1\}$ and consider $f : D \to \mathbb{R}$ defined by $f(x) := 1/(1 - x)$. Then $f'(x) = 1/(1 - x)^2$ for $x \in D$ and in particular, $f'(0) = 1$. Thus, the linear approximation to f around 0 is given by

$$L(x) = f(0) + f'(0)(x - 0) = 1 + x \quad \text{for } x \in \mathbb{R}.$$

The error bound e_1 for $f - L$ in this case can be worked out as follows. Given $b \in (-1, 1)$, $b \neq 0$, let us consider two cases. First, suppose $b > 0$ and $I_b = (0, b)$. Then

$$|f''(x)| = \frac{2}{(1 - x)^3} \leq \frac{2}{(1 - b)^3} \quad \text{for } x \in I_b.$$

Thus, in this case we may take $M_2(b) = 2/(1 - b)^3$ and by Proposition 5.11 conclude that $|e_1(b)| \leq b^2/(1 - b)^3$. For instance, if $0 < b < 0.1$, then we have $|e_1(b)| \leq (0.1)^2/(0.9)^2 < 0.014$. Next, suppose $b < 0$ and $I_b = (b, 0)$. Then

$$|f''(x)| = \frac{2}{(1 - x)^3} \leq 2 \quad \text{for } x \in I_b.$$

Thus, in this case we may take $M_2(b) = 2$ and by Proposition 5.11 conclude that $|e_1(b)| \leq b^2$. For instance, if $-0.1 < b < 0$, then we have $|e_1(b)| \leq (0.1)^2 = 0.01$.   ◇

As the above example shows, linear approximation gives a reasonable approximation to the values of a function around a point where it is differentiable. However, if one wants to do better, then one may take recourse to quadratic approximation, which is available provided the relevant second derivative exists.

Let, as before, $D \subseteq \mathbb{R}$ and c be an interior point of D. If $f : D \to \mathbb{R}$ is twice differentiable at c, then the function $Q : \mathbb{R} \to \mathbb{R}$ defined by

$$Q(x) := f(c) + (x - c)f'(c) + \frac{(x - c)^2}{2}f''(c) \quad \text{for } x \in \mathbb{R}$$

is called the **quadratic approximation** to f around c. Note that $Q(x)$ is the second Taylor polynomial of f around c. Geometrically speaking, $y = Q(x)$ represents a parabola passing through the point $(c, f(c))$ such that this

parabola and the curve $y = f(x)$ have a common tangent at $(c, f(c))$. The difference

$$e_2(x) := f(x) - Q(x) \quad \text{for } x \in D$$

is called the **error** at x in the quadratic approximation to f around c.

Proposition 5.13. *Let $D \subseteq \mathbb{R}$ and c be an interior point of D. If $f : D \to \mathbb{R}$ is twice differentiable at c, then the quadratic approximation Q to f around c is indeed an approximation to f around c, that is,*

$$\lim_{x \to c} Q(x) = f(c) \quad \text{or equivalently,} \quad \lim_{x \to c} e_2(x) = 0.$$

In fact, $e_2(x)$ approaches zero as $x \to c$ doubly rapidly in the sense that

$$\lim_{x \to c} \frac{e_2(x)}{(x - c)^2} = 0.$$

Moreover, given any $b \in D$ with $b \neq c$, if I_b denotes the open interval with c and b as its endpoints and if f'' exists and is continuous on I_b as well as at its endpoints c and b, f''' exists on I_b, and $|f'''(x)| \leq M_3(b)$ for all $x \in I_b$, then we have the following error bound:

$$|e_2(b)| \leq \frac{M_3(b)}{3!} |b - c|^3.$$

Proof. It is obvious from the definition of Q that

$$\lim_{x \to c} Q(x) = f(c) \quad \text{or equivalently,} \quad \lim_{x \to c} e_2(x) = 0.$$

The assertion about doubly rapid vanishing of $e_2(x)$ follows by noting that by L'Hôpital's Rule for $\frac{0}{0}$ indeterminate forms, we have

$$\lim_{x \to c} \frac{e_2(x)}{(x - c)^2} = \lim_{x \to c} \frac{f'(x) - f'(c) - f''(c)(x - c)}{2(x - c)} = \frac{1}{2} [f''(c) - f''(c)] = 0.$$

Finally, if $b \in D$ and $b \neq c$, then applying Taylor's Formula (Remark 4.24), with $I = I_b$ and $n = 2$, we see that there is $\eta \in I_b$ such that

$$e_2(b) = f(b) - f(c) - f'(c)(b - c) - \frac{f''(c)}{2}(b - c)^2 = \frac{f'''(\eta)}{3!}(b - c)^3.$$

This implies the desired inequality for $|e_2(b)|$.   □

Now let us revisit Example 5.12 and see what the quadratic approximation and the corresponding error bound look like.

Example 5.14. Consider $f : (-1, 1) \to \mathbb{R}$ defined by $f(x) := 1/(1 - x)$ and $c := 0$. Then $f'(x) = 1/(1 - x)^2$ and $f''(x) = 2/(1 - x)^3$ for $x \in (-1, 1)$. In

particular, $f'(0) = 1$ and $f''(0) = 2$. Thus, the quadratic approximation to f around 0 is given by

$$Q(x) = f(0) + f'(0)(x - 0) + \frac{f''(0)}{2}(x - 0)^2 = 1 + x + x^2.$$

The error bound e_2 for $f - Q$ in this case can be worked out as follows. Given $b \in (-1, 1)$, $b \neq 0$, let us consider two cases. First, suppose $b > 0$ and $I_b = (0, b)$. Then

$$|f'''(x)| = \frac{6}{(1 - x)^4} \leq \frac{6}{(1 - b)^4} \quad \text{for } x \in I_b.$$

Thus, in this case we may take $M_3(b) = 6/(1 - b)^4$ and by Proposition 5.13 conclude that $|e_2(b)| \leq |b|^3/(1 - b)^4$. For instance, if $0 < b < 0.1$, then we have $|e_2(b)| \leq (0.1)^3/(0.9)^4 < 0.0016$. Next, suppose $b < 0$ and $I_b = (b, 0)$. Then

$$|f''(x)| = \frac{6}{(1 - x)^4} \leq 6 \quad \text{for } x \in I_b.$$

Thus, in this case we may take $M_3(b) = 6$ and by Proposition 5.13 conclude that $|e_2(b)| \leq |b|^3$. For instance, if $-0.1 < b < 0$, then we have $|e_2(b)| \leq (0.1)^3 = 0.001$. $\diamond$

It may be noted that in the above example, the estimates have become sharper than those in Example 5.12. In a similar way, if we were to consider cubic approximations, quartic approximations, and so on, the estimates would become more and more sharp. These higher-degree approximations and the corresponding error bounds can be obtained in an analogous manner. See Exercise 21.

5.4 The Picard and Newton Methods

The title of this section refers to methods that can be used to obtain approximate solutions to the following two interrelated problems:

1. The problem of finding a **fixed point** of a function, namely, if $D \subseteq \mathbb{R}$ and $f : D \to D$, then the problem is to find $x \in D$ such that $f(x) = x$.
2. The problem of finding a solution of an equation, namely, if $D \subseteq \mathbb{R}$ and $f : D \to \mathbb{R}$, then the problem is to find $x \in D$ such that $f(x) = 0$.

To see that these problems are interrelated, it suffices to note that if $f : D \to D$ and if we set $F(x) = f(x) - x$, then finding a fixed point of f is equivalent to finding a solution to $F(x) = 0$. On the other hand, if $f : D \to \mathbb{R}$ and if we set $F(x) = x + h(x)f(x)$ where $h : D \to \mathbb{R}$ is so chosen that $h(x) \neq 0$ and $x + h(x)f(x) \in D$ for all $x \in D$, then finding a solution of $f(x) = 0$ is equivalent to finding a fixed point of $F : D \to D$.

Finding a Fixed Point

Let us first take up the problem of finding fixed points. For simplicity, we shall restrict ourselves to the case of functions from a closed and bounded interval of $\mathbb{R}$ into itself. In this case, the existence of a fixed point is guaranteed if the function satisfies a mild condition such as continuity.

Proposition 5.15. *If $f : [a, b] \to [a, b]$ is continuous, then f has a fixed point.*

Proof. Let $F : [a, b] \to \mathbb{R}$ be defined by $F(x) = f(x) - x$. Since $a \leq f(x) \leq b$ for all $x \in [a, b]$, we have,

$$F(a) = f(a) - a \leq 0 \quad \text{and} \quad F(b) = f(b) - b \geq 0.$$

Also, since f is continuous, so is F. Hence by Proposition 3.13, F has the IVP on $[a, b]$. So, there is $c \in [a, b]$ such that $F(c) = 0$, that is, $f(c) = c$. $\qquad\square$

Examples 5.16. (i) While a fixed point in $[a, b]$ exists for a continuous function $f : [a, b] \to [a, b]$, it need not be unique. Consider, for example, $f : [0, 1] \to [0, 1]$ defined by $f(x) := x$, where every point of $[0, 1]$ is a fixed point of f.

(ii) The condition that f be defined on a closed subset of $\mathbb{R}$ is essential for the existence of a fixed point. For example, if $f : [0, 1) \to \mathbb{R}$ is defined by $f(x) := (1 + x)/2$, then f maps $[0, 1)$ into itself, and f is continuous. But f has no fixed point in $[0, 1)$. Indeed, $(1 + x)/2 = x$ only when $x = 1$.

(iii) The condition that f be defined on a bounded subset of $\mathbb{R}$ is essential for the existence of a fixed point. For example, if $f : [1, \infty) \to \mathbb{R}$ is defined by $f(x) := x + (1/x)$, then f maps $[1, \infty)$ into itself, and f is continuous. But clearly, f has no fixed point in $[1, \infty)$.

(iv) The condition that f be defined on an interval in $\mathbb{R}$ is essential for the existence of a fixed point. For example, if $D = [-2, -1] \cup [1, 2]$ and $f : D \to \mathbb{R}$ is defined by $f(x) := -x$, then f maps D into itself, and f is continuous. But f has no fixed point in D. $\qquad\diamond$

Suppose we know that a function $f : [a, b] \to [a, b]$ has a fixed point. Then a natural question is whether we can find it. It is not easy, in general, to find it exactly. A simple and effective method given by Picard seeks to achieve what may be the next best alternative to finding a fixed point exactly, namely, to find it approximately. In geometric terms, the basic idea of the Picard method can be described as follows.

First, pick any point $P_0 = (x_0, f(x_0))$ on the curve $y = f(x)$. Project P_0 horizontally to a point Q_0 on the diagonal line $y = x$, and then, project the point Q_0 vertically onto the curve $y = f(x)$ to obtain a point $P_1 = (x_1, f(x_1))$. Again, project P_1 horizontally to Q_1 on $y = x$ and then, project Q_1 vertically onto $y = f(x)$ to obtain $P_2 = (x_2, f(x_2))$. This process can be be repeated a number of times. Often, it will weave a cobweb in which the fixed point of f, that is, the point of intersection of the curve $y = f(x)$ and the diagonal

line $y = x$, gets trapped. [See Figure 5.5 (i).] In fact, we shall see that such trapping occurs if the slopes of tangents to the curve $y = f(x)$ are smaller (in absolute value) than the slope of the diagonal line $y = x$. When the slope condition is not met, then the points $P_0, P_1, P_2, \ldots$ may move away from a fixed point. [See Figure 5.5 (ii).]
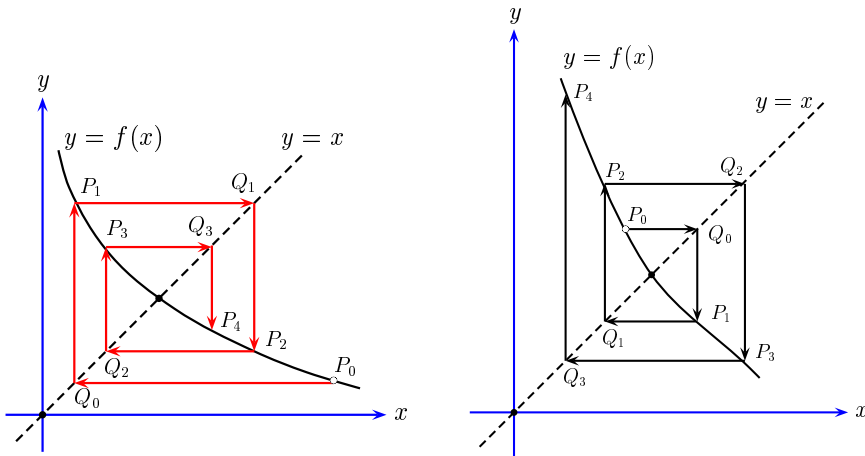
Fig. 5.5. Picard sequence that is (i) converging to a fixed point, and (ii) diverging away from a fixed point

In analytic terms, the **Picard method** can be described as follows. Given any $x_0 \in [a, b]$, we recursively define a sequence (x_n) by

$$x_n = f(x_{n-1}) \quad \text{for } n \in \mathbb{N}.$$

Such a sequence (x_n) is called a **Picard sequence** for the function f (with its initial point x_0). It is clear that if a Picard sequence (x_n) for f is convergent and f is continuous, then the limit x of (x_n) is a fixed point of f. Indeed,

$$f(x) = f\left(\lim_{n\to\infty} x_n\right) = f\left(\lim_{n\to\infty} x_{n-1}\right) = \lim_{n\to\infty} f(x_{n-1}) = \lim_{n\to\infty} x_n = x.$$

A sufficient condition for the convergence of a Picard sequence, which is a formal analogue of the geometric condition on slopes mentioned above, is given by the following result. It is to be noted here that the same condition guarantees the uniqueness of a fixed point.

Proposition 5.17 (Picard Convergence Theorem). *If $f : [a, b] \to [a, b]$ is continuous on $[a, b]$ and differentiable on (a, b) with $|f'(x)| < 1$ for all $x \in (a, b)$, then f has a unique fixed point. Furthermore, any Picard sequence for f is convergent and converges to the unique fixed point of f.*

Proof. By Proposition 5.15, f has a fixed point in $[a, b]$. If there are two fixed points $c_1, c_2 \in [a, b]$, then by the MVT, there is $c \in (a, b)$ such that

$$|c_1 - c_2| = |f(c_1) - f(c_2)| = |f'(c)||c_1 - c_2| < |c_1 - c_2|,$$

which is a contradiction. Thus, f has a unique fixed point.

Let c^* denote the unique fixed point of f. Consider any $x_0 \in [a, b]$, and let (x_n) be the Picard sequence for f with its initial point x_0. Now, given any $n \in \mathbb{N}$, by the MVT, there is c_{n-1} between x_{n-1} and c^* such that

$$x_n - c^* = f(x_{n-1}) - f(c^*) = f'(c_{n-1})(x_{n-1} - c^*).$$

As a consequence, $|x_n - c^*| \leq |x_{n-1} - c^*|$ for all $n \in \mathbb{N}$. We shall now show that $x_n \to c^*$. First, note that since $x_n \in [a, b]$ for $n \geq 0$, the sequence (x_n) is bounded. Thus, by Proposition 2.17, it suffices to show that every convergent subsequence of (x_n) has c^* as its limit. Let $x \in \mathbb{R}$ and (x_{n_k}) be a subsequence of (x_n) such that $x_{n_k} \to x$. Then

$$\left|x_{n_k+1} - c^*\right| \leq |x_{n_k+1} - c^*| \leq |x_{n_k} - c^*|.$$

But both the sequences $(|x_{n_k} - c^*|)$ and $\left(\left|x_{n_k+1} - c^*\right|\right)$ converge to $|x - c^*|$. So, by the Sandwich Theorem, $|x_{n_k+1} - c^*| \to |x - c^*|$ as $k \to \infty$. On the other hand,

$$\lim_{k\to\infty} |x_{n_k+1} - c^*| = \lim_{k\to\infty} |f(x_{n_k}) - f(c^*)| = |f(x) - f(c^*)|.$$

It follows that $|f(x) - f(c^*)| = |x - c^*|$. Now, if $x \neq c^*$, then by the MVT, there is $c \in (a, b)$ such that

$$|x - c^*| = |f(x) - f(c^*)| = |f'(c)||x - c^*| < |x - c^*|,$$

which is a contradiction. This proves that $x_n \to c^*$. $\square$

Remark 5.18. The proof of Picard's Convergence Theorem becomes simpler if instead of assuming $|f'(x)| < 1$ for all $x \in (a, b)$, we make the stronger assumption that there is $\alpha < 1$ such that $|f'(x)| < \alpha$ for all $x \in (a, b)$. In this case we can also obtain the 'rate of convergence' for the Picard sequence. [See Exercise 32.] An alternative set of conditions for the convergence of the Picard sequence is given in Exercise 28. The Picard Convergence Theorem itself admits several generalizations and extensions. Some of these are outlined in Exercise 29. $\diamond$

Examples 5.19. (i) Consider $f : [0, 2] \to \mathbb{R}$ defined by $f(x) := (1 + x)^{1/5}$. Then $0 \leq f(x) \leq 3^{1/5} < 2$ for all $x \in [0, 2]$. Thus, f maps the interval $[0, 2]$ into itself. Moreover, f is continuous on $[0, 2]$, differentiable on $(0, 2)$, and

$$|f'(x)| = \frac{1}{5(1 + x)^{4/5}} \leq \frac{1}{5} < 1 \quad \text{for } x \in [0, 2].$$

Thus, by the Picard convergence theorem, f has a unique fixed point, and successively better approximations to this fixed point are given by the successive terms of a Picard sequence for f. For example, if we take $x_0 = 0$, then the first few terms of the corresponding Picard sequence for f are roughly given by $x_1 = 1$, $x_2 = 1.148698$, $x_3 = 1.1652928$, and $x_4 = 1.1670872$. It may be noted that finding a fixed point of f is equivalent to finding the root of the quintic polynomial $x^5 - x - 1$ in the interval $[0, 2]$.

(ii) Consider $f : [0, 1] \to \mathbb{R}$ defined by $f(x) := x^2/2$. Then f maps $[0, 1]$ into itself. Moreover, f is continuous on $[0, 1]$, differentiable on $(0, 1)$ and $|f'(x)| = |x| < 1$ for all $x \in (0, 1)$. Thus, by the Picard convergence theorem, f has a unique fixed point and any Picard sequence for f will converge to this fixed point. Indeed, it is easily verified that 0 is the only fixed point of f in $[0, 1]$. Note that in this case there is no $\alpha < 1$ such that $|f'(x)| < \alpha$ for all $x \in (0, 1)$.

(iii) The condition $|f'(x)| < 1$ for all $x \in (a, b)$ is essential for the uniqueness of a fixed point. For example, if $f : [a, b] \to \mathbb{R}$ is defined by $f(x) := x$, then f maps $[a, b]$ into itself, and every point of $[a, b]$ is a fixed point of f. Here, $f'(x) = 1$ for all $x \in (a, b)$.

(iv) When the condition $|f'(x)| < 1$ for all $x \in (a, b)$ is not satisfied, a function can still have a unique fixed point c^* but the Picard sequence (x_n) with its initial point $x_0 \neq c$ may not converge to c^*. For example, if $f : [-1, 1] \to \mathbb{R}$ is defined by $f(x) := -x$, then f maps $[-1, 1]$ into itself, f is differentiable and $|f'(x)| = 1$ for all $x \in [-1, 1]$. Clearly, $c^* = 0$ is the unique fixed point of f, but if $x_0 \neq 0$, then the corresponding Picard sequence looks like $-x_0, x_0, -x_0, x_0, \ldots$; in other words, it oscillates between x_0 and $-x_0$ and never reaches the fixed point. In geometric terms, the cobweb that we hope to weave just traces out a square over and over again.   ◇

Remark 5.20. When the hypothesis of the Picard Convergence Theorem is satisfied, a Picard sequence for $f : [a, b] \to [a, b]$ with arbitrary $x_0 \in [a, b]$ as its initial point will converge to a fixed point. It is natural to expect that if x_0 is closer to the fixed point, then the convergence will be rapid. But since we do not know the fixed point to begin with, it may not be clear how one picks a 'good' initial point x_0. To this end, observe that a fixed point of f is necessarily in its range. The range is usually smaller than $[a, b]$. Thus, it is better that x_0 be picked from $f([a, b])$. For example, if $f : [0, 1] \to [0, 1]$ is given by $f(x) := (x+1)/4$, then the range of f equals $\left[\frac{1}{4}, \frac{1}{2}\right]$, and so we should choose x_0 to be this smaller subinterval. In fact, this simple observation can be extended further. A fixed point of f lies not only in the range of f but also in the ranges of the composites $f \circ f$, $f \circ f \circ f$, and so on. Thus, if R_n is the range of the n-fold composite $f \circ \cdots \circ f$, then a fixed point is in each R_n as n varies over $\mathbb{N}$. If only a single point belongs to each R_n ($n \in \mathbb{N}$), then we have found our fixed point! In fact, the Picard method amounts to starting with any $x_0 \in [a, b]$ and considering the image of x_0 under the n-fold composite $f \circ \cdots \circ f$ of f. For example, if, as before, $f : [0, 1] \to [0, 1]$ is given

by $f(x) = (x+1)/4$, then it is easy to see that the n-fold composite of f, say f_n, and its range are given by

$$f_n(x) = \frac{x + (4^n - 1)/3}{4^n} \text{ and } R_n := f_n([0, 1]) = \left[\frac{1}{3}\left(1 - \frac{1}{4^n}\right), \frac{1}{3}\left(1 + \frac{2}{4^n}\right)\right].$$

It is clear, therefore, that $\frac{1}{3}$ is the only point in each R_n ($n \in \mathbb{N}$), and this is the unique fixed point of f (as can also be verified directly from the definition of f). Of course, in general, it is not practical to determine the ranges of the n-fold composites of f for all $n \in \mathbb{N}$. So it is simpler to use the Picard method. But the Picard method will be more effective if the above observations are used to some extent in choosing the initial point. $\diamond$

Finding a Solution of an Equation

We now turn to the second problem mentioned earlier, namely, the problem of finding a solution of $f(x) = 0$, where f is a real-valued function defined on a subset of $\mathbb{R}$. For simplicity, we shall restrict ourselves to the case in which f is defined on a closed and bounded interval of $\mathbb{R}$. In this case, the existence of a solution is guaranteed if the IVP is available.

Proposition 5.21. *If $f : [a, b] \to \mathbb{R}$ is continuous and if $f(a)$ and $f(b)$ have opposite signs, then $f(x) = 0$ has a solution in $[a, b]$.*

Proof. The result is an immediate consequence of the fact that a continuous function on $[a, b]$ has the IVP. $\square$

Suppose we know that $f : [a, b] \to \mathbb{R}$ is such that the equation $f(x) = 0$ has a solution. Then a natural question is whether we can find it. It is not easy, in general, to find an exact solution.[1] A method given by Newton seeks to achieve what may be the next best alternative to finding an exact solution, namely, to find an approximate solution. In geometric terms, the basic idea of the Newton method can be described as follows.

First, pick any point $P_0 = (x_0, f(x_0))$ on the curve $y = f(x)$. Draw a tangent to this curve at P_0 and if it intersects the x-axis at $(x_1, 0)$, then

---

[1] In fact, this is a very difficult problem even for the nicest of functions, namely, polynomial functions. In the special case of linear and quadratic equations, there are simple and well-known formulas for their solutions. For the solutions of cubic and quartic equations, there are more intricate formulas, ascribed to Cardan and Ferrari, which express the solutions in terms of the coefficients of the polynomial using the basic operations of algebra, namely, addition, subtraction, multiplication, division, and extraction of roots. After several unsuccessful attempts to find a similar formula for a general polynomial equation of degree 5 or more, it was proved by Abel that no such formula exists. In other words, a general equation of degree 5 or more is not *solvable by radicals*. An elegant proof of Abel's result was given by Galois, who also gave a criterion for an equation to be solvable by radicals. For more on these topics, we refer to the book of Tignol [64].

consider the corresponding point $P_1 = (x_1, f(x_1))$. Again, draw the tangent to $y = f(x)$ at P_1 and if it intersects the x-axis at $(x_2, 0)$, then consider the corresponding point $P_2 = (x_2, f(x_2))$. This process can be repeated a number of times. Often, it will rapidly bring us near to the point of intersection of the curve $y = f(x)$ and the x-axis, that is, to the solution of $f(x) = 0$. In effect, each time we replace the curve by the tangent line approximation and utilize the fact that linear equations can be solved. It is clear, however, that the procedure will fail if at some point, the tangent is parallel to the x-axis.
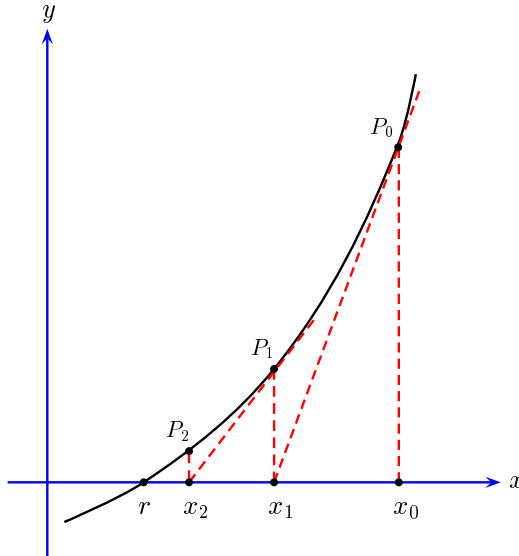
Fig. 5.6. Newton sequence approaching a solution of an equation

In analytic terms, the **Newton method** (sometimes also called the **Newton–Raphson method**) can be described as follows. Choose any $x_0 \in [a, b]$ such that $f'(x_0)$ exists and $f'(x_0) \neq 0$. Given any $n \in \mathbb{N}$ and $x_{n-1} \in [a, b]$ such that $f'(x_{n-1}) \neq 0$, we let x_n be the root of the linear approximation

$$L(x) = f(x_{n-1}) + f'(x_{n-1})(x - x_{n-1})$$

to f around x_{n-1}. In other words,

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})}.$$

Such a sequence (x_n) is called a **Newton sequence** for the function f (with its initial point x_0). It is clear that if a Newton sequence (x_n) for f is convergent and f' is bounded, then the limit x of (x_n) satisfies $f(x) = 0$. Indeed,

$$f(x) = f\left(\lim_{n\to\infty} x_{n-1}\right) = \lim_{n\to\infty} f(x_{n-1}) = \lim_{n\to\infty} f'(x_{n-1})(x_{n-1} - x_n) = 0.$$

A sufficient condition for the convergence of a Newton sequence can be derived from the Picard Convergence Theorem as follows.

Proposition 5.22 (Convergence of Newton Sequences). *Let* $f : [a,b] \to$
$\mathbb{R}$ *be differentiable with* $f'(x) \neq 0$ *for all* $x \in [a,b]$, *and*

$$x - \frac{f(x)}{f'(x)} \in [a,b] \quad \text{for all } x \in [a,b].$$

Assume that f' *is continuous on* $[a,b]$, *differentiable on* (a,b), *and*

$$\left| \frac{f(x)f''(x)}{[f'(x)]^2} \right| < 1 \quad \text{for all } x \in (a,b).$$

Then there is a unique $x^* \in [a,b]$ *such that* $f(x^*) = 0$. *Furthermore, the Newton sequence for* f *with any initial point* $x_0 \in [a,b]$ *converges to* x^*.

Proof. Define $F : [a,b] \to \mathbb{R}$ by

$$F(x) := x - \frac{f(x)}{f'(x)} \quad \text{for } x \in [a,b].$$

Then F is continuous on $[a,b]$, differentiable on (a,b), and F maps the interval $[a,b]$ into itself. Notice that $x \in [a,b]$ is a fixed point of F if and only if $f(x) = 0$. Moreover, for any $x \in [a,b]$, we have

$$|F'(x)| = \left| 1 - \frac{[f'(x)]^2 - f(x)f''(x)}{[f'(x)]^2} \right| = \left| \frac{f(x)f''(x)}{[f'(x)]^2} \right| < 1.$$

Therefore, by the Picard Convergence Theorem, F has a unique fixed point x^* in $[a,b]$, which is then the unique root of f in $[a,b]$. Furthermore, if $x_0 \in [a,b]$ is any initial point, then the Newton sequence for f is, in fact, the Picard sequence for F, and hence it converges to x^*. $\qquad\square$

Examples 5.23. (i) Consider $f : \left[\frac{5}{4}, \frac{3}{2}\right] \to \mathbb{R}$ defined by $f(x) := x^3 - 3$.
Then f is continuous on $\left[\frac{5}{4}, \frac{3}{2}\right]$ and

$$f\left(\frac{5}{4}\right) = \frac{125}{64} - 3 < 0, \quad \text{while} \quad f\left(\frac{3}{2}\right) = \frac{27}{8} - 3 > 0.$$

Hence, by Proposition 5.21, f has a root in $\left[\frac{5}{4}, \frac{3}{2}\right]$. In this case, the iterative formula for the Newton sequence is given by

$$x_n = x_{n-1} - \frac{x_{n-1}^3 - 3}{3x_{n-1}^2} = \frac{2}{3}x_{n-1} + \frac{1}{x_{n-1}^2} \quad \text{provided } x_{n-1} \neq 0.$$

Thus, if we take $x_0 = \frac{5}{4}$, then we obtain

$$x_1 = 1.473333\ldots, \quad x_2 = 1.442900\ldots, \quad x_3 = 1.442249\ldots \ .$$

On the other hand, if we take $x_0 = \frac{3}{2}$, then we obtain

$$x_1 = 1.444444\ldots, \quad x_2 = 1.442252\ldots, \quad x_3 = 1.442249\ldots \ .$$

This indicates that both the Newton sequences converge to the same limit, which is approximately $1.442249\cdots$. In fact, this is quite in accordance with the theory because

$$x - \frac{f(x)}{f'(x)} = \frac{2x^3 + 3}{3x^2} \in \left[\frac{5}{4}, \frac{3}{2}\right] \quad \text{for all } x \in \left[\frac{5}{4}, \frac{3}{2}\right],$$

since $15x^2 \leq 8x^3 + 12$ and $4x^3 + 6 \leq 9x^2$. (See Exercise 31 (ii) of Chapter 4.) Moreover, for $x \in \left(\frac{5}{4}, \frac{3}{2}\right)$, we have

$$\left|\frac{f(x)f''(x)}{[f'(x)]^2}\right| = \left|\frac{(x^3 - 3)6x}{9x^4}\right| = \frac{2}{3}\frac{|x^3 - 3|}{x^3} \leq \frac{2}{3} < 1.$$

Thus, the hypothesis of Proposition 5.22 is satisfied. Hence the equation $f(x) = 0$ has a unique solution in $\left[\frac{5}{4}, \frac{3}{2}\right]$ and any Newton sequence converges to it.

(ii) To illustrate how the Newton sequence behaves where there is more than one solution, consider $f : [-2, 4] \to \mathbb{R}$ defined by $f(x) := x^2 - 2x - 3 = (x + 1)(x - 3)$. [See Figure 5.7 (i).] Clearly $f(x) = 0$ has two solutions $x = -1$ and $x = 3$. Now, $f'(x) = 2(x - 1)$, and thus the Newton sequence for f with any initial point $x_0 \neq 1$ is given by

$$x_n = x_{n-1} - \frac{x_{n-1}^2 - 2x_{n-1} - 3}{2(x_{n-1} - 1)} = \frac{x_{n-1}^2 + 3}{2(x_{n-1} - 1)}, \quad \text{provided } x_{n-1} \neq 1.$$

It is not difficult to show that if $x_0 < 1$, then (x_n) converges to the root -1 of f, whereas if $x_0 > 1$, then (x_n) converges to the root 3 of f. [See Exercise 23.]

(iii) Consider $f : [-10, 10] \to \mathbb{R}$ defined by

$$f(x) := \begin{cases} \sqrt{x - 1} & \text{if } x \geq 1, \\ -\sqrt{1 - x} & \text{if } x < 1. \end{cases}$$

In this case, we have

$$f'(x) = \begin{cases} 1/(2\sqrt{x - 1}) & \text{if } x > 1, \\ 1/(2\sqrt{1 - x}) & \text{if } x < 1. \end{cases}$$

The Newton sequence for f with any initial point $x_0 \neq 1$ is given by

$$x_n = x_{n-1} - 2(x_{n-1} - 1) = -x_{n-1} + 2.$$

Since $x_n - 1 = -(x_{n-1} - 1)$, we have $x_n = 1 + (-1)^n(x_0 - 1)$. Thus, the Newton sequence oscillates between x_0 and $2 - x_0$. [See Figure 5.7 (ii).] ◇
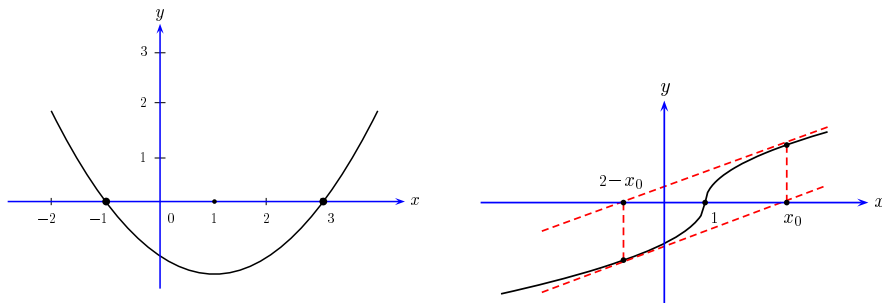
Fig. 5.7. Graphs of (i) $y = x^2 - 2x - 3$ and (ii) $y = \begin{cases} \sqrt{x-1} & \text{if } x \geq 1, \\ -\sqrt{1-x} & \text{if } x < 1 \end{cases}$

As the above examples show, a Newton sequence may not always converge to the desired root, but when it does converge, the rate of convergence is quite rapid and just a few iterations give us values that are fairly close to the desired root. The conditions for convergence given in Proposition 5.22, which is a consequence of the Picard Convergence Theorem, are rather unwieldy and difficult to check in practice. However, there are alternative sets of sufficient conditions such as those given by the following result.

Proposition 5.24. *Let $f : [a, b] \to \mathbb{R}$ be such that $f(r) = 0$ for some $r \in [a, b]$. If f' is nonzero throughout $[a, b]$ and f' is monotonic on $[a, b]$, then r is the unique solution of $f(x) = 0$ in $[a, b]$ and the Newton sequence for f with any initial point $x_0 \in [a, b]$ converges to r.*

Proof. Since $f(r) = 0$ and f' is nonzero throughout $[a, b]$, it follows from Rolle's Theorem that r is the unique solution of $f(x) = 0$ in $[a, b]$. Moreover, since f' has the IVP on $[a, b]$ (Proposition 4.14), we see that either f' is positive throughout $[a, b]$ or negative throughout $[a, b]$. Also, since f' is monotonic on $[a, b]$, there are four possible cases according as f' is positive or f' is negative, and f' is monotonically increasing or f' is monotonically decreasing.

To begin with, suppose f' is positive and monotonically increasing on $[a, b]$. Choose an arbitrary initial point $x_0 \in [a, b]$. Let (x_n) denote the Newton sequence for f with its initial point x_0. If $x_0 = r$, then clearly, $x_n = r$ for all $n \in \mathbb{N}$ and so $x_n \to r$. Now assume that $x_0 > r$. Then by the MVT, there is $c \in (r, x_0)$ such that

$$\frac{f(x_0)}{x_0 - r} = \frac{f(x_0) - f(r)}{x_0 - r} = f'(c).$$

Moreover, since f' is positive and monotonically increasing on $[a, b]$, we have $0 < f'(c) \leq f'(x_0)$. Hence $f(x_0) > 0$ and $(f(x_0)/f'(x_0)) \leq x_0 - r$. Thus,

$$r = x_0 - (x_0 - r) \leq x_0 - \frac{f(x_0)}{f'(x_0)} < x_0.$$

Since $x_1 := x_0 - (f(x_0)/f'(x_0))$, we see that $r \leq x_1 < x_0$. [See Figure 5.8 (i).] Now, if $x_1 \neq r$, then $x_1 > r$ and we can proceed as before to obtain $r \leq x_2 < x_1$. Continuing in this way, we see that the Newton sequence (x_n) has the property that either $x_n = r$ for some $n \in \mathbb{N}$ (in which case $x_m = r$ for all $m > n$) or (x_n) is strictly decreasing and bounded below by r. In the latter event, by part (ii) of Proposition 2.8, $x_n \to s$ for some $s \in [a, b]$ with $r \leq s$. But then, $f(s) = 0$ and hence $s = r$. Thus, in any event, $x_n \to r$.
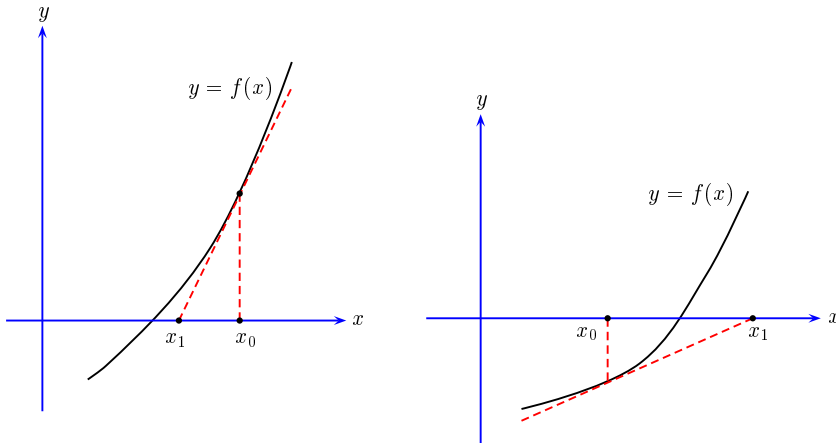
Fig. 5.8. Newton iterates for f with its initial point x_0 when f' is positive and monotonically increasing and (i) $x_0 > r$, (ii) $x_0 < r$, where $r \in \mathbb{R}$ with $f(r) = 0$

Next, assume that $x_0 < r$. [See Figure 5.8 (ii).] Using the MVT as before, we see this time that there is $d \in (x_0, r)$ such that

$$\frac{f(x_0)}{x_0 - r} = \frac{f(x_0) - f(r)}{x_0 - r} = f'(d).$$

Again, since f' is positive and monotonically increasing on $[a, b]$, we have $0 < f'(x_0) \leq f'(d)$. Hence $(f(x_0)/f'(x_0)) \leq x_0 - r$. Thus,

$$x_1 := x_0 - \frac{f(x_0)}{f'(x_0)} \geq r.$$

This means that we are in one of the previous cases, where the initial value is $\geq r$. Consequently, $x_n \to r$. This proves the proposition when f' is positive and monotonically increasing on $[a, b]$.

If f' is negative and monotonically decreasing on $[a, b]$, then it suffices to consider $-f$ and note that the Newton sequences for $-f$ and f are identical, provided both have the same initial point.
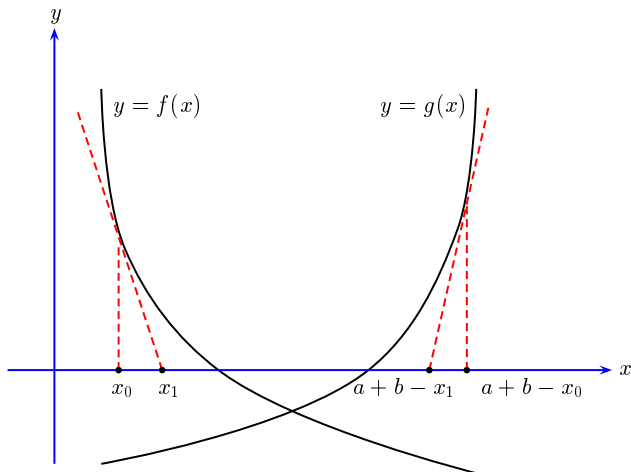
Fig. 5.9. Newton iterates for $f : [a, b] \to \mathbb{R}$ when f' is negative and monotonically increasing, and for its 'reflection' $g : [a, b] \to \mathbb{R}$ defined by $g(x) := f(a + b - x)$

If f' is negative and monotonically increasing, then it suffices to consider its 'reflection' along the vertical line $x = (b - a)/2$, that is, the function $g : [a, b] \to \mathbb{R}$ defined by $g(x) := f(a + b - x)$ for $x \in [a, b]$, and note the following. First, g is differentiable and $s := a + b - r$ is a solution of $g(x) = 0$ in $[a, b]$. Next, g' is positive and monotonically increasing on $[a, b]$. Further, if (x_n) is the Newton sequence for f with its initial point x_0, then $(a + b - x_n)$ is the Newton sequence for g with its initial point $a + b - x_0$. Finally, if $a + b - x_n \to s$, then $x_n \to r$. [See Figure 5.9.]

If f' is positive and monotonically decreasing, then it suffices to consider $-f$ and use the result of the previous paragraph.   □

Corollary 5.25. *Suppose $f : [a, b] \to \mathbb{R}$ is twice differentiable and $f(r) = 0$ for some $r \in [a, b]$. If f' is nonzero throughout $[a, b]$ and f'' does not change sign throughout $[a, b]$, then r is the unique solution of $f(x) = 0$ in $[a, b]$ and the Newton sequence for f with any initial point $x_0 \in [a, b]$ converges to r.*

Proof. Applying part (i) of Corollary 4.28 to f', we see that f' is monotonic on $[a, b]$. Now use Proposition 5.24.   □

To end this section, we remark that if $f(x)$ is a polynomial of degree ≥ 2, then $f'(x)$ and $f''(x)$ are nonzero polynomials of smaller degree, and, in particular, they have finitely many roots. Thus, the real line can be divided into finitely many intervals in each of which f' and f'' are nonzero and do not change signs. In particular, for any root r of f, we can find $a, b \in \mathbb{R}$ such that the restriction of f to $[a, b]$ satisfies the hypothesis of Corollary 5.25. In this way, we may say that the Newton method is always applicable to polynomials, provided we keep away from points at which the derivative vanishes.

Notes and Comments

The applications of differentiation to various tests for local extrema and points of inflection are bread-and-butter topics in calculus courses, so much so that many students think of these tests as definitions of the concepts such as local minimum, local maximum, and point of inflection. However, these concepts are basically of a geometric nature. In fact, this was the reason why we introduced these concepts in Chapter 1 before discussing the notion of derivative. In a similar way, many students try to use the Second Derivative Test when asked to find the absolute extrema of a real-valued function (on, say, a closed and bounded interval). The fact of the matter is that for finding absolute extrema, this test is neither necessary nor sufficient! To emphasize this point, we have arranged the discussion of absolute minima and maxima before the discussion of the Second Derivative Test, which is useful in finding local maxima and minima.

The method of Picard that we have discussed in the last section of this chapter is perhaps a starting point of an area of mathematics, known as fixed point theory, that has grown considerably over the years. Fixed point theorems such as the Picard Convergence Theorem and its generalizations are extremely useful in proving the existence and uniqueness of solutions of certain differential equations with prescribed initial conditions. For an introduction, we refer to the delightful book of Simmons [55]. The method of Newton for finding approximate solutions can be found toward the beginning of any book on numerical analysis. The fact that it converges very rapidly is almost folklore. But precise results about conditions that ensure the convergence of Newton sequences seem a bit difficult to locate. Results similar to the last proposition in this chapter can be found, for example, in the little booklet of Vilenkin [65] and the substantive book on calculus by Klambauer [40].

Exercises

Part A

1. In each of the following, find the greatest and the least value of $f : D \to \mathbb{R}$ where $D \subseteq \mathbb{R}$ and f are given by the following:
   (i) $D := [0, 2]$ and $f(x) := 4x^3 - 8x^2 + 5x$,
   (ii) $D := \mathbb{R}$ and $f(x) := (x + 2)^2/(x^2 + x + 1)$,
   (iii) $D := [-2, 5]$ and $f(x) := 1 + 12|x| - 3x^2$.
2. Given any constants $a, b \in \mathbb{R}$ with $a > b$, find the value of x at which the difference $\left(x/\sqrt{x^2 + a^2}\right) - \left(x/\sqrt{x^2 + b^2}\right)$ has the maximum value.
3. If $n \in \mathbb{N}$ is odd and the polynomial $1 + x + (1/2!)x^2 + \cdots + (1/n!)x^n$ has only one real root $x = c$, then show that

$$1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + \frac{x^{n+1}}{(n+1)!} \geq \frac{c^{n+1}}{(n+1)!} \quad \text{for all } x \in \mathbb{R}.$$

4. A window is to be made in the form of a rectangle surmounted by a semicircular portion with diameter equal to the base of the rectangle. The rectangular portion is to be of clear glass and the semicircular portion is to be of colored glass admitting only half as much light per square foot as the clear glass. If the total perimeter of the window frame is to be p feet, find the dimensions of the window which will admit the maximum amount of light.

5. The stiffness of a rectangular beam is proportional to the product of its breadth and the cube of its thickness but is not related to its length. Find the proportions of the stiffest beam that can be cut from a cylindrical log of diameter d inches.

6. A post office will accept a box for shipment only if the sum of its length and its girth (that is, distance around) does not exceed 84 inches. Find the dimensions of the largest acceptable box with a square end.

7. A wire of length ℓ inches is cut into two pieces, one being bent to form a square and the other to form an equilateral triangle. How should the wire be cut (i) if the sum of the two areas is minimum? (ii) if the sum of the two areas is maximum?

8. Let $D \subseteq \mathbb{R}$ and c be an interior point of D. If $f : D \to \mathbb{R}$ is twice differentiable at c and $f''(c) \neq 0$, then prove that f has a local extremum at c if and only if $f'(c) = 0$.

9. Let $D \subseteq \mathbb{R}$, c be an interior point of D and $f : D \to \mathbb{R}$ be differentiable at c. If c is a point of inflection for f, then is it necessarily true that $f'(c) = 0$? On the other hand, if $f'(c) = 0$, then is it necessarily true that either f has a local extremum at c or c is a point of inflection for f? (Compare Example 7.19.)

10. Find the local maxima and the local minima of $f : [0, 1] \to \mathbb{R}$ defined by $f(x) = x^m (1 - x)^n$ for $x \in [0, 1]$, where m and n are positive integers.

11. For which constants $a, b, c, d \in \mathbb{R}$ does the function $f(x) = ax^3 + bx^2 + cx + d$, $x \in \mathbb{R}$, have (i) a local maximum at -1, (ii) 1 as its point of inflection, and (iii) $f(-1) = 10$ and $f(1) = -6$?

12. Sketch the following curves after locating intervals of increase/decrease, intervals of convexity/concavity, points of local maxima/minima, and points of inflection. How many times and approximately where does the curve cross the x-axis?
    (i) $y = 2x^3 + 2x^2 - 2x - 1$ (ii) $y = x^3 - 6x^2 + 9x + 1$,
    (iii) $y = x^2/(x^2 + 1)$,   (iv) $y = 1/(1 + x^2)$,
    (v) $y = x/(x - 1)$, $x \neq 1$ (vi) $y = x/(x + 1)$, $x \neq -1$,
    (vii) $y = x^2/(x^2 - 1)$, $x \neq \pm 1$,   (viii) $y = (x^2 + 1)/x$, $x \neq 0$,
    (ix) $y = 1 + 12|x| - 3x^2$, $x \in [-2, 5]$,   (x) $y = (x^2 + x - 2)/(x - 2)$, $x \neq 2$.

13. Sketch a continuous curve $y = f(x)$ having the following properties:
    $f(-2) = 8$, $f(0) = 4$, $f(2) = 0$; $f'(2) = f'(-2) = 0$;
    $f'(x) > 0$ for $|x| > 2$, $f'(x) < 0$ for $|x| < 2$;
    $f''(x) < 0$ for $x < 0$ and $f''(x) > 0$ for $x > 0$.

14. Consider $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = (x - 2n)^3 + 2n$, where $n \in \mathbb{Z}$ is such that $x \in [2n - 1, 2n + 1)$. Show that $2n$ is a point of inflection for f, for each $n \in \mathbb{N}$. (Compare Exercise 33 of Chapter 4.)

15. Use linear approximation to find an approximate value of
    (i) $(8.01)^{4/3} + (8.01)^2 - (8.01)^{-1/3}$,   (ii) $(9.1)^{3/2} + (9.1)^{-1/2}$.

16. (i) Find an approximate value of $\sqrt{3}$ using the linear approximation to $f(x) = \sqrt{x}$ for x around 4.
    (ii) Let $f(x) = \sqrt{x} + \sqrt{x + 1} - 4$. Show that there is a unique $x_0 \in (3, 4)$ such that $f(x_0) = 0$. Using the linear approximation to f around 3, find an approximation x_1 of x_0. Find x_0 exactly and determine the error $|x_1 - x_0|$.

17. Consider the following functions:
    (i) $f(x) := \sqrt{1 + x}, \ x \geq -1$,   (ii) $f(x) := 1/\sqrt{1 - x}, \ x \leq 1$.
    For each of them, find:
    (a) The linear approximation $L(x)$ around 0.
    (b) An estimate for the error $e_1(x)$ when $x > 0$ and when $x < 0$. Also, find an upper bound for $|e_1(x)|$ that is valid for all $x \in (0, 0.1)$, and an upper bound for $|e_1(x)|$ that is valid for all $x \in (-0.1, 0)$.
    (c) The quadratic approximation $Q(x)$ around 0.
    (d) An estimate for the error $e_2(x)$ when $x > 0$ and when $x < 0$. Also, an upper bound for $|e_2(x)|$ that is valid for all $x \in (0, 0.1)$, and an upper bound for $|e_2(x)|$ that is valid for all $x \in (-0.1, 0)$.

18. Let $D \subseteq \mathbb{R}$ and c be an interior point of D. Suppose $F : D \to \mathbb{R}$ is the polynomial function defined by $F(x) = a_0 + a_1(x - c) + a_2(x - c)^2$ for $x \in D$. If a function $f : D \to \mathbb{R}$ is differentiable at c, then show that F is the linear approximation to f around c if and only if

$$f(c) = F(c) = a_0, \quad f'(c) = F'(c) = a_1, \quad \text{and} \quad a_2 = 0,$$

whereas if f is twice differentiable at c, then show that F is the quadratic approximation to f around c if and only if

$$f(c) = F(c) = a_0, \quad f'(c) = F'(c) = a_1, \quad \text{and} \quad f''(c) = F''(c) = a_2.$$

19. Let $f(x) := \sqrt{x} + (1/\sqrt{x})$ for $x > 0$. Write down the linear and the quadratic approximations $L(x)$ and $Q(x)$ to $f(x)$ around 4. Find the errors $f(4.41) - L(4.41)$ and $f(4.41) - Q(4.41)$.

20. Let $D \subseteq \mathbb{R}$ and c be an interior point of D. If $f : D \to \mathbb{R}$ is continuous at c and we let $e_0(x) = f(x) - f(c)$ for $x \in D$, then show that $e_0(x) \to 0$ as $x \to c$. Moreover, given any $b \in D$ with $b \neq c$, if I_b denotes the open interval with c and b as its endpoints, and if f' exists on I_b and $|f'(x)| \leq M_1(b)$ for all $x \in I_b$, then show that $|e_0(b)| \leq M_1(b)|b - c|$.

21. Let $D \subseteq \mathbb{R}$ and c be an interior point of D. If $f : D \to \mathbb{R}$ is n times differentiable at c and $P_n(x)$ denotes the nth Taylor polynomial of f around c and if $e_n(x) := f(x) - P_n(x)$ for $x \in D$, then show that the limit of $e_n(x)/x - c)^n$ as $x \to c$ is zero. Moreover, given any $b \in D$

with $b \neq c$, if I_b denotes the open interval with c and b as its endpoints, $f', \ldots, f^{(n)}$ are continuous on the closed interval between c and b, $f^{(n+1)}$ exists on I_b, and if $|f^{(n+1)}(x)| \leq M_{n+1}(b)$ for all $x \in I_b$, then show that

$$|e_n(b)| \leq \frac{M_{n+1}(b)}{(n+1)!}|b - c|^{n+1}.$$

(Hint: L'Hôpital's Rule for $\frac{0}{0}$ indeterminate forms and Taylor's formula.)

22. Consider $f : [0,1] \to \mathbb{R}$ defined by $f(x) = 1/(1 + x^2)$. Use the Picard Convergence Theorem to show that f has a unique fixed point in $[0,1]$ and any Picard sequence with its initial point $x_0 \in [0,1]$ will converge to this fixed point. Compute the first few values of the Picard sequence for f when $x_0 = 0$.

23. Consider $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2 - 2x - 3$. If $x_0 \neq 1$, then show that the Newton sequence with its initial point x_0 converges to -1 if $x_0 < 1$, and to 3 if $x_0 > 1$,

24. Consider $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^4 - x^3 - 75$. Show that there is a unique $r_1 \in [3, 4]$ such that $f(r_1) = 0$ and there is a unique $r_2 \in [-3, -2]$ such that $f(r_2) = 0$. Use the Newton method with initial point
    (i) $x_0 = 3$,   (ii) $x_0 = -3$,
    to find approximate values of the solutions r_1 and r_2 of $f(x) = 0$.

25. Consider $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = (x - 1)^{1/3}$. Show that the Newton sequence for f with its initial point $x_0 \neq 1$ is unbounded.

Part B

26. Let $f : \mathbb{R} \to \mathbb{R}$ be such that

$$f'(x) = 6(x - 1)(x - 2)^2(x - 3)^3(x - 4)^4 \quad \text{for all } x \in \mathbb{R}.$$

Find all the points (in $\mathbb{R}$) at which f has a local extremum. Also, find all the points of inflection for f.

27. Let I be an interval in $\mathbb{R}$, $f : I \to \mathbb{R}$, and c be an interior point of I.
    (i) Suppose there is $n \in \mathbb{N}$ such that $f^{(2n)}$ exists at c, and $f'(c) = f''(c) = \cdots = f^{(2n-1)}(c) = 0$. If $f^{(2n)}(c) < 0$, then show that f has a strict local maximum at c, whereas if $f^{(2n)}(c) > 0$, then show that f has a strict local minimum at c. (Hint: Taylor's formula.)
    (ii) Suppose there is $n \in \mathbb{N}$ such that $f^{(2n+1)}$ exists at c, and $f''(c) = f'''(c) = \cdots = f^{(2n)}(c) = 0$. If $f^{(2n+1)}(c) \neq 0$, then show that c is a strict point of inflection for f. (Hint: Taylor's formula.)
    (iii) Suppose that f is infinitely differentiable at c and $f'(c) = 0$, but $f^{(k)}(c) \neq 0$ for some $k \in \mathbb{N}$. Show that either f has a strict local extremum at c, or c is a strict point of inflection for f.

28. Let $f : [a, b] \to [a, b]$ be continuous and monotonic. Then show that for any $x_0 \in [a, b]$, the Picard sequence for f with its initial point x_0 converges to a fixed point of f. (Hint: Show that the Picard sequence (x_n) is monotonic by considering separately the cases $x_0 \leq x_1$ and $x_0 \geq x_1$.)

29. Let $D \subseteq \mathbb{R}$ and $f : D \to \mathbb{R}$ be such that $f(D) \subseteq D$. Prove the following generalizations and extensions of the Picard Convergence Theorem.

  (i) If D is closed and f is a **contraction** (that is, there is $\alpha \in \mathbb{R}$ with $0 \leq \alpha < 1$ such that $|f(x) - f(y)| \leq \alpha |x - y|$ for all $x, y \in D$), then f has a unique fixed point, and any Picard sequence will converge to this fixed point. Give an example to show that if f is a contraction but D is not closed, then f need not have a fixed point. (Hint: See Example 5.16 (ii).)

  (ii) If D is closed and bounded, and f is **contractive** (that is, $|f(x) - f(y)| < |x - y|$ for all $x, y \in D$, $x \neq y$), then f has a unique fixed point, and any Picard sequence will converge to this fixed point. Give an example to show that if f is a contractive but D is not closed and bounded, then f need not have a fixed point. (Hint: See the proof of Proposition 5.17 and Example 5.16 (iii).)

  (iii) If D is a closed and bounded interval, and f is **nonexpansive** (that is, $|f(x) - f(y)| \leq |x - y|$ for all $x, y \in D$), then f has a fixed point in D but it may not be unique. Give an example to show that if f is nonexpansive but D is not a closed and bounded interval, then f need not have a fixed point. (Hint: See the proof of Proposition 5.15 and Example 5.16 (iv).)

30. Let $f : (a, b) \to \mathbb{R}$ be a differentiable function such that f' is bounded on (a, b), and f has a root $r \in (a, b)$. For $x \in (a, b)$, $x \neq r$, let J_x denote the open interval between r and x. Assume that if $f(x) > 0$, then f is convex on J_x, while and if $f(x) < 0$, then f is concave on J_x. Show that the Newton sequence with any initial point $x_0 \in (a, b)$ converges to a root of f in (a, b).

31. Let $a, b \in \mathbb{R}$ with $a < b$ and $f : (a, b) \to \mathbb{R}$ be any function.

  (i) Suppose f is differentiable and there is $c \in (a, b)$ such that $f(c) = c$. If f' is continuous at c and $|f'(c)| < 1$, then show that there is a closed subinterval I of (a, b) with $c \in I$ such that f maps I into itself, c is the only fixed point of f in I, and the Picard sequence with any initial point $x_0 \in I$ converges to c.

  (ii) Suppose f is twice differentiable, $f'(x) \neq 0$ for all $x \in (a, b)$, and there is $r \in (a, b)$ such that $f(r) = 0$. If f'' is continuous at c, then show that there is a closed subinterval I of (a, b) with $r \in I$ such that r is the only solution of $f(x) = 0$ in I, and the Newton sequence with any initial point $x_0 \in I$ converges to r.

32. **(Linear convergence of the Picard method)** Let $f : [a, b] \to [a, b]$ be a continuous function such that f' exists and is bounded on (a, b). If f has a fixed point $c^* \in [a, b]$, then show that there is a constant $\alpha \in \mathbb{R}$ such that given any Picard sequence (x_n) for f with its initial point $x_0 \in [a, b]$, we have $|x_n - c^*| \leq \alpha |x_{n-1} - c^*|$ for all $n \in \mathbb{N}$. Deduce that $|x_n - c^*| \leq \alpha^n |x_0 - c^*|$ for all $n \in \mathbb{N}$.
[Note: In case $\alpha < 1$, the former inequality shows that the terms of the Picard sequence give a successively better approximation of the fixed point

c^* of f. This inequality can also be interpreted to say that the Picard
sequence **converges linearly**. On the other hand, the latter inequality
gives an error bound for the (nth term of the) Picard sequence.]

33. **(Quadratic convergence of the Newton method)** Let $f : [a, b] \to \mathbb{R}$
    be a differentiable function such that f' is continuous on $[a, b]$, $f'(x) \neq 0$
    for all $x \in [a, b]$, f'' exists and is bounded on (a, b). If $f(r) = 0$ for some
    $r \in [a, b]$, then show that r is the unique solution of $f(x) = 0$ in $[a, b]$
    and there is a constant $\alpha \in \mathbb{R}$ such that given any Newton sequence
    (x_n) for f with its initial point $x_0 \in [a, b]$, we have $|x_n - r| \leq \alpha |x_{n-1} - r|^2$, provided $x_{n-1}, x_n \in [a, b]$. Deduce that $|x_n - r| \leq \alpha^{2^n - 1}|x_0 - r|^{2^n}$,
    provided $x_1, \ldots, x_n \in [a, b]$.
    [Note: In case $\alpha < 1$, the former inequality shows that the terms of the
    Newton sequence give a successively better approximation of the solu-
    tion r of $f(x) = 0$. This inequality can also be interpreted to say that
    the Newton sequence **converges quadratically**. On the other hand, the
    latter inequality gives an error bound for the (nth term of the) Newton
    sequence.]

34. Let (x_n) be a sequence in $\mathbb{R}$ and $c \in \mathbb{R}$ such that $x_n \to c$. Assume that
    there is $n_0 \in \mathbb{N}$ such that $x_n \neq c$ for all $n \geq n_0$. If there is a real number
    p such that
    $$\alpha := \lim_{n \to \infty} \frac{|x_n - c|}{|x_{n-1} - c|^p}$$
    exists and is nonzero, then p is called the **order of convergence** of (x_n)
    to c and α is called the corresponding **asymptotic error constant**.
    (i) Let $f : (a, b) \to (a, b)$ and $x_0 \in (a, b)$ be such that the Picard sequence
        (x_n) for f with its initial point x_0 converges to some $x^* \in (a, b)$. If f
        is continuous at x^*, then show that x^* is a fixed point of f. Further, if
        f is p times differentiable at x^* and $f'(x^*) = \cdots = f^{(p-1)}(x^*) = 0$ but
        $f^{(p)}(x^*) \neq 0$, then show that the order of convergence of (x_n) to x^*
        is p and the corresponding asymptotic error constant is $|f^{(p)}(x^*)|/p!$.
        (Hint: Note that $x_n - x^* = f(x_{n-1}) - f(x^*) - f'(x^*)(x_{n-1} - x^*) - \cdots - f^{(p-1)}(x^*)(x_{n-1} - x^*)^{p-1}/(p-1)!$, and use L'Hôpital's Rule.)
    (ii) Let $f : (a, b) \to \mathbb{R}$ and $x_0 \in (a, b)$ be such that f is differentiable and
        $f'(x) \neq 0$ for all $x \in (a, b)$. Assume that the Newton sequence (x_n) for
        f with its initial point x_0 converges to some $r \in (a, b)$. If f' is bounded
        on (a, b), then show that r is a solution of $f(x) = 0$. Further, if f is
        twice differentiable at r and $f''(r) \neq 0$, then show that the order
        of convergence of (x_n) to r is 2, and the corresponding asymptotic
        error constant is $|f''(r)|/2|f'(r)|$. (Hint: Note that $f(x_{n-1})(x_n - r) = [f'(x_{n-1}) - (f(x_{n-1}) - f(r))/(x_{n-1} - r)](x_{n-1} - r)$ for $n \in \mathbb{N}$.)

6

Integration

In this chapter, we embark upon a project that is of a very different kind as compared to our development of calculus and analysis so far, namely the project of finding the 'area' of a planar region of a certain kind. This leads us to the theory of integration propounded by Riemann. Although this theory would seem unrelated to continuity and differentiability of functions, it has deep underlying connections. These connections manifest themselves mainly in the form of a central result known as the Fundamental Theorem of Calculus. In Section 6.1 below, we motivate and formulate a definition of Riemann integral. Later in this section we prove a useful characterization of the integrability of functions, and also a key property of the Riemann integral known as domain additivity. Next, in Section 6.2, we establish a number of basic properties of integrable functions. The Fundamental Theorem of Calculus and several of its consequences are proved in Section 6.3. In Section 6.4, we show that the Riemann integral of a function can be approximated by certain sums involving its values at more or less randomly chosen points. This approach yields an alternative definition of the Riemann integral via a result of Darboux.

6.1 The Riemann Integral

Consider a nonnegative bounded function defined on an interval $[a, b]$. Let us investigate whether we can assign a meaning to what can be called the 'area' of the region that lies under the graph of such a function, between the lines $x = a$, $x = b$ and above the x-axis. The only thing that we shall *assume* is that the area of a rectangle $[x_1, x_2] \times [y_1, y_2]$ is equal to $(x_2 - x_1)(y_2 - y_1)$. Following Archimedes, our problem can be approached by subdividing the interval $[a, b]$ into a finite number of subintervals and then finding the sum of the areas of rectangles inscribed within the region and also the sum of the areas of rectangles that circumscribe the region. [See Figure 6.1.]

While the sum of the areas of the inscribed rectangles ought to be less than the expected 'area' of the region, the sum of the areas of the circumscribing
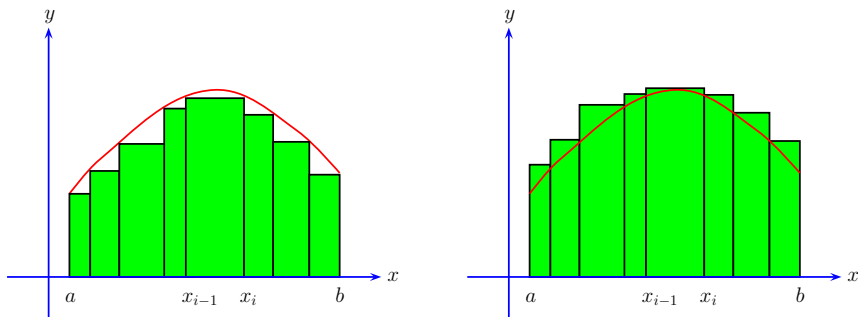
Fig. 6.1. Approximating the 'area' under a curve by means of inscribed or circumscribing rectangles

rectangles ought to be larger than it. Further, if the given function is 'well behaved', both these sums should come close to the expected 'area' of the region if the subdivision of the interval is made finer and finer.

To proceed formally, we introduce the following concept. By a **partition** of an interval $[a, b]$ (where $a, b \in \mathbb{R}$ and $a < b$) we mean a finite ordered set $\{x_0, x_1, \ldots, x_n\}$ of points in $[a, b]$ such that

$$a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b.$$

The simplest partition of $[a, b]$ is given by $\mathsf{P}_1 := \{a, b\}$. More generally, for $n \in \mathbb{N}$, the partition $\mathsf{P}_n := \{x_0, x_1, \ldots, x_n\}$, where

$$x_i = a + \frac{i(b - a)}{n} \quad \text{for } i = 0, 1, \ldots, n,$$

subdivides the interval $[a, b]$ into n subintervals, each of length $(b - a)/n$. We may refer to P_n as the partition of $[a, b]$ into n equal parts. It is clear that as n becomes larger, the subdivision of $[a, b]$ corresponding to P_n becomes uniformly finer.

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

$$m(f) := \inf\{f(x) : x \in [a, b]\} \quad \text{and} \quad M(f) := \sup\{f(x) : x \in [a, b]\}.$$

Given a partition $P = \{x_0, x_1, \ldots, x_n\}$ of $[a, b]$, let

$$m_i(f) := \inf\{f(x) : x \in [x_{i-1}, x_i]\} \quad \text{and} \quad M_i(f) = \sup\{f(x) : x \in [x_{i-1}, x_i]\}$$

for $i = 1, \ldots, n$. Clearly,

$$m(f) \leq m_i(f) \leq M_i(f) \leq M(f) \quad \text{for all } i = 1, \ldots, n.$$

We define the **lower sum** and the **upper sum** for the function f with respect to the partition P as follows:

$$L(P, f) := \sum_{i=1}^{n} m_i(f)(x_i - x_{i-1}) \quad \text{and} \quad U(P, f) := \sum_{i=1}^{n} M_i(f)(x_i - x_{i-1}).$$

We note that if f is nonnegative, then the lower sum is the sum of the areas of the inscribed rectangles and the upper sum is the sum of the areas of the circumscribing rectangles mentioned earlier.

Proposition 6.1. *Let* $f : [a, b] \to \mathbb{R}$ *be a bounded function. Then for any partition* P *of* $[a, b]$, *we have*

$$m(f)(b - a) \leq L(P, f) \leq U(P, f) \leq M(f)(b - a).$$

Proof. Let $P = \{x_0, x_1, \ldots, x_n\}$ be a partition of $[a, b]$. Since $m(f) \leq m_i(f) \leq M_i(f) \leq M(f)$ for each $i = 1, \ldots, n$ and $\sum_{i=1}^{n}(x_i - x_{i-1}) = b - a$, the desired inequalities follow. $\square$

Our goal is to look for partitions of $[a, b]$ with respect to which the lower sums are as large as possible and the upper sums are as small as possible, so that the expected 'area' will get tightly caught between the lower sums and the upper sums. With this mind, we define

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

and

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

It is natural to expect that $L(f) \leq U(f)$. To prove this, we need the following concepts. Given a partition P of $[a, b]$, we say that a partition P^* of $[a, b]$ is a **refinement** of P if every point of P is also a point of P^*. Given partitions P_1 and P_2 of $[a, b]$, the partition P^* consisting entirely of the points of P_1 and the points of P_2 is called the **common refinement** of P_1 and P_2.

Lemma 6.2. *Let* $f : [a, b] \to \mathbb{R}$ *be a bounded function.*

(i) *If* P *is partition of* $[a, b]$, *and* P^* *is a refinement of* P, *then*

$$L(P, f) \leq L(P^*, f) \quad \text{and} \quad U(P^*, f) \leq U(P, f),$$

and consequently,

$$U(P^*, f) - L(P^*, f) \leq U(P, f) - L(P, f).$$

(ii) *If* P_1 *and* P_2 *are partitions of* $[a, b]$, *then* $L(P_1, f) \leq U(P_2, f)$.
(iii) $L(f) \leq U(f)$.

Proof. (i) Let $P = \{x_0, x_1, \ldots, x_n\}$ be a partition of $[a, b]$. First let us assume that a refinement P^* of P contains just one additional point x^*. Then $x_{i-1} \leq x^* \leq x_i$ for some $i \in \{1, \ldots, n\}$. Using the abbreviations 'ℓ' and 'r' for 'left' and 'right' respectively, define

$$M_\ell^* := \sup\{f(x) : x \in [x_{i-1}, x^*]\} \quad \text{and} \quad M_r^* := \sup\{f(x) : x \in [x^*, x_i]\}.$$

Clearly, M_ℓ^* and M_r^* are both less than or equal to $M_i(f)$. Also, we have $x_i - x_{i-1} = (x_i - x^*) + (x^* - x_{i-1})$ and therefore,

$$\begin{aligned}
U(P, f) - U(P^*, f) &= M_i(f)(x_i - x_{i-1}) - M_\ell^*(x^* - x_{i-1}) - M_r^*(x_i - x^*) \\
&= (M_i(f) - M_\ell^*)(x^* - x_{i-1}) + (M_i(f) - M_r^*)(x_i - x^*) \\
&\geq 0 + 0 = 0.
\end{aligned}$$

If a refinement P^* of P contains several additional points, we repeat the above argument several times and again obtain $U(P^*, f) \leq U(P, f)$. The proof of $L(P, f) \leq L(P^*, f)$ is similar. Subtracting, we obtain $U(P^*, f) - L(P^*, f) \leq U(P, f) - L(P, f)$.

(ii) Let P^* denote the common refinement of partitions P_1 and P_2. Then in view of (i) above,

$$L(P_1, f) \leq L(P^*, f) \leq U(P^*, f) \leq U(P_2, f).$$

(iii) Let us fix a partition P_0 of $[a, b]$. By (ii) above, we have $L(P_0, f) \leq U(P, f)$ for any partition P of $[a, b]$. Hence $L(P_0, f)$ is a lower bound of the set $\{U(P, f) : P \text{ is a partition of } [a, b]\}$. Since $U(f)$ is the greatest lower bound of this set, we have $L(P_0, f) \leq U(f)$. Now, since P_0 is an arbitrary partition of $[a, b]$, we see that $U(f)$ is an upper bound of the set $\{L(P_0, f) : P_0 \text{ is a partition of } [a, b]\}$. Again, since $L(f)$ is the least upper bound of this set, we have $L(f) \leq U(f)$. $\qquad\square$

If a bounded function $f : [a, b] \to \mathbb{R}$ is nonnegative, and if we wish to define the 'area' of the region lying under the graph of f, between the lines $x = a$, $x = b$, and above the x-axis with the help of inscribed and circumscribing rectangles, then the 'area' must be at least $L(f)$ and it can be at most $U(f)$. Thus, if $L(f) = U(f)$, then it would be appropriate to define the area to be this common value. This motivates the following definition.

Let $f : [a, b] \to \mathbb{R}$ be a bounded function. Then f is said to be **integrable** (on $[a, b]$) if $L(f) = U(f)$. In this case, the common value $L(f) = U(f)$ is called the **Riemann integral** of f (on $[a, b]$) and it is denoted by

$$\int_a^b f(x)dx \quad \text{or simply by} \quad \int_a^b f.$$

The notation $\int_a^b f(x)dx$ emphasizes that f is 'integrated' as a function of the 'variable' x. While the Riemann integral of f does not depend on the name of the variable, this notation is useful when several variables are being considered. The number $U(f)$ is known as the **upper Riemann integral** of f and the number $L(f)$ as the **lower Riemann integral** of f. Thus, a bounded function on $[a, b]$ is integrable if its upper Riemann integral is equal to its lower Riemann integral.

If, in addition, f is nonnegative, then the **area** of the region under the curve given by $y = f(x)$, $x \in [a, b]$, is defined to be the Riemann integral of f:

$$\text{Area}(R_f) := \int_a^b f(x)dx, \text{ where } R_f := \{(x, y) \in \mathbb{R}^2 : a \leq x \leq b,\ 0 \leq y \leq f(x)\}.$$

We shall show later that in general, if $f : [a, b] \to \mathbb{R}$ is any integrable function, then its Riemann integral is equal to $\text{Area}(R_{f^+}) - \text{Area}(R_{f^-})$, where f^+ and f^- are the positive and the negative parts of f. (See Remark 6.19.) In this sense, the Riemann integral of f represents, in general, the 'signed area' delineated by the curve $y = f(x)$, $x \in [a, b]$.

Admittedly, the definition of a Riemann integral of a bounded function $f : [a, b] \to \mathbb{R}$ is rather involved. This is because we need to consider lower sums for the function f with respect to all possible partitions of $[a, b]$ and calculate their supremum on the one hand, and also consider the corresponding upper sums and calculate their infimum on the other. We shall presently give several examples to illustrate what it takes to decide whether a bounded function on $[a, b]$ is integrable. When the definition of a Riemann integral has been well understood, it will be relatively easy to deduce its interesting properties and use them to obtain important results.

Now we give an elementary but useful estimate for the absolute value of a Riemann integral.

Proposition 6.3 (Basic Inequality for Riemann Integrals). *Suppose $f : [a, b] \to \mathbb{R}$ is an integrable function and there are $\alpha, \beta \in \mathbb{R}$ such that $\beta \leq f \leq \alpha$. Then*

$$\beta(b - a) \leq \int_a^b f(x)dx \leq \alpha(b - a).$$

In particular, if $|f| \leq \alpha$, then

$$\left| \int_a^b f(x)dx \right| \leq \alpha(b - a).$$

Proof. Since $\beta \leq f(x) \leq \alpha$ for all $x \in \mathbb{R}$, we see that $\beta \leq m(f)$ and $M(f) \leq \alpha$. Let $P := \{a, b\}$ denote the trivial partition of $[a, b]$. Then we have

$$\beta(b - a) \leq m(f)(b - a) = L(P, f) \quad \text{and} \quad U(P, f) = M(f)(b - a) \leq \alpha(b - a).$$

Hence it follows that

$$\beta(b - a) \leq L(f) = \int_a^b f(x)dx = U(f) \leq \alpha(b - a).$$

If $|f| \leq \alpha$, then letting $\beta := -\alpha$, we obtain the desired conclusion. $\qquad \square$

We work out a few examples to show how the integrability of a function can be investigated from first principles.

Examples 6.4. (i) Consider the constant function on $[a, b]$ defined by $f(x) :=$ 1 for all $x \in [a, b]$. Then for every partition $P = \{x_0, x_1, \ldots, x_n\}$ of $[a, b]$, we have $m_i(f) = 1 = M_i(f)$ for all $i = 1, \ldots, n$ and so

$$L(P, f) = U(P, f) = \sum_{i=1}^{n} 1 \cdot (x_i - x_{i-1}) = b - a.$$

Hence $L(f) = b - a = U(f)$. Thus f is integrable and its Riemann integral is equal to $b - a$.
The above reasoning shows that if $r \in \mathbb{R}$ and $f(x) := r$ for all $x \in [a, b]$, then f is integrable on $[a, b]$ and

$$\int_a^b r \, dx = r(b - a).$$

(ii) Consider the **Dirichlet function** on $[a, b]$ defined by

$$f(x) := \begin{cases} 1 & \text{if } x \in [a, b] \text{ and } x \text{ is a rational number,} \\ 0 & \text{if } x \in [a, b] \text{ and } x \text{ is an irrational number.} \end{cases}$$

Let $P = \{x_0, x_1, \ldots, x_n\}$ be a partition of $[a, b]$. Since each $[x_{i-1}, x_i]$ contains a rational number as well as an irrational number, we see that $m_i(f) = 0$ and $M_i(f) = 1$ for all $i = 1, \ldots, n$, and so

$$L(P, f) = \sum_{i=1}^{n} 0 \cdot (x_i - x_{i-1}) = 0, \quad \text{but} \quad U(P, f) = \sum_{i=1}^{n} 1 \cdot (x_i - x_{i-1}) = b - a.$$

Hence $L(f) = 0$ and $U(f) = b - a$. Since $a < b$, we have $L(f) \neq U(f)$, that is, f is not integrable.
(iii) Consider the identity function on $[a, b]$ defined by $f(x) := x$ for all $x \in [a, b]$. Let $P = \{x_0, x_1, \ldots, x_n\}$ be a partition of $[a, b]$. Since $M_i(f) = x_i$ and $m_i(f) = x_{i-1}$ for $i = 1, \ldots, n$, we have

$$U(P, f) = \sum_{i=1}^{n} x_i(x_i - x_{i-1}) \quad \text{and} \quad L(P, f) = \sum_{i=1}^{n} x_{i-1}(x_i - x_{i-1}).$$

Hence

$$U(P, f) - L(P, f) = \sum_{i=1}^{n} (x_i - x_{i-1})^2$$

and

$$U(P, f) + L(P, f) = \sum_{i=1}^{n} (x_i^2 - x_{i-1}^2) = b^2 - a^2.$$

It follows that

$$U(P, f) = \frac{b^2 - a^2}{2} + \frac{1}{2} \sum_{i=1}^{n} (x_i - x_{i-1})^2$$

and

$$L(P, f) = \frac{b^2 - a^2}{2} - \frac{1}{2} \sum_{i=1}^{n} (x_i - x_{i-1})^2.$$

Now given $\epsilon > 0$, there is a partition $P = \{x_0, x_1, \ldots, x_n\}$ of $[a, b]$ such that $\sum_{i=1}^{n} (x_i - x_{i-1})^2 < \epsilon$. For example, let $P := \mathsf{P}_n$ be the partition of $[a, b]$ into n equal parts, where $n \in \mathbb{N}$ is so chosen that $(b - a)^2/n < \epsilon$. Then

$$\sum_{i=1}^{n} (x_i - x_{i-1})^2 = \sum_{i=1}^{n} \frac{(b-a)^2}{n^2} = \frac{(b-a)^2}{n} < \epsilon.$$

Hence we have

$$U(f) \leq \frac{b^2 - a^2}{2} + \frac{\epsilon}{2} \quad \text{and} \quad L(f) \geq \frac{b^2 - a^2}{2} - \frac{\epsilon}{2} \quad \text{for every } \epsilon > 0.$$

It follows that $U(f) \leq (b^2 - a^2)/2$ and $L(f) \geq (b^2 - a^2)/2$. But $L(f) \leq U(f)$ by part (iii) of Lemma 6.2. This shows that $L(f) = (b^2 - a^2)/2 = U(f)$, that is, f is integrable and

$$\int_a^b f(x)dx = \frac{b^2 - a^2}{2}.$$

This result will be generalized in Example 6.24 (i). $\diamond$

The foregoing examples indicate that it is not easy to determine whether a bounded function on $[a, b]$ is integrable, and when the function is in fact integrable, it may be even more difficult to evaluate its Riemann integral. We shall first take up the question of determining whether a bounded function on $[a, b]$ is integrable. We give a simple criterion for this purpose and use it extensively in Section 6.2, not only to find large classes of integrable functions, but also to obtain many important properties of the Riemann integral. The more involved question concerning the evaluation of the Riemann integral will be discussed in Sections 6.3, 6.4, and later in Section 8.6.

Proposition 6.5 (Riemann Condition). *Let $f : [a, b] \to \mathbb{R}$ be a bounded function. Then f is integrable if and only if for every $\epsilon > 0$, there is a partition P_ϵ of $[a, b]$ such that*

$$U(P_\epsilon, f) - L(P_\epsilon, f) < \epsilon.$$

Proof. Suppose that the stated condition is satisfied. Then for every $\epsilon > 0$, we have

$$0 \leq U(f) - L(f) \leq U(P_\epsilon, f) - L(P_\epsilon, f) < \epsilon.$$

Hence $L(f) = U(f)$, that is, f is integrable.

Conversely, assume that f is integrable. Let $\epsilon > 0$ be given. By the definitions of $U(f)$ and $L(f)$, there are partitions Q_ϵ and $\widetilde{Q}_\epsilon$ of $[a, b]$ such that

$$U(Q_\epsilon, f) < U(f) + \frac{\epsilon}{2} \quad \text{and} \quad L(\widetilde{Q}_\epsilon, f) > L(f) - \frac{\epsilon}{2}.$$

Let P_ϵ denote the common refinement of Q_ϵ and $\widetilde{Q}_\epsilon$. Then by part (i) of Lemma 6.2, we have

$$L(f) - \frac{\epsilon}{2} < L(\widetilde{Q}_\epsilon, f) \leq L(P_\epsilon, f) \leq U(P_\epsilon, f) \leq U(Q_\epsilon, f) < U(f) + \frac{\epsilon}{2}.$$

Since $L(f) = U(f)$, it follows that

$$U(P_\epsilon, f) - L(P_\epsilon, f) < \epsilon,$$

as desired   □

We give below a nontrivial example to illustrate the relative ease gained by the use of the Riemann condition, as compared to the difficulty we faced earlier in showing the integrability of the identity function (Example 6.4 (iii)), which is a particular case of the function considered in this nontrivial example.

Example 6.6. Let $a, b \in \mathbb{R}$ with $0 \leq a < b$, $m \in \mathbb{N}$, and $f(x) := x^m$ for all $x \in [a, b]$. Consider a partition $P = \{x_0, x_1, \ldots, x_n\}$ of $[a, b]$. Then $M_i(f) = x_i^m$ and $m_i(f) = x_{i-1}^m$ for $i = 1, \ldots, n$, and so

$$U(P, f) - L(P, f) = \sum_{i=1}^{n} (x_i^m - x_{i-1}^m)(x_i - x_{i-1}).$$

Also, we have for $i = 1, \ldots, n$,

$$(x_i^m - x_{i-1}^m) = (x_i^{m-1} + x_i^{m-2}x_{i-1} + \cdots + x_i x_{i-1}^{m-2} + x_{i-1}^{m-1})(x_i - x_{i-1})$$

and $0 \leq x_i^{m-1-j} x_{i-1}^j \leq b^{m-1}$ for $j = 0, 1, \ldots, m-1$. Hence we obtain

$$U(P, f) - L(P, f) \leq mb^{m-1} \sum_{i=1}^{n} (x_i - x_{i-1})^2.$$

Now given $\epsilon > 0$, there is a partition $P = \{x_0, x_1, \ldots, x_n\}$ of $[a, b]$ such that $\sum_{i=1}^{n} (x_i - x_{i-1})^2 < \epsilon$. For example, we may take $P := \mathsf{P}_n$ to be the partition of $[a, b]$ into n equal parts, where $n \in \mathbb{N}$ is so chosen that $(b-a)^2/n < \epsilon$. Then it follows that

$$U(P, f) - L(P, f) \leq mb^{m-1} \sum_{i=1}^{n} \left(\frac{b-a}{n}\right)^2 = mb^{m-1}\frac{(b-a)^2}{n} < mb^{m-1}\epsilon.$$

Since $\epsilon > 0$ is arbitrary, the Riemann condition is satisfied, and hence f is integrable. It may be noted that this procedure gives no clue about the evaluation of the Riemann integral of f. The evaluation will be accomplished in Example 6.24 (i). For a direct approach, see Exercise 40.   ◇

Next, we prove an important and useful result that says that if $[a, b]$ is divided into two subintervals, then the Riemann integral of $f : [a, b] \to \mathbb{R}$ is the sum of the Riemann integrals of the restrictions of f to the two subintervals. This corresponds to the geometric notion that if a region R splits into two nonoverlapping regions R_1 and R_2, then the area of R is equal to the sum of the areas of R_1 and R_2.

Proposition 6.7 (Domain Additivity of Riemann Integrals). *Let $f : [a, b] \to \mathbb{R}$ be a bounded function and let $c \in (a, b)$. Then f is integrable on $[a, b]$ if and only if f is integrable on $[a, c]$ and on $[c, b]$. In this case,*

$$\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx.$$

Proof. Assume that f is integrable on $[a, b]$. Let $\epsilon > 0$ be given. Then there is a partition P_ϵ of $[a, b]$ such that $U(P_\epsilon, f) - L(P_\epsilon, f) < \epsilon$, thanks to the Riemann condition. Adjoining c to the points of P_ϵ, if c is not already a point of P_ϵ, we obtain a refinement $P_\epsilon^* = \{x_0, x_1, \ldots, x_k, \ldots, x_n\}$ of P_ϵ, where $c = x_k$ for some $k \in \{1, \ldots, n - 1\}$. Part (i) of Lemma 6.2 shows that

$$0 \leq U(P_\epsilon^* f) - L(P_\epsilon^*, f) \leq U(P_\epsilon, f) - L(P_\epsilon, f) < \epsilon.$$

Now $Q_\epsilon^* := \{x_0, x_1, \ldots, x_k\}$ is a partition of $[a, c]$ and if g denotes the restriction of f to $[a, c]$, then we have

$$
\begin{aligned}
U(Q_\epsilon^*, g) - L(Q_\epsilon^*, g) &= \sum_{i=1}^k [M_i(g) - m_i(g)](x_i - x_{i-1}) \\
&= \sum_{i=1}^k [M_i(f) - m_i(f)](x_i - x_{i-1}) \\
&\leq \sum_{i=1}^n [M_i(f) - m_i(f)](x_i - x_{i-1}) \\
&= U(P_\epsilon^*, f) - L(P_\epsilon^*, f),
\end{aligned}
$$

which is less than ϵ. Hence the Riemann condition shows that g is integrable, that is, f is integrable on $[a, c]$. Similarly, it can be seen that f is integrable on $[c, b]$.

Conversely, assume that f is integrable on $[a, c]$ and on $[c, b]$. Let g and h denote the restrictions of f to $[a, c]$ and to $[c, b]$ respectively. Let $\epsilon > 0$ be given. By the Riemann condition, there are partitions Q_ϵ of $[a, c]$ and R_ϵ of $[c, b]$ such that

$$U(Q_\epsilon, g) - L(Q_\epsilon, g) < \frac{\epsilon}{2} \quad \text{and} \quad U(R_\epsilon, h) - L(R_\epsilon, h) < \frac{\epsilon}{2}.$$

Let P_ϵ denote the partition of $[a, b]$ obtained from the points of Q_ϵ followed by the points of R_ϵ. Then P_ϵ contains the point c. Therefore,

$$U(P_\epsilon, f) = U(Q_\epsilon, g) + U(R_\epsilon, h) \quad \text{and} \quad L(P_\epsilon, f) = L(Q_\epsilon, g) + L(R_\epsilon, h),$$

and so $U(P_\epsilon, f) - L(P_\epsilon, f) < \epsilon/2 + \epsilon/2 = \epsilon$. Thus, by the Riemann condition, f is integrable on $[a, b]$.

Let $\alpha := U(P_\epsilon, f)$ and $\beta := L(P_\epsilon, f)$. Evidently,

$$\beta \le \int_a^b f(x) dx \le \alpha.$$

Also, since $\alpha = U(Q_\epsilon, g) + U(R_\epsilon, h)$ and $\beta = L(Q_\epsilon, g) + L(R_\epsilon, h)$, we have

$$\beta \le \int_a^c f(x) dx + \int_c^b f(x) dx \le \alpha.$$

Since $\alpha - \beta < \epsilon$, it follows that

$$\left| \int_a^c f(x) dx + \int_c^b f(x) dx - \int_a^b f(x) dx \right| < \epsilon.$$

Since $\epsilon > 0$ is arbitrary, we see that

$$\int_a^b f(x) dx = \int_a^c f(x) dx + \int_c^b f(x) dx,$$

as desired.   □

We shall see in the next two sections that the domain additivity plays a crucial role in several proofs.

Remark 6.8. According to our convention stated in Chapter 1, we have assumed $a < b$ while defining the Riemann integral of $f : [a, b] \to \mathbb{R}$. In order to obtain uniformity of presentation and simplicity of notation, we adopt the following definitions: If $a = b$, then every $f : [a, b] \to \mathbb{R}$ is integrable and

$$\int_a^b f(x) dx := 0,$$

whereas if $a > b$ and $f : [b, a] \to \mathbb{R}$ is integrable, then

$$\int_a^b f(x) dx := - \int_b^a f(x) dx.$$

We emphasize that the Riemann integral of $f : [a, b] \to \mathbb{R}$ is defined over the *subset* $\{x : a \le x \le b\}$ of $\mathbb{R}$ and that we have not associated any direction or orientation with this subset. What we have mentioned above are mere *conventions*; they are not *results* that follow from our definition. In view of the domain additivity (Proposition 6.7), these conventions imply that

$$\int_c^d f(x) dx = \int_a^d f(x) dx - \int_a^c f(x) dx$$

for any points c and d in $[a, b]$.   ◇

6.2 Integrable Functions

In this section we shall use the Riemann condition to prove the integrability of a wide variety of functions. We shall also consider algebraic and order properties of the Riemann integral. Readers who wish to directly look at the fundamental connection between differentiation and Riemann integration may pass on to the next section, assuming only that every continuous function on $[a, b]$ is integrable. This is proved in part (ii) of the following proposition.

Proposition 6.9. *Let* $f : [a, b] \to \mathbb{R}$ *be a function.*

(i) *If* f *is monotonic, then it is integrable.*
(ii) *If* f *is continuous, then it is integrable.*

Proof. (i) Assume that $f : [a, b] \to \mathbb{R}$ is monotonically increasing. Then f is bounded since $f(a) \leq f(x) \leq f(b)$ for all $x \in [a, b]$. If $P = \{x_0, x_1, \ldots, x_n\}$ is any partition of $[a, b]$, then we have $M_i(f) = f(x_i)$ and $m_i(f) = f(x_{i-1})$ for $i = 1, \ldots, n$, and so

$$U(P, f) - L(P, f) = \sum_{i=1}^{n} [f(x_i) - f(x_{i-1})](x_i - x_{i-1}).$$

If $f(b) = f(a)$, then $f(x) = f(a)$ for all $x \in [a, b]$, and we see that $U(P, f) = f(a)(b-a) = L(P, f)$ for every partition P of $[a, b]$. If $f(b) > f(a)$, we proceed as follows. Let $\epsilon > 0$ be given. Consider a partition $P_\epsilon = \{x_0, x_1, \ldots, x_n\}$ of $[a, b]$ such that $x_i - x_{i-1} < \epsilon/[f(b) - f(a)]$ for $i = 1, \ldots, n$. Then we have

$$U(P_\epsilon, f) - L(P_\epsilon, f) < \sum_{i=1}^{n} [f(x_i) - f(x_{i-1})] \frac{\epsilon}{f(b) - f(a)} = \frac{[f(b) - f(a)]\epsilon}{f(b) - f(a)} = \epsilon.$$

Hence by the Riemann condition, f is integrable.

A similar proof holds if f is monotonically decreasing.

(ii) Assume that $f : [a, b] \to \mathbb{R}$ is continuous. Then f is bounded by part (i) of Proposition 3.8. Also, by Proposition 3.17, f is uniformly continuous. Let $\epsilon > 0$ be given. By Proposition 3.19, there is $\delta > 0$ such that

$$x, y \in [a, b], \ |x - y| < \delta \implies |f(x) - f(y)| < \frac{\epsilon}{b - a}.$$

Let $P_\epsilon = \{x_0, x_1, \ldots, x_n\}$ be a partition of $[a, b]$ such that $x_i - x_{i-1} < \delta$ for $i = 1, \ldots, n$. Now for each $i = 1, \ldots, n$, and $x, y \in [x_{i-1}, x_i]$, we have

$$f(x) - f(y) < \frac{\epsilon}{b - a}.$$

Taking the supremum for $x \in [x_{i-1}, x_i]$ and the infimum for $y \in [x_{i-1}, x_i]$, we obtain $M_i(f) - m_i(f) \leq \epsilon/(b - a)$. Hence

$$U(P_\epsilon, f) - L(P_\epsilon, f) = \sum_{i=1}^{n} [M_i(f) - m_i(f)](x_i - x_{i-1}) \le \frac{\epsilon}{b-a} \sum_{i=1}^{n} (x_i - x_{i-1}) = \epsilon.$$

Hence by the Riemann condition, f is integrable.   □

The above proposition shows that monotonicity or continuity of a function is a sufficient condition for its integrability. Since a monotonic function need not be continuous (for example, $f(x) := 0$ if $a \le x \le (a+b)/2$ and $f(x) := 1$ if $(a+b)/2 < x \le b$) and a continuous function need not be monotonic (for example, $f(x) := |x - (a+b)/2|$ if $x \in [a, b]$), it follows that neither monotonicity nor continuity is a necessary condition for integrability. In fact, Corollary 6.11 shows that if f is monotonic or continuous on a finite number of 'pieces' constituting the interval $[a, b]$, then it is integrable on $[a, b]$. To obtain this result, we first show that a 'piecewise' integrable function is integrable.

Proposition 6.10. *Let $f : [a, b] \to \mathbb{R}$ be a bounded function. Suppose there are $c_1 < \cdots < c_n$ in $[a, b]$ such that the function f is integrable on each of the subintervals $[a, a_1], [b_1, a_2], \ldots, [b_{n-1}, a_n], [b_n, b]$, whenever $a_1, \ldots, a_n$, $b_1, \ldots, b_n$ in $[a, b]$ are such that*

$$a_1 \le c_1 < b_1 < a_2 < c_2 < b_2 < \cdots < a_n < c_n \le b_n,$$

where the equality $a_1 = c_1$ holds only if $c_1 = a$ and the equality $c_n = b_n$ holds only if $c_n = b$. Then f is integrable on $[a, b]$ and

$$\int_a^b f(x)dx = \int_a^{c_1} f(x)dx + \int_{c_1}^{c_2} f(x)dx + \cdots + \int_{c_n}^b f(x)dx.$$

Proof. First we assume that there is only one point $c_1 \in [a, b]$ such that $c_1 \ne a$, $c_1 \ne b$, and f is integrable on the subintervals $[a, a_1]$ and $[b_1, b]$ whenever $a_1, b_1 \in [a, b]$ are such that $a_1 < c_1 < b_1$.

If f is constant on $[a, b]$, then it is integrable on $[a, b]$, as we saw in Example 6.4 (i). Assume now that f is not constant on $[a, b]$, and so $M(f) \ne m(f)$. Let $\epsilon > 0$ be given. Choose a_1 and b_1 in $[a, b]$ such that

$$a_1 < c_1 < b_1 \quad \text{and} \quad b_1 - a_1 < \frac{\epsilon}{3[M(f) - m(f)]}.$$

Let g_1 denote the restriction of f to $[a, a_1]$ and h_1 denote the restriction of f to $[b_1, b]$. Then g_1 and h_1 are given to be integrable on $[a, a_1]$ and on $[b_1, b]$ respectively. By the Riemann condition, there are partitions P_1 of $[a, a_1]$ and Q_1 of $[b_1, b]$ such that

$$U(P_1, g_1) - L(P_1, g_1) < \frac{\epsilon}{3} \quad \text{and} \quad U(Q_1, h_1) - L(Q_1, h_1) < \frac{\epsilon}{3}.$$

Let P_ϵ denote the partition of $[a, b]$ obtained from the points of P_1 followed by the points of Q_1. Thus P_ϵ contains the points a_1 and b_1. Now,

$$U(P_\epsilon, f) = U(P_1, g_1) + M_1^*(b_1 - a_1) + U(Q_1, h_1)$$

and

$$L(P_\epsilon, f) = L(P_1, g_1) + m_1^*(b_1 - a_1) + L(Q_1, h_1),$$

where

$$M_1^* = \sup\{f(x) : x \in [a_1, b_1]\} \quad \text{and} \quad m_1^* = \inf\{f(x) : x \in [a_1, b_1]\}.$$

Since $M_1^* - m_1^* \leq M(f) - m(f)$, it follows that

$$U(P_\epsilon, f) - L(P_\epsilon, f) < \frac{\epsilon}{3} + [M(f) - m(f)] \cdot \frac{\epsilon}{3[M(f) - m(f)]} + \frac{\epsilon}{3} = \epsilon.$$

By the Riemann condition, we see that f is integrable on $[a, b]$. Hence, by the domain additivity (Proposition 6.7), we have

$$\int_a^b f(x)dx = \int_a^{c_1} f(x)dx + \int_{c_1}^b f(x)dx.$$

If $c_1 = a$, then we let $a_1 := a$, and if $c_1 = b$, then we let $b_1 := b$ in the above argument and complete the proof.

In the general case involving n points $c_1, \ldots, c_n$ in $[a, b]$, the integrability of f follows by arguing as above repeatedly. $\qquad\square$

Corollary 6.11. *Let $f : [a, b] \to \mathbb{R}$ be a bounded function. Assume that there are points $c_1 < c_2 < \cdots < c_n$ in $[a, b]$ such that on each of the subintervals $[a, c_1), (c_1, c_2), \ldots, (c_{n-1}, c_n), (c_n, b]$, the function f is either monotonic or continuous. Then f is integrable on $[a, b]$.*

Proof. Consider points $a_1, b_1, \ldots, a_n, b_n \in [a, b]$ such that

$$a_1 \leq c_1 < b_1 < a_2 < c_2 < b_2 < \cdots < a_n < c_n \leq b_n,$$

where the equality $a_1 = c_1$ holds only if $c_1 = a$ and the equality $c_n = b_n$ holds only if $c_n = b$. Proposition 6.9 shows that f is integrable on each of the subintervals $[a, a_1], [b_1, a_2], \ldots, [b_{n-1}, a_n]$, and $[b_n, b]$. Hence by Proposition 6.10, f is integrable on $[a, b]$. $\qquad\square$

We shall now use Proposition 6.10 to prove an interesting property of the Riemann integral. Roughly speaking, it says that if the values of an integrable function are arbitrarily changed at a finite number of points, then the modified function is also integrable and its Riemann integral is equal to the Riemann integral of the given function.

Proposition 6.12. *Let $f : [a, b] \to \mathbb{R}$ be integrable, and $g : [a, b] \to \mathbb{R}$ be such that $\{x \in [a, b] : g(x) \neq f(x)\} = \{c_1, \ldots, c_n\}$. Then g is integrable and*

$$\int_a^b g(x)dx = \int_a^b f(x)dx.$$

Proof. Since f is bounded, and g differs from f at only a finite number of points, there is $\alpha > 0$ such that $|f(x)| \leq \alpha$ and $|g(x)| \leq \alpha$ for all $x \in [a, b]$.

To show that g is integrable, consider points $a_1, b_1, \ldots, a_n, b_n \in [a, b]$ such that $a_1 \leq c_1 < b_1 < a_2 < c_2 < b_2 < \cdots < a_n < c_n \leq b_n$, where the equality $a_1 = c_1$ holds only if $c_1 = a$ and the equality $c_n = b_n$ holds only if $c_n = b$. Now $g(x) = f(x)$ for all x in each of the subintervals $[a, a_1], [b_1, a_2], \ldots, [b_{n-1}, a_n]$, $[b_n, b]$, except if $c_1 = a$ and/or $c_n = b$. If $c_1 = a$, then $a_1 = a$ and g is clearly integrable on $[a, a_1]$, while if $c_n = b$, then $b_n = b$ and g is clearly integrable on $[b_n, b]$. Otherwise, since f is integrable on $[a, b]$, Proposition 6.7 shows that f is integrable on each of the subintervals $[a, a_1], [b_1, a_2], \ldots, [b_{n-1}, a_n], [b_n, b]$. In other words, g is integrable on each of these subintervals. Since the points $a_1, b_1, \ldots, a_n, b_n$ (satisfying the above inequalities) are arbitrary, it follows from Proposition 6.10 that g is integrable on $[a, b]$.

To show that the Riemann integral of g is equal to the Riemann integral of f, let $\epsilon > 0$ be given. Choose points $a_1, b_1, \ldots, a_n, b_n \in [a, b]$ satisfying the above-mentioned inequalities and also satisfying $(b_j - a_j) < \epsilon/2n\alpha$ for each $j = 1, \ldots, n$. By the domain additivity (Proposition 6.7), we see that

$$\int_a^b f(x)dx - \int_a^b g(x)dx = \sum_{j=1}^n \int_{a_j}^{b_j} f(x)dx - \sum_{j=1}^n \int_{a_j}^{b_j} g(x)dx.$$

By the basic inequality for Riemann integrals (Proposition 6.3), we have

$$\left| \int_{a_j}^{b_j} f(x)dx \right| \leq \alpha(b_j - a_j) \quad \text{and} \quad \left| \int_{a_j}^{b_j} g(x)dx \right| \leq \alpha(b_j - a_j) \quad \text{for } j = 1, \ldots, n.$$

Hence

$$\left| \int_a^b f(x)dx - \int_a^b g(x)dx \right| \leq \alpha \sum_{j=1}^n (b_j - a_j) + \alpha \sum_{j=1}^n (b_j - a_j) = 2n\alpha \cdot \frac{\epsilon}{2n\alpha} = \epsilon.$$

Since $\epsilon > 0$ is arbitrary, we conclude that the Riemann integral of f is equal to the Riemann integral of g. $\qquad\square$

Examples 6.13. (i) Let $f(x) := [x]$, the integral part of x, for all $x \in [a, b]$. Since f is (monotonically) increasing, it is integrable. This conclusion also follows by noting that f is bounded, and if $m_1, \ldots, m_n$ are the integers belonging to $[a, b]$, then f is continuous on each of the subintervals $[a, m_1], (m_1, m_2), \ldots, (m_{n-1}, m_n), (m_n, b]$.

(ii) Let $f(x) := |x|$, the absolute value of x, for all $x \in [a, b]$. Since f is continuous, it is integrable.

(iii) Let $f : [a, b] \to \mathbb{R}$ be a polynomial function, or more generally, a rational function whose denominator does not vanish at any point in $[a, b]$. Then f is continuous on $[a, b]$, and hence it is integrable.

(iv) Let $f : [0, 1] \to \mathbb{R}$ be the 'infinite-steps function' given by

$$f(x) := \begin{cases} 0 & \text{if } x = 0, \\ 1/n & \text{if } 1/(n+1) < x \leq 1/n \text{ for some } n \in \mathbb{N}. \end{cases}$$

[See Figure 6.2.] Since f is (monotonically) increasing on $[0, 1]$, it is integrable. Note that f is discontinuous at infinitely many points in $[0, 1]$.
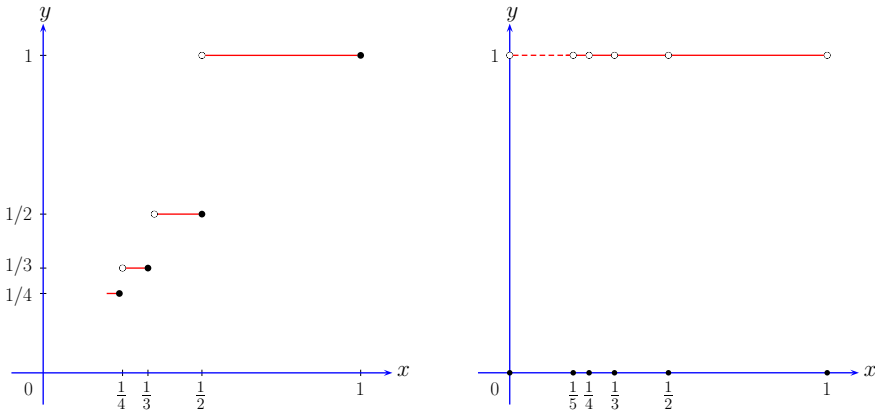
Fig. 6.2. Graphs of an 'infinite-steps function' and a 'broken-line function'

(v) Let $f : [-1, 1] \to \mathbb{R}$ be the zigzag function defined in Example 1.17 (iv). Since f is continuous (as shown in Example 3.6 (v)), it is integrable. Note that f is alternately increasing and decreasing on infinitely many subintervals of $[-1, 1]$ and thus, f is 'piecewise monotonic' on $[-1, 1]$.

(vi) Let $f : [0, 1] \to \mathbb{R}$ be the 'broken-line function' given by

$$f(x) := \begin{cases} 0 & \text{if } x = 0 \text{ or } x = 1/n \text{ for some } n \in \mathbb{N}, \\ 1 & \text{otherwise.} \end{cases}$$

[See Figure 6.2.] Here Proposition 6.10 is not directly applicable to f. Nevertheless, we can use it to show that f is integrable as follows. Let $\epsilon > 0$ be given, and choose $n_0 \in \mathbb{N}$ such that $1/n_0 < \epsilon/2$. Let $a_1 := 1/n_0$, and g denote the restriction of f to $[a_1, 1]$. Since g is bounded, and it is not continuous only at $1/n_0, 1/(n_0 - 1), \dots, 1/2, 1$, Corollary 6.11 shows that g is integrable. By the Riemann condition, there is a partition Q_ϵ of $[a_1, 1]$ such that $U(Q_\epsilon, g) - L(Q_\epsilon, g) < \epsilon$. Let P_ϵ denote the partition of $[0, 1]$ obtained by adding the point 0 to the partition Q_ϵ of $[a_1, 1]$. Since

$$\sup\{f(x) : 0 \leq x \leq a_1\} = 1 \quad \text{and} \quad \inf\{f(x) : 0 \leq x \leq a_1\} = 0,$$

we have

$$U(P_\epsilon, f) = 1 \cdot (a_1 - 0) + U(Q_\epsilon, g) \quad \text{and} \quad L(P_\epsilon, f) = 0 \cdot (a_1 - 0) + L(Q_\epsilon, g).$$

Thus

$$U(P_\epsilon, f) - L(P_\epsilon, f) = (a_1 - 0) + U(Q_\epsilon, g) - L(Q_\epsilon, g) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Hence by the Riemann condition, f is integrable. $\diamond$

Remark 6.14. The integrability of a bounded function $f : [a, b] \to \mathbb{R}$ is intimately related to the nature of the set of points in $[a, b]$ at which f is discontinuous. This connection is briefly explained in the Notes and Comments at the end of this chapter. For more details, one can consult the references cited there. See also Exercise 54 in this regard. $\diamond$

Algebraic and Order Properties

First we consider how Riemann integration behaves with respect to the algebraic operations on functions.

Proposition 6.15. *Let $f, g : [a, b] \to \mathbb{R}$ be integrable functions. Then*

(i) *$f + g$ is integrable and $\int_a^b (f + g)(x)dx = \int_a^b f(x)dx + \int_a^b g(x)dx$,*
(ii) *rf is integrable for any $r \in \mathbb{R}$ and $\int_a^b (rf)(x)dx = r \int_a^b f(x)dx$,*
(iii) *fg is integrable,*
(iv) *if there is $\delta > 0$ such that $|f(x)| \geq \delta$ and all $x \in [a, b]$, then $1/f$ is integrable,*
(v) *if $f(x) \geq 0$ for all $x \in [a, b]$, then for any $k \in \mathbb{N}$, the function $f^{1/k}$ is integrable.*

Proof. Let $\epsilon > 0$ be given. By the Riemann condition, there are partitions Q and R of $[a, b]$ such that

$$U(Q, f) - L(Q, f) < \epsilon \quad \text{and} \quad U(R, g) - L(R, g) < \epsilon.$$

Let P_ϵ denote the common refinement of Q and R. Then by part (i) of Lemma 6.2, we have

$$U(P_\epsilon, f) - L(P_\epsilon, f) \leq U(Q, f) - L(Q, f) < \epsilon$$

and

$$U(P_\epsilon, g) - L(P_\epsilon, g) \leq U(R, g) - L(R, g) < \epsilon.$$

(i) Let $P_\epsilon = \{x_0, x_1, \ldots, x_n\}$. For any $i = 1, \ldots, n$, we have

$$M_i(f + g) \leq M_i(f) + M_i(g) \quad \text{and} \quad m_i(f + g) \geq m_i(f) + m_i(g).$$

Multiplying both sides of these inequalities by $x_i - x_{i-1}$ and summing from $i = 1$ to $i = n$, we obtain

$U(P_\epsilon, f + g) \leq U(P_\epsilon, f) + U(P_\epsilon, g)$ and $L(P_\epsilon, f + g) \geq L(P_\epsilon, f) + L(P_\epsilon, g)$.

Hence

$U(P_\epsilon, f+g) - L(P_\epsilon, f+g) \leq U(P_\epsilon, f) - L(P_\epsilon, f) + U(P_\epsilon, g) - L(P_\epsilon, g) < \epsilon + \epsilon = 2\epsilon$.

Since $\epsilon > 0$ is arbitrary, the Riemann condition shows that the function $f + g$ is integrable. Further, if we let $\alpha := U(P_\epsilon, f) + U(P_\epsilon, g)$ and $\beta := L(P_\epsilon, f) + L(P_\epsilon, g)$, then we have

$$\beta \leq L(P_\epsilon, f+g) \leq L(f+g) = \int_a^b (f+g)(x)dx = U(f+g) \leq U(P_\epsilon, f+g) \leq \alpha.$$

Also, we have

$$\beta \leq L(f) + L(g) = \int_a^b f(x)dx + \int_a^b g(x)dx = U(f) + U(g) \leq \alpha.$$

Thus, we see that

$$\left| \int_a^b f(x)dx + \int_a^b g(x)dx - \int_a^b (f+g)(x)dx \right| \leq \alpha - \beta < 2\epsilon.$$

Since this is true for every $\epsilon > 0$, we obtain

$$\int_a^b (f+g)(x)dx = \int_a^b f(x)dx + \int_a^b g(x)dx.$$

(ii) Let $r \in \mathbb{R}$. If $r = 0$, then $rf(x) = 0$ for all $x \in [a, b]$ and (ii) follows easily. Now assume that $r > 0$. Then for any partition P of $[a, b]$, we see that

$$L(P, rf) = rL(P, f) \quad \text{and} \quad U(P, rf) = rU(P, f).$$

Hence
$$L(rf) = rL(f) = rU(f) = U(rf).$$

On the other hand, if $r < 0$, then for any partition P of $[a, b]$, we see that

$$L(P, rf) = rU(P, f) \quad \text{and} \quad U(P, rf) = rL(P, f),$$

and so
$$L(rf) = rU(f) = rL(f) = U(rf).$$

In both the cases, we see that rf is integrable and

$$\int_a^b (rf)(x)dx = r \int_a^b f(x)dx.$$

(iii) For any $i = 1, \ldots, n$, and $x, y \in [x_{i-1}, x_i]$, we have

$$(fg)(x) - (fg)(y) = f(x)[g(x) - g(y)] + [f(x) - f(y)]g(y)$$
$$\leq |f(x)|\,|g(x) - g(y)| + |g(y)|\,|f(x) - f(y)|$$
$$\leq M(|f|)[M_i(g) - m_i(g)] + M(|g|)[M_i(f) - m_i(f)].$$

Taking the supremum for x in $[x_{i-1}, x_i]$ and the infimum for y in $[x_{i-1}, x_i]$, we obtain

$$M_i(fg) - m_i(fg) \leq M(|f|)[M_i(g) - m_i(g)] + M(|g|)[M_i(f) - m_i(f)].$$

Multiplying both sides of this inequality by $x_i - x_{i-1}$ and summing from $i = 1$ to $i = n$, we obtain

$$U(P_\epsilon, fg) - L(P_\epsilon, fg)$$
$$\leq M(|f|)[U(P_\epsilon, g) - L(P_\epsilon, g)] + M(|g|)[U(P_\epsilon, f) - L(P_\epsilon, f)]$$
$$< [M(|f|) + M(|g|)]\epsilon.$$

Since $\epsilon > 0$ arbitrary, the Riemann condition shows that the function fg is integrable.

(iv) Let $\delta > 0$ be such that $|f(x)| \geq \delta$ for all $x \in [a, b]$. For any $i = 1, \ldots, n$ and $x, y \in [x_{i-1}, x_i]$, we have

$$\frac{1}{f(x)} - \frac{1}{f(y)} = \frac{f(y) - f(x)}{f(x)f(y)} \leq \frac{|f(x) - f(y)|}{|f(x)|\,|f(y)|} \leq \frac{1}{\delta^2}[M_i(f) - m_i(f)].$$

Taking the supremum for x in $[x_{i-1}, x_i]$ and the infimum for y in $[x_{i-1}, x_i]$, we obtain

$$M_i\left(\frac{1}{f}\right) - m_i\left(\frac{1}{f}\right) \leq \frac{1}{\delta^2}[M_i(f) - m_i(f)]$$

and consequently

$$U\left(P_\epsilon, \frac{1}{f}\right) - L\left(P_\epsilon, \frac{1}{f}\right) \leq \frac{1}{\delta^2}[U(P_\epsilon, f) - L(P_\epsilon, f)] < \frac{\epsilon}{\delta^2}.$$

Again, since $\epsilon > 0$ is arbitrary while $\delta > 0$ is fixed, the Riemann condition shows that the function $1/f$ is integrable.

(v) Let $k \in \mathbb{N}$ and write $F = f^{1/k}$ for simplicity. First we assume that there is $\delta > 0$ such that $f(x) \geq \delta$ for all $x \in [a, b]$. For any x, y in $[a, b]$, we see that $f(x) - f(y) = F(x)^k - F(y)^k$ can be written as

$$[F(x) - F(y)][F(x)^{k-1} + F(x)^{k-2}F(y) + \cdots + F(x)F(y)^{k-2} + F(y)^{k-1}].$$

Now

$$F(x)^{k-j}F(y)^{j-1} \geq \delta^{(k-j)/k}\delta^{(j-1)/k} = \delta^{(k-1)/k} > 0 \quad \text{for } j = 1, \ldots, k,$$

and so

$$F(x) - F(y) = \frac{f(x) - f(y)}{F(x)^{k-1} + F(x)^{k-2}F(y) + \cdots + F(x)F(y)^{k-2} + F(y)^{k-1}}$$
$$\leq \frac{f(x) - f(y)}{k\delta^{(k-1)/k}}.$$

If $P = \{x_0, x_1, \ldots, x_n\}$ is any partition of $[a, b]$ and $x, y \in [x_{i-1}, x_i]$ for some $i = 1, \ldots, n$, then

$$F(x) - F(y) \leq \frac{|f(x) - f(y)|}{k\delta^{(k-1)/k}} \leq \frac{M_i(f) - m_i(f)}{k\delta^{(k-1)/k}}.$$

Taking the supremum for x in $[x_{i-1}, x_i]$ and the infimum for y in $[x_{i-1}, x_i]$, we obtain

$$M_i(F) - m_i(F) \leq \frac{M_i(f) - m_i(f)}{k\delta^{(k-1)/k}} \qquad \text{for } i = 1, \ldots, n.$$

Multiplying both sides of this inequality by $x_i - x_{i-1}$ and summing from $i = 1$ to $i = n$, we obtain

$$U(P, F) - L(P, F) \leq \frac{1}{k\delta^{(k-1)/k}}[U(P, f) - L(P, f)].$$

Since f is integrable, the Riemann condition shows that F is also integrable.

Next, we consider the general case of any nonnegative integrable function f on $[a, b]$. Let $\delta > 0$ and define $g : [a, b] \to \mathbb{R}$ by $g(x) := f(x) + \delta$ and $G := g^{1/k}$. Then g is integrable by part (i) above, and $g(x) \geq \delta$ for all $x \in [a, b]$. It follows from what we have proved above that G is integrable. Moreover, since f is nonnegative, we have

$$G - \delta^{1/k} = (f + \delta)^{1/k} - \delta^{1/k} \leq f^{1/k} = F \leq (f + \delta)^{1/k} = G,$$

and therefore,
$$L(G - \delta^{1/k}) \leq L(F) \leq U(F) \leq U(G).$$

But

$$L(G - \delta^{1/k}) = L(G) - \delta^{1/k}(b - a) = \int_a^b G(x)dx - \delta^{1/k}(b - a) = U(G) - \delta^{1/k}(b - a).$$

This shows that

$$0 \leq U(F) - L(F) \leq U(G) - L(G - \delta^{1/k}) = \delta^{1/k}(b - a).$$

Since $\delta^{1/k} \to 0$ as $\delta \to 0$, we see that $F = f^{1/k}$ is integrable. $\square$

We remark that there is no simple way to express the Riemann integral of fg in terms of the Riemann integrals of f and g. In Proposition 6.25, we shall give a method of evaluating the Riemann integral of fg under additional assumptions.

With notation and hypotheses as in the above proposition, a combined application of its parts (i) and (ii) shows that the difference $f - g$ is integrable and

$$\int_a^b (f - g)(x)dx = \int_a^b f(x)dx - \int_a^b g(x)dx.$$

Further, given any $n \in \mathbb{N}$, successive applications of part (iii) of the above proposition show that the nth power f^n is integrable. Likewise, a combined application of parts (iii) and (iv) shows that if there is $\delta > 0$ such that $|g(x)| \geq \delta$ for all $x \in [a, b]$, then the quotient f/g is integrable. Also, a combined application of parts (iii) and (v) shows that if $f(x) \geq 0$ for all $x \in [a, b]$, then given any positive $r \in \mathbb{Q}$, the rth power f^r is integrable since $r = n/k$, where $n, k \in \mathbb{N}$.

The results obtained in Proposition 6.15 are in line with analogous results for continuity and differentiability of functions (Propositions 3.3 and 4.5). On the other hand, a composition of integrable functions need not be integrable, in contrast to the facts that a composition of continuous functions is continuous and a composition of differentiable functions is differentiable (Propositions 3.4 and 4.9).

Example 6.16. Let $f : [0, 1] \to \mathbb{R}$ be given by

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

Since f is continuous on $(0, 1]$, it follows from Corollary 6.11 that f is integrable. Consider **Thomae's function** $g : [0, 1] \to \mathbb{R}$ defined by

$$g(x) := \begin{cases} 1 & \text{if } x = 0 \in [0, 1], \\ 1/q & \text{if } x \in \mathbb{Q} \cap [0, 1] \text{ and } x = p/q, \text{ where } p, q \in \mathbb{N} \\ & \quad \text{are relatively prime,} \\ 0 & \text{otherwise.} \end{cases}$$

We show that g is also integrable. Let $\epsilon > 0$ be given. First note that the set $\{x \in [0, 1] : g(x) \geq \epsilon/2\}$ is finite, say $\{c_1, \ldots, c_\ell\}$. This follows by observing that if we choose $n_0 \in \mathbb{N}$ such that $1/n_0 < \epsilon/2$, then there are only finitely many rational numbers in $[0, 1]$ having denominators less than n_0. Let $P_\epsilon = \{x_0, x_1, \ldots, x_n\}$ be a partition of $[0, 1]$ such that $(x_i - x_{i-1}) < \epsilon/4\ell$ for $i = 1, \ldots, n$. Since there is an irrational number in $[x_{i-1}, x_i]$, we have $m_i(g) = 0$ for $i = 1, \ldots, n$, and so $L(P_\epsilon, g) = 0$. Also, we note the following: $g(x) \leq 1$ for all $x \in [0, 1]$, the points $c_1, \ldots, c_\ell$ belong to at most 2ℓ subintervals of P_ϵ, and if $x \in [0, 1]$ belongs to any of the remaining subintervals, then $g(x) < \epsilon/2$. Hence we have

$$U(P_\epsilon, g) = \sum_{i=1}^n M_i(g)(x_i - x_{i-1}) < 1 \cdot \frac{\epsilon}{4\ell} \cdot 2\ell + \frac{\epsilon}{2} \sum_{i=1}^n (x_i - x_{i-1}) = \epsilon.$$

Thus $U(P_\epsilon, g) - L(P_\epsilon, g) < \epsilon - 0 = \epsilon$. The Riemann condition implies that g is integrable. In fact,

$$\int_0^1 g(x)dx = \inf\{U(P, g) : P \text{ is a partition of } [a, b]\} = 0.$$

Now the composite functions $g \circ f$ and $f \circ g$ are both defined on $[0, 1]$, and

$$g \circ f(x) = 1 \quad \text{for all } x \in [0, 1], \quad \text{whereas} \quad f \circ g(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q}, \\ 0 & \text{if } x \notin \mathbb{Q}. \end{cases}$$

Thus from Examples 6.4 (i) and (ii), we see that $g \circ f$ is integrable, whereas $f \circ g$ is not integrable.   ◇

Remark 6.17. In Example 6.16, the function g is not continuous. (In fact, the set of discontinuities of g is $\mathbb{Q} \cap [0, 1]$. See Exercise 34 of Chapter 3.) It is possible to construct an integrable function $f : [0, 1] \to \mathbb{R}$ and a *continuous* function $g : [0, 1] \to [0, 1]$ such that $f \circ g$ is not integrable. [See Problem 28 of Chapter 3 in [52], or the article [45].] On the other hand, if $f : [a, b] \to \mathbb{R}$ is integrable, $\phi : [\alpha, \beta] \to [a, b]$ is differentiable, ϕ' is integrable, and $|\phi'| \geq \delta$ for some $\delta > 0$, then $f \circ \phi$ is integrable. (See part (ii) of Proposition 6.26.) Also, if $f : [a, b] \to \mathbb{R}$ is integrable and $\phi : [m(f), M(f)] \to \mathbb{R}$ is continuous, then $\phi \circ f$ is integrable. (See Exercise 42.)   ◇

Next, we consider how Riemann integration behaves with respect to the order relation on functions.

Proposition 6.18. *Let $f, g : [a, b] \to \mathbb{R}$ be integrable.*

(i) *If $f \leq g$, then $\int_a^b f(x)dx \leq \int_a^b g(x)dx$.*
(ii) *The function $|f|$ is integrable and $\left|\int_a^b f(x)dx\right| \leq \int_a^b |f|(x)dx$.*

Proof. (i) Assume that $f(x) \leq g(x)$ for all $x \in [a, b]$. Then for any partition P of $[a, b]$, we have $U(P, f) \leq U(P, g)$, and so

$$\int_a^b f(x)dx = U(f) \leq U(g) = \int_a^b g(x)dx.$$

(ii) Let $\epsilon > 0$ be given. By the Riemann condition, there is a partition P_ϵ of $[a, b]$ such that $U(P_\epsilon, f) - L(P_\epsilon, f) < \epsilon$. Let $P_\epsilon = \{x_0, x_1, \ldots, x_n\}$. For any $i = 1, \ldots, n$ and $x, y \in [x_{i-1}, x_i]$, we have

$$|f|(x) - |f|(y) \leq |f(x) - f(y)| \leq M_i(f) - m_i(f).$$

Taking the supremum for x in $[x_{i-1}, x_i]$ and the infimum for y in $[x_{i-1}, x_i]$, we obtain

$$M_i(|f|) - m_i(|f|) \leq M_i(f) - m_i(f) \quad \text{for } i = 1, \ldots, n.$$

Multiplying both sides of this inequality by $x_i - x_{i-1}$ and summing from $i = 1$ to $i = n$, we obtain

$$U(P_\epsilon, |f|) - L(P_\epsilon, |f|) \leq U(P_\epsilon, f) - L(P_\epsilon, f) < \epsilon.$$

Now the Riemann condition shows that $|f|$ is integrable. Further, since $-|f|(x) \leq f(x) \leq |f|(x)$ for all $x \in [a, b]$, by part (i) above we see that

$$\int_a^b -|f|(x)dx \leq \int_a^b f(x)dx \leq \int_a^b |f|(x)dx.$$

But $\int_a^b -|f|(x)dx = -\int_a^b |f|(x)dx$ by part (ii) of Proposition 6.15. Hence

$$\left| \int_a^b f(x)dx \right| \leq \int_a^b |f|(x)dx,$$

as desired.   □

Remark 6.19. If $f : [a, b] \to \mathbb{R}$ is any function, then

$$f = f^+ - f^-, \quad \text{where} \quad f^+ = \frac{|f| + f}{2} \quad \text{and} \quad f^- = \frac{|f| - f}{2}.$$

Note that both f^+ and f^- are nonnegative functions defined on $[a, b]$, and

$$f^+(x) = \max\{f(x), 0\} \quad \text{and} \quad f^-(x) = -\min\{f(x), 0\} \quad \text{for all } x \in [a, b].$$

The functions f^+ and f^- are known as the **positive part** and the **negative part** of f, respectively. By part (ii) of the above proposition, and parts (i) and (ii) of Proposition 6.15, we see that f is integrable if and only if f^+ and f^- are integrable, and then

$$\int_a^b f(x)dx = \int_a^b f^+(x)dx - \int_a^b f^-(x)dx = \text{Area}(R_{f^+}) - \text{Area}(R_{f^-}),$$

where $R_{f^+} := \{(x, y) \in \mathbb{R}^2 : a \leq x \leq b \text{ and } 0 \leq y \leq f(x)\}$ and $R_{f^-} := \{(x, y) \in \mathbb{R}^2 : a \leq x \leq b \text{ and } f(x) \leq y \leq 0\}$.

As remarked earlier, this suggests that the Riemann integral of f on $[a, b]$ can be interpreted as the 'signed area' of the corresponding region.   ◇

6.3 The Fundamental Theorem of Calculus

Differentiation and integration are the two most important processes in calculus and analysis. As we have remarked in the introduction of Chapter 4, differentiation is a local process, that is, the value of the derivative at a point depends only on the values of the function in a small interval about that

point. On the other hand, integration is a global process in the sense that
the integral of a function depends on the values of the function on the en-
tire interval. Further, these processes are defined in entirely different manners
without any apparent connection between them. Indeed, from a geometric
point of view, differentiation corresponds to finding (slopes of) tangents to
curves, while integration corresponds to finding areas under curves. At first
glance, there seems to be no reason for these two geometric processes to be
intimately related.

In this section, we shall obtain a wonderful result, known as the Fun-
damental Theorem of Calculus or, for short, the FTC, which says that the
processes of differentiating a function and integrating it are inverse to each
other. Roughly speaking, if one differentiates a function over an interval and
then integrates it, one gets back the original function. Also, if one first inte-
grates a function and then differentiates it, again one gets back the original
function. We remark that the proof of the FTC depends only on the Riemann
condition and the domain additivity proved in Section 6.1.

Let us recall that if $f : [a, b] \to \mathbb{R}$ is differentiable, then we obtain a new
function $f' : [a, b] \to \mathbb{R}$, called the derivative of f. Likewise, if $f : [a, b] \to \mathbb{R}$ is
integrable, then we obtain a new function $F : [a, b] \to \mathbb{R}$ defined by

$$F(x) = \int_a^x f(t)dt \quad \text{for } x \in [a, b].$$

Indeed, in view of Proposition 6.7, f is integrable on $[a, x]$ for every $x \in [a, b]$,
and $F(a) = 0$ in accord with our convention. Hence the function F is well
defined on $[a, b]$. To begin with, we shall study an important property of this
function.

Proposition 6.20. *Let $f : [a, b] \to \mathbb{R}$ be integrable and $F : [a, b] \to \mathbb{R}$ be
defined by*

$$F(x) := \int_a^x f(t)dt.$$

Then F is continuous on $[a, b]$.

*In fact, F satisfies a **Lipschitz condition** on $[a, b]$: there is $\alpha > 0$ such
that*

$$|F(x) - F(y)| \leq \alpha|x - y| \quad \text{for all } x, y \in [a, b].$$

Proof. Since f is integrable on $[a, b]$, it is bounded on $[a, b]$, that is, there is
$\alpha > 0$ such that $|f(t)| \leq \alpha$ for all $t \in [a, b]$.

Let $c \in [a, b]$. Then for $x \in [a, b]$, by the domain additivity (Proposition
6.7), we have

$$F(x) - F(c) = \int_a^x f(t)dt - \int_a^c f(t)dt = \int_c^x f(t)dt.$$

Hence by the basic inequality for Riemann integrals (Proposition 6.3),

$$|F(x) - F(c)| = \left| \int_c^x f(t)dt \right| \leq \alpha\, |x - c|.$$

This implies that F is continuous at c.

Since x and c are arbitrary points in $[a, b]$, and α does not depend on them, we see that f satisfies a Lipschitz condition on $[a, b]$.   □

The above proposition says that although an integrable function f may be discontinuous on $[a, b]$, the function $F : [a, b] \to \mathbb{R}$ obtained by integrating f from a to $x \in [a, b]$ is continuous on $[a, b]$. We shall see below in the second part of the FTC that if f happens to be continuous on $[a, b]$, then F is differentiable on $[a, b]$. Thus, integration is a smoothing process, unlike the process of differentiation (since the derivative of a differentiable function may not turn out to be differentiable).

In order to state the main result of this section in a concise form, we introduce the following concept. Let I be an interval containing more than one point and $f : I \to \mathbb{R}$ be any function. We say that f **has an antiderivative** on I if there is a differentiable function $F : I \to \mathbb{R}$ such that $f = F'$. Such a function F is called an **antiderivative** or a **primitive** of f. It follows from Corollary 4.21 that if an antiderivative of f exists, then it is unique up to addition of a constant. Just before stating Rolle's Theorem (Proposition 4.15), we showed that the integral part function is not the derivative of any function. Thus there exist functions that have no antiderivative.

Proposition 6.21 (Fundamental Theorem of Calculus). *Let $f : [a, b] \to \mathbb{R}$ be integrable.*

(i) *If f has an antiderivative F, then*

$$\int_a^x f(t)dt = F(x) - F(a) \quad \text{for all } x \in [a, b].$$

(ii) *Let $F : [a, b] \to \mathbb{R}$ be defined by*

$$F(x) = \int_a^x f(t)dt.$$

If f is continuous at $c \in [a, b]$, then F is differentiable at c and

$$F'(c) = f(c).$$

In particular, if f is continuous on $[a, b]$, then F is an antiderivative of f on $[a, b]$.

Proof. (i) Let $f : [a, b] \to \mathbb{R}$ be integrable and have an antiderivative F. If $x = a$, then in view of our convention,

$$\int_a^x f(t)dt = \int_a^a f(t)dt = 0 = F(x) - F(a).$$

Now assume that $x \in (a, b]$ and let g denote the restriction of f to $[a, x]$. Then in view of Proposition 6.7, g is integrable. Let $\epsilon > 0$ be given. By the Riemann condition, there is a partition $P_\epsilon = \{x_0, x_1, \ldots, x_n\}$ of $[a, x]$ such that $U(P_\epsilon, g) - L(P_\epsilon, g) < \epsilon$. By the MVT (Proposition 4.18) applied to F, for each $i = 1, \ldots, n$, there is $s_i \in (x_{i-1}, x_i)$ such that

$$F(x_i) - F(x_{i-1}) = F'(s_i)(x_i - x_{i-1}) = f(s_i)(x_i - x_{i-1}) = g(s_i)(x_i - x_{i-1}).$$

Hence

$$F(x) - F(a) = \sum_{i=1}^{n}[F(x_i) - F(x_{i-1})] = \sum_{i=1}^{n} g(s_i)(x_i - x_{i-1}),$$

and so

$$L(P_\epsilon, g) \leq F(x) - F(a) \leq U(P_\epsilon, g).$$

Since we also have

$$L(P_\epsilon, g) \leq \int_a^x g(t)dt \leq U(P_\epsilon, g),$$

it follows that

$$\left| F(x) - F(a) - \int_a^x g(t)dt \right| \leq U(P_\epsilon, g) - L(P_\epsilon, g) < \epsilon.$$

Because this inequality holds for every $\epsilon > 0$, we see that

$$\int_a^x f(t)dt = \int_a^x g(t)dt = F(x) - F(a),$$

as desired.

(ii) Let f be continuous at $c \in [a, b]$. Then by Proposition 3.7, for any given $\epsilon > 0$, there is $\delta > 0$ such that

$$t \in [a, b] \text{ and } |t - c| < \delta \Longrightarrow |f(t) - f(c)| < \epsilon.$$

Now if $x \in [a, b]$ and $x \neq c$, then we have

$$\frac{F(x) - F(c)}{x - c} = \frac{1}{x - c} \int_c^x f(t)dt = \frac{1}{x - c} \left(\int_c^x [f(t) - f(c)]dt \right) + f(c),$$

since $\int_c^x f(c)dx = f(c)(x - c)$ (as in Example 6.4 (i)). Next, if $0 < |x - c| < \delta$, then $|f(t) - f(c)| < \epsilon$ for all t in the closed interval between c and x, and hence, in view of the basic inequality for Riemann integrals (Proposition 6.3),

$$\left| \frac{F(x) - F(c)}{x - c} - f(c) \right| \leq \frac{1}{|x - c|} \cdot \epsilon |x - c| = \epsilon.$$

If $c \in (a, b)$, then it follows that F is differentiable at c and

$$F'(c) = \lim_{x \to c} \frac{F(x) - F(c)}{x - c} = f(c).$$

If $c = a$, then it also follows that the right (hand) derivative $F'_+(c)$ of F at c exists and equals $f(c)$, whereas if $c = b$, then the left (hand) derivative $F'_-(c)$ of F at c exists and equals $f(c)$. This proves (ii).   □

Remarks 6.22. (i) In view of part (i) of the FTC, if an integrable function $f : [a, b] \to \mathbb{R}$ has an antiderivative F, then F is called an **indefinite integral** of f, and it is denoted by $\int f(x)dx$. Note, however, that this notation is somewhat ambiguous since an indefinite integral of f is unique only up to an additive constant. For this reason, one writes

$$\int f(x)dx = F(x) + C,$$

where C denotes an arbitrary constant. Notice that in this case,

$$\int_a^b f(x)dx = F(b) - F(a),$$

where the right (hand) side is independent of the choice of an indefinite integral. The right (hand) side of the above equality is sometimes denoted by

$$\left[F(x)\right]_a^b \quad \text{or} \quad F(x)\big|_a^b.$$

With this in mind, the Riemann integral of $f : [a, b] \to \mathbb{R}$ is sometimes referred to as the **definite integral** of f over $[a, b]$.

(ii) The following stronger version of part (ii) of the FTC can be proved by slightly modifying its proof.

If $c \in [a, b)$ and $\lim_{x \to c+} f(x)$ exists, then the right (hand) derivative $F'_+(c)$ of F at c exists and equals this limit. Likewise, if $c \in (a, b]$ and $\lim_{x \to c-} f(x)$ exists, then the left (hand) derivative $F'_-(c)$ of F at c exists and equals this limit.

In proving this version, one appeals to analogues of Proposition 3.27 for right (hand) and left (hand) limits, and also to Proposition 6.12 because here the value $f(c)$ of f at c can be arbitrary.

Simple examples show that the converse of part (ii) of the FTC does not hold, that is, F may be differentiable without f being continuous. (See Exercise 14.) In fact, the converse of its stronger version also does not hold, that is, $F'_+(c)$ may exist and equal $f(c)$, even if $\lim_{x \to c+} f(x)$ does not exist. (See Proposition 7.17.)

(iii) We have seen in Part (ii) of the FTC that every continuous function on $[a, b]$ has an antiderivative. However, an integrable function may not have an antiderivative. This follows by noting that there are integrable functions

that do not have the IVP, a property possessed by all derivative functions (Proposition 4.14). On the other hand, a function on $[a, b]$ may have a derivative (i) that is not bounded (Exercise 47 of Chapter 7), or (ii) that is bounded, but not integrable (Volterra's example given on pages 56–57 of [35]), or (iii) that is integrable but not continuous (Example 7.19).   ◇

The two parts of the FTC can be combined to obtain a necessary and sufficient condition for a function to have a continuous derivative on $[a, b]$. The following result is known as the **Fundamental Theorem of Riemann Integration**.

Proposition 6.23. *Let $F : [a, b] \to \mathbb{R}$ be a function. Then F is differentiable and F' is continuous on $[a, b]$ if and only if there is a continuous function $f : [a, b] \to \mathbb{R}$ such that*

$$F(x) = F(a) + \int_a^x f(t)dt \quad \text{for all } x \in [a, b].$$

In this case, we have $F'(x) = f(x)$ for all $x \in [a, b]$.

Proof. Assume that F is differentiable and F' is continuous on $[a, b]$. Then F' is integrable and F is its antiderivative on $[a, b]$. Hence by part (i) of the FTC, we have

$$\int_a^x F'(t)dt = F(x) - F(a) \quad \text{for all } x \in [a, b].$$

Letting $f := F'$, we obtain the desired representation of F.

Conversely, assume that there is a continuous function $f : [a, b] \to \mathbb{R}$ such that

$$F(x) = F(a) + \int_a^x f(t)dt \quad \text{for all } x \in [a, b].$$

Then by part (ii) of the FTC, F is differentiable and $F'(x) = 0 + f(x) = f(x)$ for all $x \in [a, b]$, as desired.   □

As mentioned in the beginning of this section, the FTC shows that the processes of differentiation and integration are inverse to each other. The FTC is the major link between the so called 'differentiable calculus' and 'integral calculus'. Also, part (i) of the FTC provides the most widely used method of evaluating Riemann integrals. Of course, in order to employ it, one must be able to conjure up a function whose derivative is the given function f. It is not always easy to do so, but some corollaries of the FTC (Propositions 6.25 and 6.26) are useful in this regard. On the other hand, part (ii) of the FTC can be used to construct a differentiable function whose derivative is equal to a given continuous function on an interval. We shall illustrate this powerful technique in Chapter 7 while introducing the logarithmic and arctangent functions.

Examples 6.24. (i) Let r be a rational number and $r \neq -1$. Consider $a > 0$ and $f : [a, b] \to \mathbb{R}$ defined by $f(x) = x^r$. Then f is continuous on $[a, b]$, as we have seen in Example 3.6 (iii). Hence f is integrable. Also, it follows from Example 4.7 that if $F(x) := x^{r+1}/(r+1)$ for $x \in [a, b]$, then $F' = f$. Hence part (i) of the FTC shows that

$$\int_a^b x^r dx = F(b) - F(a) = \frac{b^{r+1} - a^{r+1}}{r+1}.$$

(In Corollary 7.10, this result will be generalized to the case that r is a real number.) It is easy to see that if r is a positive integer, then the above result holds even when $a \leq 0$, and if r is a negative integer $\neq -1$, then the above result also holds when $a < 0$ and $b < 0$.

(ii) Let $a, b \in \mathbb{R}$ with $a < 0 < b$, and define $f : [a, b] \to \mathbb{R}$ by $f(x) = x^2$ if $a \leq x \leq 0$ and $f(x) = x$ if $0 < x < b$. Then f is continuous on $[a, b]$, as we have seen in Example 3.6 (iv). Hence f is integrable. Let $F_1(x) = x^3/3$ for $x \in [a, 0]$ and $F_2(x) = x^2/2$ for $x \in [0, b]$. Then

$$\int_a^b f(x)dx = \int_a^0 f(x)dx + \int_0^b f(x)dx = 0 - \frac{a^3}{3} + \frac{b^2}{2} - 0 = \frac{b^2}{2} - \frac{a^3}{3}$$

by the domain additivity and by (i) above. $\diamond$

Now we consider two important consequences of the FTC that yield the two most powerful methods of evaluating integrals. The first result is about the Riemann integral of the product of two functions.

Proposition 6.25 (Integration by Parts). *Let $f : [a, b] \to \mathbb{R}$ be a differentiable function such that f' is integrable. Assume that $g : [a, b] \to \mathbb{R}$ is integrable and has an antiderivative G on $[a, b]$. Then*

$$\int_a^b f(x)g(x)dx = f(b)G(b) - f(a)G(a) - \int_a^b f'(x)G(x)dx.$$

Proof. Let $H := fG$. Then $H' = fG' + f'G = fg + f'G$, by part (iii) of Proposition 4.5. Since f and G are differentiable, they are continuous and hence integrable. Also, since f' and g are assumed to be integrable, it follows from parts (i) and (iii) of Proposition 6.15 that the function $fg + f'G$ is integrable. Hence by part (i) of the FTC,

$$\int_a^b [f(x)g(x) + f'(x)G(x)]dx = H(b) - H(a) = f(b)G(b) - f(a)G(a).$$

This together with part (i) of Proposition 6.15 proves the proposition.   □

In the notation of Remark 6.22 (i), the conclusion of the above proposition can be stated as follows:

$$\int_a^b f(x)g(x)dx = \left[f(x) \int g(x)dx \right]_a^b - \int_a^b \left(f'(x) \int g(x)dx \right) dx,$$

with the understanding that on the right (hand) side, $\int g(x)dx$ denotes the same indefinite integral of g at both places. (Recall that two indefinite integrals of g differ by an additive constant.)

Next, we consider the method of substitution for evaluating a Riemann integral. The following result has two parts; while the proofs of both are based on the FTC, each is applicable in a different situation.

Proposition 6.26 (Integration by Substitution). *Let $\phi : [\alpha, \beta] \to \mathbb{R}$ be a differentiable function such that ϕ' is integrable on $[\alpha, \beta]$, and let $\phi([\alpha, \beta]) = [a, b]$.*

(i) *If $f : [a, b] \to \mathbb{R}$ is continuous, then the function $(f \circ \phi)\phi' : [\alpha, \beta] \to \mathbb{R}$ is integrable and*

$$\int_{\phi(\alpha)}^{\phi(\beta)} f(x)dx = \int_\alpha^\beta f(\phi(t))\phi'(t)dt.$$

(ii) *If $f : [a, b] \to \mathbb{R}$ is integrable and $\phi'(t) \neq 0$ for every $t \in (\alpha, \beta)$, then the function $(f \circ \phi)|\phi'| : [\alpha, \beta] \to \mathbb{R}$ is integrable and*

$$\int_a^b f(x)dx = \int_\alpha^\beta f(\phi(t))|\phi'(t)|dt.$$

Proof. (i) Let $f : [a, b] \to \mathbb{R}$ be continuous and for $x \in [a, b]$, define $F(x) := \int_a^x f(u)du$. Part (ii) of the FTC shows that the function $F : [a, b] \to \mathbb{R}$ is differentiable and $F' = f$. Now consider the function $H : [\alpha, \beta] \to \mathbb{R}$ defined by $H := F \circ \phi$. Then by the Chain Rule (Proposition 4.9), we have

$$H'(t) = F'(\phi(t))\phi'(t) = f(\phi(t))\phi'(t) \quad \text{for all } t \in [a, b].$$

Since the function $f \circ \phi$ is continuous and the function ϕ' is integrable, the function $(f \circ \phi)\phi'$ is integrable by part (iii) of Proposition 6.15. Hence part (i) of the FTC shows that

$$\int_\alpha^\beta f(\phi(t))\phi'(t)dt = H(\beta) - H(\alpha) = \int_a^{\phi(\beta)} f(x)dx - \int_a^{\phi(\alpha)} f(x)dx.$$

Thus, if $\phi(\alpha) \leq \phi(\beta)$, then the domain additivity (Proposition 6.7) shows that

$$\int_\alpha^\beta f(\phi(t))\phi'(t)dt = \int_{\phi(\alpha)}^{\phi(\beta)} f(x)dx,$$

and if $\phi(\beta) \leq \phi(\alpha)$, then we obtain the same result since in accord with our convention, $\int_{\phi(\alpha)}^{\phi(\beta)} f(x)dx = - \int_{\phi(\beta)}^{\phi(\alpha)} f(x)dx$. This proves the desired result.

(ii) Let $f : [a, b] \to \mathbb{R}$ be integrable and $\phi'(t) \neq 0$ for all $t \in (\alpha, \beta)$ and $\psi := (f \circ \phi)|\phi'|$. We first show that $L(f) \leq L(\psi)$.

By the IVP of ϕ' (Proposition 4.14), either $\phi'(t) > 0$ for all $t \in (\alpha, \beta)$, or $\phi'(t) < 0$ for all $t \in (\alpha, \beta)$. Suppose $\phi'(t) > 0$ for all $t \in (\alpha, \beta)$. Then, by part (iii) of Proposition 4.27, ϕ is strictly increasing on $[\alpha, \beta]$, $\phi(\alpha) = a$ and $\phi(\beta) = b$. Consider a partition $P := \{x_0, x_1, \ldots, x_n\}$ of $[a, b]$ and let $t_i := \phi^{-1}(x_i)$ for $i = 0, 1, \ldots, n$. Then $\alpha = t_0 < t_1 < \cdots < t_n = \beta$ and by part (i) of the FTC, we have

$$\int_{t_{i-1}}^{t_i} |\phi'(t)|dt = \int_{t_{i-1}}^{t_i} \phi'(t)dt = \phi(t_i) - \phi(t_{i-1}) = x_i - x_{i-1} \quad \text{for } i = 1, \ldots, n.$$

Also, since $f([x_{i-1}, x_i]) = (f \circ \phi)([t_{i-1}, t_i])$ for $i = 1, \ldots, n$, we see that

$$L(P, f) = \sum_{i=1}^{n} m_i(f)(x_i - x_{i-1}) = \sum_{i=1}^{n} \int_{t_{i-1}}^{t_i} m_i(f \circ \phi)|\phi'(t)|dt.$$

For $i = 1, \ldots, n$, let ϕ_i and ψ_i denote the restrictions of ϕ and ψ to $[t_{i-1}, t_i]$ respectively. Then $|\phi_i'|$ is integrable on $[t_{i-1}, t_i]$ and $m_i(f \circ \phi)|\phi_i'| \leq \psi_i$ for $i = 1, \ldots, n$. Hence we obtain

$$L(P, f) \leq \sum_{i=1}^{n} L(m_i(f \circ \phi)|\phi_i'|) \leq \sum_{i=1}^{n} L(\psi_i).$$

Let $\epsilon > 0$ be given. For each $i = 1, \ldots, n$, there is a partition Q_i of $[t_{i-1}, t_i]$ such that

$$L(\psi_i) - \frac{\epsilon}{n} < L(Q_i, \psi_i).$$

If Q denotes the partition of $[\alpha, \beta]$ obtained from the points of $Q_1, \ldots, Q_n$, then

$$\sum_{i=1}^{n} L(\psi_i) < \sum_{i=1}^{n} L(Q_i, \psi_i) + \epsilon = L(Q, \psi) + \epsilon \leq L(\psi) + \epsilon.$$

It follows that $L(P, f) < L(\psi) + \epsilon$ for every $\epsilon > 0$, and so $L(P, f) \leq L(\psi)$. Taking the supremum over all partitions P of $[a, b]$, we have $L(f) \leq L(\psi)$.

Next, let us assume that $\phi'(t) < 0$ for all $t \in (\alpha, \beta)$. Then by part (iv) of Proposition 4.27, ϕ is strictly decreasing on $[\alpha, \beta]$, $\phi(\alpha) = b$, and $\phi(\beta) = a$. For $i = 0, 1, \ldots, n$, if we define $t_i := \phi^{-1}(x_{n-i})$, then the argument given above for proving $L(f) \leq L(\psi)$ works equally well because

$$\int_{t_{i-1}}^{t_i} |\phi'(t)|dt = -\int_{t_{i-1}}^{t_i} \phi'(t)dt = -\big[\phi(t_i) - \phi(t_{i-1})\big] = x_{n-i+1} - x_{n-i}$$

and $f([x_{n-i}, x_{n-i+1}]) = (f \circ \phi)([t_{i-1}, t_i])$, so that

$$L(P, f) = \sum_{i=1}^{n} m_{n-i+1}(f)(x_{n-i+1} - x_{n-i}) = \sum_{i=1}^{n} \int_{t_{i-1}}^{t_i} m_i(f \circ \phi)|\phi'(t)|dt.$$

Hence $L(f) \leq L(\psi)$. Similarly, we can show that $U(f) \geq U(\psi)$. Since f is integrable, that is, since $L(f) = U(f)$, we see that $L(\psi) = U(\psi)$, that is, ψ is integrable and

$$\int_a^b f(x)dx = \int_\alpha^\beta \psi(t)dt = \int_\alpha^\beta (f \circ \phi)|\phi'(t)|dt,$$

as desired.   □

While the proof of part (ii) of the above proposition is rather involved, it may be noted that the result is proved without assuming the continuity of the function f. If we let $f(x) := 1$ for all $x \in [a, b]$ in this result, then we obtain

$$b - a = \int_\alpha^\beta |\phi'(t)|dt.$$

This shows that the absolute value of the derivative of the function ϕ acts as the change of scale factor in the method of substitution. For example, if p is a nonzero real number, $q \in \mathbb{R}$, and $\phi(t) := pt + q$ for all $t \in [\alpha, \beta]$, then the change of scale factor is the constant $|p|$.

Examples 6.27. (i) To evaluate $\int_0^1 x\sqrt{1-x}\,dx$, let $f(x) := x$ and $g(x) := \sqrt{1-x}$ for $x \in [0, 1]$. Then $f'(x) = 1$ for $x \in [0, 1]$. Also, if we let $G(x) := -(2/3)(1-x)^{3/2}$, then $G'(x) = g(x)$ for $x \in [0, 1]$, that is, $G' = g$. Integrating by Parts (Proposition 6.25), we obtain

$$\int_0^1 x\sqrt{1-x}\,dx = 0 - 0 - \int_0^1 \left(-\frac{2}{3}\right)(1-x)^{3/2}dx = \frac{2}{3}\int_0^1 (1-x)^{3/2}dx.$$

If we let $F(x) := -(2/5)(1-x)^{5/2}$ for $x \in [0, 1]$, then $F'(x) = (1-x)^{3/2}$ for $x \in [0, 1]$ and hence

$$\int_0^1 x\sqrt{1-x}\,dx = \frac{2}{3}[F(1) - F(0)] = \frac{2}{3}\left[0 - \left(-\frac{2}{5}\right)\right] = \frac{4}{15}.$$

(ii) To evaluate $\int_0^1 t\sqrt{1-t^2}\,dt$, let $\phi(t) := 1 - t^2$ for $t \in [0, 1]$ and $f(x) := \sqrt{x}$ for $x \in [0, 1]$. Since $\phi(0) = 1$, $\phi(1) = 0$, and $\phi'(t) = -2t$ for all $t \in [0, 1]$. Integration by Substitution (part (ii) of Proposition 6.26) yields

$$\int_0^1 t\sqrt{1-t^2}\,dt = \frac{1}{2}\int_0^1 f(\phi(t))|\phi'(t)|dt = \frac{1}{2}\int_0^1 f(x)dx = \frac{1}{2}\int_0^1 \sqrt{x}\,dx.$$

Now if we let $F(x) := (2/3)x^{3/2}$ for $x \in [0, 1]$, then

$$\frac{1}{2}\int_0^1 \sqrt{x}\,dx = \frac{1}{2}[F(1) - F(0)] = \frac{1}{2}\left(\frac{2}{3} - 0\right) = \frac{1}{3}.$$

Thus $\int_0^1 t\sqrt{1-t^2}dt = \frac{1}{3}$.   ◇

We shall now prove an interesting application of the method of integration by substitution. If the degree of a polynomial function defined on an interval is at most one, then we give a formula for its Riemann integral in terms of the values of the function at the two endpoints of the interval, and if the degree is at most two, then we give such a formula in terms of the values of the function at the two endpoints and the midpoint of the interval. These formulas will be useful when we attempt to find approximations of the Riemann integral of an arbitrary integrable function in Section 8.6.

Proposition 6.28. *Let $f : [a, b] \to \mathbb{R}$ be a polynomial function of degree m.*
(i) *If $m \leq 1$, then*
$$\int_a^b f(x)dx = \frac{(b-a)}{2}[f(a) + f(b)].$$

(ii) *If $m \leq 2$, then*
$$\int_a^b f(x)dx = \frac{(b-a)}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right].$$

Proof. Consider the function $\phi : [0, 1] \to \mathbb{R}$ given by $\phi(t) := (b-a)t+a$. Then $\phi(0) = a$ and $\phi(1) = b$. Define $g := f \circ \phi$. By Proposition 6.26, we have
$$\int_a^b f(x)dx = \int_0^1 f(\phi(t))\phi'(t)dt = (b-a)\int_0^1 g(t)dt.$$

It is clear that g is also a polynomial function of degree m.

(i) Let $m \leq 1$. Then there are $c_1, c_0 \in \mathbb{R}$ such that $g(t) := c_1 t + c_0$ for all $t \in [0, 1]$, and hence
$$\int_0^1 g(t)dt = \frac{c_1}{2} + c_0.$$

But since $g(0) = c_0$ and $g(1) = c_1 + c_0$, we see that
$$\int_0^1 g(t)dt = \frac{1}{2}[g(1) + g(0)].$$

Further, since $g(0) = f(a)$ and $g(1) = f(b)$, we conclude that
$$\int_a^b f(x)dx = \frac{(b-a)}{2}[f(a) + f(b)],$$

as desired.

(ii) Let $m \leq 2$. Then there are $c_2, c_1, c_0 \in \mathbb{R}$ such that $g(t) := c_2 t^2 + c_1 t + c_0$ for all $t \in [0, 1]$, and hence
$$\int_0^1 g(t)dt = \frac{c_2}{3} + \frac{c_1}{2} + c_0.$$

But since $g(0) = c_0$, $g(1) = c_2 + c_1 + c_0$, and $g\left(\frac{1}{2}\right) = \frac{c_2}{4} + \frac{c_1}{2} + c_0$, we see that

$$\int_0^1 g(t)dt = \frac{1}{6}(2c_2 + 3c_1 + 6c_0)$$

$$= \frac{1}{6}\left[g(1) + 4g\left(\frac{1}{2}\right) - 5g(0) + 6g(0)\right]$$

$$= \frac{1}{6}\left[g(0) + 4g\left(\frac{1}{2}\right) + g(1)\right].$$

Further, since $g(0) = f(a)$, $g(\frac{1}{2}) = f\left(\frac{a+b}{2}\right)$, and $g(1) = f(b)$, we conclude that

$$\int_a^b f(x)dx = \frac{(b-a)}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right],$$

as desired. $\square$

6.4 Riemann Sums

Let $f : [a, b] \to \mathbb{R}$ be a bounded function. In view of the Riemann condition, f is integrable if and only if there is a sequence (P_n) of partitions of $[a, b]$ such that $U(P_n, f) - L(P_n, f) \to 0$. Although we have made good use of the Riemann condition to prove several interesting results in the previous section, there are a number of difficulties in employing it to test the integrability of an arbitrary bounded function and, if such a function is found to be integrable, then to compute its Riemann integral. First, the calculation of $U(P, f)$ and $L(P, f)$, for a given partition P, involves finding the absolute maxima and minima of f over several subintervals of $[a, b]$. This task is rarely easy. Second, it is not clear how to go about choosing a partition P_n, $n \in \mathbb{N}$, so as to obtain $U(P_n, f) - L(P_n, f) \to 0$. Finally, when f is known to be integrable, how does one actually find at least an approximate value of its integral? In this section, we shall address these questions.

To overcome the first difficulty mentioned above, namely, of having to calculate several maxima and minima of f, we give an alternative approach. While calculating maxima and minima of f over several subintervals of $[a, b]$ may be difficult, evaluating f at several points of $[a, b]$ is relatively easy. With this in mind, we introduce the following concept. Let $P = \{x_0, x_1, \ldots, x_n\}$ be a partition of $[a, b]$, and let s_i be a point in the ith subinterval $[x_{i-1}, x_i]$ for $i = 1, \ldots, n$. Then

$$S(P, f) := \sum_{i=1}^n f(s_i)(x_i - x_{i-1})$$

is called a **Riemann sum** for f corresponding to P. Note that the upper sum $U(P, f)$ and the lower sum $L(P, f)$ are determined by P and f, whereas a Riemann sum $S(P, f)$ depends on P and f, and also on the choice of the points

$s_i \in [x_{i-1}, x_i]$ for $i = 1, \ldots, n$. Now we give a criterion for the integrability of f in terms of Riemann sums.

Proposition 6.29 (Cauchy Condition). *Let $f : [a, b] \to \mathbb{R}$ be a bounded function. Then f is integrable on $[a, b]$ if and only if for every $\epsilon > 0$, there is a partition P_ϵ of $[a, b]$ such that*

$$|S(P_\epsilon, f) - T(P_\epsilon, f)| < \epsilon$$

for any Riemann sums $S(P_\epsilon, f)$ and $T(P_\epsilon, f)$ for f corresponding to P_ϵ.

Proof. Suppose f is integrable. Let $\epsilon > 0$ be given. Then by the Riemann condition, there is a partition P_ϵ of $[a, b]$ such that $U(P_\epsilon, f) - L(P_\epsilon, f) < \epsilon$. If $S(P_\epsilon, f)$ and $T(P_\epsilon, f)$ are any Riemann sums for f corresponding to P_ϵ, then

$$L(P_\epsilon, f) \leq S(P_\epsilon, f) \leq U(P_\epsilon, f) \quad \text{and} \quad L(P_\epsilon, f) \leq T(P_\epsilon, f) \leq (P_\epsilon, f).$$

It follows that

$$|S(P_\epsilon, f) - T(P_\epsilon, f)| \leq U(P_\epsilon, f) - L(P_\epsilon, f) < \epsilon.$$

Conversely, assume that the condition stated in the proposition holds. Given $\epsilon > 0$, let $P = \{x_0, x_1, \ldots, x_n\}$ be a partition of $[a, b]$ such that the difference between any two Riemann sums for f corresponding to P is less than $\epsilon/3$. Now for $i = 1, \ldots, n$, there is $s_i \in [x_{i-1}, x_i]$ such that

$$M_i(f) < f(s_i) + \frac{\epsilon}{3(b-a)}.$$

Let

$$S(P, f) := \sum_{i=1}^{n} f(s_i)(x_i - x_{i-1}).$$

From our choice of s_i, $i = 1, \ldots, n$, it follows that

$$U(P, f) < S(P, f) + \frac{\epsilon}{3}.$$

Similarly, for $i = 1, \ldots, n$, there is $t_i \in [x_{i-1}, x_i]$ such that

$$m_i(f) > f(t_i) - \frac{\epsilon}{3(b-a)}.$$

Let

$$T(P, f) := \sum_{i=1}^{n} f(t_i)(x_i - x_{i-1}).$$

From our choice of t_i, $i = 1, \ldots, n$, it follows that

$$L(P, f) > T(P, f) - \frac{\epsilon}{3}.$$

Further, since $S(P, f) - T(P, f) < \epsilon/3$, we have

$$U(P, f) - L(P, f) < S(P, f) + \frac{\epsilon}{3} - T(P, f) + \frac{\epsilon}{3} < \epsilon.$$

Hence by the Riemann condition, f is integrable. $\qquad\qquad\qquad\qquad\square$

Let us now take up the second question regarding the choice of a partition P so as to make $U(P, f) - L(P, f)$ small. The discussion at the beginning of this chapter suggests that we may start with any partition of $[a, b]$ and refine it successively. An important point to note here is the following: Mere addition of new points would not make the difference between the corresponding upper and lower sums tend to zero; the new points need to be so chosen that the length of the largest subinterval of the refined partition is smaller than the length of the largest subinterval of the given partition. This consideration leads us to the following notion.

For a partition P of $[a, b]$, we define the **mesh** of P to be the length of the largest subinterval of P. Thus, if $P = \{x_0, x_1, \ldots, x_n\}$, then

$$\mu(P) := \max\{x_i - x_{i-1} : i = 1, \ldots, n\}.$$

Now we shall prove an important result, which, roughly speaking, says that upper sums $U(P, f)$ approximate the upper integral $U(f)$ and lower sums approximate the lower integral $L(f)$ if the mesh $\mu(P)$ of P is made small.

Lemma 6.30. *Let $f : [a, b] \to \mathbb{R}$ be a bounded function. Given any $\epsilon > 0$, there is $\delta > 0$ such that for every partition P of $[a, b]$ with $\mu(P) < \delta$, we have*

$$0 \le U(P, f) - U(f) < \epsilon \quad and \quad 0 \le L(f) - L(P, f) < \epsilon,$$

and consequently

$$L(f) - \epsilon < S(P, f) < U(f) + \epsilon,$$

where $S(P, f)$ is any Riemann sum for f corresponding to P.

Proof. Let $\epsilon > 0$ be given. Since $U(f)$ is the infimum of the set of all upper sums for f and $L(f)$ is the supremum of the set of all lower sums for f, there are partitions P_1 and P_2 of $[a, b]$ such that $U(P_1, f) < U(f) + \epsilon/2$ and $L(P_2, f) > L(f) - \epsilon/2$. Let P_0 denote the common refinement of P_1 and P_2. Then by part (i) of Lemma 6.2, we have

$$U(P_0, f) < U(f) + \frac{\epsilon}{2} \quad and \quad L(P_0, f) > L(f) - \frac{\epsilon}{2}.$$

Let $\alpha > 0$ be such that $|f(x)| \le \alpha$ for all $x \in [a, b]$. If the partition P_0 contains n_0 points, define $\delta := \epsilon/4\alpha n_0$. Consider any partition $P = \{x_0, x_1, \ldots, x_n\}$ of $[a, b]$ such that $\mu(P) < \delta$. Let P^* denote the common refinement of P and P_0. Again, by part (i) of Lemma 6.2, we have

$$U(P^*, f) \le U(P, f) \quad and \quad U(P^*, f) \le U(P_0, f).$$

We observe that positive contributions to the difference $U(P, f) - U(P^*, f)$ can arise only when a point x^* of the partition P_0 lies in an open interval (x_{i-1}, x_i) induced by the partition P. Further, any such contribution is at most $2\alpha\mu(P)$. This follows by noting that if $x^* \in (x_{i-1}, x_i)$, and if

$$M_\ell^* = \sup\{f(x) : x \in [x_{i-1}, x^*]\} \quad \text{and} \quad M_r^* = \sup\{f(x) : x \in [x^*, x_i]\},$$

then the contribution to $U(P, f) - U(P^*, f)$ arising from the point x^* is

$$
\begin{aligned}
&M_i(f)(x_i - x_{i-1}) - M_\ell^*(x^* - x_{i-1}) - M_r^*(x_i - x^*) \\
&= (M_i(f) - M_\ell^*)(x^* - x_{i-1}) + (M_i(f) - M_r^*)(x_i - x^*) \\
&\leq 2\alpha[(x^* - x_{i-1}) + (x_i - x^*)] \\
&= 2\alpha(x_i - x_{i-1}) \\
&\leq 2\alpha\mu(P).
\end{aligned}
$$

Since the total number of points in the partition P_0 is n_0, we immediately see that

$$U(P, f) - U(P^*, f) \leq n_0 \cdot 2\alpha\mu(P) < 2\alpha n_0 \delta = \frac{\epsilon}{2}.$$

Thus for every partition P of $[a, b]$ with $\mu(P) < \delta$, we have

$$U(P, f) < U(P^*, f) + \frac{\epsilon}{2} \leq U(P_0, f) + \frac{\epsilon}{2} < U(f) + \frac{\epsilon}{2} + \frac{\epsilon}{2} = U(f) + \epsilon.$$

In a similar manner, we can show that for every partition P of $[a, b]$ with $\mu(P) < \delta$, we have

$$L(P, f) > L(f) - \epsilon.$$

Finally, if $S(P, f)$ is any Riemann sum for f corresponding to a partition P with $\mu(P) < \delta$, then

$$L(f) - \epsilon < L(P, f) \leq S(P, f) \leq U(P, f) < U(f) + \epsilon,$$

as desired.   □

Proposition 6.31 (Theorem of Darboux). *Let $f : [a, b] \to \mathbb{R}$ be a bounded function. If f is integrable, then given any $\epsilon > 0$, there is $\delta > 0$ such that for every partition P of $[a, b]$ with $\mu(P) < \delta$, we have*

$$\left| S(P, f) - \int_a^b f(x)dx \right| < \epsilon,$$

where $S(P, f)$ is any Riemann sum for f corresponding to P.

Conversely, assume that there is $r \in \mathbb{R}$ satisfying the following condition: Given $\epsilon > 0$, there is a partition P of $[a, b]$ such that

$$|S(P, f) - r| < \epsilon,$$

where $S(P, f)$ is any Riemann sum for f corresponding to P. Then f is integrable and its Riemann integral is equal to r.

Proof. Suppose f is integrable. Let $\epsilon > 0$ be given. It follows from Lemma 6.30 that there is $\delta > 0$ such that for every partition P of $[a, b]$ with $\mu(P) < \delta$ and any Riemann sum $S(P, f)$ for f corresponding to P, we have

$$\int_a^b f(x)dx - \epsilon = L(f) - \epsilon < S(P, f) < U(f) + \epsilon = \int_a^b f(x)dx + \epsilon,$$

as desired.

Conversely, let $r \in \mathbb{R}$ satisfy the stated condition. Let $\epsilon > 0$ be given and $P := \{x_0, x_1, \ldots, x_n\}$ denote a partition of $[a, b]$ such that $|S(P, f) - r| < \epsilon$, where $S(P, f)$ is any Riemann sum for f corresponding to P. If $S(P, f)$ and $T(P, f)$ are Riemann sums for f corresponding to P, then

$$|S(P, f) - T(P, f)| \leq |S(P, f) - r| + |r - T(P, f)| < \epsilon + \epsilon = 2\epsilon.$$

Since $\epsilon > 0$ is arbitrary, the Cauchy condition (Proposition 6.29) shows that f is integrable. To show that the Riemann integral of f is equal to r, we note, as in the proof of (Proposition 6.29), that there are $s_i, t_i \in [x_{i-1}, x_i]$ for $i = 1, \ldots, n$ such that if we let

$$S(P, f) := \sum_{i=1}^n f(s_i)(x_i - x_{i-1}) \quad \text{and} \quad T(P, f) := \sum_{i=1}^n f(t_i)(x_i - x_{i-1}),$$

then

$$U(P, f) < S(P, f) + \epsilon \quad \text{and} \quad T(P, f) - \epsilon < L(P, f).$$

Since $L(P, f) \leq \int_a^b f(x)dx \leq U(P, f)$, we see that

$$r - 2\epsilon < T(P, f) - \epsilon < \int_a^b f(x)dx < S(P, f) + \epsilon < r + 2\epsilon.$$

Since this holds for every $\epsilon > 0$, the Riemann integral of f is equal to r. □

Remark 6.32. As an immediate consequence of the above result, we note that if $f : [a, b] \to \mathbb{R}$ is integrable and if (P_n) is a sequence of partitions of $[a, b]$ such that $\mu(P_n) \to 0$, then

$$L(P_n, f) \to \int_a^b f(x)dx \quad \text{and} \quad U(P_n, f) \to \int_a^b f(x)dx,$$

and moreover, if $S(P_n, f)$ is any Riemann sum for f corresponding to P_n, then

$$S(P_n, f) \to \int_a^b f(x)dx.$$

It may be emphasized that the only requirement here is that $\mu(P_n) \to 0$; the actual partition points and the points in the subintervals at which f is evaluated can be chosen with sole regard to the convenience of summation.

This enables us to find approximations of the Riemann integral of f when we are not able to evaluate it exactly. For example, if f does not have an antiderivative, or if we are simply not able to think of an antiderivative of f, or if the evaluation of an antiderivative of f at a and b is impossible, then part (i) of the FTC (Proposition 6.21) becomes inoperative as far as the evaluation of the Riemann integral of f is concerned, and we may resort to calculating it approximately. On the other hand, if the Riemann integral of f can be evaluated by employing part (i) of the FTC, then limits of Riemann sums for f can be found.   ◇

Examples 6.33. (i) Let $f(x) := 1/x$ for $x \in [a, b]$, where either $a > 0$ or
$b < 0$. Then f is integrable, because it is continuous on $[a, b]$. Since we
have not introduced any function F such that $F' = f$, it is not possible at
this stage to evaluate the Riemann integral of f. For simplicity, consider
$a = 1$ and $b = 2$. Let $n \in \mathbb{N}$ and $\mathsf{P}_n := \{1, 1 + (1/n), \ldots, 1 + (n/n)\}$
be the partition of $[1, 2]$ into n equal parts. Consider the left endpoints
$s_{n,i} := 1 + (i - 1)/n$ for $i = 1, \ldots, n$. Then $\mu(\mathsf{P}_n) = 1/n \to 0$ and

$$S(\mathsf{P}_n, f) = \sum_{i=1}^{n} \frac{1}{1 + (i-1)/n} \left(\frac{i}{n} - \frac{i-1}{n} \right) = \sum_{i=1}^{n} \frac{1}{n + i - 1}.$$

Hence by Proposition 6.31,

$$\frac{1}{n} + \frac{1}{n+1} + \cdots + \frac{1}{2n} = \sum_{i=1}^{n} \frac{1}{n+i-1} \to \int_1^2 \frac{1}{x} dx \quad \text{as } n \to \infty.$$

(ii) Let $f(x) := 1/(1 + x^2)$ for $x \in [a, b]$. Again, f is integrable, because it is
continuous on $[a, b]$. As in the previous example, we have not introduced
any function F such that $F' = f$ so far. This time, for simplicity, consider
$a = 0$ and $b = 1$. Let $n \in \mathbb{N}$ and $\mathsf{P}_n := \{0, 1/n, \ldots, n/n\}$ be the partition
of $[0, 1]$ into n equal parts and consider the right endpoints $s_{n,i} := i/n$ for
$i = 1, \ldots, n$. Then $\mu(P_n) = 1/n \to 0$ and

$$S(\mathsf{P}_n, f) = \sum_{i=1}^{n} \frac{1}{1 + (i/n)^2} \left(\frac{i}{n} - \frac{i-1}{n} \right) = \sum_{i=1}^{n} \frac{n^2}{n^2 + i^2} \cdot \frac{1}{n} = \sum_{i=1}^{n} \frac{n}{n^2 + i^2}.$$

Hence by Proposition 6.31,

$$\frac{n}{n^2 + 1^2} + \frac{n}{n^2 + 2^2} + \cdots + \frac{n}{n^2 + n^2} = \sum_{i=1}^{n} \frac{n}{n^2 + i^2} \to \int_0^1 \frac{1}{1 + x^2} dx \quad \text{as } n \to \infty.$$

(iii) Consider

$$a_n := \sum_{i=1}^{n} \frac{1}{\sqrt{n^2 + in}} = \frac{1}{\sqrt{n^2 + n}} + \frac{1}{\sqrt{n^2 + 2n}} + \cdots + \frac{1}{\sqrt{n^2 + n^2}} \quad \text{for } n \in \mathbb{N}.$$

Then

$$a_n = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sqrt{1+(i/n)}} = \sum_{i=1}^{n} \frac{1}{\sqrt{1+(i/n)}} \left(\frac{i}{n} - \frac{i-1}{n} \right) \qquad \text{for all } n \in \mathbb{N}.$$

We observe that if we consider $f : [0,1] \to \mathbb{R}$ defined by $f(x) := 1/\sqrt{1+x}$ and let $P_n := \{0, 1/n, \ldots, n/n\}$ and $s_{n,i} := i/n$ for $n \in \mathbb{N}$ and $i = 1, \ldots, n$, then $a_n = S(P_n, f)$. In this case, f clearly has an antiderivative, namely $F : [0,1] \to \mathbb{R}$ given by $F(x) = 2\sqrt{1+x}$. Since $\mu(P_n) = 1/n \to 0$, we have by Proposition 6.31 and part (i) of the FTC (Proposition 6.21),

$$\sum_{i=1}^{n} \frac{1}{\sqrt{n^2 + in}} \to \int_0^1 \frac{1}{\sqrt{1+x}} dx = F(1) - F(0) = 2(\sqrt{2} - 1) \qquad \text{as } n \to \infty.$$

(iv) Let r be a nonnegative rational number and consider

$$a_n := \sum_{i=1}^{n} \frac{i^r}{n^{r+1}} = \frac{1^r + 2^r + \cdots + n^r}{n^{r+1}} \qquad \text{for } n \in \mathbb{N}.$$

Then

$$a_n = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{i}{n} \right)^r = \sum_{i=1}^{n} \left(\frac{i}{n} \right)^r \left(\frac{i}{n} - \frac{i-1}{n} \right) \qquad \text{for all } n \in \mathbb{N}.$$

As in the previous example, it follows that

$$\sum_{i=1}^{n} \frac{i^r}{n^{r+1}} \to \int_0^1 x^r \, dx = \frac{1}{r+1} \qquad \text{as } n \to \infty. \qquad \diamond$$

Notes and Comments

Given a bounded function $f : [a,b] \to \mathbb{R}$, we have produced two candidates, namely $U(f)$ and $L(f)$, for being designated the integral of f; when they coincide, we say that f is integrable and the common value is called its Riemann integral. While this approach demands patience and careful attention on the part of the reader to begin with, it is a natural way to formulate a plausible definition of 'area'. Hence we have preferred it to an alternative approach of defining integrability in terms of the existence of a 'limit' of Riemann sums. Actually, such an alternative approach goes beyond the concept of a limit of a sequence or of a function of a real variable introduced earlier. Indeed, it involves the limit of a 'net' of real numbers.

We have deduced all the essential features of the set of integrable functions from a single criterion called the Riemann condition. It is simple to state and easy to use. It does not involve the concept of a mesh of a partition. While

we have given a number of sufficient conditions for a bounded function f on $[a, b]$ to be integrable, we have not discussed a characterization of integrability in terms of the nature of the set of points of $[a, b]$ at which f is discontinuous. This characterization involves the notion of (Lebesgue) measure, or at least the notion of a subset of $\mathbb{R}$ having (Lebesgue) measure zero. It can be stated as follows: A bounded function $f : [a, b] \to \mathbb{R}$ is integrable if and only if the set of discontinuities of f has (Lebesgue) measure zero. We refer the reader to Theorem 11.33 of Rudin [53] or to Theorem 7.34 of Goldberg [28] for this result. It can be used to derive some of the main properties of the Riemann integral rather neatly. (See, for example, Section 7.3 of Goldberg [28].) A weaker condition involving the notion of a subset of $\mathbb{R}$ having content zero is discussed in Exercises 53–57.

The Fundamental Theorem of Calculus (FTC) has two parts. The first says in essence that the integral of the derivative of a function gives back the function and the second says that the derivative of the integral of a function again gives back the function. These two parts are variously known as the First Fundamental Theorem of Calculus and the Second Fundamental Theorem of Calculus. We have given proofs of these two parts that are independent of each other. When the given function is continuous, the first part can be easily deduced from the second. (See Exercise 16.) The methods of Integration by Parts and Integration by Substitution are derived from the FTC under minimal hypotheses.

In the section on Riemann sums, we have introduced the concept of the mesh of a partition and used it to obtain approximations of a Riemann integral on the one hand, and also to calculate limits of certain sequences, each term of which is a Riemann sum. Again, we have carefully avoided any mention of 'limits' as the mesh of a partition approaches zero since that would involve a more general notion of 'limit', which we do not wish to discuss in this book. The reader may consult the description of the 'Riemann net' given on page 230 of the book of Joshi [38].

Exercises

Part A

1. Let $c \in [a, b]$ and $f : [a, b] \to \mathbb{R}$ be given by

$$f(x) := \begin{cases} 0 & \text{if } a \leq x \leq c, \\ 1 & \text{if } c < x \leq b. \end{cases}$$

   Show from first principles that f is integrable on $[a, b]$. Also, prove that this follows from Proposition 6.10.

2. Let $c \in (a, b)$ and $f : [a, b] \to \mathbb{R}$ be given by

$$f(x) := \begin{cases} (x-c)/(a-c) & \text{if } a \le x \le c, \\ (x-c)/(b-c) & \text{if } c < x \le b. \end{cases}$$

Show from first principles that f is integrable on $[a,b]$. Also, prove that this follows from Proposition 6.10. (Hint: For $n \in \mathbb{N}$, consider the partition $P_n := \{a, a + (c-a)/n, \ldots, a + (c-a)(n-1)/n, c, c + (b-c)/n, \ldots, c + (b-c)(n-1)/n, b\}$.)

3. Let $f : [0,1] \to \mathbb{R}$ be given by

$$f(x) := \begin{cases} 1+x & \text{if } x \text{ is rational}, \\ 0 & \text{if } x \text{ is irrational}. \end{cases}$$

Is f integrable?

4. Let $f : [a,b] \to \mathbb{R}$ be integrable. Show that the Riemann integral of f is the unique real number r satisfying the following condition: For every $\epsilon > 0$, there is a partition P_ϵ of $[a,b]$ such that

$$r - \epsilon < L(P_\epsilon, f) \le r \le U(P_\epsilon, f) < r + \epsilon.$$

5. Let $f : [0,3] \to \mathbb{R}$ be defined by

$$f(x) := \begin{cases} 0 & \text{if } 0 \le x \le 1, \\ 2 & \text{if } 1 < x \le 2, \\ -1 & \text{if } 2 < x \le 3. \end{cases}$$

Show that f is neither monotonic nor continuous on $[0,3]$, but f is integrable on $[0,3]$. Find the Riemann integral of f.

6. Let $f, g : [a,b] \to \mathbb{R}$ be bounded functions. Show that

$$L(f) + L(g) \le L(f+g) \quad \text{and} \quad U(f+g) \le U(f) + U(g).$$

Hence conclude that if f and g are integrable, then so is $f + g$, and the Riemann integral of $f + g$ is equal to the sum of the Riemann integrals of f and g.

7. Let $f, g : [a,b] \to \mathbb{R}$ be integrable. Show that the functions $\max(f, g) : [a,b] \to \mathbb{R}$ and $\min(f,g) : [a,b] \to \mathbb{R}$ given by

$$\max(f,g)(x) = \max\{f(x), g(x)\} \quad \text{and} \quad \min(f,g)(x) = \min\{f(x), g(x)\}$$

are integrable. (Hint: $\max(f,g) = (f + g + |f - g|)/2$ and $\min(f,g) = (f + g - |f - g|)/2$.)

8. Give examples of bounded functions $f, g : [a,b] \to \mathbb{R}$ that are not integrable, but $|f|$, $f + g$, and fg are all integrable.

9. Let $f : [a,b] \to \mathbb{R}$ be a function. Show that f is integrable if (i) rf is integrable for some nonzero $r \in \mathbb{R}$, or (ii) if f is bounded, $f(x) \ne 0$ for all $x \in [a,b]$, and $1/f$ is integrable.

10. Let $f : [a, b] \to \mathbb{R}$ be any function. Suppose there is $r \in \mathbb{R}$ and for each $n \in \mathbb{N}$, there are integrable functions $g_n, h_n : [a, b] \to \mathbb{R}$ with $g_n \leq f \leq h_n$ such that $\int_a^b g_n(x)dx \to r$ and $\int_a^b h_n(x)dx \to r$ as $n \to \infty$. Show that f is integrable and the Riemann integral of f is equal to r.

11. Let $f : [a, b] \to \mathbb{R}$ be integrable and $f(x) \geq 0$ for all $x \in [a, b]$. Show that $\int_a^b f(x)dx \geq 0$. If, in addition, f is continuous and $\int_a^b f(x)dx = 0$, then show that $f(x) = 0$ for all $x \in [a, b]$. Give an example of an integrable function on $[a, b]$ such that $f(x) \geq 0$ for all $x \in [a, b]$ and $\int_a^b f(x)dx = 0$, but $f(x) \neq 0$ for some $x \in [a, b]$.

12. Evaluate the following limits.

(i) $\displaystyle\lim_{h \to 0} \frac{1}{h} \int_x^{x+h} \frac{du}{u + \sqrt{u^2 + 1}}$,   (ii) $\displaystyle\lim_{x \to 0} \frac{1}{x^3} \int_0^x \frac{t^2 dt}{t^4 + 1}$,

(iii) $\displaystyle\lim_{x \to 0} \frac{1}{x^6} \int_0^{x^2} \frac{t^2 dt}{t^6 + 1}$,   (iv) $\displaystyle\lim_{x \to x_0} \frac{x}{x - x_0} \int_{x_0}^x f(t)dt$,

(v) $\displaystyle\lim_{x \to x_0} \frac{x}{x^2 - x_0^2} \int_{x_0}^x f(t)dt$, provided f continuous at x_0.

13. If $x := \displaystyle\int_0^y \frac{dt}{\sqrt{1 + t^2}}$, find $\dfrac{d^2 y}{dx^2}$.

14. Let $a, b, c \in \mathbb{R}$ with $a < c < b$ and for $j = 1, 2, 3$, consider $f_j : [a, b] \to \mathbb{R}$ given by

(i) $f_1(x) := \begin{cases} 0 & \text{if } x \leq c, \\ 1 & \text{if } c < x, \end{cases}$

(ii) $f_2(x) := \begin{cases} 0 & \text{if } x \neq c, \\ 1 & \text{if } x = c, \end{cases}$

(iii) $f_3(x) := \begin{cases} (x - c)/(a - c) & \text{if } x \leq c, \\ (x - c)/(b - c) & \text{if } c < x. \end{cases}$

For $j = 1, 2, 3$, let $F_j(x) := \int_a^x f_j(t)dt, x \in [a, b]$. Find F_j for $j = 1, 2, 3$. Verify that f_1 is discontinuous at c, F_1 is continuous but not differentiable at c, f_2 is discontinuous at c, F_2 is differentiable at c but $F_2'(c) \neq f_2(c)$, f_3 is continuous at c but not differentiable at c, F_3 is differentiable at c and $F_3'(c) = f_3(c)$.

[**Note**: There is an integrable function $f : [a, b] \to \mathbb{R}$ such that f is discontinuous at c, but the corresponding function $F : [a, b] \to \mathbb{R}$ is differentiable at c and $F'(c) = f(c)$. See Proposition 7.17.]

15. Let $n \in \mathbb{N}$. Find a function $f : [-1, 1] \to \mathbb{R}$ for which $f^{(n)}(0)$ exists, but $f^{(n+1)}(0)$ does not. (Hint: Begin with the absolute value function and use part (ii) of Proposition 6.20 repeatedly.)

16. If $f : [a, b] \to \mathbb{R}$ is continuous, then prove part (i) of the FTC using part (ii) of the FTC. (Hint: Two antiderivatives of f differ by an additive constant.)

17. Let $f : [a, b] \to \mathbb{R}$ be continuous and consider the function $F : [a, b] \to \mathbb{R}$ given by $F(x) := \int_a^x f(t)dt$ for $x \in [a, b]$. If $f(x) \geq 0$ for all $x \in [a, b]$, then show that F is monotonically increasing on $[a, b]$, and if f monotonically

increasing on $[a, b]$, then F is convex on $[a, b]$. (Hint: Part (i) of Proposition 4.27 and Part (i) of Proposition 4.31.)

18. Let $f : [a, \infty) \to \mathbb{R}$ be a bounded function such that f is integrable on $[a, x]$ for every $x \geq a$. Let $F(x) := \int_a^x f(t)dt$ for $x \geq a$. Show that F is uniformly continuous on $[a, \infty)$.

19. Let $f : [0, \infty) \to \mathbb{R}$ be continuous and $f(x) \geq 0$ for all $x \in [0, \infty)$. If for each $b > 0$, the area bounded by the x-axis, the lines $x = 0, x = b$, and the curve $y = f(x)$ is given by $\sqrt{b^2 + 1} - 1$, determine the function f.

20. Let p be a real number and let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function such that $f(x + p) = f(x)$ for all $x \in \mathbb{R}$. (Such a function is said to be **periodic**.) Show that the integral $\int_a^{a+p} f(t)dt$ has the same value for every real number a. (Hint: Part (ii) of Proposition 6.21.)

21. Let $f : [a, b] \to \mathbb{R}$ be continuous. Show that for every $x \in [a, b]$,
$$\int_a^x \left[\int_a^u f(t)dt \right] du = \int_a^x (x - u)f(u)du.$$

22. Let $f : [a, b] \to \mathbb{R}$ be integrable. Define $G : [a, b] \to \mathbb{R}$ by
$$G(x) := \int_x^b f(t)dt.$$

Show that G is continuous on $[a, b]$. Further, show that if f is continuous at $c \in [a, b]$, then G is differentiable at c and $G'(c) = -f(c)$. (Hint: Propositions 6.7, 6.20, and 6.21.)

23. Let $g : [c, d] \to \mathbb{R}$ be such that $g([c, d]) \subseteq [a, b]$, and let $f : [a, b] \to \mathbb{R}$ be integrable. Define $G : [c, d] \to \mathbb{R}$ by
$$G(y) := \int_a^{g(y)} f(t)dt.$$

If g is differentiable at $y_0 \in [c, d]$ and f is continuous at $g(y_0)$, then show that G is differentiable at y_0 and $G'(y_0) = f(g(y_0))g'(y_0)$.

24. **(Leibniz's Rule for Integrals)** Let f be a continuous function on $[a, b]$ and u, v be differentiable functions on $[c, d]$. If the ranges of u and v are contained in $[a, b]$, prove that
$$\frac{d}{dx} \int_{u(x)}^{v(x)} f(t)dt = \left[f(v(x)) \frac{dv}{dx} - f(u(x)) \frac{du}{dx} \right].$$

25. For $x \in \mathbb{R}$, let $F(x) := \int_1^{2x} \frac{1}{1 + t^2} dt$ and $G(x) := \int_0^{x^2} \frac{1}{1 + \sqrt{|t|}} dt$. Find F' and G'.

26. Let $f : [0, \infty) \to \mathbb{R}$ be continuous. Find $f(2)$ if for all $x \geq 0$,
(i) $\int_0^x f(t)dt = x^2(1 + x)$, (ii) $\int_0^{f(x)} t^2 dt = x^2(1 + x)$,
(iii) $\int_0^{x^2} f(t)dt = x^2(1 + x)$, (iv) $\int_0^{x^2(1+x)} f(t)dx = x$.

27. Let $n, m \in \mathbb{N}$. Find $\lim\limits_{m \to \infty} \int_0^1 \dfrac{x^n}{(1+x)^m} dx$ and $\lim\limits_{n \to \infty} \int_0^1 \dfrac{x^n}{(1+x)^m} dx$.

28. Find $\lim\limits_{n \to \infty} \int_0^1 \dfrac{nx^{n-1}}{1+x} dx$. (Hint: Proposition 6.25.)

29. Let $f : [a, b] \to \mathbb{R}$ be a differentiable function. If F is an antiderivative of f on $[a, b]$, then show that

$$\int_a^b f^2(x)dx = F(b)F'(b) - F(a)F'(a) - \int_a^b F(x)F''(x)dx.$$

30. Evaluate (i) $\int_0^{1/4} \dfrac{x}{\sqrt{1-4x^2}} dx$, (ii) $\int_1^8 x^{1/3}(x^{4/3} - 1)^{1/2} dx$.
    (Hint: Proposition 6.26.)

31. Let $f : [a, b] \to \mathbb{R}$ be a differentiable function such that f' is continuous on $[a, b]$ and $f'(x) \neq 0$ for all $x \in [a, b]$. If $f([a, b]) = [c, d]$, then show that $f^{-1} : [c, d] \to \mathbb{R}$ is integrable and

$$\int_c^d f^{-1}(y)dy = f^{-1}(d)d - f^{-1}(c)c - \int_{f^{-1}(c)}^{f^{-1}(d)} f(x)dx.$$

(Hint: Propositions 6.25 and 6.26.)

32. Let $f : [a, b] \to \mathbb{R}$ be a bounded function and define $g : [-b, -a] \to \mathbb{R}$ by $g(t) := f(-t)$. Show that $L(g) = L(f)$ and $U(g) = U(f)$. Deduce that g is integrable on $[-b, -a]$ if and only if f is integrable on $[a, b]$ and in that case the Riemann integral of g is equal to the Riemann integral of f. (Compare the proof of part (ii) of Proposition 6.26.)

33. Let $f : [a, b] \to \mathbb{R}$ be integrable and for $n \in \mathbb{N}$, let P_n be a partition of $[a, b]$ such that $U(P_n, f) - L(P_n, f) \to 0$. Show that $U(P_n, f) \to \int_a^b f(x)dx$, $L(P_n, f) \to \int_a^b f(x)dx$, and also $S(P_n, f) \to \int_a^b f(x)dx$, where $S(P_n, f)$ is a Riemann sum for f corresponding to P_n. (Compare Proposition 6.5 and Lemma 6.30.)

34. Let $f : [a, b] \to \mathbb{R}$ be an integrable function. If (P_n) is a sequence of partitions of $[a, b]$ such that $\mu(P_n) \to 0$, then show that $U(P_n, f) - L(P_n, f) \to 0$. Is the converse true?

35. Let $f : [a, b] \to \mathbb{R}$ be a bounded function. Without using Lemma 6.30, show that f is Riemann integrable if and only if there is $r \in \mathbb{R}$ satisfying the following condition: Given $\epsilon > 0$, there is a partition P_ϵ of $[a, b]$ such that $|S(P, f) - r| < \epsilon$, where P is any refinement of P_ϵ and $S(P, f)$ is any Riemann sum for f corresponding to P.

36. Assuming that f is integrable on $[0, 1]$, show that

$$\lim_{n \to \infty} \frac{1}{n} \left[f\left(\frac{1}{n}\right) + f\left(\frac{2}{n}\right) + \cdots + f\left(\frac{n}{n}\right) \right] = \int_0^1 f(x)dx.$$

37. Consider the sequence whose nth term is given by the following. In each case, determine the limit of the sequence by expressing the nth term as a Riemann sum for a suitable function.

(i) $\dfrac{1}{n^{17}} \displaystyle\sum_{i=1}^{n} i^{16}$,   (ii) $\dfrac{1}{n^{5/2}} \displaystyle\sum_{i=1}^{n} i^{3/2}$,   (iii) $\displaystyle\sum_{i=1}^{n} \dfrac{1}{\sqrt{in + n^2}}$,

(iv) $\dfrac{1}{n} \left\{ \displaystyle\sum_{i=1}^{n} \left(\dfrac{i}{n}\right) + \displaystyle\sum_{i=n+1}^{2n} \left(\dfrac{i}{n}\right)^{3/2} + \displaystyle\sum_{i=2n+1}^{3n} \left(\dfrac{i}{n}\right)^{2} \right\}$.

38. Do $\displaystyle\lim_{n\to\infty} \sum_{i=1}^{n} \dfrac{1}{\sqrt{i+n}}$ and $\displaystyle\lim_{n\to\infty} \dfrac{1}{n^{18}} \sum_{i=1}^{n} i^{16}$ exist? If yes, find them.

39. Find an approximate value of $1^{1/3} + 2^{1/3} + \cdots + 1000^{1/3}$.

Part B

40. Let $a, b \in \mathbb{R}$ with $0 \leq a < b$ and $m \in \mathbb{N}$, and let $f : [a, b] \to \mathbb{R}$ be defined by $f(x) := x^m$. Show from the first principles that

$$\int_a^b f(x)dx = \frac{b^{m+1} - a^{m+1}}{m + 1}.$$

(Hint: If $P = \{x_0, x_1, \ldots, x_n\}$ is a partition of $[a, b]$, then for each $j = 0, 1, \ldots, m$, we have $L(P, f) \leq \sum_{i=1}^{n} x_i^{m-j} x_{i-1}^{j}(x_i - x_{i-1}) \leq U(P, f)$. Also, $\sum_{j=0}^{m} \left[\sum_{i=1}^{n} x_i^{m-j} x_{i-1}^{j}(x_i - x_{i-1}) \right] = b^{m+1} - a^{m+1}$.)

41. Let $f : [a, b] \to \mathbb{R}$ be a bounded function. For $c \in (a, b)$, let f_1 and f_2 denote the restrictions of f to the subintervals $[a, c]$ and $[c, b]$ respectively. Prove the following:

(i) $L(f) = L(f_1) + L(f_2)$,   (ii) $U(f) = U(f_1) + U(f_2)$.

[Note: The results in (i) and (ii) are refined versions of Proposition 6.7, and may be referred to as **Domain Additivity of Lower Riemann Integrals** and **Domain Additivity of Upper Riemann Integrals**, respectively.]

42. Let $f : [a, b] \to \mathbb{R}$ be integrable and $\phi : [m(f), M(f)] \to \mathbb{R}$ be continuous. Show that $\phi \circ f : [a, b] \to \mathbb{R}$ is integrable. (Hint: Given $\epsilon > 0$, find $\delta > 0$ using the uniform continuity of ϕ. There is a partition P of $[a, b]$ such that $U(P, f) - L(P, f) < \delta^2$. Divide the sum in $U(P, f) - L(P, f)$ into two classes depending on whether $M_i(f) - m_i(f)$ is less than δ, or greater than or equal to δ. Use the Riemann condition for $\phi \circ f$.)

43. Let $f_1, \ldots, f_m : [a, b] \to \mathbb{R}$ be integrable functions and let $r_j := \int_a^b f_j(x)dx$ for $j = 1, \ldots, m$. Show that the function $\sqrt{f_1^2 + \cdots + f_m^2}$ is integrable and

$$\sqrt{r_1^2 + \cdots + r_m^2} \leq \int_a^b \sqrt{f_1^2(x) + \cdots + f_m^2(x)} \, dx.$$

(Hint: Note that $\sum_{j=1}^{m} r_j^2 = \sum_{j=1}^{m} r_j \int_a^b f_j(x)dx = \int_a^b \left(\sum_{j=1}^{m} r_j f_j(x) \right) dx$ and use Proposition 1.12.)

44. Let $m, n \in \mathbb{Z}$ with $m, n \geq 0$. Show that

$$\int_0^1 x^m (1-x)^n \, dx = \frac{m! \, n!}{(m+n+1)!}.$$

(Hint: If $n \in \mathbb{N}$ and $I_{m,n}$ denotes the given integral, then using Integration by Parts, $I_{m,n} = [n/(m+1)]I_{m+1,n-1}$, and $I_{m+n,0} = 1/(m+n+1)$.)

45. Let $a \in \mathbb{R}$ and $n \in \mathbb{Z}$ with $n \geq 0$. Show that

$$\int_0^a (a^2 - x^2)^n \, dx = \frac{(2^n n!)^2}{(2n+1)!} \cdot a^{2n+1}.$$

Deduce that

$$1 - \frac{1}{3}\binom{n}{1} + \frac{1}{5}\binom{n}{2} - \frac{1}{7}\binom{n}{3} + \cdots + \frac{(-1)^n}{2n+1}\binom{n}{n} = \frac{(2^n n!)^2}{(2n+1)!}.$$

(Hint: If $n \in \mathbb{N}$ and I_n denotes the given integral, then $I_n = a^2 I_{n-1} - \int_0^a x[x(a^2-x^2)^{n-1}]dx$, and using Integration by Parts, $I_n = a^2[2n/(2n+1)]I_{n-1}$, and $I_0 = a$.)

46. **(Taylor's Theorem with Integral Remainder)** Let n be a nonnegative integer and let $f : [a, b] \to \mathbb{R}$ be such that $f', f'', \ldots, f^{(n+1)}$ exist and $f^{(n+1)}$ is continuous on $[a, b]$. Show that

$$f(b) = f(a) + f'(a)(b-a) + \cdots + \frac{f^{(n)}(a)}{n!}(b-a)^n + \frac{1}{n!}\int_a^b (b-t)^n f^{(n+1)}(t)dt.$$

Further, show that the remainder is equal to

$$\frac{(b-a)^{n+1}}{n!} \int_0^1 (1-s)^n f^{(n+1)}(a + s(b-a))ds.$$

(Hint: Induction on n and Integration by Parts.)
[Note: The integral remainder does not involve an undetermined number $c \in (a, b)$.]

47. **(Taylor's Theorem for Integrals)** Let $n \in \mathbb{N}$ and $f : [a, b] \to \mathbb{R}$ be such that $f', f'', \ldots, f^{(n-1)}$ exist on $[a, b]$, and further, $f^{(n-1)}$ is continuous on $[a, b]$ and differentiable on (a, b). Show that there is $c \in (a, b)$ such that

$$\int_a^b f(x)dx = f(a)(b-a) + \cdots + \frac{f^{(n-1)}(a)}{n!}(b-a)^n + \frac{f^{(n)}(c)}{(n+1)!}(b-a)^{n+1}.$$

(Hint: For $x \in [a, b]$, define $F(x) := \int_a^x f(t)dt$ and apply Proposition 4.23.)

48. **(Theorem of Bliss)** Let $f, g : [a, b] \to \mathbb{R}$ be integrable. For each $n \in \mathbb{N}$, consider a partition $P_n := \{x_{n,0}, x_{n,1}, \ldots, x_{n,k_n}\}$ of $[a, b]$, and for $i = 1, \ldots, k_n$, let $s_{n,i}, t_{n,i} \in [x_{n,i-1}, x_{n,i}]$, and let

$$\widetilde{S}(P_n, fg) := \sum_{i=1}^{k_n} f(s_{n,i})g(t_{n,i})(x_{n,i} - x_{n,i-1}).$$

If $\mu(P_n) \to 0$ as $n \to \infty$ and if g is continuous on $[a, b]$, then show that $\widetilde{S}(P_n, fg) \to \int_a^b f(x)g(x)dx$ as $n \to \infty$. (Hint: Note that $S(P_n, fg) :=$ $\sum_{i=1}^{k_n} f(s_{n,i})g(s_{n,i})(x_{n,i} - x_{n,i-1}) \to \int_a^b f(x)g(x)dx$ and use the uniform continuity of g.)

49. Let $f : [a, b] \to \mathbb{R}$ be a monotonic function. If $G : [a, b] \to \mathbb{R}$ is differentiable and G' is continuous, then show that there is $c \in [a, b]$ such that

$$\int_a^b f(x)G'(x)dx = f(b)G(b) - f(a)G(a) - G(c)[f(b) - f(a)].$$

(Hint: Given any partition $P = \{x_0, x_1, \ldots, x_n\}$ of $[a, b]$, consider the sum $\sum_{i=1}^n f(x_i)[G(x_i) - G(x_{i-1})]$. Write it as $f(b)G(b) - f(a)G(a) - \sum_{i=1}^n G(x_{i-1})[f(x_i) - f(x_{i-1})]$ and also as $\sum_{i=1}^n f(x_i)G'(s_i)(x_i - x_{i-1})$ for some $s_i \in [x_{i-1}, x_i]$. Use the Theorem of Bliss (Exercise 48) and the inequalities $m(g)[f(b) - f(a)] \leq \sum_{i=1}^n G(x_{i-1})[f(x_i) - f(x_{i-1})] \leq M(g)[f(b) - f(a)]$.)

50. **(First Mean Value Theorem for Integrals)** Let $f : [a, b] \to \mathbb{R}$ be a continuous function and $g : [a, b] \to \mathbb{R}$ be a nonnegative integrable function. Use the IVP of f to show that there is $c \in [a, b]$ such that

$$\int_a^b f(x)g(x)dx = f(c) \int_a^b g(x)dx.$$

Give examples to show that neither the continuity of f nor the nonnegativity of g can be omitted.

[Note: For another version of this result, see Exercise 72.]

51. **(Second Mean Value Theorem for Integrals)** Let $f : [a, b] \to \mathbb{R}$ be a monotonic function and $g : [a, b] \to \mathbb{R}$ be either a nonnegative integrable function or a continuous function. Show that there is $c \in [a, b]$ such that

$$\int_a^b f(x)g(x)dx = f(a) \int_a^c g(x)dx + f(b) \int_c^b g(x)dx.$$

Give an example to show that the monotonicity of f cannot be omitted. (Hint: Without loss of generality, suppose f is (monotonically) increasing. Let $G(x) := \int_a^x g(t)dt$ for $x \in [a, b]$. If g is a nonnegative integrable function, then $f(a)G(b) \leq \int_a^b f(x)g(x)dx \leq f(b)G(b)$. If g is continuous, use Exercise 49.)

52. Let D be a bounded subset of $\mathbb{R}$ and $f : D \to \mathbb{R}$ be a bounded function. Suppose $D \subseteq [a, b]$ for $a, b \in \mathbb{R}$ and $f^* : [a, b] \to \mathbb{R}$ is defined by

$$f^*(x) := \begin{cases} f(x) & \text{if } x \in D, \\ 0 & \text{otherwise.} \end{cases}$$

The function f is said to be **integrable** (on D) if the function f^* is integrable (on $[a, b]$). In this case, we define the **Riemann integral** of f (on D) by

$$\int_D f(x)dx := \int_a^b f^*(x)dx.$$

(i) Show that the above definition is independent of the interval $[a, b]$ containing D.

(ii) Show that analogues of Propositions 6.15 and 6.18 hold for integrable functions on D.

53. A bounded subset E of $\mathbb{R}$ is said to be of **(one-dimensional) content zero** if the following condition holds: For every $\epsilon > 0$, there is a finite number of closed intervals whose union contains E and the sum of whose lengths is less than ϵ. Prove the following statements:

(i) A subset of a set of content zero is of content zero.

(ii) A finite union of sets of content zero is of content zero.

(iii) If E is of content zero and ∂E denotes the boundary of E, then $E \cup \partial E$ is of content zero.

(iv) A set E is of content zero if and only if the interior of E is empty and ∂E is of content zero.

(v) Every finite subset of $\mathbb{R}$ is of content zero.

(vi) The infinite set $\{1/n : n \in \mathbb{N}\}$ is of content zero.

(vii) The infinite set $\mathbb{Q} \cap [0, 1]$ is not of content zero.

54. Let D be a bounded subset of $\mathbb{R}$ and $f : D \to \mathbb{R}$ be a bounded function. If the boundary ∂D of D is of content zero and if the set of discontinuities of f is also of content zero, then show that f is integrable. In particular, if D is of content zero, then show that f is integrable and its Riemann integral is equal to zero. (Compare Remark 6.8.)

55. Let $f : [a, b] \to \mathbb{R}$ be a bounded function. If the set of discontinuities of f is of content zero, show that f is integrable. Is the converse true? (Hint: Exercise 34 of Chapter 3 and Example 6.16.)

56. Let D be a bounded subset of $\mathbb{R}$. Let $1_D : D \to \mathbb{R}$ be defined by $1_D(x) := 1$ for all $x \in D$. Prove the following statements:

(i) 1_D is integrable if and only if ∂D is of content zero.
[Note: If 1_D is integrable, then $\int_D 1_D(x)dx$ is called the **length** of the set D.]

(ii) The length of D is zero if and only if D is of content zero.

(iii) If $f : D \to \mathbb{R}$ is a bounded function and $D_0 \subseteq D$ is such that ∂D is of content zero, then f is integrable on D_0.

57. Let $f : [a, b] \to \mathbb{R}$ be integrable, and $g : [a, b] \to \mathbb{R}$ be a bounded function such that the set $\{x \in [a, b] : g(x) \neq f(x)\}$ is of content zero. Show that g is integrable and

$$\int_a^b g(x)dx = \int_a^b f(x)dx.$$

(Compare Proposition 6.12.)

7

Elementary Transcendental Functions

In this chapter we shall use the theory of Riemann integration developed in Chapter 6 to introduce some classical functions, known as the logarithmic, exponential, and trigonometric functions. Collectively, these are called the **elementary transcendental functions**. In Sections 7.1 and 7.2 below we give formal definitions of these functions and derive several of their interesting properties. In this process, the important real numbers e and π will also be formally defined.

In the earlier chapters, we have scrupulously avoided any mention of the logarithmic, exponential, and trigonometric functions since their very definitions had to be postponed. As a result, several interesting examples and counterexamples could not be given earlier. Many of these arise from the function obtained by taking the sine of the reciprocal of the identity function. These are discussed in Section 7.3.

The trigonometric functions enable us to introduce polar coordinates of a point in the plane other than the origin. This is done in Section 7.4 and in this context, we also give a formal definition of the angle between two line segments emanating from a point as well as of the angle between two intersecting curves.

In the final section of this chapter, we show that the elementary transcendental functions are indeed transcendental, that is, they are not algebraic functions.

In the section on exercises, we have given problems of theoretical importance as well as of problems of practical use. The latter include several trigonometric results that are listed for ready reference. In addition to this section of exercises, we include a section devoted to revision exercises in which the reader will revisit many concepts considered earlier in this book in relation to the new supply of functions that is made available in this chapter.

7.1 Logarithmic and Exponential Functions

We have seen in Example 4.7 that if $r \in \mathbb{Q}$ and $g : (0, \infty) \to \mathbb{R}$ is the rth-power function defined by $g(x) := x^r$, then $g'(x) = rx^{r-1}$ for all $x \in (0, \infty)$. This implies that for every rational number s, except $s = -1$, the sth-power function can be integrated in terms of a similar function. In fact,

$$\frac{d}{dx}\left(\frac{x^{s+1}}{s+1}\right) = x^s \quad \text{provided } s \neq -1.$$

To deal with the exceptional case $s = -1$, a new function has to be introduced, and we shall do so in this section. This will in fact enable us to define and study real powers of positive real numbers.

The function $f : (0, \infty) \to \mathbb{R}$ defined by $f(t) = 1/t$ is continuous. Hence it is Riemann integrable on every closed and bounded subinterval of $(0, \infty)$. For $x \in (0, \infty)$, we define the **natural logarithm** of x by

$$\ln x := \int_1^x \frac{1}{t} dt.$$

The function $\ln : (0, \infty) \to \mathbb{R}$ is known as the **logarithmic function**. We write $\ln x$, rather than $\ln(x)$, for the value of the logarithmic function at $x \in (0, \infty)$.

Clearly, $\ln 1 = 0$. Moreover, since $1/t \geq 0$ for all $t \in (0, \infty)$, we have $\ln x \geq 0$ if $x > 1$, while $\ln x \leq 0$ if $0 < x < 1$.

Proposition 7.1 (Properties of the Logarithmic Function).

(i) $\ln$ *is a differentiable function on* $(0, \infty)$ *and*

$$(\ln)'x = \frac{1}{x} \quad \text{for every } x \in (0, \infty).$$

(ii) $\ln$ *is strictly increasing as well as strictly concave on* $(0, \infty)$.
(iii) $\ln x \to \infty$ *as* $x \to \infty$, *whereas* $\ln x \to -\infty$ *as* $x \to 0^+$.
(iv) *For every* $y \in \mathbb{R}$, *there is a unique* $x \in (0, \infty)$ *such that* $\ln x = y$. *In other words,* $\ln : (0, \infty) \to \mathbb{R}$ *is a bijective function.*

Proof. (i) Since the function $g : (0, \infty) \to \mathbb{R}$ given by $g(t) = 1/t$ is continuous, the Fundamental Theorem of Calculus (Proposition 6.21) shows that the function $\ln$ is differentiable on $(0, \infty)$ and $(\ln)'x = g(x) = 1/x$ at every $x \in (0, \infty)$.

(ii) Since the derivative of $\ln$ is positive on $(0, \infty)$, it follows from part (iii) of Proposition 4.27 that $\ln$ is strictly increasing on $(0, \infty)$. Further, since

$$(\ln)''x = -\frac{1}{x^2} < 0 \quad \text{for every } x \in (0, \infty),$$

it follows from part (iv) of Proposition 4.32 that ln is strictly concave on $(0, \infty)$.

(iii) Given any positive integer $n \geq 2$, we have

$$\ln n = \int_1^n \frac{1}{t} dt = \sum_{k=2}^n \int_{k-1}^k \frac{1}{t} dt \geq \sum_{k=2}^n \int_{k-1}^k \frac{1}{k} dt = \sum_{k=2}^n \frac{1}{k}$$

since $(1/t) \geq (1/k)$ for all $0 < t \leq k$, whereas

$$\ln \frac{1}{n} = -\int_{1/n}^1 \frac{1}{t} dt = -\sum_{k=2}^n \int_{1/k}^{1/(k-1)} \frac{1}{t} dt$$

$$\leq -\sum_{k=2}^n \int_{1/k}^{1/(k-1)} (k-1) dt = -\sum_{k=2}^n \frac{1}{k}$$

since $-(1/t) \leq -(k-1)$ for all $0 < t \leq 1/(k-1)$. Because $\sum_{k=2}^n (1/k) \to \infty$ as $n \to \infty$ by part (ii) of Example 2.13, we see that the function ln is neither bounded above nor bounded below on $(0, \infty)$. Also, since ln is (strictly) increasing on $(0, \infty)$, it follows from Proposition 3.35 that $\ln x \to \infty$ as $x \to \infty$ and $\ln x \to -\infty$ as $x \to 0^+$.

(iv) The function ln is one-one since it is strictly increasing. Now, let $y \in \mathbb{R}$. Since $\ln x \to -\infty$ as $x \to 0^+$, there is some $x_0 > 0$ such that $\ln x_0 < y$ and since $\ln x \to \infty$ as $x \to \infty$, there is some $x_1 > 0$ such that $y < \ln x_1$. But by part (i) above, the function ln is continuous on the interval $[x_0, x_1]$. So, the IVP (Proposition 3.13) shows that there is some $x \in (x_0, x_1)$ such that $\ln x = y$. Thus the function $\ln : (0, \infty) \to \mathbb{R}$ is bijective. $\qquad \square$

An immediate consequence of part (iv) of Proposition 7.1 is that there is a unique positive real number e such that $\ln e = 1$. This number e plays a significant role in analysis.

An elementary estimate for the number e can be obtained by noting that

$$\ln 2 = \int_1^2 \frac{1}{t} dt \leq \int_1^2 1 \, dt = 1$$

and

$$\ln 4 = \int_1^4 \frac{1}{t} dt = \int_1^2 \frac{1}{t} dt + \int_2^4 \frac{1}{t} dt \geq \int_1^2 \frac{1}{2} dt + \int_2^4 \frac{1}{4} dt = \frac{1}{2} + \frac{1}{2} = 1.$$

Since $\ln : (0, \infty) \to \mathbb{R}$ is increasing, it follows that

$$2 \leq e \leq 4.$$

Better lower and upper bounds for the number e can be obtained. (See, for example, Exercise 7.) The first few digits in the decimal expansion of e are given by

$$e = 2.71828182\ldots.$$

In fact, e is an irrational number (Exercise 20 of Chapter 2), and furthermore, e is a transcendental number, that is, it is not a root of any nonzero polynomial with rational coefficients. (See [7] for a proof.)

The geometric properties of the logarithmic function proved in Proposition 7.1 can be used to draw its graph as in Figure 7.1.
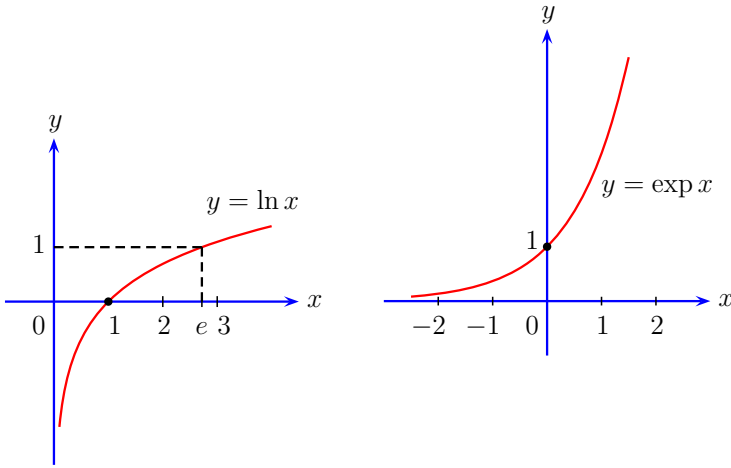
Fig. 7.1. Graphs of the logarithmic and the exponential functions

It follows from part (i) of Proposition 7.1 that the function ln is infinitely differentiable on $(0, \infty)$ and for $k = 1, 2, \ldots$, we have

$$(\ln)^{(k)} x = (-1)^{k-1}(k-1)! \, x^{-k}, \quad x \in (0, \infty).$$

Hence the nth Taylor polynomial for ln about 1 is given by

$$P_n(x) = \ln 1 + \sum_{k=1}^{n} \frac{(\ln)^{(k)} 1}{k!}(x-1)^k = \sum_{k=1}^{n} \frac{(-1)^{k-1}}{k}(x-1)^k, \quad x \in \mathbb{R}.$$

In particular, the linear and the quadratic approximations of ln around 1 are given by

$$L(x) = P_1(x) = x - 1 \quad \text{and} \quad Q(x) = P_2(x) = (x-1) - \frac{(x-1)^2}{2}, \quad x \in \mathbb{R}.$$

Let us now turn to the inverse of the logarithmic function. The inverse of the bijective function $\ln : (0, \infty) \to \mathbb{R}$ is known as the **exponential function** and is denoted by $\exp : \mathbb{R} \to (0, \infty)$. We write $\exp x$, rather than $\exp(x)$, for

the value of the exponential function at $x \in \mathbb{R}$. Thus for any $x \in \mathbb{R}$ and any $y \in (0, \infty)$, we have

$$\exp x = y \iff \ln y = x.$$

Note that by definition, $\exp x > 0$ for all $x \in \mathbb{R}$. Moreover, since $\ln 1 = 0$ and $\ln e = 1$, we have $\exp 0 = 1$ and $\exp 1 = e$.

Proposition 7.2 (Properties of the Exponential Function).

(i) exp *is a differentiable function on* $\mathbb{R}$ *and*

$$(\exp)'x = \exp x \quad \text{for every } x \in \mathbb{R}.$$

(ii) exp *is strictly increasing as well as strictly convex on* $\mathbb{R}$.
(iii) $\exp x \to \infty$ *as* $x \to \infty$, *whereas* $\exp x \to 0$ *as* $x \to -\infty$.

Proof. (i) Let $x \in \mathbb{R}$ and $c \in (0, \infty)$ be such that $\ln c = x$. Since the function $\ln : (0, \infty) \to \mathbb{R}$ is differentiable at c and $(\ln)'c = 1/c \neq 0$, Proposition 4.11 shows that the inverse function $\exp : \mathbb{R} \to (0, \infty)$ is differentiable at $x = \ln c$ and

$$(\exp)'(x) = \exp'(\ln c) = \frac{1}{(\ln)'c} = c = \exp x.$$

(ii) Since the derivative of exp is positive on $\mathbb{R}$, it follows from part (iii) of Proposition 4.27 that exp is strictly increasing on $\mathbb{R}$.
    Further, since

$$(\exp)''x = (\exp)'x = \exp x > 0 \quad \text{for all } x \in \mathbb{R},$$

it follows from part (iii) of Proposition 4.32 that exp is strictly convex on $\mathbb{R}$.
    (iii) Since the range of the function exp is the domain of the function ln, namely, the interval $(0, \infty)$, we see that exp is not bounded above on $\mathbb{R}$, whereas $\inf\{\exp x : x \in \mathbb{R}\} = 0$. Also, since exp is (strictly) increasing on $\mathbb{R}$, it follows from Proposition 3.35 that $\exp x \to \infty$ as $x \to \infty$, whereas $\exp x \to 0$ as $x \to -\infty$. $\qquad\square$

    The geometric properties of the exponential function proved in Proposition 7.2 can be used to draw its graph as in Figure 7.1.
    It follows from part (i) of Proposition 7.2 that the function exp is infinitely differentiable on $\mathbb{R}$ and for $k = 1, 2, \ldots$, we have

$$(\exp)^{(k)}x = \exp x \quad \text{for all } x \in \mathbb{R}.$$

Hence the nth Taylor polynomial for exp about 0 is given by

$$P_n(x) = \exp 0 + \sum_{k=1}^{n} \frac{(\exp)^{(k)}(0)}{k!}(x - 0)^k = 1 + \sum_{k=1}^{n} \frac{x^k}{k!}, \quad x \in \mathbb{R}.$$

In particular, the linear and the quadratic approximations of exp around 0 are given by

$$L(x) = P_1(x) = 1 + x \quad \text{and} \quad Q(x) = P_2(x) = 1 + x + \frac{x^2}{2}, \quad x \in \mathbb{R}.$$

The logarithmic and the exponential functions have interesting behavior with respect to the multiplication and addition of real numbers. This is made precise in the following result.

Proposition 7.3. (i) *For any positive real numbers x_1 and x_2, we have*

$$\ln x_1 x_2 = \ln x_1 + \ln x_2.$$

(ii) *For any real numbers x_1 and x_2, we have*

$$\exp(x_1 + x_2) = (\exp x_1)(\exp x_2).$$

Proof. (i) Fix $x_2 \in (0, \infty)$ and consider the function $f : (0, \infty) \to \mathbb{R}$ given by $f(x) = \ln x x_2 - \ln x$. Then

$$f'(x) = \frac{1}{x x_2} x_2 - \frac{1}{x} = 0 \quad \text{for all } x \in (0, \infty).$$

Hence $f(x) = f(1) = \ln x_2 - \ln 1 = \ln x_2$ for all $x \in (0, \infty)$. In particular, $f(x_1) = \ln x_1 x_2 - \ln x_1 = \ln x_2$. This proves (i).

(ii) Let x_1 and x_2 be real numbers. Define $y_1 := \exp x_1$ and $y_2 = \exp x_2$. Then y_1 and y_2 are positive real numbers and by (i) above, we see that $\ln y_1 y_2 = \ln y_1 + \ln y_2 = x_1 + x_2$. Consequently, $\exp(x_1 + x_2) = y_1 y_2 = (\exp x_1)(\exp x_2)$.   □

In the examples below, we consider some important limits involving the functions ln and exp.

Examples 7.4. 1.
(i) By the definition of derivative and the earlier observations that $\ln 1 = 0$, $(\ln)'1 = 1$, $\exp 0 = 1$, and $(\exp)'0 = 1$, we obtain

$$\lim_{h \to 0} \frac{\ln(1 + h)}{h} = 1 \quad \text{and} \quad \lim_{h \to 0} \frac{\exp h - 1}{h} = 1.$$

(ii) Since $(\ln)'x = 1/x$ for $x \in (0, \infty)$ and $\ln x \to \infty$ as $x \to \infty$, while $(\exp)'x = x$ for $x \in \mathbb{R}$ and $\exp x \to \infty$ as $x \to \infty$, L'Hôpital's Rule for $\frac{\infty}{\infty}$ indeterminate forms (Proposition 4.40) shows that

$$\lim_{x \to \infty} \frac{\ln x}{x} = 0 \quad \text{and} \quad \lim_{x \to \infty} \frac{x}{\exp x} = 0.$$
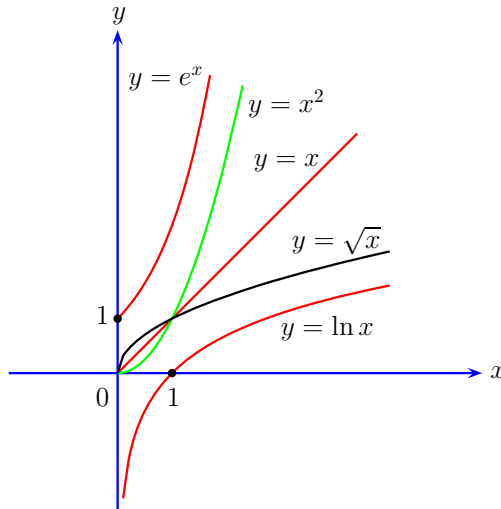
Fig. 7.2. Illustration of the growth rate: Graphs of $f(x) = \ln x$, $f(x) = \sqrt{x}$, $f(x) = x$, $f(x) = x^2$, and $f(x) = e^x$ (for $x > 0$)

In a similar manner, it can be easily seen that for any $k \in \mathbb{N}$, we have

$$\lim_{x \to \infty} \frac{\ln x}{x^{1/k}} = 0 \quad \text{and} \quad \lim_{x \to \infty} \frac{x^k}{\exp x} = 0.$$

These limits show that the growth rate of $\ln x$ is less than that of any root of x, while the growth rate of $\exp x$ is more than that of any positive integral power of x as x tends to ∞ (Remark 3.31). This can be illustrated by the graphs of the curves $y = \ln x$, $y = \sqrt{x}$, $y = x$, $y = x^2$ and $y = e^x$ (for $x > 0$) drawn in Figure 7.2. $\qquad\qquad\diamond$

Real Powers of Positive Numbers

The logarithmic and exponential functions enable us to define the bth power of a, where a is any positive real number and b is any real number. Recall that if r is any rational number, then we have defined in Chapter 1 the rth power a^r of any $a \in (0, \infty)$. We observe that

$$\ln a^r = r \ln a \quad \text{for all } a \in (0, \infty) \text{ and } r \in \mathbb{Q}.$$

To see this, consider the function $f : (0, \infty) \to \mathbb{R}$ given by

$$f(x) = \ln x^r - r \ln x.$$

Then by part (i) of Proposition 7.1 and the Chain Rule (Proposition 4.9), we have

$$f'(x) = \frac{1}{x^r}rx^{r-1} - r\frac{1}{x} = 0 \quad \text{for all } x \in (0, \infty),$$

and so $f(x) = f(1) = \ln 1^r - r\ln 1 = 0 - 0 = 0$ for all $x \in (0, \infty)$. In particular, the equation $f(a) = 0$ gives $\ln a^r = r\ln a$. Thus we have

$$a^r = \exp(r\ln a) \quad \text{for all } a \in (0, \infty) \text{ and } r \in \mathbb{Q}.$$

Since the number $\exp(r\ln a)$ is well defined for any real (and not just rational) number r, we are naturally led to the following definition. Let a be a positive number and b be a real number. The bth **power** of a is defined by

$$a^b := \exp(b\ln a).$$

Here a is called the **base** and b is called the **exponent**. If b is a rational number, this definition of a^b coincides with our earlier definition as we have just seen. Clearly, the equality $a^b := \exp(b\ln a)$ is equivalent to the equality $\ln a^b = b\ln a$ for $a > 0$ and $b \in \mathbb{R}$. We note that $a^b > 0$ for all $a \in (0, \infty)$ and $b \in \mathbb{R}$.

Let us consider the special case $a = e$, that is, when the base is the unique positive real number satisfying $\ln e = 1$. Then we have

$$e^x = \exp(x\ln e) = \exp x \quad \text{for all } x \in \mathbb{R}.$$

We have thus found a short notation for $\exp x$, namely e^x, where $x \in \mathbb{R}$. From now on, we may employ this notation.

The following result gives an alternative way of determining e^x for $x \in \mathbb{R}$.

Proposition 7.5. *For any $x \in \mathbb{R}$, we have*

$$\lim_{h \to 0}(1 + xh)^{1/h} = e^x.$$

In particular, we have

$$e = \lim_{h \to 0}(1 + h)^{1/h}.$$

Proof. The first assertion is obvious if $x = 0$. Suppose $x \in \mathbb{R}$ and $x \neq 0$. Now,

$$\lim_{h \to 0}\frac{\ln(1 + xh)}{xh} = \lim_{h \to 0}\frac{\ln(1 + h)}{h} = 1,$$

as we have noted before. Consequently,

$$\lim_{h \to 0}\ln(1 + xh)^{1/h} = \lim_{h \to 0}\frac{\ln(1 + xh)}{h} = x\left[\lim_{h \to 0}\frac{\ln(1 + xh)}{xh}\right] = x.$$

Now since the exponential function is continuous at x, we obtain

$$\lim_{h \to 0}(1+xh)^{1/h} = \lim_{h \to 0}\exp\left(\ln(1 + xh)^{1/h}\right) = \exp\left(\lim_{h \to 0}\ln(1 + xh)^{1/h}\right) = \exp x.$$

This proves the first assertion. The particular case is obtained by considering $x = 1$. □

In Examples 2.10 (i) and (ii), we have stated that the sequences (a_n) and (b_n) defined by

$$a_n := \sum_{k=0}^{n} \frac{1}{k!} \quad \text{and} \quad b_n := \left(1 + \frac{1}{n}\right)^n \quad \text{for } n \in \mathbb{N}$$

are convergent and have the same limit. Now we shall show that this common limit is equal to e.

Corollary 7.6. *For any $x \in \mathbb{R}$, we have*

$$\lim_{n \to \infty} \left(1 + \frac{x}{n}\right)^n = e^x.$$

In particular, we have

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

Proof. Consider the sequence (h_n) defined by $h_n = 1/n$ for $n \in \mathbb{N}$. Then $h_n \to 0$. Hence the desired results follow from Proposition 7.5. $\square$

Returning to the general power a^b with base $a \in (0, \infty)$ and exponent $b \in \mathbb{R}$, let us consider the following two functions. For a fixed $a \in (0, \infty)$, the **power function $f_a : \mathbb{R} \to \mathbb{R}$ with base** a, and for a fixed $b \in \mathbb{R}$, the **power function $g_b : (0, \infty) \to \mathbb{R}$ with exponent** b, are defined by

$$f_a(x) := a^x \quad \text{and} \quad g_b(x) := x^b.$$

Note that $f_a(x) > 0$ for all $x \in \mathbb{R}$ and $g_b(x) > 0$ for all $x \in (0, \infty)$. We study the basic properties of these functions in Propositions 7.7 and 7.9.

Proposition 7.7 (Properties of Power Function with Fixed Base).
Let a be a positive real number and $f_a : \mathbb{R} \to (0, \infty)$ be the power function with base a given by $f_a(x) := a^x$. Then we have the following:

(i) *f_a is a differentiable function on $\mathbb{R}$ and*

$$f_a'(x) = (\ln a)a^x = (\ln a)f_a(x) \quad \text{for every } x \in \mathbb{R}.$$

(ii) *If $a > 1$, then f_a is strictly increasing as well as strictly convex on $\mathbb{R}$. If $a < 1$, then f_a is strictly decreasing as well as strictly convex on $\mathbb{R}$. If $a = 1$, then $f_a(x) = 1$ for all $x \in \mathbb{R}$.*

(iii) *If $a \neq 1$, then f_a is not bounded above on $\mathbb{R}$. More precisely, if $a > 1$, then $f_a(x) \to \infty$ as $x \to \infty$, whereas $f_a(x) \to 0$ as $x \to -\infty$, and if $a < 1$, then $f_a(x) \to \infty$ as $x \to -\infty$, whereas $f_a(x) \to 0$ as $x \to \infty$.*

(iv) *If $a \neq 1$, then the function $f_a : \mathbb{R} \to (0, \infty)$ is bijective and its inverse $f_a^{-1} : (0, \infty) \to \mathbb{R}$ is given by*

$$f_a^{-1}(x) = \frac{\ln x}{\ln a} \quad \text{for } x \in (0, \infty).$$

(v) *For any x_1 and x_2 in $\mathbb{R}$, we have*

$$f_a(x_1 + x_2) = f_a(x_1)f_a(x_2), \quad \text{that is,} \quad a^{x_1+x_2} = a^{x_1}a^{x_2}.$$

(vi) *For any x_1 and x_2 in $\mathbb{R}$, we have*

$$[f_a(x_1)]^{x_2} = f_a(x_1x_2) = [f_a(x_2)]^{x_1}, \quad \text{that is,} \quad (a^{x_1})^{x_2} = a^{x_1x_2} = (a^{x_2})^{x_1}.$$

Proof. (i) Since $f_a(x) = \exp(x \ln a)$ for $x \in \mathbb{R}$ and $(\exp)' = \exp$, the Chain Rule (Proposition 4.9) shows that

$$f_a'(x) = (\ln a)\exp(x \ln a) = (\ln a)a^x = (\ln a)f_a(x).$$

(ii) Let $a > 1$. Then $\ln a > 0$ and hence the derivative f_a' of f_a is positive on $\mathbb{R}$. This shows that f_a is strictly increasing on $\mathbb{R}$. Further, since

$$f_a''(x) = (\ln a)f_a'(x) = (\ln a)^2 f_a(x) > 0 \quad \text{for all } x \in \mathbb{R},$$

it follows that f_a is strictly convex on $\mathbb{R}$. Similar arguments yield the desired results for f_a if $a < 1$. If $a = 1$, then $f_1(x) = 1^x = e^{x \ln 1} = e^0 = 1$ for all $x \in \mathbb{R}$.

(iii) If $a > 1$, then $\ln a > 0$ and so $f_a(x) = e^{(\ln a)x} \to \infty$ as $x \to \infty$, whereas $f_a(x) \to 0$ as $x \to -\infty$. Similarly, if $a < 1$, then $\ln a < 0$ and so $f_a(x) = e^{(\ln a)x} \to \infty$ as $x \to -\infty$, whereas $f_a(x) \to 0$ as $x \to \infty$. This shows that if $a \neq 1$, then f_a is not bounded above on $\mathbb{R}$.

(iv) Let $a \neq 1$. The bijectivity of $f_a : \mathbb{R} \to (0, \infty)$ follows from the bijectivity of the function $\exp : \mathbb{R} \to (0, \infty)$ and the fact that $\ln a \neq 0$. For $x \in (0, \infty)$, we have

$$f_a(\ln x / \ln a) = a^{\ln x / \ln a} = \exp([\ln x / \ln a] \ln a) = \exp(\ln x) = x.$$

Hence $f_a^{-1}(x) = \ln x / \ln a$ for $x \in (0, \infty)$.

(v) For any x_1 and x_2 in $\mathbb{R}$, we have

$$\exp((x_1 + x_2) \ln a) = \exp(x_1 \ln a + x_2 \ln a) = \exp(x_1 \ln a)\exp(x_2 \ln a),$$

and thus, $f_a(x_1 + x_2) = f_a(x_1)f_a(x_2)$, as desired.

(vi) For any x_1 and x_2 in $\mathbb{R}$, we have

$$[f_a(x_1)]^{x_2} = (a^{x_1})^{x_2} = \exp(x_2 \ln a^{x_1}) = \exp(x_2 x_1 \ln a) = a^{x_2 x_1} = f_a(x_2 x_1).$$

Interchanging x_1 and x_2, we have

$$[f_a(x_2)]^{x_1} = f_a(x_1 x_2).$$

Since $x_2 x_1 = x_1 x_2$, we obtain the desired result. $\square$
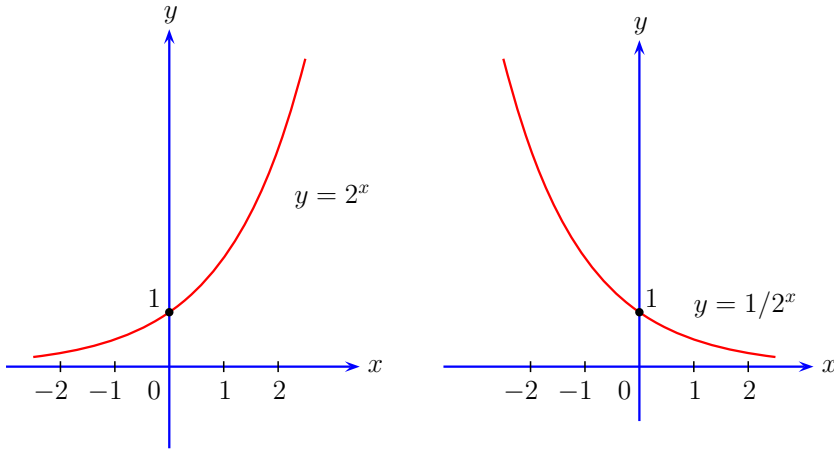
Fig. 7.3. Graphs of the power function f_a with base (i) $a = 2$, and (ii) $a = 1/2$

The geometric properties of the function f_a proved in Proposition 7.7 can be used to draw its graph as in Figure 7.3, when $a = 2$ and $a = \frac{1}{2}$.

Remark 7.8. Let $a > 0$ and $f_a(x) := a^x$ for $x \in \mathbb{R}$. If $a \neq 1$, then the inverse $f_a^{-1} : (0, \infty) \to \mathbb{R}$ of the function $f_a : \mathbb{R} \to (0, \infty)$ exists and it is denoted by $\log_a$; this is known as the **logarithmic function with base** a. Thus we have

$$\log_a x = \frac{\ln x}{\ln a}, \quad x \in (0, \infty).$$

Clearly, if $a = e$, then $\log_a = \log_e = \ln$. For this reason, the number e is often referred to as the **base of the natural logarithm**. Another commonly used base is $a = 10$. The function $\log_{10}$ is often written simply as log. Thus

$$\log x = \frac{\ln x}{\ln 10}, \quad x \in (0, \infty).$$

The first few digits in the decimal expansion of $\ln 10$ are given by

$$\ln 10 = 2.30258509\ldots.$$

This enables us to compute the value of $\log x$ if we know $\ln x$, and vice versa for $x \in (0, \infty)$. $\diamond$

Proposition 7.9 (Properties of the Power Function with Fixed Exponent). Let b be a real number and $g_b : (0, \infty) \to (0, \infty)$ be the power function with exponent b given by $g_b(x) := x^b$. Then we have the following:

(i) g_b is a differentiable function on $(0, \infty)$ and

$$g_b'(x) = bx^{b-1} = bg_{b-1}(x) \quad \text{for every } x \in (0, \infty).$$

(ii) *If $b > 0$, then g_b is strictly increasing, and if $b < 0$, then g_b is strictly decreasing on $(0, \infty)$. Further, if $b > 1$ or $b < 0$, then g_b is strictly convex, and if $0 < b < 1$, then g_b is strictly concave on $(0, \infty)$. If $b = 0$, then $g_b(x) = 1$, and if $b = 1$, then $g_b(x) = x$ for all $x \in (0, \infty)$.*

(iii) *If $b \neq 0$, then g_b is not bounded above on $(0, \infty)$. More precisely, if $b > 0$, then $g_b(x) \to \infty$ as $x \to \infty$, whereas $g_b(x) \to 0$ as $x \to 0$, and if $b < 0$, then $g_b(x) \to \infty$ as $x \to 0$, whereas $g_b(x) \to 0$ as $x \to \infty$.*

(iv) *If $b \neq 0$, then the function $g_b : (0, \infty) \to (0, \infty)$ is bijective and the function $g_{1/b} : (0, \infty) \to (0, \infty)$ is the inverse of g_b.*

(v) *For any x_1 and x_2 in $(0, \infty)$, we have*

$$g_b(x_1 x_2) = g_b(x_1) g_b(x_2), \quad \text{that is,} \quad (x_1 x_2)^b = x_1^b x_2^b.$$

Proof. (i) Since $g_b(x) = \exp(b \ln x)$ for $x \in (0, \infty)$, the Chain Rule (Proposition 4.9) shows that

$$g_b'(x) = \frac{b}{x} \exp(b \ln x) = b x^{b-1} = b g_{b-1}(x).$$

(ii) We note that for each $b \in \mathbb{R}$, the function g_b is positive on $(0, \infty)$. Hence by (i) above, the derivative g_b' of g_b is positive on $(0, \infty)$ or negative on $(0, \infty)$ according as $b > 0$ or $b < 0$. This shows that g_b is strictly increasing on $(0, \infty)$ or strictly decreasing on $(0, \infty)$ according as $b > 0$ or $b < 0$. Further,

$$g_b''(x) = b g_{b-1}'(x) = b(b-1) g_b(x) \quad \text{for all } x \in (0, \infty).$$

Hence it follows that g_b is strictly convex on $(0, \infty)$ or strictly concave on $(0, \infty)$ according as $b \in (-\infty, 0) \cup (1, \infty)$ or $b \in (0, 1)$. If $b = 0$, then $g_0(x) = e^{0 \ln x} = e^0 = 1$, and if $b = 1$, then $g_1(x) = e^{\ln x} = x$ for all $x \in (0, \infty)$.

(iii) If $b > 0$, then by the properties of the function $\ln$, $g_b(x) = e^{b \ln x} \to \infty$ as $x \to \infty$, whereas $g_b(x) \to 0$ as $x \to 0$. Similarly, we obtain the desired results for g_b if $b < 0$. This shows that if $b \neq 0$, then g_b is not bounded above on $(0, \infty)$.

(iv) Let $b \neq 0$. The bijectivity of $g_b : (0, \infty) \to (0, \infty)$ follows from the bijectivity of the functions $\ln : (0, \infty) \to \mathbb{R}$ and $\exp : \mathbb{R} \to (0, \infty)$. Also,

$$g_{1/b}(g_b(x)) = g_{1/b}(x^b) = \exp[(\ln x^b)/b] = \exp(\ln x) = x \quad \text{for all } x \in (0, \infty).$$

Similarly, $g_b(g_{1/b}(x)) = x$ for all $x \in (0, \infty)$. Thus $g_{1/b}$ is the inverse of g_b.

(v) For any $x_1, x_2 \in (0, \infty)$, we have

$$\exp(b \ln x_1 x_2) = \exp b(\ln x_1 + \ln x_2) = \exp(b \ln x_1) \exp(b \ln x_2),$$

and thus, $g_b(x_1 x_2) = g_b(x_1) g_b(x_2)$, as desired. $\square$

The geometric properties of the function g_b proved in Proposition 7.9 can be used to draw its graph when $b = \sqrt{2}$ and $b = 1/\sqrt{2}$ as in Figure 7.4.

The following result is a generalization of Example 6.24 (i).
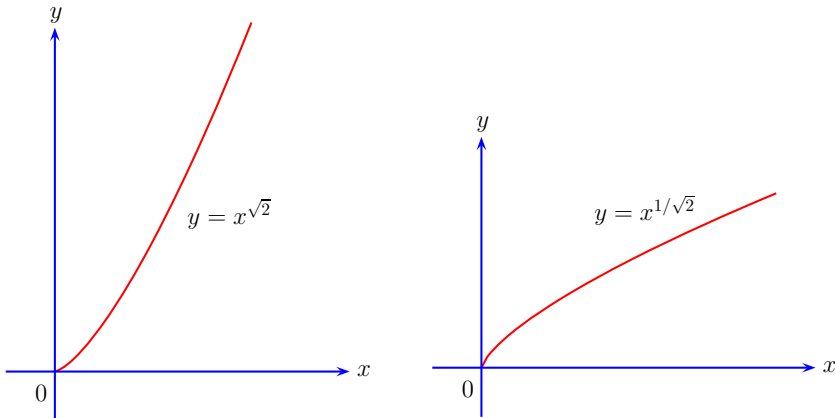
Fig. 7.4. Graphs of the power function g_b with exponent: (i) $b = \sqrt{2}$, and (ii) $b = 1/\sqrt{2}$

Corollary 7.10. *Let r be a real number such that $r \neq -1$. Let a and b be real numbers such that $0 < a < b$, and $f : [a, b] \to \mathbb{R}$ be given by $f(x) := x^r$. Then f is integrable and*

$$\int_a^b f(x)dx = \frac{b^{r+1} - a^{r+1}}{r+1}.$$

Proof. Since f is continuous on $[a, b]$, it is integrable. Define $F : [a, b] \to \mathbb{R}$ by $F(x) := x^{r+1}/(r+1)$. By part (i) of Proposition 7.9, F is differentiable and $F'(x) = x^r$ for all $x \in [a, b]$. Hence part (i) of the FTC (Proposition 6.21) shows that

$$\int_a^b f(x)dx = F(b) - F(a) = \frac{b^{r+1} - a^{r+1}}{r+1},$$

as desired. □

Remark 7.11. Parts (v) and (vi) of Proposition 7.7 and and part (v) of Proposition 7.9 are known as the '**laws of exponents**' or the '**laws of indices**'. We list them below:

(i) $a^{r+s} = a^r a^r$ for all $a \in (0, \infty)$ and r, $s \in \mathbb{R}$,
(ii) $(a^r)^s = a^{rs}$ for all $a \in (0, \infty)$ and r, $s \in \mathbb{R}$,
(iii) $(a_1 a_2)^r = (a_1)^r (a_2)^r$ for all a_1, $a_2 \in (0, \infty)$ and $r \in \mathbb{R}$.

For integral powers, we have stated these earlier in Section 1.1. ◇

Remark 7.12. Let $D \subseteq \mathbb{R}$ and $f, g : D \to \mathbb{R}$ be functions such that $f(x) > 0$ for all $x \in D$. Consider the function $h : D \to \mathbb{R}$ defined by

$$h(x) := f(x)^{g(x)}.$$

Properties of the function h can be studied by considering the function $k :$ $D \to \mathbb{R}$ defined by

$$k(x) := \ln h(x) = g(x) \ln f(x).$$

For example, let $D := (0, \infty)$, and $f, g : D \to \mathbb{R}$ be defined by $f(x) := x$ and $g(x) := 1/x$. As we have seen in Example 7.4 (ii), $k(x) := (\ln x)/x \to 0$ as $x \to \infty$, and hence

$$\lim_{x \to \infty} x^{1/x} = 1.$$

We note that the indeterminate forms 0^0, ∞^0, and 1^∞ involving the functions f and g can be reduced to the indeterminate form $0 \cdot \infty$ (treated in Remark 4.44) involving the functions $\ln f$ and g, since

 (i) $f(x) \to 0$ and $g(x) \to 0 \Longrightarrow \ln f(x) \to -\infty$ and $g(x) \to 0$,
 (ii) $f(x) \to \infty$ and $g(x) \to 0 \Longrightarrow \ln f(x) \to \infty$ and $g(x) \to 0$,
(iii) $f(x) \to 1$ and $g(x) \to \infty \Longrightarrow \ln f(x) \to 0$ and $g(x) \to \infty$.

Revision Exercise 22 given at the end of this chapter can be worked out using these considerations.   ◇

7.2 Trigonometric Functions

Using the logarithmic function defined in Section 7.1, the reciprocal of any linear polynomial can be integrated. Indeed, up to a constant multiple, such a function is given by $1/(x - \alpha)$, where $\alpha \in \mathbb{R}$, and we have

$$\frac{d}{dx} \big(\ln(x - \alpha) \big) = \frac{1}{x - \alpha} \quad \text{for } x \in \mathbb{R}, \ x > \alpha.$$

The next question that naturally arises is whether we can integrate the reciprocal of a quadratic polynomial, say $x^2 + ax + b$, where $a, b \in \mathbb{R}$. If this quadratic happens to be the square of a linear polynomial, say $(x - \alpha)^2$, then the answer is easy because

$$\frac{d}{dx} \left(\frac{-1}{x - \alpha} \right) = \frac{1}{(x - \alpha)^2} \quad \text{for } x \in \mathbb{R}, \ x \neq \alpha.$$

Further, if the quadratic factors into distinct linear factors, that is, if

$$x^2 + ax + b = (x - \alpha)(x - \beta) \quad \text{for some } \alpha, \beta \in \mathbb{R}, \ \alpha > \beta,$$

then we have

$$\frac{1}{x^2 + ax + b} = \frac{1}{\alpha - \beta} \left[\frac{1}{x - \alpha} - \frac{1}{x - \beta} \right] = \frac{1}{\alpha - \beta} \frac{d}{dx} \left(\ln \frac{x - \alpha}{x - \beta} \right) \quad \text{for } x > \alpha.$$

If, however, the quadratic $x^2 + ax + b$ has no real root, then we face a difficulty. The simplest example of this kind is the quadratic $x^2 + 1$. To be able to

integrate the reciprocal of this polynomial, a new function has to be introduced and we shall do it in this section. In the 'Notes and Comments' at the end of this chapter, we shall indicate how every rational function can be integrated using only this new function, the logarithmic function, and, of course, the rational functions themselves.

The function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(t) := 1/(1+t^2)$ is continuous. Hence it is Riemann integrable on every closed and bounded interval. For $x \in \mathbb{R}$, we define the **arctangent** of x by

$$\arctan x := \int_0^x \frac{1}{1+t^2} dt.$$

The function $\arctan : \mathbb{R} \to \mathbb{R}$ is known as the **arctangent function**. The reason for this terminology will be explained in Section 8.4 when we discuss the 'length' of an arc of a circle. We write $\arctan x$, rather than $\arctan(x)$, for the value of the arctangent function at $x \in \mathbb{R}$.

Clearly, $\arctan 0 = 0$, and since $1/(1+t^2) \geq 0$ for all $t \in \mathbb{R}$, we have $\arctan x \geq 0$ if $x > 0$, while $\arctan x \leq 0$ if $x < 0$.

Proposition 7.13 (Properties of the Arctangent Function and Definition of π).

(i) arctan *is a differentiable function on $\mathbb{R}$ and*

$$(\arctan)'x = \frac{1}{1+x^2} \quad \text{for every } x \in \mathbb{R}.$$

(ii) arctan *is strictly increasing on $\mathbb{R}$. Also, it is strictly convex on $(-\infty, 0)$ and strictly concave on $(0, \infty)$.*

(iii) arctan *is an odd function. Also, it is bounded on $\mathbb{R}$. We define*

$$\pi := 2 \sup\{\arctan x : x \in (0, \infty)\}.$$

Then $\arctan x \to \pi/2$ as $x \to \infty$, whereas $\arctan x \to -\pi/2$ as $x \to -\infty$.

(iv) *For every $y \in (-\pi/2, \pi/2)$, there is a unique $x \in \mathbb{R}$ such that $\arctan x = y$. In other words, $\arctan : \mathbb{R} \to (-\pi/2, \pi/2)$ is a bijective function.*

Proof. (i) Since the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(t) = 1/(1+t^2)$ is continuous, the Fundamental Theorem of Calculus (Proposition 6.21) shows that the function arctan is differentiable at every $x \in \mathbb{R}$ and $(\arctan)'x = f(x) = 1/(1+x^2)$ for every $x \in \mathbb{R}$.

(ii) Since the derivative of arctan is positive on $\mathbb{R}$, it follows from part (iii) of Proposition 4.27 that arctan is strictly increasing on $\mathbb{R}$. Further, since

$$(\arctan)''x = -\frac{2x}{1+x^2},$$

which is positive for all $x \in (-\infty, 0)$ and negative for all $x \in (0, \infty)$, it follows from parts (iii) and (iv) of Proposition 4.32 that arctan is strictly convex on $(-\infty, 0)$ and strictly concave on $(0, \infty)$.

(iii) For $x \in \mathbb{R}$, we have

$$\arctan(-x) = \int_0^{-x} \frac{1}{1+t^2} dt = - \int_0^x \frac{1}{1+s^2} ds = -\arctan x,$$

by employing the substitution $s = -t$. Hence arctan is an odd function.

Next, we show that arctan is a bounded function. By the Mean Value Theorem (Proposition 4.18), we have

$$\arctan 1 - \arctan 0 = (\arctan)'c = \frac{1}{1+c^2} \qquad \text{for some } c \in (0,1).$$

Since $1/(1+c^2) < 1$ for every $c \in \mathbb{R}$ and $\arctan 0 = 0$, we have

$$\arctan 1 < 1.$$

Now let $x \in (1, \infty)$. By Cauchy's Mean Value Theorem (Proposition 4.36) for the function arctan restricted to the interval $[1, x]$ and the function $g : [1, x] \to \mathbb{R}$ given by $g(t) = -1/t$, we obtain

$$\frac{\arctan x - \arctan 1}{g(x) - g(1)} = \frac{(\arctan)'c}{g'(c)} = \frac{c^2}{1+c^2} \qquad \text{for some } c \in (1, x).$$

This shows that

$$\arctan x = \arctan 1 + \left(-\frac{1}{x} + 1 \right) \frac{c^2}{1+c^2} < 1 + 1 - \frac{1}{x} = 2 - \frac{1}{x}.$$

Hence $0 < \arctan x < 2$ for every $x \in (0, \infty)$. Since arctan is an odd function, we see that $0 > \arctan x > -2$ for every $x \in (-\infty, 0)$. Thus arctan is bounded above and bounded below on $\mathbb{R}$. Consequently, there is a well-defined real number π such that $\pi = 2 \sup\{\arctan x : x \in (0, \infty)\}$, that is, $\pi/2 = \sup\{\arctan x : x \in (0, \infty)\}$. Now since arctan is (strictly) increasing, it follows from Proposition 3.35 that $\arctan x \to \pi/2$ as $x \to \infty$. Again, since arctan is an odd function, $\arctan x \to -\pi/2$ as $x \to -\infty$.

(iv) The function arctan is one-one since it is strictly increasing. Let $y \in (-\pi/2, \pi/2)$. Since $\arctan x \to -\infty$ as $x \to -\pi/2$, there is some $x_0 \in \mathbb{R}$ such that $\arctan x_0 < y$ and since $\arctan x \to \infty$ as $x \to \pi/2$, there is some $x_1 \in \mathbb{R}$ such that $y < \arctan x_1$. But by part (i) above, the function arctan is continuous on the interval $[x_0, x_1]$. So, the IVP (Proposition 3.13) shows that there is some $x \in (x_0, x_1)$ such that $\arctan x = y$. Thus the function $\arctan : \mathbb{R} \to (-\pi/2, \pi/2)$ is bijective. $\square$

The geometric properties of arctan obtained in the above proposition can be used to draw its graph as follows.

Let us estimate the number $\pi = 2 \sup\{\arctan x : x \in (0, \infty)\}$. We have seen in the proof of part (iii) of the above proposition that $\arctan x < 2$ for every $x \in (0, \infty)$. Hence $\pi \leq 4$. Further, this proof shows that
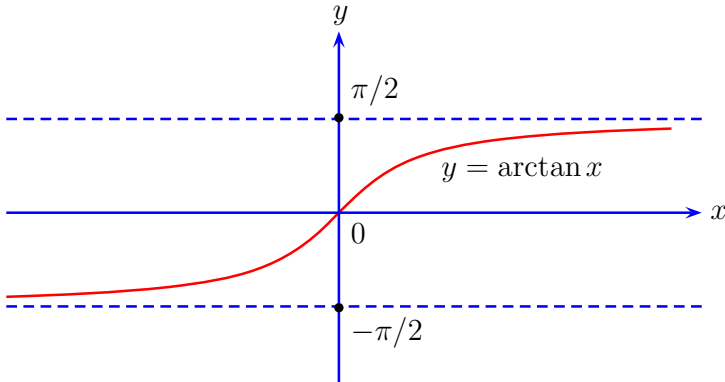
Fig. 7.5. Graph of the arctangent function

$$\arctan 1 = \frac{1}{1 + c^2} \quad \text{for some } c \in (0, 1)$$

and so $\arctan 1 > \frac{1}{2}$. Also, we have shown that for $x \in (1, \infty)$,

$$\arctan x = \arctan 1 + \left(1 - \frac{1}{x}\right) \frac{c^2}{1 + c^2} \quad \text{for some } c \in (1, x).$$

Since $c^2/(1 + c^2) > \frac{1}{2}$ for any $c \in (1, \infty)$, we see that

$$\arctan x > \frac{1}{2} + \left(1 - \frac{1}{x}\right) \frac{1}{2} = 1 - \frac{1}{2x}, \quad x \in (1, \infty).$$

Hence $2 \arctan x > 2 - (1/x)$ for each $x \in (0, \infty)$, showing that $\pi \geq 2$. Thus

$$2 \leq \pi \leq 4.$$

The number π is traditionally 'defined' as the area of a circular disk of radius 1 or as half the perimeter of a circle of radius 1. But these definitions presuppose the notions of 'area' or 'length'. In Chapter 8, we shall reconcile our definition of π with the traditional definitions after giving precise definitions of 'area' and 'length'.

Better lower and upper bounds for the number π can be obtained. (See, for example, Exercise 16.) The first few digits in the decimal expansion of π are given by

$$\pi = 3.14159265\ldots.$$

In fact, π is an irrational number (Exercise 57), and furthermore, π is transcendental, that is, it is not a root of any nonzero polynomial with rational coefficients. (See [7] for a proof.)

Let us now turn to the inverse of the arctangent function. The inverse of the bijective function $\arctan : \mathbb{R} \to (-\pi/2, \pi/2)$ is known as the **tangent**

function and is denoted by $\tan : (-\pi/2, \pi/2) \to \mathbb{R}$. We write $\tan x$, rather than $\tan(x)$, for the value of the tangent function at $x \in (-\pi/2, \pi/2)$. The function tan is characterized by the following:

$$x \in (-\pi/2, \pi/2) \text{ and } \tan x = y \Longleftrightarrow y \in \mathbb{R} \text{ and } \arctan y = x.$$

Note that $\tan x > 0$ for $x \in (0, \pi/2)$, while $\tan x < 0$ for $x \in (-\pi/2, 0)$. Moreover, since $\arctan 0 = 0$, we have $\tan 0 = 0$.
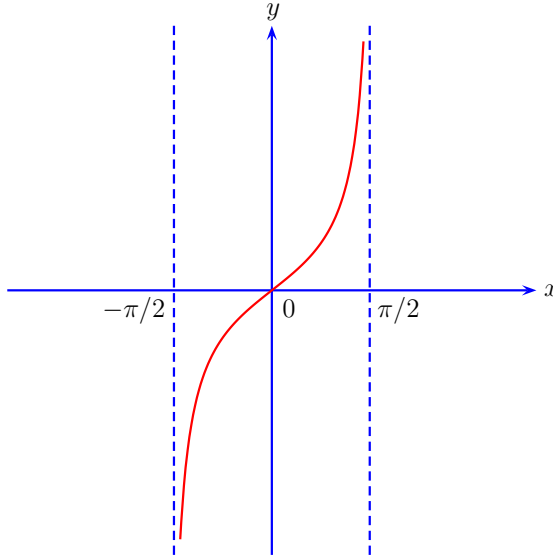
Fig. 7.6. Graph of the tangent function on $(-\pi/2, \pi/2)$

Proposition 7.14 (Properties of the Tangent Function).

(i) tan *is a differentiable function on* $(-\pi/2, \pi/2)$ *and*

$$(\tan)' x = 1 + \tan^2 x \quad \textit{for every } x \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right).$$

(ii) tan *is strictly increasing on* $(-\pi/2, \pi/2)$. *Also, it is strictly concave on* $(-\pi/2, 0)$ *and strictly convex on* $(0, \pi/2)$.

(iii) tan *is an odd function on* $(-\pi/2, \pi/2)$. *Also,* $\tan x \to \infty$ *as* $x \to \pi/2$, *whereas* $\tan x \to -\infty$ *as* $x \to -\pi/2$.

Proof. (i) Let $x \in (-\pi/2, \pi/2)$ and $c \in \mathbb{R}$ be such that $\arctan c = x$. Since the function $\arctan : \mathbb{R} \to (-\pi/2, \pi/2)$ is differentiable at c and $(\arctan)' c = 1/(1 + c^2) \neq 0$, Proposition 4.11 shows that the inverse function $\tan : (-\pi/2, \pi/2) \to \mathbb{R}$ is differentiable at $x = \arctan c$ and

$$(\tan)'x = \frac{1}{(\arctan)'c} = 1 + c^2 = 1 + \tan^2 x.$$

(ii) Since the derivative of tan is positive on $(-\pi/2, \pi/2)$, it follows from part (iii) of Proposition 4.27 that tan is strictly increasing on $(-\pi/2, \pi/2)$. Further, since

$$(\tan)''x = \frac{d}{dx}\left(1 + \tan^2 x\right) = 2\tan x \left(1 + \tan^2 x\right) \quad \text{for } x \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right),$$

and since $\tan x > 0$ for $x \in (0, \pi/2)$, while $\tan x < 0$ for $x \in (-\pi/2, 0)$, it follows from parts (iii) and (iv) of Proposition 4.32 that tan is strictly concave on $(-\pi/2, 0)$ and it is strictly convex on $(0, \pi/2)$.

(iii) Let $x \in (-\pi/2, \pi/2)$ and $y = \tan x$. Then

$$\tan(-x) = \tan(-\arctan y) = \tan(\arctan(-y)) = -y = -\tan x.$$

Thus tan is an odd function on $(-\pi/2, \pi/2)$.

Since the range of the function tan is the domain of the function arctan, namely $\mathbb{R}$, we see that tan is neither bounded above nor bounded below on $(-\pi/2, \pi/2)$. Also, since tan is (strictly) increasing on $(-\pi/2, \pi/2)$, it follows from Proposition 3.35 that $\tan x \to \infty$ as $x \to \pi/2$ and $\tan x \to -\infty$ as $x \to -\pi/2$. $\qquad\square$

The geometric properties of tan obtained in the above proposition can be used to draw its graph as in Figure 7.6.

Sine and Cosine Functions

To begin with, we define the **sine function** and the **cosine function** on the interval $(-\pi/2, \pi/2)$ by

$$\sin x := \frac{\tan x}{\sqrt{1 + \tan^2 x}} \quad \text{and} \quad \cos x := \frac{1}{\sqrt{1 + \tan^2 x}} \quad \text{for } x \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right).$$

It is clear from the definition that

$$\tan x = \frac{\sin x}{\cos x} \quad \text{for } x \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right).$$

Further, the properties of the tangent function (Proposition 7.14) yield the following:

1. $\sin 0 = 0$ and $\cos 0 = 1$.
2. $\sin(-x) = -\sin x$ and $\cos(-x) = \cos x$ for all $x \in (-\pi/2, \pi/2)$.
3. $0 < \sin x < 1$ for all $x \in (0, \pi/2)$ and $-1 < \sin x < 0$ for all $x \in (-\pi/2, 0)$, while $0 < \cos x < 1$ for all $x \in (-\pi/2, \pi/2)$.
4. $\sin x \to 1$ as $x \to (\pi/2)^-$ and $\sin x \to -1$ as $x \to (-\pi/2)^+$, while $\cos x \to 0$ as $x \to (\pi/2)^-$ or as $x \to (-\pi/2)^+$.

5. Both sin and cos are differentiable on $(-\pi/2, \pi/2)$ and satisfy

$$(\sin)'x = \cos x \quad \text{and} \quad (\cos)'x = -\sin x \quad \text{for all } x \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right).$$

It follows that sin is strictly increasing on $(-\pi/2, \pi/2)$, while cos is strictly increasing on $(-\pi/2, 0)$ and strictly decreasing on $(0, \pi/2)$. Also, since

$$(\sin)''x = (\cos)'x = -\sin x \text{ and } (\cos)''x = (-\sin)'x = -\cos x \text{ for } x \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right),$$

we see that sin is strictly convex on $(-\pi/2, 0)$ and strictly concave on $(0, \pi/2)$, while cos is strictly concave on $(-\pi/2, \pi/2)$.

The geometric properties of sin and cos obtained above can be used to draw their graphs as in Figure 7.7.
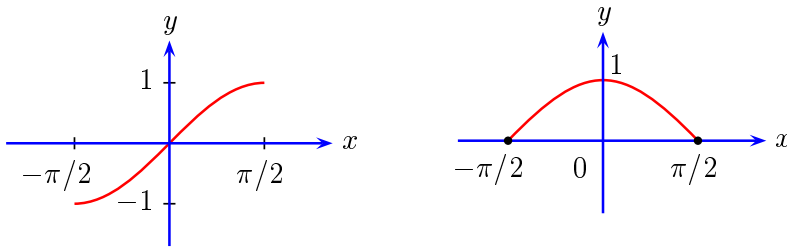
Fig. 7.7. Graphs of the sine function and the cosine function on $(-\pi/2, \pi/2)$

By the definition of derivative and the earlier observations that $\sin 0 = 0$, $(\sin)'0 = 1$, $\cos 0 = 1$, and $(\cos)'0 = 0$, we obtain the following important limits:

$$\lim_{h \to 0} \frac{\sin h}{h} = 1 \quad \text{and} \quad \lim_{h \to 0} \frac{\cos h - 1}{h} = 0.$$

We now define sin and cos at $\pi/2$ as follows:

$$\sin \frac{\pi}{2} := 1 \quad \text{and} \quad \cos \frac{\pi}{2} := 0.$$

Next, we extend sin and cos to $\mathbb{R}$ by requiring

$$\sin(x + \pi) := -\sin x \quad \text{and} \quad \cos(x + \pi) := -\cos x \quad \text{for } x \in \mathbb{R}.$$

It follows that

1. $\sin(x + 2\pi) = \sin x$ and $\cos(x + 2\pi) = \cos x$ for all $x \in \mathbb{R}$.
2. $\sin(-x) = \sin x$ and $\cos(-x) = \cos x$ for all $x \in \mathbb{R}$, that is, sin is an odd function and cos is an even function on $\mathbb{R}$.
3. $\sin x = 0$ if and only if $x = k\pi$ for some $k \in \mathbb{Z}$, and $\cos x = 0$ if and only if $x = (2k + 1)\pi/2$ for some $k \in \mathbb{Z}$.

Recalling that $\tan x = \sin x / \cos x$ for $x \in (-\pi/2, \pi/2)$, we extend the function tan to $\mathbb{R} \setminus \{(2k + 1)\pi/2 : k \in \mathbb{Z}\}$ as follows:

$$\tan x := \frac{\sin x}{\cos x} \quad \text{for } x \in \mathbb{R} \setminus \{(2k + 1)\pi/2 : k \in \mathbb{Z}\}.$$

Then we have

$$\tan(x+\pi) = \frac{\sin(x + \pi)}{\cos(x + \pi)} = \frac{-\sin x}{-\cos x} = \tan x \quad \text{for } x \in \mathbb{R} \setminus \{(2k+1)\pi/2 : k \in \mathbb{Z}\}.$$

Hence for each $k \in \mathbb{Z}$,

$$\tan x \to \infty \text{ as } x \to \frac{(2k + 1)\pi}{2}^{-} \quad \text{and} \quad \tan x \to -\infty \text{ as } x \to \frac{(2k + 1)\pi}{2}^{+}.$$

In view of the above remarks, the graphs of the (extended) sine, cosine, and tangent functions can be drawn as in Figures 7.8, 7.9, and 7.10, respectively.
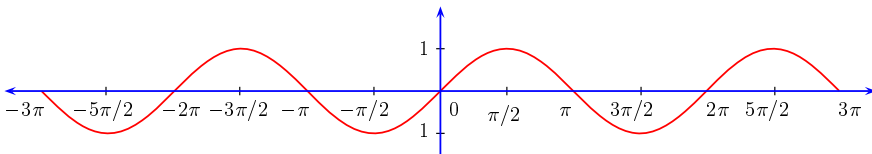
Fig. 7.8. Graph of the sine function on $\mathbb{R}$
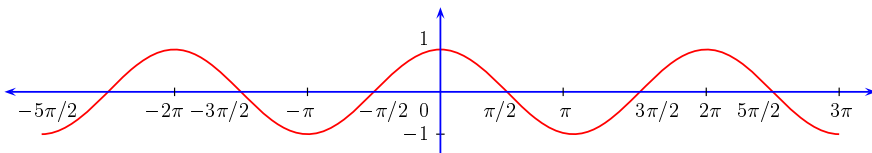
Fig. 7.9. Graph of the cosine function on $\mathbb{R}$

We shall now consider the differentiability of the functions sin, cos, and tan.

Proposition 7.15. *The functions* sin *and* cos *are differentiable on* $\mathbb{R}$ *and satisfy*

$$(\sin)' x = \cos x \quad \text{and} \quad (\cos)' x = -\sin x \quad \text{for all } x \in \mathbb{R}.$$

Also, the function tan *is differentiable on* $\mathbb{R} \setminus \{(2k + 1)\pi/2 : k \in \mathbb{Z}\}$ *and*

$$(\tan)' x = 1 + \tan^2 x \quad \text{for all } \mathbb{R} \setminus \{(2k + 1)\pi/2 : k \in \mathbb{Z}\}.$$

Proof. We have mentioned before that both sin and cos are differentiable on $(-\pi/2, \pi/2)$ and their derivatives satisfy the relations stated in the proposition. Also, from our extension of sin and cos to $\mathbb{R}$, it is clear that this holds at every $x \in \mathbb{R}$, $x \neq (2k+1)\pi/2$ for any $k \in \mathbb{Z}$.

Let us now prove that sin is differentiable at $\pi/2$ and its derivative at $\pi/2$ is 0. First, note that $\sin x \to 1$ and $\cos x \to 0$ as $x \to (\pi/2)^-$, and $\sin \pi/2 = 1$. Hence by L'Hôpital's Rule for $\frac{0}{0}$ indeterminate forms 4.37, we have

$$\lim_{x \to (\pi/2)^-} \frac{\sin x - \sin(\pi/2)}{x - (\pi/2)} = \lim_{x \to (\pi/2)^-} \frac{\cos x}{1} = 0.$$

Next, let $u = x - \pi$. Then as $x \to (\pi/2)^+$, we have $u \to (-\pi/2)^+$ and $\sin x = \sin(u+\pi) = -\sin u$. Also, $\sin u \to -1$ and $\cos u \to 0$ as $u \to (-\pi/2)^+$. Hence again by L'Hôpital's Rule, we have

$$\lim_{x \to (\pi/2)^+} \frac{\sin x - \sin(\pi/2)}{x - (\pi/2)} = \lim_{u \to (-\pi/2)^+} \frac{-\sin u - 1}{u + (\pi/2)} = \lim_{u \to (-\pi/2)^+} \frac{-\cos u}{1} = 0.$$

This proves $(\sin)'(\pi/2) = 0$. Similarly, it can be shown that for each $k \in \mathbb{Z}$,

$$(\sin)'\left((2k+1)\frac{\pi}{2}\right) = \quad 0 = \quad \cos\left((2k+1)\frac{\pi}{2}\right),$$

$$(\cos)'\left((4k+1)\frac{\pi}{2}\right) = -1 = -\sin\left((4k+1)\frac{\pi}{2}\right),$$

$$(\cos)'\left((4k-1)\frac{\pi}{2}\right) = \quad 1 = -\sin\left((4k-1)\frac{\pi}{2}\right).$$

Thus sin and cos are are differentiable on $\mathbb{R}$ and their derivatives satisfy the relations stated in the proposition.

The differentiability of the function tan and the formula for its derivative follow from parts (iii) and (iv) of Proposition 4.5. □

The above proposition implies that sin and cos are infinitely differentiable on $\mathbb{R}$, and for $k \in \mathbb{N}$, their kth derivatives are given by

$$(\sin)^{(k)} x = \begin{cases} (-1)^{k/2} \sin x & \text{if } k \text{ is even,} \\ (-1)^{(k-1)/2} \cos x & \text{if } k \text{ is odd,} \end{cases}$$

and

$$(\cos)^{(k)} x = \begin{cases} (-1)^{k/2} \cos x & \text{if } k \text{ is even,} \\ (-1)^{(k+1)/2} \sin x & \text{if } k \text{ is odd.} \end{cases}$$

Hence the nth Taylor polynomial for sin about 0 is given by

$$P_n(x) = \begin{cases} x - \dfrac{x^3}{3!} + \dfrac{x^5}{5!} - \cdots + (-1)^{(n-1)/2} \dfrac{x^n}{n!} & \text{if } n \text{ is odd,} \\[3mm] x - \dfrac{x^3}{3!} + \dfrac{x^5}{5!} - \cdots + (-1)^{(n-2)/2} \dfrac{x^{n-1}}{(n-1)!} & \text{if } n \text{ is even.} \end{cases}$$
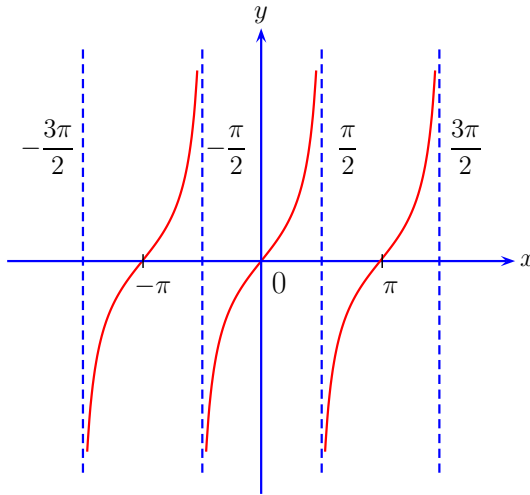
Fig. 7.10. Graph of the tangent function on $\mathbb{R}$

In particular, the linear as well as the quadratic approximation of sin around 0 is given by
$$L(x) = Q(x) = x, \quad x \in \mathbb{R}.$$

Similarly, the nth Taylor polynomial for cos about 0 is given by

$$P_n(x) = \begin{cases} 1 - \dfrac{x^2}{2!} + \dfrac{x^4}{4!} - \cdots + (-1)^{n/2} \dfrac{x^n}{n!} & \text{if } n \text{ is even,} \\[3mm] 1 - \dfrac{x^2}{2!} + \dfrac{x^4}{4!} - \cdots + (-1)^{(n-1)/2} \dfrac{x^{n-1}}{(n-1)!} & \text{if } n \text{ is odd.} \end{cases}$$

In particular, the linear and the quadratic approximations of cos around 0 are given by
$$L(x) = 1 \quad \text{and} \quad Q(x) = 1 - \frac{x^2}{2}, \quad x \in \mathbb{R}.$$

We now prove an important identity and the addition formulas for the functions sin and cos.

Proposition 7.16.

(i) *For all $x \in \mathbb{R}$, we have the identity*

$$\sin^2 x + \cos^2 x = 1.$$

(ii) *For all x_1 and x_2 in $\mathbb{R}$, we have the addition formulas*

$$\sin(x_1 + x_2) = \sin x_1 \cos x_2 + \cos x_1 \sin x_2$$

and

$$\cos(x_1 + x_2) = \cos x_1 \cos x_2 - \sin x_1 \sin x_2.$$

Proof. (i) Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) := \sin^2 x + \cos^2 x \quad \text{for } x \in \mathbb{R}.$$

We have $f'(x) = 2 \sin x \cos x + 2 \cos x(-\sin x) = 0$ for all $x \in \mathbb{R}$ and hence $f(x) = f(0) = \sin^2 0 + \cos^2 0 = 0 + 1 = 1$ for all $x \in \mathbb{R}$. This proves (i).

(ii) To derive the addition formulas, fix $x_2 \in \mathbb{R}$ and define $g, h : \mathbb{R} \to \mathbb{R}$ by

$$g(x) := \cos x \sin(x + x_2) - \sin x \cos(x + x_2) \quad \text{for } x \in \mathbb{R}$$

and

$$h(x) := \sin x \sin(x + x_2) + \cos x \cos(x + x_2) \quad \text{for } x \in \mathbb{R}.$$

Then it can be easily checked that $g'(x) = 0 = h'(x)$ for all $x \in \mathbb{R}$ and hence

$$g(x) = g(-x_2) = \sin x_2 \quad \text{and} \quad h(x) = h(-x_2) = \cos x_2.$$

Putting $x = x_1$ in these equations, we obtain

$$\cos x_1 \sin(x_1 + x_2) - \sin x_1 \cos(x_1 + x_2) = \sin x_2$$

and

$$\sin x_1 \sin(x_1 + x_2) + \cos x_1 \cos(x_1 + x_2) = \cos x_2.$$

Solving these two linear equations for $\sin(x_1 + x_2)$ and $\cos(x_1 + x_2)$, we obtain

$$\sin(x_1 + x_2) = \sin x_1 \cos x_2 + \cos x_1 \sin x_2$$

and

$$\cos(x_1 + x_2) = \cos x_1 \cos x_2 - \sin x_1 \sin x_2,$$

as desired. □

We now consider the reciprocals of the functions sin, cos, and tan. The **cosecant function** and the **secant function** are defined by

$$\csc x := \frac{1}{\sin x} \quad \text{if } x \in \mathbb{R}, \ x \neq k\pi \text{ for any } k \in \mathbb{Z},$$

and

$$\sec x := \frac{1}{\cos x} \quad \text{if } x \in \mathbb{R}, \ x \neq (2k + 1)\frac{\pi}{2} \text{ for any } k \in \mathbb{Z}.$$

The **cotangent function** is defined by

$$\cot x := \frac{\cos x}{\sin x} \quad \text{if } x \in \mathbb{R},\ x \neq k\pi \text{ for any } k \in \mathbb{Z}.$$

Thus $\cot x$ is the reciprocal of $\tan x$ if $x \neq k\pi/2$ for any $k \in \mathbb{Z}$.

The functions sin, cos, tan, csc, sec, and cot are known as the **trigono-metric functions**. Several elementary results concerning these functions are given in Exercises 27–33. They follow from their definitions and from Proposition 7.16.

Let us now consider the **inverse trigonometric functions**. The function $f : (-\pi/2, \pi/2) \to \mathbb{R}$ defined by $f(x) = \tan x$ is bijective. Its inverse is the function arctan $: \mathbb{R} \to (-\pi/2, \pi/2)$ with which we started our discussion in this section. This function is also denoted by $\tan^{-1}$. Thus

$$\tan^{-1} : \mathbb{R} \to (-\pi/2, \pi/2)$$

is the function characterized by the following:

$$y \in \mathbb{R} \text{ and } \tan^{-1} y = x \iff x \in (-\pi/2, \pi/2) \text{ and } \tan x = y.$$

Also, as we have seen in part (i) of Proposition 7.13,

$$(\tan^{-1})'y = \frac{1}{1+y^2} \quad \text{for all } y \in \mathbb{R}.$$

The function $g : [-\pi/2, \pi/2] \to [-1, 1]$ defined by $g(x) := \sin x$ is bijective. Its inverse is denoted by $\sin^{-1}$ or by arcsin. Thus

$$\sin^{-1} : [-1, 1] \to [-\pi/2, \pi/2]$$

is the function characterized by the following:

$$y \in [-1, 1] \text{ and } \sin^{-1} y = x \iff x \in [-\pi/2, \pi/2] \text{ and } \sin x = y.$$

By the Continuous Inverse Theorem (Proposition 3.14), the function $\sin^{-1}$ is continuous on $[-1, 1]$. Also, the derivative formula for the inverse function (Proposition 4.11) shows that for $y \in (-1, 1)$ and $y = \sin x$ with $x \in (-\pi/2, \pi/2)$, we have

$$(\sin^{-1})'y = \frac{1}{f'(x)} = \frac{1}{\cos x} = \frac{1}{\sqrt{1 - \sin^2 x}} = \frac{1}{\sqrt{1 - y^2}}.$$

Thus $\sin^{-1}$ is differentiable on $(-1, 1)$.

Similarly, the function $h : [0, \pi] \to [-1, 1]$ defined by $h(x) := \cos x$ is bijective. Its inverse is denoted by $\cos^{-1}$ or by arccos. Thus

$$\cos^{-1} : [-1, 1] \to [0, \pi]$$

is the function characterized by the following:

$$y \in [-1, 1] \text{ and } \cos^{-1} y = x \Longleftrightarrow x \in [0, \pi] \text{ and } \cos x = y.$$

By Proposition 3.14, the function $\cos^{-1}$ is continuous on $[-1, 1]$. Also, as before, for $y \in (-1, 1)$ and $y = \cos x$ with $x \in (0, \pi)$, we have

$$(\cos^{-1})' y = \frac{1}{g'(x)} = \frac{1}{-\sin x} = \frac{-1}{\sqrt{1 - \cos^2 x}} = \frac{-1}{\sqrt{1 - y^2}}.$$

Thus $\cos^{-1}$ is differentiable on $(-1, 1)$.

The graphs of the inverse trigonometric functions $\sin^{-1} : [-1, 1] \rightarrow [-\pi/2, \pi/2]$ and $\cos^{-1} : [-1, 1] \rightarrow [0, \pi]$ are shown in Figure 7.11.
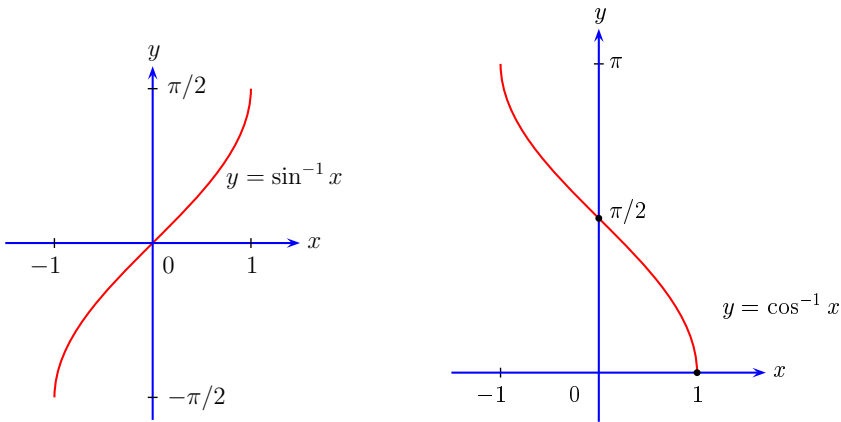
Fig. 7.11. Graphs of $\sin^{-1} : [-1, 1] \rightarrow [-\pi/2, \pi/2]$ and $\cos^{-1} : [-1, 1] \rightarrow [0, \pi]$

The function $f_1 : (0, \pi) \rightarrow \mathbb{R}$ defined by $f_1(x) := \cot x$ is bijective. Its inverse is denoted by $\cot^{-1}$. Thus

$$\cot^{-1} : \mathbb{R} \rightarrow (0, \pi)$$

is the function characterized by the following:

$$y \in \mathbb{R} \text{ and } \cot^{-1} y = x \Longleftrightarrow x \in (0, \pi) \text{ and } \cot x = y.$$

The function $g_1 : [0, \pi/2) \cup (\pi/2, \pi] \rightarrow (-\infty, -1] \cup [1, \infty)$ defined by $g_1(x) := \csc x$ is bijective. Its inverse is denoted by $\csc^{-1}$. Thus

$$\csc^{-1} : (-\infty, -1] \cup [1, \infty) \rightarrow [0, \pi/2) \cup (\pi/2, \pi]$$

is the function characterized by the following:

$$y \in \mathbb{R}, \ |y| \geq 1 \text{ and } \csc^{-1} y = x \Longleftrightarrow x \in \mathbb{R}, \ 0 < |x| \leq \pi/2 \text{ and } \csc x = y.$$

The function $h_1 : [0, \pi/2) \cup (\pi/2, \pi] \to (-\infty, -1] \cup [1, \infty)$ defined by $h_1(x) := \sec x$ is bijective. Its inverse is denoted by $\sec^{-1}$. Thus

$$\sec^{-1} : (-\infty, -1] \cup [1, \infty) \to [0, \pi/2) \cup (\pi/2, \pi]$$

is the function characterized by the following:

$$y \in \mathbb{R}, \ |y| \geq 1 \text{ and } \sec^{-1} y = x \iff x \in [0, \pi], \ x \neq \pi/2 \text{ and } \sec x = y.$$

The formulas for the derivatives of the functions $\cot^{-1}$, $\csc^{-1}$, and $\sec^{-1}$ are given in Exercise 40. For various relations involving the inverse trigonometric functions, see Exercises 34–39.

7.3 Sine of the Reciprocal

In this section we study the composition of the reciprocal function $x \mapsto 1/x$ and the sine function. We also study some related functions. To begin with, consider the function $f : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ defined by

$$f(x) = \sin \frac{1}{x}, \quad x \neq 0.$$

The reason for paying special attention to this function is that it has many interesting properties and it provides, along with other associated functions, several examples and counterexamples for various statements in calculus and analysis. We have deferred these examples till now since the trigonometric functions were introduced only in Section 7.2.

Properties of the sine function developed earlier yield the following:

1. f is an odd function.
2. f is a bounded function. In fact, $-1 \leq f(x) \leq 1$ for all $x \in \mathbb{R} \setminus \{0\}$.
3. $f(x) = 0$ if and only if $x = 1/k\pi$ for some nonzero $k \in \mathbb{Z}$, while $f(x) = 1$ if and only if $x = 2/(4k + 1)\pi$ for some $k \in \mathbb{Z}$, and $f(x) = -1$ if and only if $x = 2/(4k - 1)\pi$ for some $k \in \mathbb{Z}$.
   The graph of the function f drawn in Figure 7.12 shows that it oscillates between 1 and -1 more and more rapidly as we approach 0 from the left or from the right.
4. f is continuous on $\mathbb{R} \setminus \{0\}$ since the reciprocal function $x \mapsto 1/x$ is continuous on $\mathbb{R} \setminus \{0\}$ and the sine function is continuous on $\mathbb{R}$ (Proposition 3.4).
5. f is not uniformly continuous on $(0, \delta)$ for any $\delta > 0$. This follows by noting that if $x_n = 1/n\pi$ and $y_n = 2/(4n + 1)\pi$ for $n \in \mathbb{N}$, then for all large n, we have $x_n, y_n \in (0, \delta)$ and $x_n - y_n \to 0$, but since $f(x_n) = 0$, $f(y_n) = 1$, we see that $f(x_n) - f(y_n) \not\to 0$. Similarly, f is not uniformly continuous on $(-\delta, 0)$ for any $\delta > 0$.
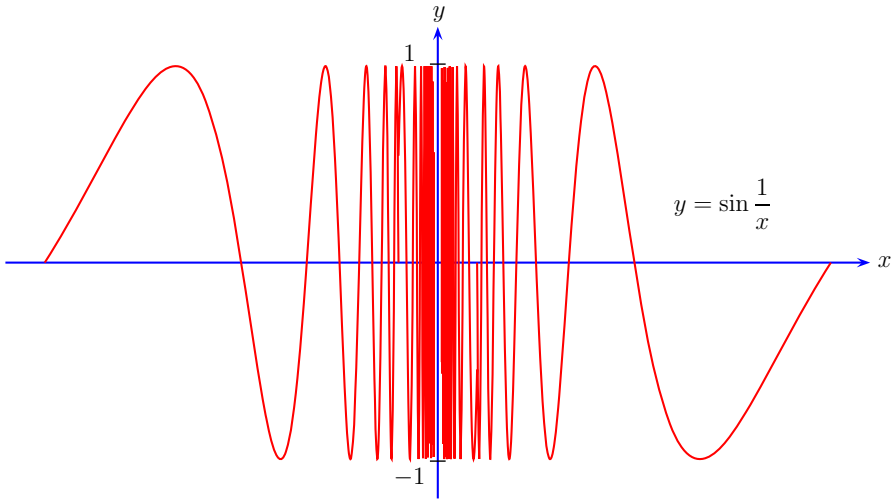
Fig. 7.12. Graph of $f : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ given by $f(x) = \sin \frac{1}{x}$

However, if $D \subseteq \mathbb{R}$ and there is $\delta > 0$ such that $D \subseteq (-\infty, -\delta] \cup [\delta, \infty)$, then f is uniformly continuous on D. This can be seen as follows. Let (x_n) and (y_n) be any sequences in D such that $x_n - y_n \to 0$. Then

$$\sin \frac{1}{x_n} - \sin \frac{1}{y_n} = 2 \cos \frac{1}{2} \left(\frac{1}{x_n} + \frac{1}{y_n} \right) \sin \frac{1}{2} \left(\frac{1}{x_n} - \frac{1}{y_n} \right)$$

$$= -2 \cos \left(\frac{x_n + y_n}{2 x_n y_n} \right) \sin \left(\frac{x_n - y_n}{2 x_n y_n} \right) \quad \text{for all } n \in \mathbb{N}.$$

(See Exercise 28.) Since $|x_n| \geq \delta$ and $|y_n| \geq \delta$ for all $n \in \mathbb{N}$, we see that $(x_n - y_n)/2 x_n y_n \to 0$ and hence

$$|f(x_n) - f(y_n)| = \left| \sin \frac{1}{x_n} - \sin \frac{1}{y_n} \right| \leq 2 \left| \sin \left(\frac{x_n - y_n}{2 x_n y_n} \right) \right| \to 0 \quad \text{as} \quad n \to \infty.$$

6. f is infinitely differentiable at every nonzero $x \in \mathbb{R}$, thanks to the Chain Rule (Proposition 4.9). In particular, we have

$$f'(x) = -\frac{1}{x^2} \cos \frac{1}{x} \quad \text{and} \quad f''(x) = \frac{2}{x^3} \cos \frac{1}{x} - \frac{1}{x^4} \sin \frac{1}{x} \quad \text{for } x \in \mathbb{R} \setminus \{0\}.$$

It is clear that for any $\delta > 0$, f' and f'' are not bounded on $(0, \delta)$ as well as on $(-\delta, 0)$.

7. For any $\delta > 0$, f is not monotonic on $(0, \delta)$ as well as on $(-\delta, 0)$. To see this, let $x_k := 1/k\pi$ for nonzero $k \in \mathbb{Z}$; note that $f'(x_k) = (-1)^{k+1} k^2 \pi^2$ and apply part (i) of 4.28.

8. For any $\delta > 0$, f is neither convex nor concave on $(0, \delta)$ as well as on $(-\delta, 0)$. To see this, let $x_k := 1/k\pi$ for nonzero $k \in \mathbb{Z}$; note that $f''(x_k) = (-1)^k k^3 \pi^3$ and apply part (i) of Corollary 4.33.

We remark that similar properties are possessed by the real-valued function $g : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ defined by

$$g(x) = \cos\frac{1}{x}, \quad x \neq 0.$$

(See Exercise 43.)

Let $r_0 \in \mathbb{R}$, and consider the function $f_0 : \mathbb{R} \to \mathbb{R}$ defined by

$$f_0(x) = \begin{cases} f(x) & \text{if } x \neq 0, \\ r_0 & \text{if } x = 0. \end{cases}$$

It follows from Corollary 6.11 that for any a, $b \in \mathbb{R}$ with $a < b$, f_0 is Riemann integrable on $[a, b]$. Consider now the function $F_0 : \mathbb{R} \to \mathbb{R}$ defined by

$$F_0(x) := \int_0^x f_0(t)dt.$$

Observe that this function does not depend on the choice of $r_0 \in \mathbb{R}$, thanks to Proposition 6.12. Let $c \in \mathbb{R}$, $c \neq 0$. Since f_0 is continuous at c, part (ii) of the Fundamental Theorem of Calculus (Proposition 6.21) shows that the function F_0 is differentiable at c and $F_0'(c) = f_0(c) = f(c)$. However, it is not clear how the functions f_0 and F_0 behave near 0. We now analyze this case separately. This analysis will show that the converse of a stronger version of part (ii) of the FTC stated in Remark 6.22 (ii) does not hold.

Proposition 7.17. *Let f_0 and F_0 be as defined above.*

(i) *f_0 is not continuous at 0. In fact, neither $\lim_{x \to 0+} f_0(x)$ nor $\lim_{x \to 0-} f_0(x)$ exists.*

(ii) *F_0 is differentiable at 0 and $F_0'(0) = 0$, that is,*

$$\lim_{x \to 0} \frac{1}{x} \int_0^x \sin\frac{1}{t}dt = 0.$$

Proof. (i) Let $x_n = 1/n\pi$ and $y_n = 2/(4n + 1)\pi$ for $n \in \mathbb{N}$. Then (x_n) and (y_n) are sequences of positive real numbers such that $x_n \to 0$ and $y_n \to 0$, but $f(x_n) = 0$ and $f(y_n) = 1$ for all $n \in \mathbb{N}$, and so $f(x_n) \to 0$, whereas $f(y_n) \to 1$. Thus it follows that $\lim_{x \to 0+} f(x)$ does not exist. Since f is an odd function, $\lim_{x \to 0-} f(x)$ does not exist. Finally, since $f_0(x) = f(x)$ for all nonzero $x \in \mathbb{R}$, (i) is proved.

(ii) Let $x \in \mathbb{R}$, $x > 0$. By Proposition 6.20, we have

$$F_0(x) - F_0(0) = \int_0^x \sin\frac{1}{t}dt = \lim_{r \to 0+} \int_r^x \sin\frac{1}{t}dt.$$

Fix $r \in \mathbb{R}$ such that $0 < r \le x$. Substituting $t = 1/u$ and then integrating by parts (that is, using Propositions 6.26 and 6.25), we obtain

$$\int_r^x \sin\frac{1}{t}dt = \int_{1/x}^{1/r}(\sin u)\frac{1}{u^2}du = x^2\cos\frac{1}{x} - r^2\cos\frac{1}{r} - 2\int_{1/x}^{1/r}\frac{\cos u}{u^3}du.$$

Since

$$0 \le \int_{1/x}^{1/r}\left|\frac{\cos u}{u^3}\right|du \le \int_{1/x}^{1/r}\frac{1}{u^3}du = \frac{1}{2}(x^2 - r^2),$$

we see that

$$\left|\int_0^x \sin\frac{1}{t}dt\right| \le \lim_{r\to 0+}\left|\int_r^x \sin\frac{1}{t}dt\right| \le \lim_{r\to 0+}(x^2 + r^2 + x^2 - r^2) = 2x^2.$$

Thus for every $x \in (0, \infty)$, we have

$$\left|\frac{1}{x}\int_0^x \sin\frac{1}{t}dt\right| \le 2x \quad\text{and so}\quad \lim_{x\to 0+}\frac{F_0(x) - F_0(0)}{x - 0} = \lim_{x\to 0+}\frac{1}{x}\int_0^x \sin\frac{1}{t}dt = 0.$$

Replacing x by $-x$ and noting that $\sin$ is an odd function, we also obtain

$$\lim_{x\to 0-}\frac{F_0(x) - F_0(0)}{x - 0} = \lim_{x\to 0-}\frac{1}{x}\int_0^x \sin\frac{1}{t}dt = 0.$$

Hence the desired result follows by Proposition 3.29.   □

We remark that a similar result holds for the cosine of the reciprocal. (See Exercise 44.)

We shall now study some functions associated with the function f_0. They are obtained by multiplying f_0 by the identity function and by the square of the identity function.

Example 7.18. Consider the function $f_1 : \mathbb{R} \to \mathbb{R}$ defined by

$$f_1(x) = \begin{cases} x\sin(1/x) & \text{if } x \ne 0, \\ 0 & \text{if } x = 0. \end{cases}$$

Properties of the function f developed earlier yield the following:

1. f_1 is an even function.
2. f_1 is a bounded function. In fact, $-1 < f_1(x) < 1$ for all $x \in \mathbb{R}$. To see this, note that $f_1(x) = \sin(1/x)/(1/x)$ if $x \ne 0$, and $-y < \sin y < y$ for all nonzero $y \in \mathbb{R}$ (as can be seen by a simple application of the MVT). Since $\lim_{h\to 0}\sin h/h = 1$, it follows that $f_1(x) \to 1$ as $x \to \infty$ or as $x \to -\infty$.
3. The oscillations of the function f_1, inherited from the function f, are 'damped' near 0, because $|f_1(x)| \le |x|$ for all $x \in \mathbb{R}$. This behavior of f_1 is shown in Figure 7.13.
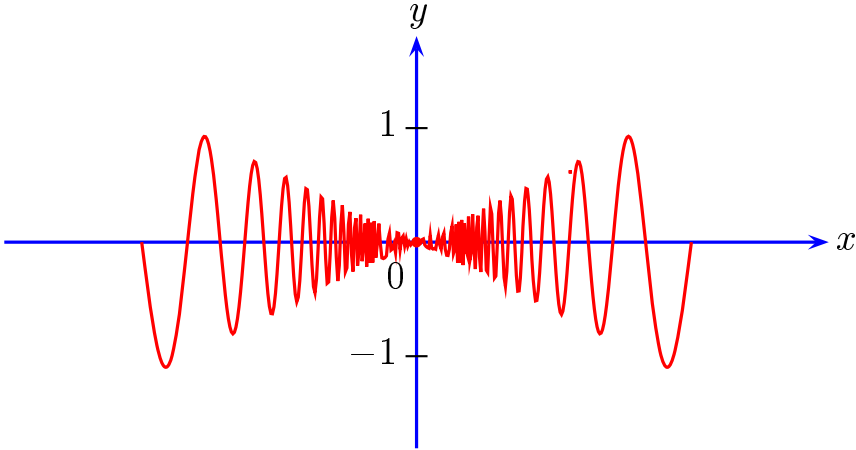
Fig. 7.13. Illustration of damped oscillations: Graph of $f_1 : \mathbb{R} \to \mathbb{R}$ given by $f_1(0) = 0$ and $f_1(x) = x \sin \frac{1}{x}$ for $x \neq 0$

4. f_1 is continuous on $\mathbb{R}$. To see this, we note that f_1 is a product of two functions each of which is continuous on $\mathbb{R} \setminus \{0\}$, and moreover, f_1 is continuous at 0, because if (x_n) is any sequence such that $x_n \to 0$, then we have $|f(x_n)| \leq |x_n|$, and so $f(x_n) \to 0$.
5. f_1 is infinitely differentiable at every nonzero $x \in \mathbb{R}$, thanks to part (iv) of Proposition 4.5. In particular, we have

$$f_1'(x) = \sin \frac{1}{x} - \frac{1}{x} \cos \frac{1}{x} \quad \text{and} \quad f_1''(x) = -\frac{1}{x^3} \sin \frac{1}{x} \quad \text{for } x \in \mathbb{R} \setminus \{0\}.$$

It is clear that for any $\delta > 0$, f_1' and f_1'' are not bounded on $(0, \delta)$ as well as on $(-\delta, 0)$.
6. For any $\delta > 0$, f_1 is not monotonic on $(0, \delta)$ as well as on $(-\delta, 0)$. To see this, let $x_k := 1/k\pi$ for nonzero $k \in \mathbb{Z}$; note that $f_1'(x_k) = (-1)^{k+1}k\pi$ and apply part (i) of Corollary 4.28.
7. For any $\delta > 0$, f_1 is neither convex nor concave on $(0, \delta)$ as well as on $(-\delta, 0)$. To see this, let $y_k := 2/(2k + 1)\pi$ for nonzero $k \in \mathbb{Z}$; note that $f_1''(y_k) = (-1)^{k+1}(k + (1/2))^3\pi^3$ and apply part (i) of Corollary 4.33.
8. The right (hand) and the left (hand) derivatives of f_1 at 0 do not exist, that is, the limits

$$\lim_{x \to 0+} \frac{f_1(x) - f_1(0)}{x - 0} = \lim_{x \to 0+} \sin \frac{1}{x} \quad \text{and} \quad \lim_{x \to 0-} \frac{f_1(x) - f_1(0)}{x - 0} = \lim_{x \to 0-} \sin \frac{1}{x}$$

do not exist, as we have seen in part (i) of Proposition 7.17.

The function $|f_1|$ has an absolute minimum (although it is not a strict minimum) at 0. However, there is no $\delta > 0$ such that f is decreasing on

$(-\delta, 0]$ and f is increasing on $[0, \delta)$, because $|f_1|(1/k\pi) = 0$ for all nonzero $k \in \mathbb{Z}$, while $|f_1|(2/(2k + 1)\pi) = 2/|2k + 1|\pi$ for all $k \in \mathbb{Z}$. This phenomenon was earlier illustrated in Example 1.18 of infinitely many zigzags.

The function f_1 can be used to conclude that the converse of L'Hôpital's rule for $\frac{\infty}{\infty}$ indeterminate forms is not true. For this purpose, consider functions $h_1, g_1 : (0, \infty) \to \mathbb{R}$ defined by

$$h_1(x) := \frac{1 + f_1(x)}{x} \quad \text{and} \quad g_1(x) := \frac{1}{x}.$$

Then $g_1(x) \to \infty$ as $x \to 0^+$, $g'(x) = -1/x^2 \neq 0$ for all $x \in (0, \infty)$, and

$$\lim_{x \to 0^+} \frac{h_1(x)}{g_1(x)} == \lim_{x \to 0^+} [1 + f_1(x)] = 1 + 0 = 0,$$

but

$$\lim_{x \to 0^+} \frac{h_1'(x)}{g_1'(x)} = \lim_{x \to 0^+} \frac{[-1 - \cos(1/x)]/x^2}{-1/x^2} = \lim_{x \to 0^+} [1 + \cos(1/x)]$$

does not exist because $\lim_{x \to 0^+} \cos(1/x)$ does not exist. $\diamond$

Example 7.19. Consider the function $f_2 : \mathbb{R} \to \mathbb{R}$ defined by

$$f_2(x) = \begin{cases} x^2 \sin(1/x) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

Properties of the functions f and f_1 developed earlier yield the following.

1. f_2 is an odd function.
2. For any $a \in \mathbb{R}$, f_2 is not bounded above on (a, ∞). This follows by noting that $\sin(1/x)/(1/x) \to 1$ as $x \to \infty$, so that

$$f_2(x) = x \frac{\sin(1/x)}{(1/x)} \to \infty \quad \text{as } x \to \infty.$$

   Being an odd function, f_2 is not bounded below on $(-\infty, b)$ for any $b \in \mathbb{R}$.
3. The oscillations of the function f_2, inherited from the function f, are doubly damped near 0, because $|f_2(x)| \leq |x|^2$ for all $x \in \mathbb{R}$. This behavior of f_2 is shown in Figure 7.14.
4. f_2 is continuous on $\mathbb{R}$, since $f_2(x) = xf_1(x)$ for all $x \in \mathbb{R}$ and the function f_1 is continuous on $\mathbb{R}$.
5. f_2 is infinitely differentiable at every nonzero $x \in \mathbb{R}$. In particular, for any $x \in \mathbb{R} \setminus \{0\}$, we have

$$f_2'(x) = 2x \sin \frac{1}{x} - \cos \frac{1}{x} \quad \text{and} \quad f_2''(x) = 2 \sin \frac{1}{x} - \frac{2}{x} \cos \frac{1}{x} - \frac{1}{x^2} \sin \frac{1}{x}.$$

   It is clear that f_2' is bounded on $x \in \mathbb{R} \setminus \{0\}$, but for any $\delta > 0$, f_2'' is not bounded on $(0, \delta)$ as well as on $(-\delta, 0)$.

$$y = x^2 \sin \frac{1}{x}$$
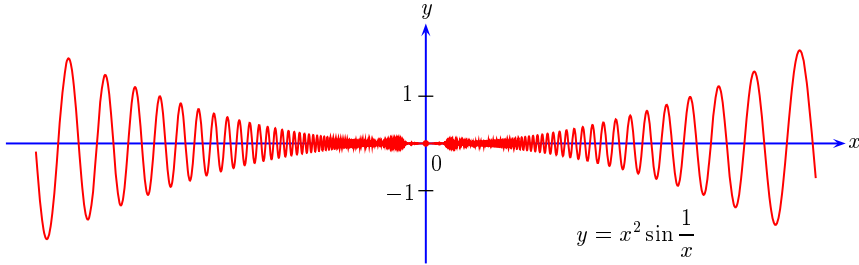
Fig. 7.14. Illustration of doubly damped oscillations: Graph of $f_2 : \mathbb{R} \to \mathbb{R}$ given by $f_2(0) = 0$ and $f_2(x) = x^2 \sin \frac{1}{x}$ for $x \neq 0$

6. For any $\delta > 0$, f_2 is not monotonic on $(0, \delta)$ as well as on $(-\delta, 0)$. To see this, let $x_k := 1/k\pi$ for nonzero $k \in \mathbb{Z}$; note that $f_2'(x_k) = (-1)^{k+1}$ and apply part (i) of Corollary 4.28.
7. For any $\delta > 0$, f_2 is neither convex nor concave on $(0, \delta)$ as well as on $(-\delta, 0)$. To see this, let $y_k := 1/k\pi$ for nonzero $k \in \mathbb{Z}$; note that $f_2''(y_k) = (-1)^{k+1}2k\pi$ and apply part (i) of Corollary 4.33.
8. f_2 is differentiable at 0. In fact,

$$f_2'(0) = \lim_{x \to 0} \frac{f_2(x) - f_2(0)}{x - 0} = \lim_{x \to 0} x \sin \frac{1}{x} = 0.$$

But f_2' is not continuous at 0, because $\lim_{x \to 0} f_2'(x)$ does not exist. This follows because $\lim_{x \to 0} \cos(1/x)$ does not exist.
9. Although $f_2'(0) = 0$, neither does f_2 have a local extremum at 0 nor is 0 a point of inflection for f because of the oscillatory nature of f_2 and f_2' around 0.

The function f_2 can be used to conclude that the converse of L'Hôpital's rule for $\frac{0}{0}$ indeterminate forms is not true. For this purpose, let $g_2(x) := \sin x$ for $x \in \mathbb{R}$. Then $\lim_{x \to 0} f_2(x) = 0 = \lim_{x \to 0} g_2(x)$ and

$$\lim_{x \to 0} \frac{f_2(x)}{g_2(x)} = \left(\lim_{x \to 0} \frac{x}{\sin x} \right) \left(\lim_{x \to 0} x \sin \frac{1}{x} \right) = (1)(0) = 0,$$

but

$$\lim_{x \to 0} \frac{f_2'(x)}{g_2'(x)} = \lim_{x \to 0} \frac{2x \sin(1/x) - \cos(1/x)}{\cos x}$$

does not exist, because $\lim_{x \to 0} \cos(1/x)$ does not exist, but on the other hand, $\lim_{x \to 0} 2x \sin(1/x) = 0$ and $\lim_{x \to 0} \cos x = 1$. $\qquad \diamond$

For further examples, see Exercises 45, 48, and 60.

7.4 Polar Coordinates

Having defined the trigonometric functions and the number π, we are in a position to describe an alternative and useful way of representing points in the plane $\mathbb{R}^2$ by their polar coordinates. Roughly speaking, the polar coordinates of a point $P = (x, y) \in \mathbb{R}^2$ are the numbers r and θ satisfying the equations

$$x = r\cos\theta \quad \text{and} \quad y = r\sin\theta.$$

Geometrically speaking, the number r represents the distance from P to the origin $O = (0, 0)$, whereas θ can be interpreted as the 'angle' from the positive x-axis to the line segment OP. However, there is a certain ambiguity if we define r and θ simply by the above equations. Indeed, if (r, θ) satisfy these equations, then so do $(r, \theta + 2\pi)$, $(r, \theta - 2\pi)$, $(-r, \theta + \pi)$, etc.; the special case $P = O$ is even worse because we can take $r = 0$ and θ to be any real number. To avoid such ambiguities and to enable us to give a precise definition of polar coordinates, we first prove the following proposition. In the sequel, we shall also give a formal definition of the notion of *angle*, and clarify the geometric interpretation of polar coordinates.

Proposition 7.20. *If $x, y \in \mathbb{R}$ are such that $(x, y) \neq (0, 0)$, then r and θ defined by*

$$r := \sqrt{x^2 + y^2} \quad \text{and} \quad \theta := \begin{cases} \cos^{-1}\left(\dfrac{x}{r}\right) & \text{if } y \geq 0, \\[3mm] -\cos^{-1}\left(\dfrac{x}{r}\right) & \text{if } y < 0, \end{cases}$$

satisfy the following properties:

$$r, \theta \in \mathbb{R}, \quad r > 0, \quad \theta \in (-\pi, \pi], \quad x = r\cos\theta, \quad \text{and} \quad y = r\sin\theta.$$

Conversely, if $r, \theta \in \mathbb{R}$ are such that $r > 0$ and $\theta \in (-\pi, \pi]$, then $x := r\cos\theta$ and $y := r\sin\theta$ are real numbers such that $(x, y) \neq (0, 0)$, $r = \sqrt{x^2 + y^2}$ and θ equals $\cos^{-1}(x/r)$ or $-\cos^{-1}(x/r)$ according as $y \geq 0$ or $y < 0$.

Proof. Let $x, y \in \mathbb{R}$ with $(x, y) \neq (0, 0)$ be given. Define r and θ by the formulas displayed above. Then $(x, y) \neq (0, 0)$ implies that $r > 0$. Also, since $|x/r| \leq 1$ and since $\cos^{-1}$ is a map from $[-1, 1]$ to $[0, \pi]$, we see that θ is well defined and $\theta \in [-\pi, \pi]$. Further, if $y < 0$, then $|x/r| < 1$, and so $\cos^{-1}(x/r) \neq \pi$. Thus $\theta \in (-\pi, \pi]$. Moreover, since $\cos(-\theta) = \cos\theta$, it follows that $\cos\theta = x/r$, that is, $x = r\cos\theta$. Consequently, $y^2 = r^2(1 - \cos^2\theta)$, and hence $y = \pm r\sin\theta$. But from the definition of θ, it is clear that $y \geq 0$ if and only if $0 \leq \theta \leq \pi$. So we must have $y = r\sin\theta$. This proves that r and θ satisfy the desired properties.

Conversely, let $r, \theta \in \mathbb{R}$ be given such that $r > 0$ and $\theta \in (-\pi, \pi]$. Define $x := r\cos\theta$ and $y := r\sin\theta$. Since $r > 0$ and $\cos^2\theta + \sin^2\theta = 1$, it is clear

that $r = \sqrt{x^2 + y^2}$, and, in particular, $(x, y) \neq (0, 0)$. Also, since $\theta \in (-\pi, \pi]$ and $x/r = \cos \theta$, it follows that if $\theta \in [0, \pi]$, then $\theta = \cos^{-1}(x/r)$, whereas if $\theta \in (-\pi, 0)$, then $-\theta \in (0, \pi)$ and $\cos(-\theta) = \cos \theta = x/r$, and consequently, $-\theta = \cos^{-1}(x/r)$. Moreover, $y = r \sin \theta \geq 0$ if and only if $\theta \in [0, \pi]$, and thus we see that x and y satisfy the desired properties. $\qquad\square$

In view of the above proposition, we define the **polar coordinates** of a point $P = (x, y)$ in $\mathbb{R}^2$, different from the origin, to be the pair (r, θ) defined by the formulas displayed above. Equivalently, r and θ are real numbers determined by the conditions $r > 0$, $\theta \in (-\pi, \pi]$, $x = r \cos \theta$, and $y = r \sin \theta$.

For example, the polar coordinates of the points $(1, 0)$, $(3, 4)$, $(0, 1)$, $(-1, 0)$, $(0, -1)$, and $(3, -4)$ are $(1, 0)$, $(5, \cos^{-1}(3/5))$, $(1, \pi/2)$, $(1, \pi)$, $(1, -\pi/2)$, and $(5, -\cos^{-1}(3/5))$, respectively. The polar coordinates of the origin $(0, 0)$ are not defined.

For a point $P = (x, y) \in \mathbb{R}^2$, we sometimes call the pair (x, y) the **Cartesian coordinates** or the **rectangular coordinates** of P.

Remarks 7.21. (i) In the definition of polar coordinates, we have required that θ should lie in the interval $(-\pi, \pi]$. This is actually a matter of convention. Alternative conditions are possible and can sometimes be found in books on calculus. For example, a commonly used alternative is to require that θ should lie in the interval $[0, 2\pi)$. In this case, we have to change $-\cos^{-1}(x/r)$ to $2\pi - \cos^{-1}(x/r)$ in the formula for θ in Proposition 7.20. Yet another alternative is to let r take positive as well as negative values but restrict θ to the interval $[0, \pi)$. In this case, we have to set r equal to $\sqrt{x^2 + y^2}$ or $-\sqrt{x^2 + y^2}$, according as $y \geq 0$ or $y < 0$, and set θ equal to $\cos^{-1}(x/r)$ (regardless of the sign of y) in Proposition 7.20. In any case, the key equations remain $x = r \cos \theta$ and $y = r \sin \theta$. In fact, some books disregard the questions of uniqueness and define the polar coordinates of the point (x, y) to be *any* pair (r, θ) of real numbers satisfying $x = r \cos \theta$ and $y = r \sin \theta$. We shall, however, prefer that a change of coordinates be determined unambiguously and adhere to the definition above.

(ii) It is more common to describe the 'inverse formula' for $\theta \in (-\pi, \pi]$ satisfying $x = r \cos \theta$ and $y = r \sin \theta$ in terms of the arctangent function, namely, $\theta = \tan^{-1}(y/x)$. However, this is correct only when $x > 0$. For a comprehensive 'inverse formula', one has to consider four other cases separately. Indeed, $\theta = \tan^{-1}(y/x) + \pi$ if $x < 0$ and $y \geq 0$; $\theta = \tan^{-1}(y/x) - \pi$ if $x < 0$ and $y < 0$; $\theta = \pi/2$ if $x = 0$ and $y > 0$; finally, $\theta = -\pi/2$ if $x = 0$ and $y < 0$. To avoid this, we have used $\cos^{-1}$ in Proposition 7.20, and as a result, it suffices to consider only two cases.

(iii) In classical geometry, the polar coordinates are described as follows. In the plane choose a point O, called the pole, and a ray emanating from O, called the polar axis. Now, the polar coordinates of a point P are (r, θ), where r is the distance of P from the pole O, and θ is (any) angle from the polar axis to the line joining O and P. In our approach, the plane comes equipped

with a (Cartesian) coordinate system, and we have fixed the pole O to be the origin and the polar axis to be the positive x-axis. $\diamondsuit$

We have seen earlier that an equation in x and y determines a curve in the plane. Similarly, an equation in r and θ determines a curve in the plane, namely, the curve consisting of points in the plane whose polar coordinates satisfy this equation. In case the equation is satisfied when $r = 0$, we regard the origin as a point on the curve. Frequently, the equations we come across are of the form $r = p(\theta)$, where p is a real-valued function defined on some subset of $(-\pi, \pi]$. If no domain for p is specified, then this may be assumed to be $(-\pi, \pi]$. For ease of reference, we may use the terms **Cartesian equation** and **polar equation** to mean an equation (of a curve in the plane) in Cartesian coordinates and in polar coordinates, respectively.
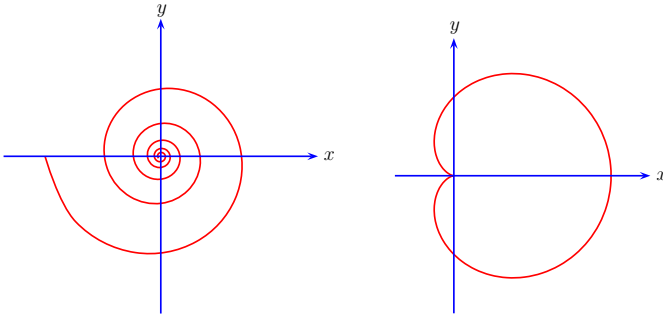
Fig. 7.15. Spiral of Archimedes $r = \theta$, and the cardioid $r = 2(1 + \cos \theta)$

Often, a curve can be described by a Cartesian equation as well as by a polar equation. Sometimes, the latter can be simpler. For example, a circle of radius 2 centered at the origin can be described by the Cartesian equation $x^2 + y^2 = 4$, whereas its polar equation is simply $r = 2$. On the other hand, to see how the curve given by the polar equation $r = 2 \sin \theta$ might look like, it may be easier to first convert it to a Cartesian equation. To do so, note that the polar equation is equivalent to $r^2 = 2r \sin \theta$, and hence the Cartesian equation is given by $x^2 + y^2 = 2y$, that is, $x^2 + (y - 1)^2 = 1$. Thus, the curve with the polar equation $r = 2 \sin \theta$ is a circle of radius 1 centered at the point $(0, 1)$ on the y-axis.

We describe below a few classical examples of curves that admit a nice description in polar coordinates.

Examples 7.22.   1. [**Spiral**] The graph of an equation of the type $r = a\theta$ looks like a curve that winds around the origin, and is known as a spiral (of Archimedes). The graph of $r = \theta$ is shown in Figure 7.15.
   2. [**Cardioid**] A polar equation of the type $r = a(1 + \cos \theta)$ gives rise to a heart-shaped curve, known as a cardioid. This curve can also be described

as the locus of a point on the circumference of a circle rolling round the circumference of another circle of equal radius. A sketch of $r = 2(1+\cos\theta)$ is shown in Figure 7.15.
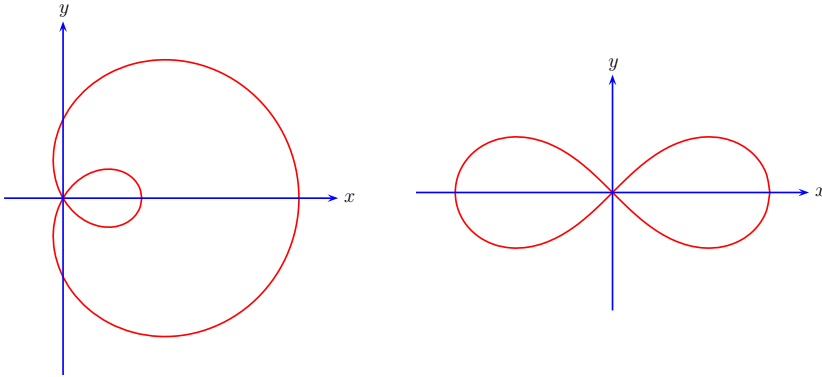
Fig. 7.16. Limaçon $r = 1 + 2\cos\theta$, and the lemniscate $r^2 = 2\cos 2\theta$

3. [**Limaçon**] A polar equation of the form $r = b + a\cos\theta$, which is more general than that of a cardioid, traces a curve known as a limaçon[1] (of Pascal). It looks similar to a cardioid except that instead of a cusp, it has an inner loop (provided $b < a$). A picture of the limaçon $r = 1 + 2\cos\theta$ is shown in Figure 7.16.
4. [**Lemniscate**] A polar equation of the form $r^2 = 2a^2\cos 2\theta$ gives rise to a curve shaped like a figure 8 or a bow tied in a ribbon, which is called a lemniscate[2] (of Bernoulli). A graph of a lemniscate with $a = 1$ is shown in Figure 7.16, and it may be observed that it displays a great deal of symmetry.
5. [**Rose**] Polar equations of the type $r = a\cos n\theta$ or $r = a\sin n\theta$ give rise to floral-shaped curves, known as **rhodonea curves**, or simply roses. Typically, if n is an odd integer, then it has n petals, whereas if n is an even integer, then it has $2n$ petals. Graphs of roses with $a = 1$ and with $n = 4, 5$ are shown in Figure 7.17. The configurations for which n is not an integer are also interesting, albeit more complicated. For example, if n is irrational, then there are infinitely many petals. Varying the values of a, we can obtain different petal lengths.   ◇

---

[1] The name *limaçon* comes from the Latin word limax, which means a snail.
[2] The name *lemniscate* comes from the Latin word lemniscus, meaning a ribbon.
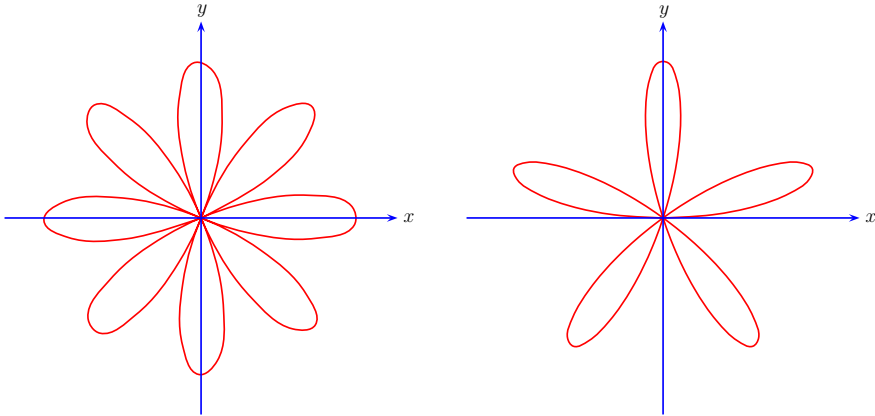
Fig. 7.17. Roses $r = \cos n\theta$ for $n = 4$ (with $2n$ petals) and $n = 5$ (with n petals)

Notion of an Angle

In this subsection, we define the basic notion of an angle in various contexts, and also relate it to polar coordinates discussed above. The formal definition will use the inverse trigonometric functions that are defined in Section 7.2.

To begin with, we consider line segments OP_1 and OP_2 emanating from a common point $O = (x_0, y_0)$. If $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$ are different from O, then we define the **angle** between OP_1 and OP_2 to be the real number

$$\cos^{-1}\left(\frac{(x_1 - x_0)(x_2 - x_0) + (y_1 - y_0)(y_2 - y_0)}{\left(\sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2}\right)\left(\sqrt{(x_2 - x_0)^2 + (y_2 - y_0)^2}\right)}\right).$$

This angle is denoted by $\angle(OP_1, OP_2)$ or by $\angle P_1 O P_2$. Note that by the Cauchy–Schwarz inequality (Proposition 1.12),

$$\left|\frac{(x_1 - x_0)(x_2 - x_0) + (y_1 - y_0)(y_2 - y_0)}{\left(\sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2}\right)\left(\sqrt{(x_2 - x_0)^2 + (y_2 - y_0)^2}\right)}\right| \leq 1.$$

Thus, $\angle(OP_1, OP_2)$ is a well defined real number and it lies between 0 and π. We shall say that the angle between OP_1 and OP_2 is: (i) **acute** if $0 \leq \angle(OP_1, OP_2) < \pi/2$, (ii) **obtuse** if $\pi/2 < \angle(OP_1, OP_2) \leq \pi$, and (iii) a **right angle** if $\angle(OP_1, OP_2) = \pi/2$. Note that the angle between OP_1 and OP_2 is acute, obtuse, or a right angle according as the number $(x_1 - x_0)(x_2 - x_0) + (y_1 - y_0)(y_2 - y_0)$ is positive, negative, or zero, respectively.

Remark 7.23. In classical geometry, the notion of angle is regarded as self-evident and synonymous with the configuration formed by two line segments

emanating from a point. One assigns the degree measure to angles in such a way that the degree measure of the angle between the line segments OP_1 and OP_2 is 180° [read 180 degrees] if O, P_1, and P_2 are collinear and O lies between P_1 and P_2. With this approach, it is far from obvious that every 'angle' is capable of a precise measurement (by real numbers). Such an assumption is also implicit when one 'defines' the trigonometric functions by drawing right-angled triangles and looking at ratios of sides. We have bypassed these difficulties by opting to define the trigonometric functions and the notion of angle by analytic means. In our set up, the degree measure is also easy to define. One simply identifies 180° with π, so that 1° becomes equivalent to the real number $\pi/180$. Thus, the **degree** measure of the angle between the line segments OP_1 and OP_2, denoted $\angle P_1OP_2$, is $(180\alpha/\pi)°$ if $\alpha = \angle(OP_1, OP_2)$. To make a distinction, one sometimes says that α is the **radian**[3] measure of $\angle P_1OP_2$. For example, $\pi/2$, $\pi/3$, $\pi/4$, and $\pi/6$ correspond, in the degree measure, to 90°, 60°, 45°, and 30°, respectively.   ◇

To relate the notion of angle with polar coordinates, let us consider the special case in which O is the origin $(0,0)$, P_1 is the point $A := (1,0)$ on the x-axis, and P_2 is an arbitrary point $P = (x, y)$ other than the origin. Let (r, θ) be the polar coordinates of P. We have seen that $r = \sqrt{x^2 + y^2}$ represents the distance from P to the origin O. On the other hand, the angle between OA and OP is given by

$$\angle(OA, OP) = \cos^{-1}\left(\frac{x \cdot 1 + y \cdot 0}{\left(\sqrt{x^2 + y^2}\right)\left(\sqrt{1^2 + 0^2}\right)}\right) = \cos^{-1}\left(\frac{x}{r}\right).$$

It follows that θ can be interpreted, in analogy with 'signed area' defined in Remark 6.19, as the '**signed angle**' from OA to OP, namely,

$$\theta = \begin{cases} \angle(OA, OP) & \text{if } P \text{ is in the upper half-plane or the } x\text{-axis } (y \geq 0), \\ -\angle(OA, OP) & \text{if } P \text{ is in the lower half-plane } (y < 0). \end{cases}$$

The notion of 'signed angle' is illustrated in Figure 7.18. It may be noted that the sign depends on the 'orientation', that is, it is positive if we move from A to P in the counterclockwise direction (when P is above the x-axis) or negative if we move from A to P in the clockwise direction (when P is below the x-axis).

Now, we shall consider a variant of the notion of angle, which enables us to talk about the angle between two lines rather than two line segments emanating from a common point. Intuitively, it is clear that two intersecting lines give rise to two distinct angles, which are *complementary* in the sense

---

[3] The word *radian*, derived from radius, has the following dictionary meaning: an angle subtended at the center of a circle by an arc whose length is equal to the radius. We can reconcile our current usage of the word with this meaning when the notion of length of an arc is formally defined in Chapter 8.
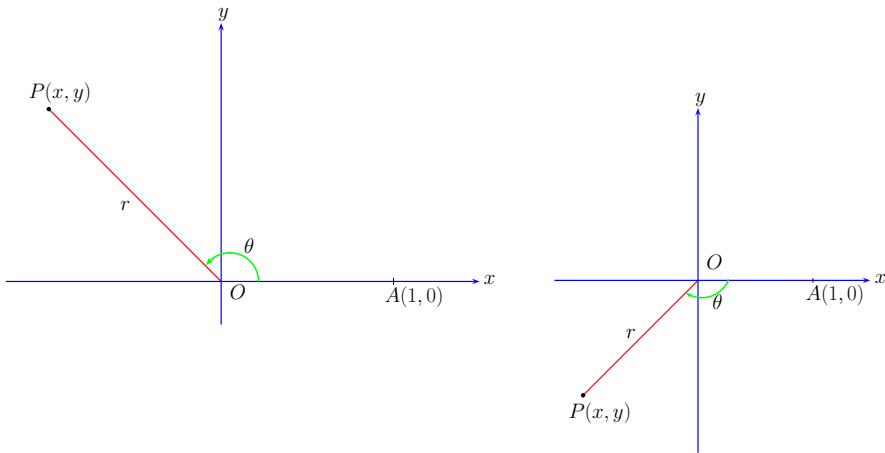
Fig. 7.18. Illustration of the 'signed angle' θ from OA to OP

that their sum is π. As a convention, we shall give preference to the acute angle among these two angles. Formally, we proceed as follows.

Let L_1 and L_2 be any lines in the plane $\mathbb{R}^2$. If $L_1 \parallel L_2$, that is, if L_1 and L_2 are parallel (in particular, if $L_1 = L_2$), then we define the **(acute) angle** between L_1 and L_2, denoted by $\angle(L_1, L_2)$, to be 0. If $L_1 \nparallel L_2$, that is, if L_1 and L_2 are not parallel, then they intersect in a unique point, say $O = (x_0, y_0)$. Now, pick up any point $P_1 = (x_1, y_1)$ on L_1 such that $P_1 \neq O$ and any point $P_2 = (x_2, y_2)$ on L_2 such that $P_2 \neq O$. Define

$$\angle(L_1, L_2) := \cos^{-1}\left(\frac{|(x_1 - x_0)(x_2 - x_0) + (y_1 - y_0)(y_2 - y_0)|}{\left(\sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2}\right)\left(\sqrt{(x_2 - x_0)^2 + (y_2 - y_0)^2}\right)} \right).$$

From the Cauchy–Schwarz inequality (and the conditions when equality holds), it follows that the fraction in the above expression is < 1, and in view of the absolute value in the numerator of this fraction, we see that $0 < \angle(L_1, L_2) \leq \pi/2$. In general, that is, regardless of whether or not $L_1 \parallel L_2$, we have $\angle(L_1, L_2) \in [0, \pi/2]$; also, since $\cos(\pi - \alpha) = -\cos\alpha$ for all $\alpha \in \mathbb{R}$, we have

$$\angle(L_1, L_2) = \begin{cases} \angle(OP_1, OP_2) & \text{if } \angle(OP_1, OP_2) \text{ is acute,} \\ \pi - \angle(OP_1, OP_2) & \text{if } \angle(OP_1, OP_2) \text{ is obtuse.} \end{cases}$$

However, it remains to be seen that when $L_1 \nparallel L_2$, then the above definition of $\angle(L_1, L_2)$ does not depend on the choice of the points P_1, P_2, other than O, on L_1, L_2, respectively. To this end, let us first note that L_i is either the vertical line $x = x_0$, or else, it has a well-defined slope, say m_i, for $i = 1, 2$. Now, if neither L_1 nor L_2 is vertical, then $x_i \neq x_0$ and $m_i = (y_i - y_0)/(x_i - x_0)$

for $i = 1, 2$. So, in this case, dividing the numerator and the denominator of the fraction in the definition of $\angle(L_1, L_2)$ by $|(x_1 - x_0)(x_2 - x_0)|$, we obtain

$$\angle(L_1, L_2) = \cos^{-1}\left(\frac{|1 + m_1 m_2|}{\left(\sqrt{1 + m_1^2}\right)\left(\sqrt{1 + m_2^2}\right)}\right).$$

In case m_1 is not defined, that is, if $x_1 = x_0$, then $y_1 \neq y_0$ (since $P_1 \neq O$) and $x_2 \neq x_0$ (since $L_1 \nparallel L_2$), and hence $|y_1 - y_0| \neq 0$ and m_2 is defined; thus, dividing the numerator and the denominator of the fraction in the definition of $\angle(L_1, L_2)$ by $|x_2 - x_0|$, we obtain

$$\angle(L_1, L_2) = \cos^{-1}\left(\frac{|m_2|}{\left(\sqrt{1 + m_2^2}\right)}\right).$$

Similarly, if m_2 is not defined, that is, if $x_2 = x_0$, then $y_2 \neq y_0$ and $x_1 \neq x_0$, and hence $|y_2 - y_0| \neq 0$ and m_1 is defined; thus, in this case

$$\angle(L_1, L_2) = \cos^{-1}\left(\frac{|m_1|}{\left(\sqrt{1 + m_1^2}\right)}\right).$$

This proves that our definition of $\angle(L_1, L_2)$ is independent of the choice of P_1, P_2, different from O, on L_1, L_2, respectively. In the process, we also obtained alternative expressions for $\angle(L_1, L_2)$. These show in particular that

$$\angle(L_1, L_2) = \pi/2 \iff (x_1 - x_0)(x_2 - x_0) + (y_1 - y_0)(y_2 - y_0) = 0$$
$$\iff \text{(i) } m_1 \text{ and } m_2 \text{ are both defined and } m_1 m_2 = -1, \text{ or}$$
$$\text{(ii) } m_1 \text{ is not defined and } m_2 = 0, \text{ or vice versa.}$$

If any of these equivalent conditions hold, then we shall say that L_1 and L_2 are **perpendicular lines** and write $L_1 \perp L_2$. As usual, we may write $L_1 \not\perp L_2$ to indicate that the lines L_1 and L_2 are not perpendicular.

In a special case, we can obtain another expression for $\angle(L_1, L_2)$ as described below.

Proposition 7.24. *Suppose L_1 and L_2 are nonvertical lines in the plane with slopes m_1 and m_2, respectively. Assume that $L_1 \not\perp L_2$ (so that $m_1 m_2 \neq -1$). Then*

$$\angle(L_1, L_2) = \tan^{-1}\left|\frac{m_1 - m_2}{1 + m_1 m_2}\right|.$$

Proof. If $L_1 \parallel L_2$, then $\angle(L_1, L_2) = 0$ and $m_1 = m_2$. So, the desired equality is clearly true in this case. Now assume that $L_1 \nparallel L_2$. Then L_1 and L_2 intersect in a unique point, say $O = (x_0, y_0)$, and we may choose points $P_i = (x_i, y_i)$ on L_i such that $P_i \neq O$, for $i = 1, 2$. Let $\alpha = \angle(L_1, L_2)$. Then

$$\cos\alpha = \frac{|(x_1 - x_0)(x_2 - x_0) + (y_1 - y_0)(y_2 - y_0)|}{\left(\sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2}\right)\left(\sqrt{(x_2 - x_0)^2 + (y_2 - y_0)^2}\right)}.$$

Now, an easy computation shows that

$$\sin^2\alpha = 1 - \cos^2\alpha = \frac{[(x_1 - x_0)(y_2 - y_0) - (x_2 - x_0)(y_1 - y_0)]^2}{[(x_1 - x_0)^2 + (y_1 - y_0)^2][(x_2 - x_0)^2 + (y_2 - y_0)^2]}.$$

Since $\alpha \in (0, \pi/2]$, it follows that $\sin\alpha > 0$, and thus

$$\sin\alpha = \frac{|(x_1 - x_0)(y_2 - y_0) - (x_2 - x_0)(y_1 - y_0)|}{\left(\sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2}\right)\left(\sqrt{(x_2 - x_0)^2 + (y_2 - y_0)^2}\right)}.$$

Since $L_1 \not\perp L_2$, we see that $\alpha \neq \pi/2$, and so $\cos\alpha \neq 0$. Thus,

$$\tan\alpha = \frac{\sin\alpha}{\cos\alpha} = \frac{|(x_1 - x_0)(y_2 - y_0) - (x_2 - x_0)(y_1 - y_0)|}{|(x_1 - x_0)(x_2 - x_0) + (y_1 - y_0)(y_2 - y_0)|}.$$

In other words,

$$\alpha = \angle(L_1, L_2) = \tan^{-1}\left|\frac{(x_1 - x_0)(y_2 - y_0) - (x_2 - x_0)(y_1 - y_0)}{(x_1 - x_0)(x_2 - x_0) + (y_1 - y_0)(y_2 - y_0)}\right|.$$

Since both L_1 and L_2 are nonvertical, we can divide the numerator and the denominator in the last fraction by $(x_1 - x_0)(x_2 - x_0)$ to obtain the desired equality. $\qquad\square$

The notion of angle between lines can be extended to curves as follows. Suppose C_1 and C_2 are curves in the plane that intersect at a point P. Assume that the tangent, say L_i, to C_i at P is defined for each $i = 1, 2$. Then the **angle at P between the curves C_1 and C_2**, denoted by $\angle(C_1, C_2; P)$, is defined to be $\angle(L_1, L_2)$. In case $\angle(C_1, C_2; P) = \pi/2$, the curves C_1 and C_2 are said to intersect **orthogonally** at the point P.

Examples 7.25. (i) Consider the curves C_1 and C_2 defined by the equations $y = x^2$ and $y = 2 - x^3$. These intersect at the point $P = (1, 1)$. Considering the derivatives at $x = 1$, we see that the slopes of tangents to C_1 and C_2 at P are given by $m_1 = 2$ and $m_2 = -3$, respectively. Hence using Proposition 7.24, we obtain

$$\angle(C_1, C_2; P) = \tan^{-1}\left|\frac{2 - (-3)}{1 + 2(-3)}\right| = \tan^{-1}|-1| = \tan^{-1}1 = \frac{\pi}{4}.$$

Thus the angle between the two curves at P is $\pi/4$.
(ii) Consider the curves C_1 and C_2 defined by the equations $y = e^x$ and $y^2 - 2y + 1 - x = 0$, respectively. These intersect at the point $P = (0, 1)$. The tangent L_1 to C_1 at P is given by the line $y - 1 = x$, whereas the tangent

L_2 to C_2 at P is given by the vertical line $x = 0$. Note that to determine the latter, it is more convenient to look at the derivatives with respect to y than with respect to x. Thus, the slope m_1 of L_1 equals 1, whereas the slope m_2 of L_2 is not defined. Hence the formula in Proposition 7.24 cannot be used. But we can directly use the definition of $\angle(L_1, L_2)$ or the formula $\cos^{-1}\left(|m_1|/\sqrt{1 + m_1^2}\right)$ applicable when m_2 is not defined, to conclude that $\angle(C_1, C_2; P) = \cos^{-1}\left(1/\sqrt{2}\right) = \pi/4$. $\diamond$

7.5 Transcendence

The functions discussed in this chapter, namely, logarithmic, exponential, and trigonometric functions, are often called **elementary transcendental functions**. As we have seen in Chapter 1, the term **transcendental function** has a definite meaning attached to it. It is therefore natural that we should justify this terminology and show that the logarithmic, exponential, and trigonometric functions are indeed transcendental. To do so is the aim of this section.

Let us begin by recalling that given a subset D of $\mathbb{R}$, a function $f : D \to \mathbb{R}$ is said to be a **transcendental function** if it is not an algebraic function, that is, if there is no polynomial

$$P(x, y) = p_n(x)y^n + p_{n-1}(x)y^{n-1} + \cdots + p_1(x)y + p_0(x),$$

where $n \in \mathbb{N}$ and $p_0(x), p_1(x), \ldots, p_n(x)$ are polynomials in x with real coefficients, such that $p_n(x)$ is a nonzero polynomial, and

$$P(c, f(c)) = 0 \quad \text{for all } c \in D.$$

In this case, we refer to $P(x, y)$ as a **polynomial satisfied by** $y = f(x)$ and the positive integer n as the **y-degree** of $P(x, y)$.

Our first goal is to show that the logarithmic function is transcendental. We shall achieve this in two steps. First, we prove a simpler result that the logarithmic function is not a rational function. Next, we will use this to prove that the logarithmic function is not an algebraic function.

Lemma 7.26. *The logarithmic function* $\ln : (0, \infty) \to \mathbb{R}$ *is not a rational function. More precisely, there do not exist polynomials* $p(x), q(x)$ *and an open interval* $I \subseteq (0, \infty)$ *such that*

$$q(x) \neq 0 \quad \text{for all } x \in I \quad \text{and} \quad \ln x = \frac{p(x)}{q(x)} \quad \text{for all } x \in I.$$

Proof. Suppose to the contrary that there are polynomials $p(x), q(x)$ and an open interval $I \subseteq (0, \infty)$ such that $q(x) \neq 0$ for all $x \in I$ and $\ln x = p(x)/q(x)$ for all $x \in I$. Canceling common factors, if any, we may assume that the

polynomials $p(x)$ and $q(x)$ are not divisible by any nonconstant polynomial in x. Since $q(x) \neq 0$ for all $x \in I$, taking derivatives on both sides of the equation

$$\ln x = \frac{p(x)}{q(x)},$$

we obtain for all $x \in I$,

$$\frac{1}{x} = \frac{p'(x)q(x) - p(x)q'(x)}{q(x)^2}, \quad \text{that is, } q(x)^2 = x[p'(x)q(x) - p(x)q'(x)].$$

Both sides of the last equation are polynomials, and the equation is satisfied at infinitely many points; hence it is an identity of polynomials. Consequently, the polynomial x divides the polynomial $q(x)$. Now let us write $q(x) = x^k q_1(x)$, where $k \in \mathbb{N}$ and $q_1(x)$ is a polynomial in x that is not divisible by x, that is, $q_1(0) \neq 0$. Then, $q'(x) = kx^{k-1}q_1(x) + x^k q_1'(x)$, and therefore,

$$x^{2k}q_1(x)^2 = x^{k+1}p'(x)q_1(x) - kx^k p(x)q_1(x) - x^{k+1}p(x)q_1'(x).$$

Dividing throughout by x^k and rearranging terms, we obtain the identity

$$kp(x)q_1(x) = x[p'(x)q_1(x) - p(x)q_1'(x) - x^{k-1}q_1(x)^2].$$

This implies that the polynomial x divides the polynomial $p(x)$, which is a contradiction since $p(x)$ and $q(x)$ were assumed to have no nonconstant common factor. This completes the proof.   □

Proposition 7.27. *The logarithmic function* $\ln : (0, \infty) \to \mathbb{R}$ *is a transcendental function.*

Proof. Assume the contrary, that is, suppose $\ln$ is an algebraic function. Let

$$P(x, y) = p_n(x)y^n + p_{n-1}(x)y^{n-1} + \cdots + p_1(x)y + p_0(x)$$

be a polynomial of y-degree n satisfied by $y = \ln x$ such that $n \in \mathbb{N}$ is the least among the y-degrees of all polynomials satisfied by $y = \ln x$. Let us write

$$q_j(x) := \frac{p_j(x)}{p_n(x)} \quad \text{for } j = 0, 1, \ldots, n-1 \quad \text{and} \quad Q(x, y) := y^n + \sum_{j=0}^{n-1} q_j(x)y^j.$$

Further, let $D := \{c \in (0, \infty) : p_n(c) \neq 0\}$. It is clear that D contains all except finitely many points of $(0, \infty)$, each $q_j(x)$ is defined on D, and $Q(c, \ln c) = 0$ for all $c \in D$. Also, note that every point of D is its interior point. Moreover, each $q_j(x)$ is differentiable on D and its derivative $q_j'(x)$ is a rational function defined on D. Thus, using the Chain Rule (Proposition 4.9), we see that the derivative of $Q(x, \ln x)$ is equal to

$$n(\ln x)^{n-1}\left(\frac{1}{x}\right) + \sum_{j=0}^{n-1}\left[q_j'(x)(\ln x)^j + jq_j(x)(\ln x)^{j-1}\left(\frac{1}{x}\right)\right]$$

$$= \left(q_{n-1}'(x) + \frac{n}{x}\right)(\ln x)^{n-1} + \sum_{j=0}^{n-2}\left(q_j'(x) + \frac{j+1}{x}q_{j+1}(x)\right)(\ln x)^j.$$

Since $q_j'(x) = [p_j'(x)p_n(x) - p_j(x)p_n'(x)]/p_n(x)^2$, taking common denominators, we see that the derivative of $Q(x, \ln x)$ is equal to $\widetilde{P}(x, \ln x)/xp_n(x)^2$, where $\widetilde{P}(x, y)$ is a polynomial in y whose coefficients are polynomials in x. Since $Q(c, \ln c) = 0$ for all $c \in D$, we have $\widetilde{P}(c, \ln c) = 0$ for all $c \in D$. Moreover, since $\widetilde{P}(x, \ln x)$ is defined at every $x \in (0, \infty)$ and gives a continuous function from $(0, \infty)$ to $\mathbb{R}$, which vanishes on D, it follows that $\widetilde{P}(c, \ln c) = 0$ for all $c \in (0, \infty)$. Also, the leading coefficient of $\widetilde{P}(x, y)$, that is, the coefficient of y^{n-1} in $\widetilde{P}(x, y)$, is a nonzero polynomial (in x). For if this leading coefficient were zero, then $q_{n-1}'(t) = -n/t$ for all $t \in D$. Now, since D misses only finitely many points of $(0, \infty)$, in view of Proposition 6.12 and the FTC, we may integrate both sides from $t = 1$ to $t = x$ and obtain $q_{n-1}(x) - q_{n-1}(1) = -n\ln x$ for all $x \in D$, and consequently, $\ln x$ is a rational function (on D), which is impossible by Lemma 7.26. Thus $\widetilde{P}(x, y)$ would be a polynomial satisfied by $y = \ln x$ and its y-degree is $n-1$. This contradicts the minimality of n. Hence $\ln : (0, \infty) \to \mathbb{R}$ is a transcendental function.   □

Corollary 7.28. *The exponential function* $\exp : \mathbb{R} \to \mathbb{R}$ *is a transcendental function.*

Proof. Assume the contrary, that is, suppose exp is an algebraic function. Let

$$P(x, y) = p_n(x)y^n + p_{n-1}(x)y^{n-1} + \cdots + p_1(x)y + p_0(x)$$

be a polynomial satisfied by $y = \exp x$, where $n \in \mathbb{N}$ and $p_n(x)$ is a nonzero polynomial. Let us write $P(x, y)$ as a polynomial in x whose coefficients are polynomials in y:

$$P(x, y) = \widetilde{p}_m(y)x^m + \widetilde{p}_{m-1}(y)x^{m-1} + \cdots + \widetilde{p}_1(y)x + \widetilde{p}_0(y),$$

where m is a nonnegative integer so chosen that $\widetilde{p}_m(y)$ is a nonzero polynomial. Note that m is, in fact, positive because otherwise $P(x, y) = \widetilde{p}_0(y)$ would be a nonzero polynomial in one variable with infinitely many roots, namely, $y = \exp c$ for every $c \in \mathbb{R}$. Now, let $\widehat{P}(x, y) := P(y, x)$. Then $\widehat{P}(x, y)$ is a polynomial in two variables with positive y-degree. Moreover, since $P(c, \exp c) = 0$ for all $c \in \mathbb{R}$, and also since $\exp : \mathbb{R} \to (0, \infty)$ is bijective with its inverse given by $\ln : (0, \infty) \to \mathbb{R}$, it follows that $P(\ln d, d) = 0$ for all $d \in (0, \infty)$. In other words, $\widehat{P}(d, \ln d) = 0$ for all $d \in (0, \infty)$. Thus, $\ln : (0, \infty) \to \mathbb{R}$ would be an algebraic function, which contradicts Proposition 7.27.   □

Now let us turn to trigonometric functions. As we shall see below, it is easier to prove that these are transcendental. The key property used in the proof is that trigonometric functions have infinitely many zeros.

Proposition 7.29. *The trigonometric functions* $\sin : \mathbb{R} \to \mathbb{R}$, $\cos : \mathbb{R} \to \mathbb{R}$, *and* $\tan : \mathbb{R} \setminus \{(2m+1)\pi/2 : m \in \mathbb{Z}\} \to \mathbb{R}$ *are transcendental functions.*

Proof. Assume the contrary, that is, suppose any one of them, say, $\sin : \mathbb{R} \to \mathbb{R}$, is an algebraic function. Then there is a polynomial

$$P(x, y) = p_n(x)y^n + p_{n-1}(x)y^{n-1} + \cdots + p_1(x)y + p_0(x)$$

that is satisfied by $y = \sin x$, and we may assume that $n \in \mathbb{N}$ is the least possible y-degree of such a polynomial. Now, we claim that $p_0(x) = P(x, 0)$ is necessarily a nonzero polynomial in x. To see this, suppose $p_0(x)$ is the zero polynomial. Then we must have $n > 1$. Indeed, were $n = 1$, then $p_1(x) \neq 0$, and since $p_0(x) = 0$, we have $P(x, y) = p_1(x)y$, and hence $p_1(c) \sin c = 0$ for all $c \in \mathbb{R}$. Consequently, $p_1(c) = 0$ for all $c \in \mathbb{R}$ for which $\sin c \neq 0$, that is, for all $c \in \mathbb{R} \setminus \{m\pi : m \in \mathbb{Z}\}$. Hence $p_1(x)$ is the zero polynomial, which is a contradiction. Thus, $n > 1$ and so if we let

$$P_1(x, y) = p_n(x)y^{n-1} + p_{n-1}(x)y^{n-2} + \cdots + p_2(x)y + p_1(x),$$

then the polynomial $P_1(x, y)$ has positive y-degree. Also, $P(x, y) = yP_1(x, y)$, and hence $(\sin c)P_1(c, \sin c) = 0$ for all $c \in \mathbb{R}$. Consequently, $P_1(c, \sin c) = 0$ for all for all $c \in \mathbb{R} \setminus \{m\pi : m \in \mathbb{Z}\}$. But the function from $\mathbb{R}$ to $\mathbb{R}$ defined by $P_1(x, \sin x)$ is continuous. So, it follows that $P_1(c, \sin c) = 0$ for all $c \in \mathbb{R}$. Thus, $y = \sin x$ satisfies the polynomial $P_1(x, y)$ of y-degree $n - 1 \in \mathbb{N}$, which contradicts the minimality of n. Thus, $p_0(x)$ is a nonzero polynomial in x, and consequently it has only finitely many roots. But $p_0(m\pi) = P(m\pi, 0) = P(m\pi, \sin m\pi) = 0$ for all $m \in \mathbb{Z}$, and so we obtain a contradiction. This proves that $\sin : \mathbb{R} \to \mathbb{R}$ is transcendental.

The proof in the case of cosine and tangent functions is similar, since each of them has infinitely many zeros (namely, $(2m+1)\pi/2$ for $m \in \mathbb{Z}$ and $m\pi$ for $m \in \mathbb{Z}$, respectively). □

Remark 7.30. Having justified the term *transcendental* in 'elementary transcendental function', one may wonder whether the term *elementary* should also be justified in the same way. To this effect, we remark that no intrinsic definition of the term *elementary function* appears to be known. In fact, an **elementary function** is usually 'defined' as a function built up from algebraic, exponential, logarithmic, and trigonometric functions and their inverses by a finite combination of the operations of addition, multiplication, division, and root extraction (which are called the elementary operations), and the operation of repeated compositions. ◇

Notes and Comments

The idea of integration, which can be traced back to the work of Archimedes around 225 BC, is one of the oldest and the most fundamental in calculus. The quest for evaluating integrals of known functions is a fruitful way of inventing new functions. When a known function is Riemann integrable (for example, if it is continuous), we can abstractly define its antiderivative. If this cannot be determined in terms of known functions, we obtain, nevertheless, a nice new function waiting to be understood better! As explained in Sections 7.1 and 7.2, $1/x$ and $1/(1 + x^2)$ are the simplest of rational functions whose integrals pose such a problem. This leads to the introduction of the logarithmic function $\ln$ and the arctangent function $\arctan$. With these at our disposal, we can integrate every rational function! This follows from the method of partial fractions, using which any rational function can be decomposed as a sum of simpler rational functions of the form $A/(ax+b)^m$ or $(Bx+C)/(ax^2+bx+c)^m$, and these can be integrated in terms of rational functions and the functions $\ln$ and $\arctan$.

Inverses of logarithmic and arctangent functions lead to even nicer functions, namely, the exponential function, and the tangent function. The other classical trigonometric functions can be easily defined using the tangent function.

The approach outlined above gives not only a precise definition of the logarithmic, exponential and the trigonometric functions, but also a genuine motivation for introducing the same. In most texts on calculus, the trigonometric functions are 'defined' by drawing triangles and mentioning that angles are 'measured' in radians. The main problem with this approach is succinctly described by Hardy [31, §163], who writes: "The whole difficulty lies in the question, what is the x which occurs in $\cos x$ and $\sin x$." Hardy also describes different methods to develop an analytic theory (cf. [31, §224]) and the approach we have chosen is one of them.

The logarithmic and exponential functions also help us to give a 'natural' definition of the important number e. Likewise, the trigonometric functions help us give a precise definition of the important number π. The numbers e and π, and to a lesser extent, Euler's constant γ (defined in Exercise 2 below), have fascinated mathematicians and amateurs alike for centuries. For more on these, see the books of Maor [47], Arndt and Haenel [6], and Havil [34], which are devoted to e, π, and γ, respectively.

Failure to be able to integrate a function has often led to interesting developments in mathematics. For example, a rich and fascinating theory of the so-called elliptic integrals and elliptic functions arises in this way. We will comment more on this in the next chapter, where the notion of arc length will be defined. As another example, we cite the theory of differential equations. Indeed, seeking an antiderivative of a function f may be viewed as the problem of finding a solution $y = F(x)$ of the equation $y' = f$. A differential equation is, more generally, an equation such as $y' = f$ with y' replaced by a combina-

tion of $y, y', y'', \ldots, y^{(n)}$. Attempts to 'solve' differential equations have led to newer classes of functions. To wit, functions known as the Legendre function, Bessel function, and Gauss hypergeometric function arise in this way. These functions enrich the realm of functions beyond algebraic and elementary transcendental functions, and they are sometimes called special functions or higher transcendental functions. For an introduction to these topics, see the books of Simmons [55] and Forsyth [26].

Exercises

Part A

1. For every $x \in \mathbb{R}$ with $x > 1$, show that
$$\sum_{k=1}^{[x]} \frac{1}{k} - \frac{[x]}{x} \leq \ln x \leq \sum_{k=2}^{[x]-1} \frac{1}{k} + \frac{[x]}{x},$$
where $[x]$ denotes the integral part of $[x]$. In particular, show that
$$\frac{13}{22} \leq \ln 2.2 \leq \frac{11}{10} \quad \text{and} \quad 1 \leq \ln 3.6 \leq \frac{17}{10}.$$

2. Consider the sequence (c_n) defined by
$$c_n := 1 + \frac{1}{2} + \cdots + \frac{1}{n} - \ln n \quad \text{for } n \in \mathbb{N}.$$
Show that (c_n) is convergent. (Hint: (c_n) is monotonically decreasing and $c_n \geq 0$ for all $n \in \mathbb{N}$.)
[Note: The limit of the sequence (c_n) is known as **Euler's constant**. It is usually denoted by γ. Approximately, $\gamma = 0.5772156649\ldots$, but it is not known whether γ is rational or irrational.]

3. Let $a > 0$ and $r \in \mathbb{Q}$. Show that $\ln ax^r = \ln a + r \ln x$ for all $x \in (0, \infty)$, assuming only that $(\ln)'x = 1/x$ for all $x \in (0, \infty)$.

4. Show that for all $x > 0$,
$$x - \frac{x^2}{2} < \ln(1 + x) < x - \frac{x^2}{2} + \frac{x^3}{3}.$$

5. Let $\alpha \in \mathbb{R}$ and $f : (0, \infty) \to \mathbb{R}$ be a differentiable function such that $f'(x) = \alpha/x$ for all $x \in (0, \infty)$ and $f(1) = 0$. Show that $f(x) = \alpha \ln x$ for all $x \in (0, \infty)$. (Compare Exercise 4 of Chapter 4.)

6. Let $f : (0, \infty) \to \mathbb{R}$ be continuous and satisfy
$$\int_1^{xy} f(t)dt = y \int_1^x f(t)dt + x \int_1^y f(t)dt \quad \text{for all } x, y \in (0, \infty).$$
Show that $f(x) = f(1)(1+\ln x)$ for all $x \in (0, \infty)$. (Hint: Consider $F(x) := \left(\int_1^x f(t)dt\right)/x$ for $x \in (0, \infty)$ and use Exercise 5.)

7. Show that $2.5 < e < 3$. (Hint: Divide $[1, 2.5]$ and $[1, 3]$ into subintervals of length $\frac{1}{4}$.)

8. Show that

   (i) $\displaystyle\int_a^b \ln x \; dx = b(\ln b - 1) - a(\ln a - 1)$ for all $a, b \in (0, \infty)$,

   (ii) $\displaystyle\int_a^b \exp x \; dx = \exp b - \exp a$ for all $a, b \in \mathbb{R}$.

9. Let $\alpha \in \mathbb{R}$ and $f : \mathbb{R} \to \mathbb{R}$ be a differentiable function such that $f' = \alpha f$ and $f(0) = 1$. Show that $f(x) = e^{\alpha x}$ for all $x \in \mathbb{R}$. (Compare Exercise 5 of Chapter 4.)

10. The **hyperbolic sine** and **hyperbolic cosine** functions from $\mathbb{R}$ to $\mathbb{R}$ are defined by

$$\sinh x := \frac{e^x - e^{-x}}{2} \quad \text{and} \quad \cosh x := \frac{e^x + e^{-x}}{2} \quad \text{for } x \in \mathbb{R}.$$

Show that for any $t \in \mathbb{R}$, the point $(\cosh t, \sinh t)$ is on the hyperbola $x^2 - y^2 = 1$. Also, show that

(i) $\sinh 0 = 0$, $\cosh 0 = 1$ and $\cosh^2 - \sinh^2 = 1$ for all $x \in \mathbb{R}$.

(ii) $(\sinh)' x = \cosh x$ and $(\cosh)' x = \sinh x$ for all $x \in \mathbb{R}$.

(iii) $\sinh(x + y) = \sinh x \cosh y + \cosh x \sinh y$ and
   $\cosh(x + y) = \cosh x \cosh y + \sinh x \sinh y$ for all $x, y \in \mathbb{R}$.

Sketch the graphs of the functions sinh and cosh.

11. Let $a, b \in (0, \infty)$.

(i) Consider the functions $f, g : (0, \infty) \to \mathbb{R}$ defined by $f(x) := \log_a x$ and $g(x) := \log_b x$. Show that f and g have the same rate as $x \to \infty$.

(ii) Consider the functions $f, g : \mathbb{R} \to \mathbb{R}$ defined by $f(x) := a^x$ and $g(x) := b^x$. Show that the growth rate of f is less than that of g as $x \to \infty$ if and only if $a < b$.

12. For $b \in \mathbb{R}$, consider the function $g_b : (0, \infty) \to (0, \infty)$ defined by $g_b(x) = x^b$. Show that $g_{b_1} \circ g_{b_2} = g_{b_1 b_2} = g_{b_2} \circ g_{b_1}$ for all $b_1, b_2 \in \mathbb{R}$.

13. Let $f : (0, \infty) \to \mathbb{R}$ satisfy $f(xy) = f(x)f(y)$ for all $x, y \in (0, \infty)$. If f is continuous at 1, show that either $f(x) = 0$ for all $x \in (0, \infty)$, or there is $r \in \mathbb{R}$ such that $f(x) = x^r$ for all $x \in (0, \infty)$. (Hint: If $f(1) \neq 0$, then $f(x) > 0$ for all $x \in (0, \infty)$, and so we can consider $g = \ln \circ f \circ \exp : \mathbb{R} \to \mathbb{R}$ and use Exercise 4 of Chapter 3.) (Compare Exercise 19 (ii) of Chapter 1 and Exercise 6 of Chapter 3.)

14. Let $r \in \mathbb{R}$ be positive and consider the function $f : (0, \infty) \to \mathbb{R}$ defined by $f(x) = x^r$. Show that the growth rate of $\ln x$ is less than that of f, while the growth rate of $\exp x$ is more than that of f as $x \to \infty$.

15. Show that

$$\frac{x}{1 + x^2} < \arctan x < x \quad \text{for all } x \in (0, 1]$$

and

$$1 - \frac{1}{2x} < \arctan x < 2 - \frac{1}{x} \quad \text{for all } x \in (1, \infty).$$

16. Prove that
$$\lim_{x \to \infty} \int_1^x \frac{1}{1+t^2}\,dt = \int_0^1 \frac{1}{1+t^2}\,dt = \frac{\pi}{4},$$

that is, $\lim_{x \to \infty} \arctan x = \arctan 1 = \pi/4$. Deduce that $2.88 < \pi < 3.39$.
(Hint: Substitute $t = 1/s$ and use Proposition 6.20. Divide $[0,1]$ into
subintervals of length $\frac{1}{4}$.)

17. Let D and E be the unions of open intervals defined as follows.
$$D = \bigcup_{k \in \mathbb{Z}} \left(\frac{(4k-1)\pi}{2}, \frac{(4k+1)\pi}{2} \right) \quad \text{and} \quad E = \bigcup_{k \in \mathbb{Z}} \left(\frac{(4k-3)\pi}{2}, \frac{(4k-1)\pi}{2} \right).$$

Show that
$$\sin x = \begin{cases} \dfrac{\tan x}{\sqrt{1+\tan^2 x}} & \text{if } x \in D, \\[2ex] -\dfrac{\tan x}{\sqrt{1+\tan^2 x}} & \text{if } x \in E, \end{cases} \qquad \cos x = \begin{cases} \dfrac{1}{\sqrt{1+\tan^2 x}} & \text{if } x \in D, \\[2ex] -\dfrac{1}{\sqrt{1+\tan^2 x}} & \text{if } x \in E. \end{cases}$$

18. Show from first principles that the function cos is differentiable at $\pi/2$
and its derivative at $\pi/2$ is -1.

19. Show that $0 < x \cos x < \sin x$ for all $x \in (0, \pi/2)$ and $\sin x < x \cos x < 0$
for all $x \in (-\pi/2, 0)$. Hence or otherwise prove that $x < \tan x$ for all
$x \in (0, \pi/2)$ and $\tan x < x$ for all $x \in (-\pi/2, 0)$.

20. Show that for $x \in (0, \pi/2)$,
$$\frac{2x}{\pi} < \sin x < \min\{1, x\} \quad \text{and} \quad 1 - \frac{2x}{\pi} < \cos x < \min\left\{1, \frac{\pi}{2} - x\right\},$$

whereas for $x \in (-\pi/2, 0)$,
$$\max\{-1, x\} < \sin x < \frac{2x}{\pi} \quad \text{and} \quad 1 + \frac{2x}{\pi} < \cos x < \min\left\{1, \frac{\pi}{2} + x\right\}.$$

21. Prove that $|\sin x - \sin y| \le |x - y|$ and $|\cos x - \cos y| \le |x - y|$ for all
$x, y \in \mathbb{R}$.

22. Show that
$$\int_a^b \sin x\,dx = \cos a - \cos b \quad \text{and} \quad \int_a^b \cos x\,dx = \sin b - \sin a \quad \text{for all } a, b \in \mathbb{R}.$$

23. Let $\beta \in \mathbb{R}$. Suppose $f, g : \mathbb{R} \to \mathbb{R}$ are differentiable functions such that
$$f' = \beta g, \quad g' = -\beta f, \quad f(0) = 0, \quad \text{and} \quad g(0) = 1.$$

Show that $f(x) = \sin \beta x$ and $g(x) = \cos \beta x$ for all $x \in \mathbb{R}$. (Hint: Consider
$h : \mathbb{R} \to \mathbb{R}$ given by $h(x) := (f(x) - \sin \beta x)^2 + (g(x) - \cos \beta x)^2$. Find h'.)
(Compare Exercise 7 of Chapter 4.)

24. Let $\alpha, \beta \in \mathbb{R}$. Suppose $f, g : \mathbb{R} \to \mathbb{R}$ are differentiable functions such that

$$f' = \alpha f + \beta g, \quad g' = \alpha g - \beta f, \quad f(0) = 0, \quad \text{and} \quad g(0) = 1.$$

Show that $f(x) = e^{\alpha x} \sin \beta x$ and $g(x) = e^{\alpha x} \cos \beta x$ for all $x \in \mathbb{R}$. (Hint: Consider $h : \mathbb{R} \to \mathbb{R}$ given by $h(x) := (f(x) - e^{\alpha x} \sin \beta x)^2 + (g(x) - e^{\alpha x} \cos \beta x)^2$. Find h'.) (Compare Exercise 6 of Chapter 4.)

25. Let $\alpha, \beta \in \mathbb{R}$. Suppose $f, g : \mathbb{R} \to \mathbb{R}$ are differentiable functions such that

$$f' = \alpha f + \beta g, \quad g' = \alpha g + \beta f, \quad f(0) = 0, \quad \text{and} \quad g(0) = 1.$$

Show that $f(x) = e^{\alpha x} \sinh \beta x$ and $g(x) = e^{\alpha x} \cosh \beta x$ for all $x \in \mathbb{R}$.

26. Show that $\lim_{x \to 0} (\sin x)/|x|$ does not exist.

27. Prove the following for all $x \in \mathbb{R}$.
$\sin(\pi - x) = \sin x, \quad \sin((\pi/2) - x) = \cos x, \quad \sin((\pi/2) + x) = \cos x,$
$\cos(\pi - x) = -\cos x, \quad \cos((\pi/2) - x) = \sin x, \quad \cos((\pi/2) + x) = -\sin x.$

28. Prove the following for all $x_1, x_2 \in \mathbb{R}$:
(i) $\sin x_1 + \sin x_2 = 2 \sin((x_1 + x_2)/2) \cos((x_1 - x_2)/2)$,
(ii) $\sin x_1 - \sin x_2 = 2 \cos((x_1 + x_2)/2) \sin((x_1 - x_2)/2)$,
(iii) $\cos x_1 + \cos x_2 = 2 \cos((x_1 + x_2)/2) \cos((x_1 - x_2)/2)$,
(iv) $\cos x_1 - \cos x_2 = 2 \sin((x_1 + x_2)/2) \sin((x_2 - x_1)/2)$.

29. Prove the following for all $x \in \mathbb{R}$:
(i) $\sin 2x = 2 \sin x \cos x$,
(ii) $\cos 2x = \cos^2 x - \sin^2 x = 2 \cos^2 x - 1 = 1 - 2 \sin^2 x$,
(iii) $\sin 3x = 3 \sin x - 4 \sin^3 x$,
(iv) $\cos 3x = 4 \cos^3 x - 3 \cos x$.
Deduce that

$$\sin \frac{\pi}{4} = \frac{1}{\sqrt{2}} = \cos \frac{\pi}{4}, \quad \sin \frac{\pi}{3} = \frac{\sqrt{3}}{2} = \cos \frac{\pi}{6}, \quad \cos \frac{\pi}{3} = \frac{1}{2} = \sin \frac{\pi}{6}.$$

30. Prove the following for all $x_1, x_2 \in \mathbb{R}$:
(i) $\sin x_1 = \sin x_2 \iff x_2 = m\pi + (-1)^m x_1$, where $m \in \mathbb{Z}$.
(ii) $\cos x_1 = \cos x_2 \iff x_2 = 2m\pi \pm x_1$, where $m \in \mathbb{Z}$.
(iii) $\sin x_1 = \sin x_2$ and $\cos x_1 = \cos x_2 \iff x_2 = 2m\pi + x_1$, where $m \in \mathbb{Z}$.
(Hint: Exercise 28 and solutions of the equations $\sin x = 0$, $\cos x = 0$.)

31. If $x \in \mathbb{R}$ with $x \neq (2k+1)\pi/2$ for any $k \in \mathbb{Z}$, then show that

$$1 + \tan^2 x = \sec^2 x, \quad (\tan)'x = \sec^2 x, \quad \text{and} \quad (\sec)'x = \sec x \tan x.$$

32. If $x \in \mathbb{R}$ with $x \neq k\pi$ for any $k \in \mathbb{Z}$, then show that

$$1 + \cot^2 x = \csc^2 x, \quad (\cot)'x = -\csc^2 x, \quad \text{and} \quad (\csc)'x = -\csc x \cot x.$$

33. If $x_1, x_2 \in \mathbb{R}$ are such that none of x_1, x_2, and $x_1 + x_2$ equals $(2k+1)\pi/2$ for any $k \in \mathbb{Z}$, then show that

$$\tan(x_1 + x_2) = \frac{\tan x_1 + \tan x_2}{1 - \tan(x_1 + x_2)}.$$

34. Prove the following for all $y_1, y_2 \in \mathbb{R}$:
   (i) $\tan^{-1} y_1 + \tan^{-1} y_2 = \tan^{-1} \left(\dfrac{y_1 + y_2}{1 - y_1 y_2} \right)$ if $y_1 y_2 < 1$,
   (ii) $\tan^{-1} |y_1| + \tan^{-1} |y_2| = \dfrac{\pi}{2}$ if $y_1 y_2 = 1$,
   (iii) $\tan^{-1} |y_1| + \tan^{-1} |y_2| = \tan^{-1} \left(\dfrac{|y_1| + |y_2|}{1 - y_1 y_2} \right)$ if $y_1 y_2 > 1$.

35. Prove the following:
   (i) $\sin \left(\sin^{-1} y \right) = y$ for all $y \in [-1, 1]$ and
   $$\sin^{-1}(\sin x) = \begin{cases} x & \text{if } x \in [-\pi/2, \pi/2], \\ \pi - x & \text{if } x \in (\pi/2, 3\pi/2]. \end{cases}$$
   (ii) $\cos \left(\cos^{-1} y \right) = y$ for all $y \in [-1, 1]$ and $\cos^{-1}(\cos x) = |x|$ for all $x \in [-\pi, \pi]$.

36. If $y \in (-1, 1)$, then show that
   $$\sin^{-1} y = \int_0^y \frac{1}{\sqrt{1 - t^2}} \, dt \quad \text{and} \quad \cos^{-1} y = \frac{\pi}{2} - \int_0^y \frac{1}{\sqrt{1 - t^2}} \, dt.$$
   Deduce that
   $$\lim_{y \to 1^-} \int_0^y \frac{1}{\sqrt{1 - t^2}} \, dt = \frac{\pi}{2}.$$

37. If $y \in (1, \infty)$, then show that
   $$\sec^{-1} y = \lim_{a \to 1^+} \int_a^y \frac{1}{t\sqrt{t^2 - 1}} \, dt \quad \text{and} \quad \csc^{-1} y = \frac{\pi}{2} - \lim_{a \to 1^+} \int_a^y \frac{1}{t\sqrt{t^2 - 1}} \, dt.$$

38. Prove the following:
   (i) $\cot^{-1} y = \dfrac{\pi}{2} - \tan^{-1} y$ for all $y \in \mathbb{R}$,
   (ii) $\csc^{-1} y = \sin^{-1} \dfrac{1}{y}$ for all $y \in \mathbb{R}$ with $|y| \geq 1$,
   (iii) $\sec^{-1} y = \cos^{-1} \dfrac{1}{y}$ for all $y \in \mathbb{R}$ with $|y| \geq 1$.
   (Hint: $\tan^{-1} |y| + \tan^{-1} |1/y| = \pi/2$ for all $y \in \mathbb{R}$ with $y \neq 0$.)

39. For all $y \in [-1, 1]$, show that
   $$\sin^{-1} y + \sin^{-1}(-y) = 0, \ \cos^{-1} y + \cos^{-1}(-y) = \pi, \ \sin^{-1} y + \cos^{-1}(y) = \frac{\pi}{2}$$
   and for all $y \in \mathbb{R}$ with $|y| \geq 1$, show that
   $$\csc^{-1} y + \sec^{-1} y = \frac{\pi}{2}.$$

40. Prove the following:
   (i) $(\cot^{-1})' y = -\dfrac{1}{1 + y^2}$ for all $y \in \mathbb{R}$,

(ii) $(\csc^{-1})'y = -\dfrac{1}{|y|\sqrt{y^2-1}}$ for all $y \in \mathbb{R}$ with $|y| > 1$,

(iii) $(\sec^{-1})'y = \dfrac{1}{|y|\sqrt{y^2-1}}$ for all $y \in \mathbb{R}$ with $|y| > 1$.

41. Let $r_0 \in \mathbb{R}$, and consider the function $f_0 : \mathbb{R} \to \mathbb{R}$ defined by

$$f_0(x) = \begin{cases} \sin(1/x) & \text{if } x \neq 0, \\ r_0 & \text{if } x = 0. \end{cases}$$

(i) Show that f_0 is not continuous at 0. Conclude that the function $x \mapsto \sin(1/x)$ for $x \in \mathbb{R} \setminus \{0\}$ cannot be extended to $\mathbb{R}$ as a continuous function.

(ii) (ii) If I is an interval and $I \subset \mathbb{R} \setminus \{0\}$, then show that f_0 has the IVP on I. If I an interval such that $0 \in I$, then show that f_0 has the IVP on I if and only if $|r_0| \leq 1$.

42. Consider the function $h : \mathbb{R} \to \mathbb{R}$ defined by

$$h(x) := \begin{cases} |x| + |x\sin(1/x)| & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

Show that h has a strict absolute minimum at 0, but for any $\delta > 0$, h is neither decreasing on $(-\delta, 0)$ nor increasing on $(0, \delta)$.

43. Consider the function $g : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ defined by $g(x) := \cos(1/x)$. Prove the following:

(i) g is an even function.

(ii) $\lim_{x \to 0} g(x)$ does not exist, but $\lim_{x \to 0+}[g(x) - g(-x)]$ exists. Also, g cannot be extended to $\mathbb{R}$ as a continuous function.

(iii) For any $\delta > 0$, g is not uniformly continuous on $(0, \delta)$ as well as on $(-\delta, 0)$, but it is uniformly continuous on $(\infty, -\delta] \cup [\delta, \infty)$.

(iv) For any $\delta > 0$, g is not monotonic, not convex, and not concave on $(0, \delta)$ as well as on $(-\delta, 0)$.

44. Let $r_0 \in \mathbb{R}$ and consider the function $g_0 : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ defined by

$$g_0(x) := \begin{cases} \cos(1/x) & \text{if } x \neq 0, \\ r_0 & \text{if } x = 0. \end{cases}$$

Show that g_0 is not continuous at 0. Define $G_0 : \mathbb{R} \to \mathbb{R}$ by $G_0(x) := \int_0^x \cos(1/t)dt$. Show that G_0 is differentiable at 0 and $G_0'(0) = 0$, that is,

$$\lim_{x \to 0} \frac{1}{x} \int_0^x \cos\frac{1}{t}dt = 0.$$

45. Consider the functions $g_1,\ g_2 : \mathbb{R} \to \mathbb{R}$ defined by

$$g_1(x) := \begin{cases} x\cos(1/x) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0, \end{cases} \quad \text{and} \quad g_2(x) := \begin{cases} x^2\cos(1/x) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

Establish properties of g_1 and g_2 similar to those of the functions f_1 and f_2 given in Example 7.18 and Example 7.19, respectively.

46. Consider the function $f : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ defined by $f(x) = (\sin(1/x))/x$. Show that the amplitude of the oscillation of the function f increases without any bound as x tends to 0.

47. Consider the function $f : \mathbb{R} \to \mathbb{R}$ defined by

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

Show that f is differentiable on $\mathbb{R}$, but for any $\delta > 0$, f' is not bounded on $[-\delta, \delta]$. Thus f' has an antiderivative on the interval $[-1, 1]$, but it is not Riemann integrable on $[-1, 1]$.

48. Let $n \in \mathbb{N}$ and consider the function $f_n : \mathbb{R} \to \mathbb{R}$ defined by

$$f_n(x) := \begin{cases} x^n \sin(1/x) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

Prove the following: (i) If n is odd and $k := (n-1)/2$, then $f_n^{(k)}$ exists and is continuous on $\mathbb{R}$, but $f_n^{(k+1)}$ does not exist at 0. (ii) If n is even and $k := n/2$, then $f_n^{(k)}$ exists on $\mathbb{R}$, but it is not continuous at 0. (Compare Exercise 12 of Chapter 4.)

49. Find the polar coordinates of the points in $\mathbb{R}^2$ whose Cartesian coordinates are as follows:
(i) $(1, 1)$,   (ii) $(0, 3)$,   (iii) $(2, 2\sqrt{3})$,   (iv) $(2\sqrt{3}, 2)$.

50. If $x, y \in \mathbb{R}$ are not both zero and (r, θ) are the polar coordinates of (x, y), then determine the polar coordinates of (i) (y, x), and (ii) (tx, ty), where t is any positive real number.

51. Let r be a positive real number and $\theta \in (-\pi, \pi]$ and $\alpha \in \mathbb{R}$ be such that $\theta + \alpha \in (-\pi, \pi]$. If P and P_α denote the points with polar coordinates (r, θ) and $(r, \theta + \alpha)$, respectively, then find the Cartesian coordinates of P_α in terms of the Cartesian coordinates of P.
[Note: The transformation $P \mapsto P_\alpha$ corresponds to a rotation of the plane by the angle α.]

52. Find the angle(s) between the curves $x^2 + y^2 = 16$ and $y^2 = 6x$ at their point(s) of intersection.

53. Determine whether the following functions are algebraic or transcendental:
(i) $f(x) = \pi x^{11} + \pi^2 x^5 + 9$ for $x \in \mathbb{R}$,
(ii) $f(x) = \dfrac{ex^2 + \pi}{\pi x^2 + e}$ for $x \in \mathbb{R}$,
(iii) $f(x) = \ln_{10} x$ for $x > 0$,
(iv) $f(x) = x^\pi$ for $x > 0$.

54. Is it possible that

$$\ln x = \left(\sqrt[3]{ex^2 + (\pi - 2e)x + e - \pi} + \sqrt{\pi x^2 + (\sqrt{2} - 2\pi)x + \pi - \sqrt{2}} \right)^{1/17}$$

for all $x > 0$? Justify your answer.

Part B

55. Let p, $q \in (1, \infty)$ be such that $(1/p) + (1/q) = 1$.

(i) If $f : [0, \infty) \to \mathbb{R}$ is defined by $f(x) := (1/q) + (1/p)\, x - x^{1/p}$, then show that $f(x) \geq f(1)$ for all $x \in [0, \infty)$.

(ii) Show that $ab \leq (a^p/p) + (b^q/q)$ for all $a, b \in [0, \infty)$. (Hint: If $b \neq 0$, let $x := a^p/b^q$ in (i).)

(iii) **(Hölder Inequality for Sums)** Given any $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ in $\mathbb{R}$, prove that

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

Deduce the Cauchy–Schwarz inequality as a special case.

(iv) **(Hölder Inequality for Integrals)** Given any continuous functions $f, g : [a, b] \to \mathbb{R}$, prove that

$$\int_a^b |f(x)g(x)| dx \leq \left(\int_a^b |f(x)|^p dx \right)^{1/p} \left(\int_a^b |g(x)|^q dx \right)^{1/q}.$$

(v) **(Minkowski Inequality for Sums)** Given any $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ in $\mathbb{R}$, prove that

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

(Hint: The pth power of the expression on the left can be written as $\sum_{i=1}^{n} |a_i| \left(|a_i + b_i| \right)^{p-1} + \sum_{i=1}^{n} |b_i| \left(|a_i + b_i| \right)^{p-1}$; now use (iii).)

(vi) **(Minkowski Inequality for Integrals)** Given any continuous functions $f, g : [a, b] \to \mathbb{R}$, prove that

$$\left(\int_a^b |f(x) + g(x)|^p dx \right)^{1/p} \leq \left(\int_a^b |f(x)|^p dx \right)^{1/p} + \left(\int_a^b |g(x)|^p dx \right)^{1/p}.$$

56. Let $n \in \mathbb{N}$. By applying L'Hôpital's rule n times, prove the following:

(i) $\displaystyle \lim_{x \to 0} \frac{\exp x - \sum_{k=0}^{n} x^k/k!}{x^{n+1}} = \frac{1}{(n+1)!}$,

(ii) $\displaystyle \lim_{x \to 1} \frac{\ln x - \sum_{k=1}^{n} (-1)^k (x-1)^k/k}{(x-1)^{n+1}} = \frac{(-1)^n}{(n+1)}$,

(iii) $\displaystyle \lim_{x \to 0} \frac{\sin x - \sum_{k=0}^{\lceil (n-2)/2 \rceil} (-1)^k x^{2k+1}/(2k+1)!}{x^{n+1}} = \begin{cases} \dfrac{(-1)^{n/2}}{(n+1)!} & \text{if } n \text{ is even}, \\[2ex] 0 & \text{if } n \text{ is odd}, \end{cases}$

$$\text{(iv)} \lim_{x \to 0} \frac{\cos x - \sum_{k=0}^{\lfloor n/2 \rfloor} (-1)^k x^{2k}/(2k)!}{x^{n+1}} = \begin{cases} \dfrac{(-1)^{(n+1)/2}}{(n+1)!} & \text{if } n \text{ is odd,} \\ \\ 0 & \text{if } n \text{ is even.} \end{cases}$$

57. Let $p, q \in \mathbb{N}$. For $n \in \mathbb{N}$, consider the function $f_n : [0, p/q] \to \mathbb{R}$ defined by $f_n(x) := x^n (p - qx)^n/n!$. Prove the following results:
   (i) $f_n(0) = 0 = f_n(p/q)$. Also, $f_n^{(k)}(0) = -f_n^{(k)}(p/q) \in \mathbb{Z}$ for each $k \in \mathbb{N}$; in fact, $f_n^{(k)}(0) = 0 = f_n^{(k)}(p/q)$ if $k \le n$ or $k > 2n$.
   (ii) $\max\{f_n(x) : x \in [0, p/q]\} = f_n(p/2q)$, and $f_n(p/2q) \to 0$ as $n \to \infty$.
   (iii) Let, if possible, $\pi = p/q$, and consider $a_n := \int_0^\pi f_n(x) \sin x \, dx$. Then $a_n \in \mathbb{Z}$ for each $n \in \mathbb{N}$ (by repeated use of Integration by Parts), whereas $0 < a_n < 1$ for all large $n \in \mathbb{N}$.
   (iv) π is irrational.

58. (i) Show that for any $n \in \mathbb{N}$, $\displaystyle \int_0^{\pi/2} \sin^n x \, dx = \frac{n-1}{n} \int_0^{\pi/2} \sin^{n-2} x \, dx$.
   (ii) Show that for any $k \in \mathbb{N}$,

   $$\int_0^{\pi/2} \sin^{2k} x \, dx = \frac{(2k-1)(2k-3)\cdots 3 \cdot 1}{(2k)(2k-2)\cdots 4 \cdot 2} \cdot \frac{\pi}{2} = \frac{(2k)!}{[2^k k!]^2} \cdot \frac{\pi}{2}$$

   and

   $$\int_0^{\pi/2} \sin^{2k+1} x \, dx = \frac{2k(2k-2)\cdots 4 \cdot 2}{(2k+1)(2k-1)\cdots 3 \cdot 1} = \frac{[2^k k!]^2}{(2k+1)!}.$$

   (iii) For $k \in \mathbb{N}$, let

   $$\mu_k := \frac{\int_0^{\pi/2} \sin^{2k} x \, dx}{\int_0^{\pi/2} \sin^{2k+1} x \, dx}.$$

   Show that $1 \le \mu_k \le (2k+1)/2k$ for each $k \in \mathbb{N}$ and consequently that $\mu_k \to 1$ as $k \to \infty$. Deduce that

   $$\sqrt{\pi} = \lim_{k \to \infty} \frac{(k!)^2 2^{2k}}{(2k)! \sqrt{k}}.$$

   Thus, $\pi \sim (k!)^4 2^{4k}/[(2k)!]^2 \, k$. (Hint: $\sin^{2k+1} x \le \sin^{2k} x \le \sin^{2k-1} x$ for all $x \in [0, \pi/2]$.)
   [Note: This result is known as the **Wallis formula**.]

59. (i) Show that for any $n \in \mathbb{N}$,

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

   (ii) Let (a_n) be the sequence defined by $a_n := n! e^n / n^n \sqrt{n}$ for $n \in \mathbb{N}$. Show that

$$\ln\left(\frac{a_n}{a_{n+1}}\right) = \left(n + \frac{1}{2}\right)\ln\left(1 + \frac{1}{n}\right) - 1,$$

and hence

$$1 \le \frac{a_n}{a_{n+1}} \le \exp\left(\frac{1}{4}\left[\frac{1}{n} - \frac{1}{n+1}\right]\right) \qquad \text{for all } n \in \mathbb{N}.$$

Deduce that (a_n) is a monotonically decreasing sequence of positive real numbers and it is convergent. Let $\alpha := \lim_{n\to\infty} a_n$.

(iii) Use the inequalities in (ii) to show that

$$1 \le \frac{a_n}{a_{n+k}} \le \exp\left(\frac{1}{4}\left[\frac{1}{n} - \frac{1}{n+k}\right]\right) \qquad \text{for all } n, k \in \mathbb{N}.$$

Taking the limit as $k \to \infty$, deduce that $\alpha > 0$ and furthermore, $1 \le (a_n/\alpha) \le \exp(1/4n)$ for all $n \in \mathbb{N}$.

(iv) Show that the Wallis formula given in Exercise 58 can be written as $\sqrt{2\pi} = \lim_{n\to\infty} a_n^2/a_{2n}$. Deduce that $\alpha = \sqrt{2\pi}$.

(v) Use (iii) and (iv) to show that for all $n \in \mathbb{N}$,

$$(\sqrt{2\pi})n^{n+\frac{1}{2}}e^{-n} \le n! \le (\sqrt{2\pi})n^{n+\frac{1}{2}}e^{-n+(1/4n)}$$

and conclude that

$$\lim_{n\to\infty} \frac{n!}{(\sqrt{2\pi n})n^n e^{-n}} = 1.$$

Thus, $n! \sim (\sqrt{2\pi n})n^n e^{-n}$.

[Note: This result is known as **Stirling's Formula**.]

60. Let $r, s \in \mathbb{R}$ and consider the function $F : [0, 1] \to \mathbb{R}$ defined by

$$F(x) := \begin{cases} x^r \sin(1/x^s) & \text{if } x \ne 0, \\ 0 & \text{if } x = 0. \end{cases}$$

Prove the following:

(i) F is continuous $\iff r > 0$.

(ii) F is differentiable $\iff r > 1$.

(iii) F' is bounded $\iff r \ge 1 + s$.

(iv) F' is continuous $\iff r > 1 + s$.

(v) F is twice differentiable $\iff r > 2 + s$.

(vi) F'' is bounded $\iff r \ge 2 + 2s$.

(vii) F'' is continuous $\iff r > 2 + 2s$.

61. Prove that the secant function, the cosecant function, and the cotangent function are transcendental.

Revision Exercises

1. Consider the functions $f, g : \mathbb{R} \to \mathbb{R}$ defined by $f(x) := x \sin x$ and $g(x) := x + \sin x$. State whether f and g are bounded.
2. Consider the sequence whose nth term is given below. Examine whether it is convergent. In case it is convergent, find its limit.
   (i) $\dfrac{n!}{10^n}$,   (ii) $\left(\dfrac{n}{n+1}\right)^n$,   (iii) $\dfrac{\ln n}{n^{1/n}}$.
3. Show that
$$\lim_{m\to\infty} \lim_{n\to\infty} |(\cos m!\pi x)^n| = \begin{cases} 1 & \text{if } x \in \mathbb{Q}, \\ 0 & \text{if } x \notin \mathbb{Q}. \end{cases}$$
4. Suppose (a_n) is a sequence of positive real numbers such that $a_n \to a$. Show that $(a_1 \cdots a_n)^{1/n} \to a$. Here $a \in \mathbb{R}$ or $a = \infty$. Give an example to show that the converse does not hold. (Hint: Exercise 21 of Chapter 2.)
5. Consider the function $f : (-\pi/2, \pi/2) \to \mathbb{R}$ defined by $f(x) = \tan x$. Show that f is not uniformly continuous on $[0, \pi/2)$, but for any $\delta > 0$, f is uniformly continuous on $[-(\pi/2) + \delta, (\pi/2) - \delta]$.
6. For $x \in \mathbb{R}$, let $f(x) := x(\sin x + 2)$ and $g(x) := x(\sin x + 1)$. Show that $f(x) \to \infty$, but $g(x) \not\to \infty$ as $x \to \infty$.
7. Find $f'(x)$ if (i) $f(x) := x^x$ for $x > 0$,   (ii) $f(x) := (x^x)^x$ for $x > 0$, (iii) $f(x) := x^{(x^x)}$ for $x > 0$,   (iv) $f(x) := (\ln x)^x / x^{\ln x}$ for $x > 1$.
8. Let $a > 0$. Show that $-x^a \ln x < 1/ae$ for all $x \in (0,1), x \neq e^{-1/a}$.
9. Let $r, s, t \in \mathbb{R}$ and $x \in (0, \infty)$. If $r > 1$, then show that $(1 + x)^r > 1 + x^r$. Deduce that if $0 < s < t$, then $(1 + x^s)^t > (1 + x^t)^s$.
10. Let $f : [0, \pi/2] \to \mathbb{R}$ be a continuous function.
    (i) If f satisfies $f'(x) = 1/(1 + \cos x)$ for all $x \in (0, \pi/2)$ and if $f(0) = 3$, then find an estimate for $f(\pi/2)$.
    (ii) If f satisfies $f'(x) = 1/(1 + x \sin x)$ for all $x \in (0, \pi/2)$ and if $f(0) = 1$, then find an estimate for $f(\pi/2)$.
11. Prove that $(\pi/15) < \tan(\pi/4) - \tan(\pi/5) < (\pi/10)$. Hence conclude that
$$\frac{10 - \pi}{10} < \tan\frac{\pi}{5} < \frac{15 - \pi}{15}.$$
12. Consider the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) := x + \sin x$. Show that f is strictly increasing on $\mathbb{R}$ although f' vanishes at infinitely many points. Find intervals of convexity/concavity, and points of inflection for f.
13. Consider the function $g : \mathbb{R} \to \mathbb{R}$ defined by $g(x) := x^2 - 2\cos x$. Show that g is strictly convex on $\mathbb{R}$ although g'' vanishes at infinitely many points. Find intervals of increase/decrease and local extrema of g. Does g have an absolute minimum?
14. Locate intervals of increase/decrease, intervals of convexity/concavity, local maxima/minima, and the points of inflection for $f : (0, \infty) \to \mathbb{R}$ defined by $f(x) := (\ln x)/x$. Sketch the curve $y = f(x)$.

15. Locate intervals of increase/decrease of the following functions:

    (i) $f(x) := x^{1/x}$, $x \in (0, \infty)$,   (ii) $f(x) := \left(1 + \dfrac{1}{x}\right)^x$, $x \in (0, \infty)$.

16. Determine which of the two numbers e^π and π^e is greater. (Hint: Find the absolute minimum of the function defined by $f(x) := x^{1/x}$ for $x \in (0, \infty)$ and put $x = \pi$; alternatively, find the absolute minimum of the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) := e^x - 1 - x$ and put $x := (\pi/e) - 1$.)

17. Consider the functions $f, g : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ defined by $f(x) := \sin(1/x)$ and $g(x) := \cos(1/x)$. Locate intervals of increase/decrease, intervals of convexity/concavity, local maxima/minima, and the points of inflection for f and g.

18. Evaluate the following limits:

    (i) $\lim\limits_{x \to 0+} x \ln x$,   (ii) $\lim\limits_{x \to 0} \dfrac{\ln x}{x}$,   (iii) $\lim\limits_{x \to \infty} (x - \ln x)$,   (iv) $\lim\limits_{x \to \infty} \dfrac{\ln(\ln x)}{\ln x}$,

    (v) $\lim\limits_{x \to \infty} \dfrac{x^5}{e^x}$,   (vi) $\lim\limits_{x \to \infty} \dfrac{2^x - 1}{2^x + 3}$,   (vii) $\lim\limits_{x \to 0} \dfrac{3^{\sin x} - 1}{x}$.

19. Evaluate the following limits:

    (i) $\lim\limits_{x \to 0} \dfrac{x - \sin^{-1} x}{\sin^3 x}$,   (ii) $\lim\limits_{x \to \pi/2} (\sec x - \tan x)$,   (iii) $\lim\limits_{x \to 0} \dfrac{x - \tan x}{x - \sin x}$,

    (iv) $\lim\limits_{x \to 0} \dfrac{x \cot x - 1}{x^2}$,   (v) $\lim\limits_{x \to \pi/2} \dfrac{\tan 3x}{\tan x}$,   (vi) $\lim\limits_{x \to 1} (1 - x) \tan(\pi x/2)$,

    (vii) $\lim\limits_{x \to 0} \sin^{-1} x \cot x$,   (viii) $\lim\limits_{x \to 0} \dfrac{\cos x - 1 + (x^2/2)}{x^4}$,   (ix) $\lim\limits_{x \to 0} \dfrac{\tan x}{\sec x}$,

    (x) $\lim\limits_{x \to 0} \left(\dfrac{1}{x} - \dfrac{1}{\sin x}\right)$,   (xi) $\lim\limits_{x \to 0} \dfrac{\sin 2x}{2x^2 + x}$,   (xii) $\lim\limits_{x \to 0} \dfrac{\sin x - x}{x}$,

    (xiii) $\lim\limits_{x \to 0} \dfrac{\sin x - x}{x^2}$,   (xiv) $\lim\limits_{x \to 0} \dfrac{\sin x - x}{x^3}$.

20. Discuss whether $\lim_{x \to c} f(x)/g(x)$ and $\lim_{x \to c} f'(x)/g'(x)$ exist if
    (i) $c := 0$, $f(x) := x^2 \sin(1/x)$, $g(x) := \sin x$ for $x \in \mathbb{R}$, $x \neq 0$,
    (ii) $c := 0$, $f(x) := x \sin(1/x)$, $g(x) := \sin x$ for $x \in \mathbb{R}$, $x \neq 0$,
    (iii) $c := \infty$, $f(x) := x(2 + \sin x)$, $g(x) := x^2 + 1$ for $x \in \mathbb{R}$,
    (iv) $c := \infty$, $f(x) := x(2 + \sin x)$, $g(x) := x + 1$ for $x \in \mathbb{R}$.

21. Let $a > 0$ and $f : [a, \infty) \to \mathbb{R}$ be a differentiable function. Assume that $f(x) + f'(x) \to \ell$ as $x \to \infty$, where $\ell \in \mathbb{R}$ or $\ell = \infty$ or $\ell = -\infty$. Show that $f(x) \to \ell$ as $x \to \infty$, In the case $\ell \in \mathbb{R}$, show that $f'(x) \to 0$ as $x \to \infty$. (Hint: Use Proposition 4.40 for the functions $g, h : [a, \infty) \to \mathbb{R}$ defined by $g(x) := f(x)e^x$ and $h(x) := e^x$.)

22. Evaluate the following limits:

    (i) $\lim\limits_{x \to 0+} (\sin x)^{\tan x}$,   (ii) $\lim\limits_{x \to \infty} (x^2)^{1/\sqrt{x}}$,   (iii) $\lim\limits_{x \to (\pi/2)-} (\sin x)^{\tan x}$.

23. For $x \in \mathbb{R}$, let $f(x) := 2x + \sin 2x$ and $g(x) := f(x)/(2 + \sin x)$. Do

$$\lim_{x \to \infty} \frac{f(x)}{g(x)} \quad \text{and} \quad \lim_{x \to \infty} \frac{f'(x)}{g'(x)}$$

exist? Explain in view of L'Hôpital's Rules.

24. For $x \in (0, \infty)$, let $f(x) := \ln x$ and $g(x) := x$. Show that

$$\lim_{x \to 0+} \frac{f(x)}{g(x)} = -\infty \quad \text{and} \quad \lim_{x \to 0+} \frac{f'(x)}{g'(x)} = \infty.$$

Explain in view of L'Hôpital's Rules.

25. Arrange the following functions in descending order of their growth rates as $x \to \infty$:

$$2^x, e^x, x^x, (\ln x)^x, e^{x/2}, x^{1/2}, \log_2 x, \ln(\ln x), (\ln x)^2, x^e, x^2, \ln x, (2x)^x, x^{2x}.$$

26. Let $n \in \mathbb{N}$ and $a_1, \ldots, a_n$ be positive real numbers. Prove that

$$\lim_{x \to 0} \left(\frac{a_1^x + \cdots + a_n^x}{n} \right)^{1/x} = (a_1 \cdots a_n)^{1/n}.$$

(Hint: Apply the logarithm and use L'Hôpital's Rule.)

27. Let $n \in \mathbb{N}$ and $a_1, \ldots, a_n$ be positive real numbers. For any $p \in \mathbb{R}$ such that $p \neq 0$, define

$$M_p = \left(\frac{a_1^p + \cdots + a_n^p}{n} \right)^{1/p}.$$

In view of Exercise 26 above, define $M_0 = (a_1 \cdots a_n)^{1/n}$. Prove that if $p, q \in \mathbb{R}$ are such that $p < q$, then $M_p \leq M_q$, and the equality holds if and only if $a_1 = \cdots = a_n$. (Hint: Use part (ii) of Proposition 7.9 and Jensen's inequality stated in Exercise 34 (ii) of Chapter 1.)
[Note: As mentioned in Exercise 57 of Chapter 1, the above inequality is called the **power mean inequality** and it includes the A.M.-G.M. inequality and the G.M.-H.M. inequality as special cases. This inequality is also valid for $p = -\infty$ and $q = \infty$ if we set $M_{-\infty} := \min\{a_1, \ldots, a_n\}$ and $M_\infty := \max\{a_1, \ldots, a_n\}$.]

28. Let $D \subseteq \mathbb{R}$ and c be an interior point of D. Let $f : D \to \mathbb{R}$ be a function that is differentiable at c. If $f''(c)$ exists, then show that there is a function $f_2 : D \to \mathbb{R}$ such that f_2 is continuous at c and

$$f(x) = f(c) + (x - c)f'(c) + (x - c)^2 f_2(x) \quad \text{for all } x \in D,$$

and then $f_2(c) = f''(c)/2$. Give an example to show that the converse is not true. (Hint: Use L'Hôpital's Rule and the function $f_n : \mathbb{R} \to \mathbb{R}$ with $n = 3$ given in Exercise 48 of Chapter 7.) (Compare Proposition 4.2.)

29. Consider the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) := e^{-1/x^2}$ if $x \neq 0$ and $f(0) = 0$. Show that for each $n \in \mathbb{N}$, the nth Taylor polynomial around 0 for f is the zero polynomial. (Hint: By mathematical induction, prove that for every $n \in \mathbb{N}$, $f^{(n)}(x) = f(x)p_n(x)/x^{k_n}$ for every $x \in \mathbb{R} \setminus \{0\}$, where p_n is a polynomial and $k_n \in \mathbb{N}$, and then use L'Hôpital's Rule.)

30. Find the absolute maximum of the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) := (\sin x - \cos x)^2$.

31. Prove the following estimates for the errors $e_1(x) := \ln x - P_1(x)$ and $e_2(x) := \ln x - P_2(x)$, $x \in (0, 2)$, in the linear and the quadratic approximations of the function ln around 1:

$$|e_1(x)| \leq \frac{(x-1)^2}{2}, \quad |e_2(x)| \leq \frac{(x-1)^3}{3} \quad \text{if } 1 < x < 2,$$

and

$$|e_1(x)| \leq \frac{1}{2}\left(\frac{1}{x} - 1\right)^2, \quad |e_2(x)| \leq \frac{1}{3}\left(\frac{1}{x} - 1\right)^3 \quad \text{if } 0 < x < 1.$$

32. Prove the following estimates for the errors $e_1(x) := \exp x - P_1(x)$ and $e_2(x) := \exp x - P_2(x)$, $x \in (-1, 1)$, in the linear and the quadratic approximations of the function exp around 0:

$$|e_1(x)| \leq \frac{e\, x^2}{2} \quad \text{and} \quad |e_2(x)| \leq \frac{e\, x^3}{6} \quad \text{if } 0 < x < 1,$$

and

$$|e_1(x)| \leq \frac{x^2}{2} \quad \text{and} \quad |e_2(x)| \leq -\frac{x^3}{6} \quad \text{if } -1 < x < 0.$$

33. Show that each of the following functions maps the given interval I into itself and has a unique fixed point in that interval. Also, show that if x_0 belongs to this interval, then the Picard sequence with initial point x_0 converges to the unique fixed point of the function.
    (i) $g(x) := \sqrt{\sin x}, \ I = [\pi/4, \pi/2]$, \quad (ii) $g(x) := 1 + (\sin x)/2, \ I = [0, 2]$.

34. (i) Show that 0 is the only fixed point of the function $\sin : \mathbb{R} \to \mathbb{R}$.
    (ii) Show that the function $\cos : \mathbb{R} \to \mathbb{R}$ has a unique fixed point c^*. Assuming $\pi > 3$, show that the function cos maps the interval $[\pi/8, 1]$ into itself. Deduce that $0.375 < c^* < 0.925$.
    [Note: When a calculator is in the radian mode, if we key in any number and press the 'sin' key repeatedly, then eventually we reach 0, and if we press the 'cos' key repeatedly, then eventually we reach 0.7390851. A similar phenomenon occurs when a calculator is in the degree mode.]

35. For each of the following functions, show that the equation $f(x) = 0$ has a unique solution in the given interval I. Use Newton's method with the given initial point x_0 to find an approximate value of this root.
    (i) $f(x) := x - \cos x, \quad I = [\cos 1, 1], \quad x_0 = 1$,
    (ii) $f(x) := x - 1 - (\sin x)/2, \quad I = [0, 2], \quad x_0 = 1.5$. (Compare these iterates with those of the Picard method obtained in Exercise 33 (ii).)

36. For all $h \in \mathbb{R}$ and $n = 1, 2, \ldots$, show that

$$2 \sin \frac{h}{2} [\sin h + \sin 2h + \cdots + \sin nh] = \cos \frac{h}{2} - \cos \left(n + \frac{1}{2}\right)h.$$

Hence find $\int_0^{\pi/2} \sin x \, dx$ without using the FTC.

37. Let $n \in \mathbb{N}$ and $a_0, a_1, \ldots, a_n, b_1, \ldots, b_n$ be real numbers and consider the function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) := a_0 + \sum_{k=1}^{n} a_k \cos kx + b_k \sin kx.$$

Show that

$$a_0 = \frac{1}{2\pi} \int_0^{2\pi} f(x) \, dx$$

and for $k = 1, \ldots, n$,

$$a_k = \frac{1}{\pi} \int_0^{2\pi} f(x) \cos kx \, dx, \quad b_k = \frac{1}{\pi} \int_0^{2\pi} f(x) \sin kx \, dx.$$

38. Let $f : [0, \pi] \to \mathbb{R}$ be defined by

$$f(x) := \begin{cases} \dfrac{[\sin(2n+1)x/2]}{\sin(x/2)} & \text{if } x \neq 0, \\[2mm] 2n+1 & \text{if } x = 0. \end{cases}$$

Show that $\int_0^\pi f(x)dx = \pi$. (Hint: The integrand equals $1 + 2\sum_{k=1}^{n} \cos kx$.)

39. Let $f : [a, b] \to \mathbb{R}$ be differentiable and assume that f' is Riemann integrable on $[a, b]$. If $f(x) > 0$ for all $x \in [a, b]$, then show that

$$\int_a^b \frac{f'(x)}{f(x)} dx = \ln f(b) - \ln f(a).$$

(Hint: Apply part (i) of the FTC to $\ln f$.)

40. Prove the following:

(i) $\displaystyle\int_a^b \frac{1}{x - \alpha} dx = \ln \frac{b - \alpha}{a - \alpha}$, provided $a, b > \alpha$,

(ii) $\displaystyle\int_a^b \frac{2x + \alpha}{x^2 + \alpha x + \beta} dx = \ln \frac{b^2 + \alpha b + \beta}{a^2 + \alpha a + \beta}$, provided $\alpha^2 < 4\beta$.

41. Prove the following:

(i) $\displaystyle\int_0^b \tan x \, dx = \ln \sec b$, provided $b \in (-(\pi/2), (\pi/2))$,

(ii) $\displaystyle\int_0^b \sec x \, dx = \ln(\sec b + \tan b)$, provided $b \in (-(\pi/2), (\pi/2))$,

(iii) $\displaystyle\int_b^{\pi/2} \cot x \, dx = \ln \csc b$, provided $b \in (0, \pi)$,

(iv) $\displaystyle\int_b^{\pi/2} \csc x \, dx = \ln(\csc b + \cot b)$, provided $b \in (0, \pi)$.

In particular, show that

$$\int_0^{\pi/4} \tan x \, dx = \ln \sqrt{2}, \quad \int_0^{\pi/4} \sec x \, dx = \ln(1 + \sqrt{2}),$$

$$\int_{\pi/4}^{\pi/2} \cot x \, dx = \ln \sqrt{2}, \quad \int_{\pi/4}^{\pi/2} \csc x \, dx = \ln(1 + \sqrt{2}).$$

42. Let $\alpha, \beta \in \mathbb{R}$ be such that $-\pi < \alpha < \beta \le \pi$ and $P(x, y)$, $Q(x, y)$ be polynomials such that $Q(\sin\theta, \cos\theta) \ne 0$ for any $\theta \in [\alpha, \beta]$. Show that

$$\int_{\alpha}^{\beta} \frac{P(\sin\theta, \cos\theta)}{Q(\sin\theta, \cos\theta)} d\theta = \int_{\tan(\alpha/2)}^{\tan(\beta/2)} \frac{P\left(2t/(1+t^2), (1-t^2)(1+t^2)\right)}{Q\left(2t/(1+t^2), (1-t^2)(1+t^2)\right)} \frac{2}{1+t^2} dt.$$

43. Evaluate the following integrals:

(i) $\displaystyle\int_{0}^{\pi/2} \frac{1}{2 + \cos\theta} d\theta$, (ii) $\displaystyle\int_{\pi/2}^{2\pi/3} \frac{\cot\theta}{1 + \cos\theta} d\theta$, (iii) $\displaystyle\int_{\alpha}^{\beta} \sec\theta \, d\theta$.

(Hint: Integrate by substituting $t = \tan(\theta/2)$.)

[Note: The substitution $t = \tan(\theta/2)$ converts the integral of any rational function in trigonometric functions (in the parameter θ) to an integral of a rational function (in the variable t). The latter can, in general, be evaluated using the method of partial fractions. Therefore, the integral of any rational function in trigonometric functions can be evaluated.]

44. Let $f : \mathbb{R} \to \mathbb{R}$ be continuous and $\lambda \in \mathbb{R}$, $\lambda \ne 0$. For $x \in \mathbb{R}$, let

$$g(x) := \frac{1}{\lambda} \int_{0}^{x} f(t) \sin\lambda(x - t) dt.$$

Show that $g''(x) + \lambda^2 g(x) = f(x)$ for all $x \in \mathbb{R}$ and $g(0) = 0 = g'(0)$.

45. Find the linear and quadratic approximations of $f : [0, \infty) \to \mathbb{R}$ defined by

$$f(x) := 1 + \int_{1}^{x} \frac{10}{1 + \sqrt{t}} dt$$

for x around 1.

46. Prove the following.

(i) For $x \in \mathbb{R}$, $\displaystyle\int_{0}^{x} \frac{1}{\sqrt{1+t^2}} dt = \ln\left(x + \sqrt{1+x^2}\right)$,

(ii) For $x \in \mathbb{R}$, $\displaystyle\int_{0}^{x} \sqrt{1+t^2} \, dt = \frac{1}{2}\left(x\sqrt{1+x^2} + \ln\left(x + \sqrt{1+x^2}\right)\right)$,

(iii) For $x \in [-1, 1]$, $\displaystyle\int_{0}^{x} \sqrt{1-t^2} \, dt = \frac{1}{2}\left(x\sqrt{1-x^2} + \sin^{-1} x\right)$.

47. Find (i) $\displaystyle\int_{4}^{9} \frac{1}{x - \sqrt{x}} dx$, (ii) $\displaystyle\int_{1}^{3} \frac{1}{\sqrt{x}(x+1)} dx$, (iii) $\displaystyle\int_{1}^{1/\sqrt{2}} \frac{1}{x\sqrt{4x^2 - 1}} dx$.

48. Evaluate the following limits:

(i) $\displaystyle\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \cos\frac{i\pi}{n}$, (ii) $\displaystyle\lim_{n\to\infty} \sum_{i=1}^{n} \frac{n}{i^2 + n^2}$, (iii) $\displaystyle\lim_{n\to\infty} \sin\frac{1}{n} \sum_{i=1}^{n} \frac{n^2}{i^2 + n^2}$.

8

Applications and Approximations of Riemann Integrals

In this chapter we shall consider some geometric applications of Riemann integrals. They deal with defining and finding the areas of certain planar regions, volumes of certain solid bodies including solid bodies generated by revolving planar regions about a line, lengths of 'piecewise smooth' curves, and areas of surfaces generated by revolving such planar curves about a line. Subsequently, we show how to find the 'centroids' of the geometric objects considered earlier. The coordinates of a centroid are in some sense the averages of the coordinate functions. In the last section of this chapter, we give a number of methods for evaluating Riemann integrals approximately. We also establish error estimates for these approximations. This procedure would be useful, in particular, if we needed to find approximations of arc lengths, areas, and volumes of various geometric objects whenever exact evaluation of the Riemann integrals involved therein is either difficult or impossible.

8.1 Area of a Region Between Curves

In this section we shall show how 'areas' of certain planar regions that lie between two curves can be found using Riemann integrals. It may be remarked that the general concept of the area of a planar region is usually defined using double integrals, which are studied in a course in multivariate calculus. The definitions of areas of special planar regions given in this section can be reconciled with the general definition.

Recall that in Section 6.1, we began our discussion of a Riemann integral by *assuming* that the area of a rectangle $[x_1, x_2] \times [y_1, y_2]$ is $(x_2 - x_1)(y_2 - y_1)$. Let $[a, b]$ be an interval in $\mathbb{R}$ and $f : [a, b] \to \mathbb{R}$ be a bounded nonnegative function. The concept of a Riemann integral was motivated by an attempt to give a meaning to the 'area' of the region lying under the graph of f. We have defined the area of the region $R := \{(x, y) \in \mathbb{R}^2 : a \le x \le b \text{ and } 0 \le y \le f(x)\}$ to be

$$\text{Area } (R) := \int_a^b f(x)dx.$$

This naturally leads us to the following definition. Let $f_1, f_2 : [a, b] \to \mathbb{R}$ be integrable functions such that $f_1 \leq f_2$. Then the **area** of the region between the curves given by $y = f_1(x)$, $y = f_2(x)$ and between the (vertical) lines given by $x = a$, $x = b$, that is, of the region

$$R := \{(x, y) \in \mathbb{R}^2 : a \leq x \leq b \text{ and } f_1(x) \leq y \leq f_2(x)\},$$

is defined to be

$$\text{Area } (R) := \int_a^b [f_2(x) - f_1(x)]dx.$$
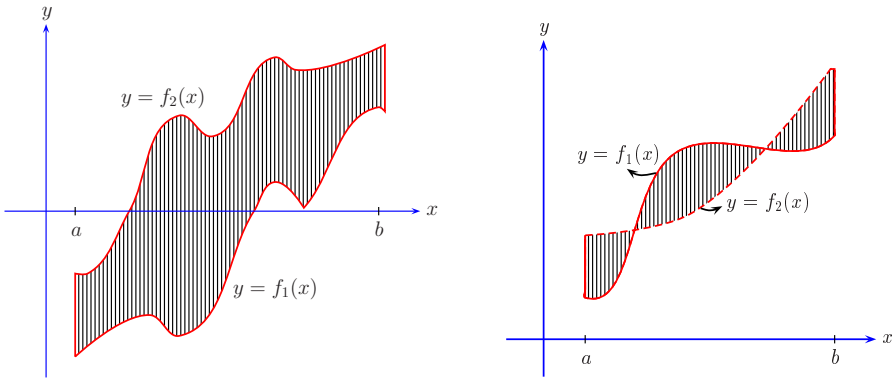
Fig. 8.1. Region between the curves $y = f_1(x)$, $y = f_2(x)$ and the lines $x = a$, $x = b$ when the curves do not cross each other, and when they cross each other

If a planar region R can be divided into a finite number of nonoverlapping subregions of the types considered above, then the area of R is defined to be the sum of the areas of these subregions. For example, if curves given by $y = f_1(x)$, $y = f_2(x)$, where $f_1, f_2 : [a, b] \to \mathbb{R}$ are continuous functions, cross each other at a finite number of points, then the area of the region bounded by these curves and the lines given by $x = a$, $x = b$ turns out to be equal to

$$\int_a^b |f_2(x) - f_1(x)|dx.$$

Similarly, if $g_1, g_2 : [c, d] \to \mathbb{R}$ are integrable functions such that $g_1 \leq g_2$, then the **area** of the region between the curves given by $x = g_1(y)$, $x = g_2(y)$ and between the (horizontal) lines given by $y = c$, $y = d$, that is, of the region

$$R := \{(x, y) \in \mathbb{R}^2 : c \leq y \leq d \text{ and } g_1(y) \leq x \leq g_2(y)\},$$
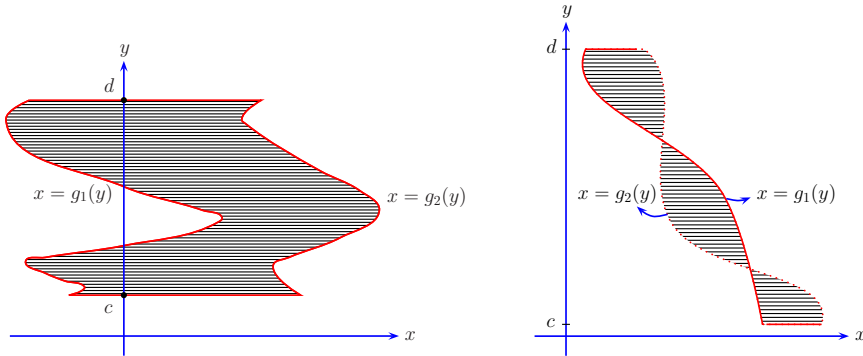
Fig. 8.2. Region between the curves $x = g_1(y)$, $x = g_2(y)$ and the lines $y = c$, $y = d$, when the curves do not cross each other, and when they cross each other

is defined to be

$$\text{Area } (R) := \int_c^d [g_2(y) - g_1(y)]dy.$$

Also, if curves given by $x = g_1(y)$, $x = g_2(y)$, where $g_1, g_2 : [c, d] \to \mathbb{R}$ are continuous functions, cross each other at a finite number of points, then the area of the region bounded by these curves and the lines given by $y = c$, $y = d$ turns out to be equal to

$$\int_c^d |g_2(y) - g_1(y)|dy.$$

Examples 8.1. (i) Let $0 < a < b$ and consider the triangular region enclosed by the lines given by $y = hx/a$, $y = h(x-b)/(a-b)$, and the x-axis. These lines form a triangle with base b and height h. We show that the area of this region is equal to $bh/2$. The perpendicular from the vertex (a, h) to the x-axis divides the triangular region into two triangular subregions having bases a and $b - a$, and both having height h. The area of the given triangular region is then equal to the sum of the areas of these subregions. [See Figure 8.3.] The first subregion is the region between the curves $y = hx/a$, $y = 0$ and between the lines given by $x = 0$, $x = a$. Hence its area is equal to

$$\int_0^a \left(\frac{hx}{a} - 0\right)dx = \frac{h}{a} \cdot \frac{a^2}{2} = \frac{ha}{2}.$$

Likewise, the area of the second subregion is equal to $h(b - a)/2$. Hence the required area is $(ha/2) + (h(b - a)/2) = hb/2$.

(ii) The region enclosed by the loop of the curve given by $y^2 = x(1 - x)^2$ is the region between the curves given by $y = \sqrt{x}(1 - x)$, $y = -\sqrt{x}(1 - x)$ and between the lines given by $x = 0$, $x = 1$. Hence its area is equal to
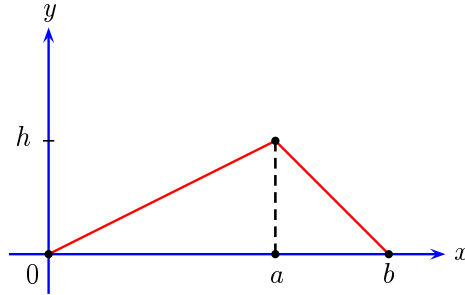
Fig. 8.3. Triangular region in Example 8.1 (i) with its two triangular subregions

$$\int_0^1 \left[\sqrt{x}(1-x) - \left(-\sqrt{x}(1-x)\right) \right] dx = 2\int_0^1 [x^{1/2} - x^{3/2}]dx = \frac{8}{15}.$$

(iii) The area of the region bounded by the curves $x = y^3$, $x = y^5$ and the lines given by $y = -1$, $y = 1$ is equal to

$$\int_{-1}^1 |y^5 - y^3|dy = \int_{-1}^0 (y^5 - y^3)dy + \int_0^1 (y^3 - y^5)dy = \frac{1}{6}.$$

(iv) To determine the area of the region bounded by the parabolas $x = -2y^2$ and $x = 1 - 3y^2$, we first find their points of intersection. Now $-2y^2 = 1 - 3y^2$ implies $y = \pm 1$ and $1 - 3y^2 \geq -2y^2$ for all $y \in [-1, 1]$. Hence

$$\int_{-1}^1 [(1 - 3y^2) - (-2y^2)]dy = \int_{-1}^1 [1 - y^2]dy = \frac{4}{3}$$

is the required area. ◇

We shall now calculate the area enclosed by an ellipse. As a special case, this will give us the area enclosed by a circle and lead us to an important classical formula for π.

Proposition 8.2. *Let a, b be positive real numbers.*

(i) *The area of the region enclosed by an ellipse given by $(x^2/a^2)+(y^2/b^2) = 1$ is equal to πab.*

(ii) *The area of a circular disk enclosed by a circle given by $x^2 + y^2 = a^2$ is equal to πa^2. In other words, if D denotes this disk, then*

$$\pi = \frac{Area\ of\ D}{(Radius\ of\ D)^2}.$$

(iii) *For $\varphi \in [0, \pi]$, the area of the sector of a disk of radius a which subtends an angle φ at the center, that is, the area of the planar region given by $\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq a^2$ and $0 \leq \theta(x, y) \leq \varphi\}$ is equal to $a^2\varphi/2$.*

Proof. (i) The area enclosed by the given ellipse is four times the area between the curves given by $y = b\sqrt{a^2 - x^2}/a$, $y = 0$ and between the lines given by $x = 0$, $x = a$. Hence it is equal to

$$4\frac{b}{a} \int_0^a \sqrt{a^2 - x^2}\, dx = \frac{4b}{a} \cdot a^2 \int_0^{\pi/2} \cos^2 \theta\, d\theta = 4ab \int_0^{\pi/2} \frac{1 + \cos 2\theta}{2}\, d\theta = \pi ab.$$

(ii) Letting $b = a$ in (i) above, we see that the area of a disk of radius a is equal to πa^2. The desired formula for π is then immediate.
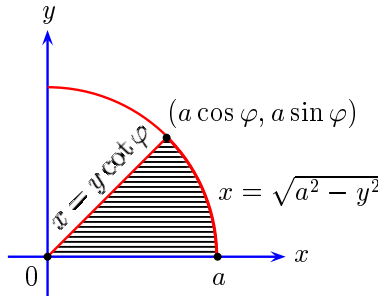
Fig. 8.4. Sector marked by the points $(0,0)$, $(a,0)$, and $(a \cos \varphi, a \sin \varphi)$

(iii) If $\varphi = 0$, then the sector reduces to a line segment, and its area is clearly equal to 0. Now let $\varphi \in (0, \pi/2]$. The sector marked by the points $(0,0)$, $(a,0)$, and $(a \cos \varphi, a \sin \varphi)$ is the region between the curves $x = (\cot \varphi)y$, $x = \sqrt{a^2 - y^2}$ and between the lines given by $y = 0$, $y = a \sin \varphi$. Hence its area is equal to

$$\int_0^{a \sin \varphi} \left[\sqrt{a^2 - y^2} - (\cot \varphi)y \right] dy = \left(a^2 \int_0^\varphi \cos^2 t\, dt \right) - \cot \varphi \frac{a^2 \sin^2 \varphi}{2} = \frac{a^2 \varphi}{2}.$$

By symmetry, the formula holds for $\varphi \in (\pi/2, \pi]$ as well. This can be seen as follows. Let $\psi := \pi - \varphi$. Then $\psi \in [0, \pi/2)$, and by what we have already proved, the area of the desired sector is equal to

$$\frac{\pi a^2}{2} - \frac{a^2 \psi}{2} = \frac{\pi a^2}{2} - \frac{a^2 (\pi - \varphi)}{2} = \frac{a^2 \varphi}{2},$$

as before. □

The formulas given in the above proposition are of fundamental importance. In part (iii) of Proposition 7.13, we have defined π as two times the supremum of the set $\{\arctan x : x \in (0, \infty)\}$. Now the same real number turns out to be the area of a circular disk divided by the square of the radius of

the disk. This formula for π makes it plain that the ratio of the area of a (circular) disk to the square of its radius is independent of the radius. We have thus proved a fact that is usually taken for granted when π is introduced in high-school geometry.

Curves Given by Polar Equations

The formula for the area of a sector of a disk given in part (iii) of Proposition 8.2 enables us to define areas of planar regions between curves given by certain polar equations.

Let us consider a curve given by a polar equation of the form $r = p(\theta)$. Let $\alpha, \beta \in \mathbb{R}$ be such that either $-\pi < \alpha < \beta < \pi$ or $\alpha = -\pi$, $\beta = \pi$. Consider a nonnegative integrable function $p : [\alpha, \beta] \to \mathbb{R}$ and assume that $p(\pi) = p(-\pi)$ if $\alpha = -\pi$, $\beta = \pi$. Let

$$R := \{(r\cos\theta, r\sin\theta) \in \mathbb{R}^2 : \alpha \leq \theta \leq \beta \text{ and } 0 \leq r \leq p(\theta)\}$$

denote the region bounded by the curve given by $r = p(\theta)$ and the rays given by $\theta = \alpha$, $\theta = \beta$. Let $(r(x, y), \theta(x, y))$ denote the polar coordinates of $(x, y) \in \mathbb{R}^2 \setminus \{(0, 0)\}$. By Proposition 7.20, it follows that

$$R \setminus \{(0,0)\} = \left\{ (x, y) \in \mathbb{R}^2 \setminus \{(0,0)\} : \alpha \leq \theta(x, y) \leq \beta \text{ and } r(x, y) \leq p(\theta(x, y)) \right\}.$$

If $\{\theta_0, \theta_1, \ldots, \theta_n\}$ is a partition of $[\alpha, \beta]$, then the planar region R gets divided into n subregions

$$\{(0,0)\} \cup \left\{ (x, y) \in \mathbb{R}^2 \setminus \{(0,0)\} : \theta_{i-1} \leq \theta(x, y) \leq \theta_i \text{ and } r(x, y) \leq p(\theta(x, y)) \right\},$$

where $i = 1, \ldots, n$. [See Figure 8.5.] For each i, let us choose $\gamma_i \in [\theta_{i-1}, \theta_i]$ and replace the ith subregion by the sector

$$\{(0,0)\} \cup \left\{ (x, y) \in \mathbb{R}^2 : \theta_{i-1} \leq \theta(x, y) \leq \theta_i \text{ and } r(x, y) \leq p(\gamma_i) \right\}$$

of the disk of radius $p(\gamma_i)$ with center at $(0,0)$. By part (iii) of Proposition 8.2, the area of this sector is equal to

$$p(\gamma_i)^2 (\theta_i - \theta_{i-1})/2, \quad i = 1, \ldots, n.$$

With this in view, the area of the region R is defined to be

$$\text{Area}\,(R) := \frac{1}{2} \int_\alpha^\beta p(\theta)^2 d\theta.$$

Further, if $p_1, p_2 : [\alpha, \beta] \to \mathbb{R}$ are integrable functions such that $0 \leq p_1 \leq p_2$ and $p_i(\pi) = p_i(-\pi)$, $i = 1, 2$, in case $\alpha = -\pi$, $\beta = \pi$, then the area of the region R between the curves given by $r = p_1(\theta)$, $r = p_2(\theta)$ and between the rays given by $\theta = \alpha$, $\theta = \beta$ is defined to be

$$\text{Area}\,(R) := \frac{1}{2} \int_\alpha^\beta \left[p_2(\theta)^2 - p_1(\theta)^2 \right] d\theta.$$
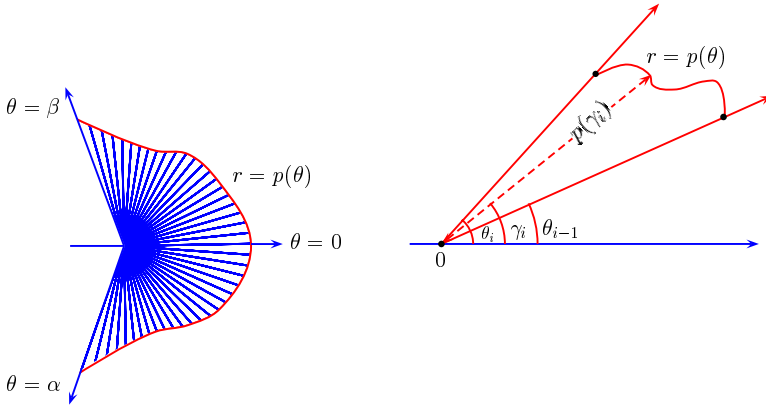
Fig. 8.5. Region bounded by the polar equation $r = p(\theta)$ and rays $\theta = \alpha$, $\theta = \beta$, and its 'ith subregion'

Examples 8.3. (i) Let $a, \alpha, \beta \in \mathbb{R}$ be such that $a > 0$ and $-\pi < \alpha < \beta \leq \pi$. Consider $p_1, p_2 : [\alpha, \beta] \to \mathbb{R}$ given by $p_1(\theta) := 0$ and $p_2(\theta) := a$. Then the area of the sector

$$\{(x, y) \in \mathbb{R}^2 : \alpha \leq \theta(x, y) \leq \beta, \, 0 \leq r(x, y) \leq a\}$$

of the disk of radius a is equal to

$$\frac{1}{2} \int_\alpha^\beta p(\theta)^2 d\theta = \frac{1}{2} \int_\alpha^\beta a^2 d\theta = \frac{a^2(\beta - \alpha)}{2},$$

as it should be in view of part (iii) of Proposition 8.2.

(ii) Let $a \in \mathbb{R}$ with $a > 0$. The area of the region enclosed by the cardioid $r = a(1 + \cos\theta)$ is equal to

$$\frac{1}{2} \int_{-\pi}^\pi [a(1 + \cos\theta)]^2 d\theta = \frac{a^2}{2} \int_{-\pi}^\pi \left(1 + 2\cos\theta + \frac{1 + \cos 2\theta}{2}\right) d\theta = \frac{3a^2\pi}{2}.$$

(iii) The area of the region between the circle given by $r = 2$ and the spiral given by $r = \theta$ that lies between the rays given by $\theta = 0$, $\theta = \pi/2$ is equal to

$$\frac{1}{2} \int_0^{\pi/2} [2^2 - \theta^2] d\theta = \pi - \frac{\pi^3}{48}.$$

Note that $\theta \leq 2$ for all $\theta \in [0, \pi/2]$. $\diamond$

Area between curves given by polar equations of the form $\theta = \alpha(r)$ is treated in Exercises 16 and 17.

We conclude this section by mentioning again that the definitions of areas of various kinds of regions discussed here can be unified with the help of double integrals. This would also show that the area of a region calculated using two different definitions given in this section must turn out to be the same!

8.2 Volume of a Solid

In this section we shall show how volumes of certain solid bodies can be found using Riemann integrals. It may be remarked that the general concept of the volume of a solid body is usually introduced in a course in multivariate calculus with the help of triple integrals. The definitions of volumes of special solid bodies given in this section can be reconciled with the general definitions.

Let us consider volumes of solid bodies that can be thought to be made up of cross-sections taken in one of the following ways:

1. Cross-sections by planes perpendicular to a fixed line,
2. Cross-section by right circular cylinders having a fixed axis.

Slicing by Planes Perpendicular to a Fixed Line

Let D be a bounded subset of $\mathbb{R}^3 := \{(x, y, z) : x, y, z \in \mathbb{R}\}$ lying between two parallel planes and let L denote a line perpendicular to these planes. A cross-section of D by a plane is called a **slice** of D. Let us assume that we are able to determine the 'area' of a slice of D by any plane perpendicular to L.

For the sake of concreteness, let the line L be the x-axis and assume that D lies between the planes given by $x = a$ and $x = b$, where $a, b \in \mathbb{R}$ with $a < b$. Further, for $s \in [a, b]$, let $A(s)$ denote the area of the slice $\{(x, y, z) \in D : x = s\}$ obtained by intersecting D with the plane given by $x = s$. If $\{x_0, x_1, \ldots, x_n\}$ is a partition of $[a, b]$, then the solid D gets divided into n subsolids
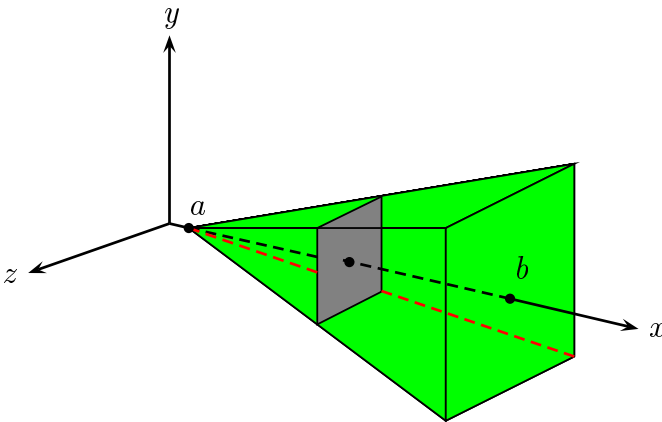
Fig. 8.6. Slicing a solid by planes perpendicular to a fixed line

$$\{(x, y, z) \in D : x_{i-1} \le x \le x_i\}, \quad i = 1, \ldots, n.$$

Let us choose $s_i \in [x_{i-1}, x_i]$ and replace the ith subsolid by a rectangular slab having volume equal to $A(s_i)(x_i - x_{i-1})$ for $i = 1, \ldots, n$. Then it is natural to consider

$$\sum_{i=1}^{n} A(s_i)(x_i - x_{i-1})$$

as an approximation of the desired volume of D. We therefore define the volume of D to be

$$\text{Vol}\,(D) := \int_a^b A(x)dx,$$

provided the 'area function' $A : [a, b] \to \mathbb{R}$ is integrable.

Similarly, if there are $c, d \in \mathbb{R}$ with $c < d$ such that $D \subseteq \{(x, y, z) \in \mathbb{R}^3 : c \le y \le d\}$, and for $t \in [c, d]$, $A(t)$ denotes the area of the slice $\{(x, y, z) \in D : y = t\}$ obtained by intersecting D with the plane given by $y = t$, then we define the volume of D to be

$$\text{Vol}\,(D) := \int_c^d A(y)dy,$$

provided the 'area function' $A : [c, d] \to \mathbb{R}$ is integrable.

Likewise, if there are $p, q \in \mathbb{R}$ with $p < q$ such that $D \subseteq \{(x, y, z) \in \mathbb{R}^3 : p \le z \le q\}$, and for $u \in [p, q]$, $A(u)$ denotes the area of the slice $\{(x, y, z) \in D : z = u\}$ obtained by intersecting D with the plane given by $z = u$, then we define the volume of D to be

$$\text{Vol}\,(D) := \int_p^q A(z)dz,$$

provided the 'area function' $A : [p, q] \to \mathbb{R}$ is integrable.

Examples 8.4. (i) Let $a, b, c, d, p, q \in \mathbb{R}$ and

$$D := \{(x, y, z) \in \mathbb{R}^3 : a \le x \le b, \ c \le y \le d, \ p \le z \le q\}$$

be a cuboid. Then for each fixed $s \in [a, b]$, the area of the slice $\{(x, y, z) \in D : x = s\}$ of D is $A(s) := (d - c)(q - p)$ and hence the volume of D is equal to

$$\int_a^b A(x)dx = (d - c)(q - p)(b - a).$$

Alternatively, for each fixed $t \in [c, d]$, we may consider the area $A(t) := (b - a)(q - p)$ of the slice $\{(x, y, z) \in D : y = t\}$ of D, or for each fixed $u \in [p, q]$, we may consider the area $A(u) := (b - a)(d - c)$ of the slice $\{(x, y, z) \in D : z = u\}$ of D for finding the volume of D. At any rate, this simple example shows that our definition of the volume of a solid as the integral of an 'area function' is consistent with our assumption regarding the volume of a cuboid.
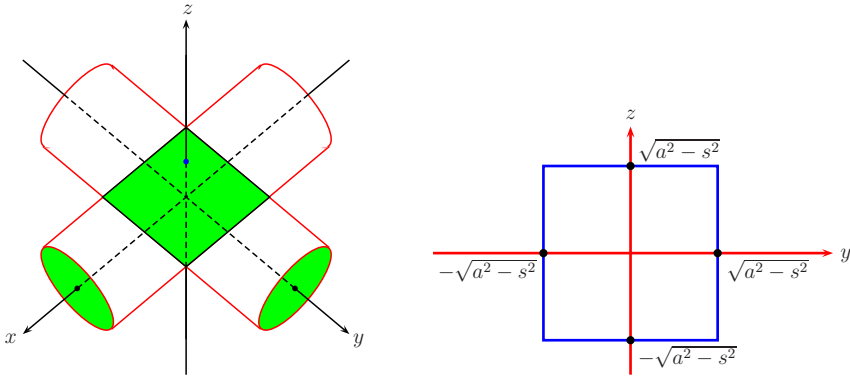
Fig. 8.7. Solid enclosed by two cylinders and a slice resulting in a square region

(ii) Let $a \in \mathbb{R}$ with $a > 0$. Let us find the volume of the solid D enclosed by the cylinders $x^2 + y^2 = a^2$ and $x^2 + z^2 = a^2$. [See Figure 8.7.] The solid D lies between the planes $x = -a$ and $x = a$, and for a fixed $s \in [-a, a]$, the slice $\{(x, y, z) \in D : x = s\}$ is given by

$$\left\{ (s, y, z) \in \mathbb{R}^3 : |y| \leq \sqrt{a^2 - s^2} \text{ and } |z| \leq \sqrt{a^2 - s^2} \right\}.$$

This slice is a square region of side $2\sqrt{a^2 - s^2}$, and its area is equal to

$$A(s) := (2\sqrt{a^2 - s^2})^2 = 4(a^2 - s^2).$$

Hence

$$\int_{-a}^{a} A(x)dx = 4 \int_{-a}^{a} (a^2 - x^2)dx = 8 \int_{0}^{a} (a^2 - x^2)dx = 8\left(a^3 - \frac{a^3}{3}\right) = \frac{16a^3}{3}$$

is the required volume. $\diamond$

We shall now calculate the volume enclosed by an ellipsoid, and as a special case, the volume enclosed by a sphere. It will lead us to another important classical formula for π.

Proposition 8.5. (i) *The volume of a solid enclosed by an ellipsoid given by* $(x^2/a^2) + (y^2/b^2) + (z^2/c^2) = 1$, *where* $a, b, c > 0$, *is equal to* $4\pi abc/3$.
(ii) *The volume of a spherical ball enclosed by the sphere given by* $x^2 + y^2 + z^2 = a^2$ *is equal to* $4\pi a^3/3$. *In other words, if* B *denotes a spherical ball, then*

$$\pi = \frac{3}{4} \frac{Volume \ of \ B}{(Radius \ of \ B)^3}.$$

(iii) *Let* $a > 0$. *For* $\varphi \in [0, \pi]$, *the volume of the (solid) spherical cone*

$$\left\{ (x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 \leq a^2 \ \text{and} \ 0 \leq \cos^{-1}\left(x/\sqrt{x^2 + y^2 + z^2}\right) \leq \varphi \right\}$$

is equal to $2\pi a^3 (1 - \cos \varphi)/3$.

Proof. (i) The given ellipsoid lies between the planes given by $x = -a$ and $x = a$. Also, for $s \in (-a, a)$, the area $A(s)$ of its slice

$$\left\{ (s, y, z) \in \mathbb{R}^3 : \frac{y^2}{b^2} + \frac{z^2}{c^2} \leq 1 - \frac{s^2}{a^2} \right\}$$

by the plane given by $x = s$ is the area enclosed by the ellipse

$$\frac{y^2}{b^2[1 - (s^2/a^2)]} + \frac{z^2}{c^2[1 - (s^2/a^2)]} = 1,$$

and hence by part (i) of Proposition 8.2, we have

$$A(s) = \pi \left(b\sqrt{1 - (s^2/a^2)} \right) \left(c\sqrt{1 - (s^2/a^2)} \right) = \pi b c \left(1 - \frac{s^2}{a^2} \right).$$

Thus the volume enclosed by the ellipsoid is equal to

$$\int_{-a}^{a} A(x)dx = \pi b c \int_{-a}^{a} \left(1 - \frac{x^2}{a^2} \right)dx = \pi b c \left(2a - \frac{2a^3}{3a^2} \right) = \frac{4}{3}\pi abc.$$

(ii) Letting $b = a$ and $c = a$ in (i) above, we see that the volume of the spherical ball of radius a is equal to $4\pi a^3/3$. The desired formula for π is then immediate.

(iii) If $\varphi = 0$, then the (solid) spherical cone reduces to the line segment $\{(x, 0, 0) \in \mathbb{R}^3 : 0 \leq x \leq a\}$, and its volume is clearly equal to 0. Also, if $\varphi = \pi/2$, then the (solid) spherical cone is the half spherical ball $\{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 \leq a^2 \ \text{and} \ x \geq 0\}$ and by (ii) above, its volume is equal to $2\pi a^3/3$. Now let $\varphi \in (0, \pi/2)$. For $s \in [0, a\cos\varphi]$, the slice of the (solid) spherical cone by the plane given by $x = s$ is a disk of radius $s\tan\varphi$ and so its area $A(s)$ is equal to $\pi s^2 \tan^2 \varphi$, whereas for $t \in (a\cos\varphi, a]$, the slice of the (solid) spherical cone by the plane given by $x = t$ is a disk of radius $\sqrt{a^2 - t^2}$ and so its area $A(t)$ is equal to $\pi(t^2 - a^2)$. [See Figure 8.8.] Hence the volume of the (solid) spherical cone is equal to

$$\int_{0}^{a\cos\varphi} \pi x^2 \tan^2 \varphi \, dx + \int_{a\cos\varphi}^{a} \pi(a^2 - x^2)dx$$

$$= \pi \tan^2 \varphi \frac{a^3 \cos^3 \varphi}{3} + \pi \left(a^3 - \frac{a^3}{3} - a^3 \cos\varphi + \frac{a^3 \cos^3 \varphi}{3} \right)$$

$$= \frac{\pi a^3}{3} \left(\sin^2 \varphi \cos\varphi + 2 - 3\cos\varphi + \cos^3 \varphi \right) = \frac{2\pi a^3}{3}(1 - \cos\varphi).$$

By symmetry, the formula holds for $\varphi \in (\pi/2, \pi]$ as well. This can be seen as follows. Let $\psi := \pi - \varphi$. Then $\psi \in [0, \pi/2)$, and by what we have already proved, the volume of the desired spherical cone is equal to

$$\frac{4\pi a^3}{3} - \frac{2\pi a^3}{3}(1 - \cos \psi) = \frac{2\pi a^3}{3} + \frac{2\pi a^3}{3}\cos(\pi - \varphi) = \frac{2\pi a^3}{3}(1 - \cos \varphi),$$

as before.   □
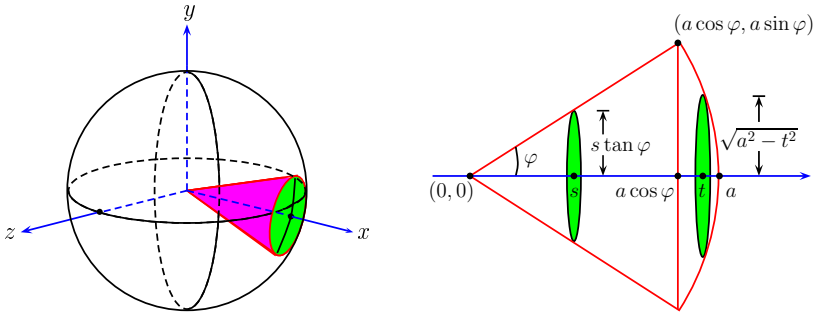
Fig. 8.8. A solid spherical cone inside a sphere and its slices by planes $x = s$, $x = t$

The formula for π given in part (ii) of the above proposition makes it plain that the ratio of the volume of a spherical ball to the cube of its radius is independent of the radius.

Slivering by Coaxial Right Circular Cylinders

Suppose that a bounded solid D lies between two cylinders having a given line L as their common axis. A cross-section of D by a cylinder is called a **sliver** of D. Let us assume that we are able to determine the 'surface area' of a sliver of D by any cylinder having L as its axis.

For the sake of concreteness, let the given line L be the z-axis, let $p, q \in \mathbb{R}$ with $0 \le p < q$, and assume that D lies between the cylinders given by $x^2 + y^2 = p^2$ and $x^2 + y^2 = q^2$, and for $r \in [p, q]$, let $A(r)$ denote the surface area of the sliver

$$\{(x, y, z) \in D : x^2 + y^2 = r^2\}$$

of D obtained by intersecting it with the cylinder given by $x^2 + y^2 = r^2$. If $\{r_0, r_1, \ldots, r_n\}$ is a partition of $[p, q]$, then the solid D gets divided into n subsolids

$$\left\{(x, y, z) \in D : r_{i-1} \le \sqrt{x^2 + y^2} \le r_i\right\}, \quad i = 1, \ldots, n.$$

Let us choose $s_i \in [r_{i-1}, r_i]$ and replace the ith subsolid by a cylindrical solid having volume equal to $A(s_i)(r_i - r_{i-1})$ for $i = 1, \ldots, n$. Then it is natural to consider
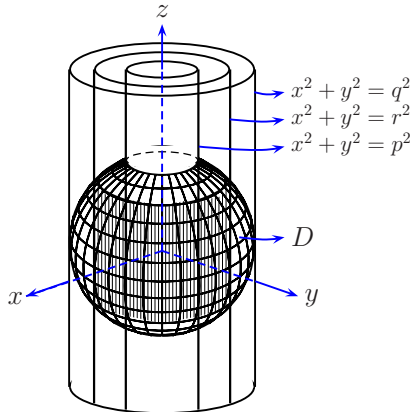
Fig. 8.9. Slivering a solid lying between the cylinders $x^2 + y^2 = p^2$ and $x^2 + y^2 = q^2$ by right coaxial cylinders $x^2 + y^2 = r^2$ for $r \in [p, q]$

$$\sum_{i=1}^{n} A(s_i)(r_i - r_{i-1})$$

as an approximation of the desired volume of D. We therefore define the volume of D to be

$$\mathrm{Vol}\,(D) := \int_p^q A(r) dr,$$

provided the 'surface area function' $A : [p, q] \to \mathbb{R}$ is integrable.

We now address the question of finding the surface area $A(r)$ of the sliver

$$\{(x, y, z) \in D : x^2 + y^2 = r^2\}$$

of D for a fixed $r \in [p, q]$. Let

$$E_r := \{(\theta, z) \in [-\pi, \pi] \times \mathbb{R} : (r \cos \theta, r \sin \theta, z) \in D\}.$$

denote the **parameter domain** for the sliver. Then the **surface area** $A(r)$ of this sliver is defined to be r times the area $B(r)$ of the planar region E_r. Thus the volume of D is equal to

$$\mathrm{Vol}\,(D) = \int_p^q r B(r) dr,$$

where $B(r)$ is the planar area of the parameter domain E_r given above for each $r \in [p, q]$.

Similar considerations hold if the given line L is the y-axis and there are $a, b \in \mathbb{R}$ with $0 \le a < b$ such that D lies between the cylinders given by $z^2 + x^2 = a^2$ and $z^2 + x^2 = b^2$, or if the given line L is the x-axis and there are $c, d \in \mathbb{R}$ with $0 \le c < d$ such that D lies between the cylinders given by $y^2 + z^2 = c^2$ and $y^2 + z^2 = d^2$.

Examples 8.6. (i) Let $p, q, h \in \mathbb{R}$ with $0 < p < q$ and $h > 0$, and consider the cylindrical shell

$$D := \{(x, y, z) \in \mathbb{R}^3 : p \leq \sqrt{x^2 + y^2} \leq q \text{ and } 0 \leq z \leq h\}.$$

For a fixed $r \in [p, q]$, the sliver

$$\{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 = r^2 \text{ and } 0 \leq z \leq h\}$$

of D, obtained by intersecting D with the cylinder given by $x^2 + y^2 = r^2$, has the parameter domain

$$E_r := \{(\theta, z) \in \mathbb{R}^2 : -\pi \leq \theta \leq \pi \text{ and } 0 \leq z \leq h\}.$$

Since the area $B(r)$ of the rectangular region E_r is equal to

$$[\pi - (-\pi)] \cdot [h - 0] = 2\pi h$$

for each $r \in [p, q]$, we see that the volume of D is equal to

$$\int_p^q r B(r) dr = \int_p^q r(2\pi h) dr = \pi h(q^2 - p^2).$$
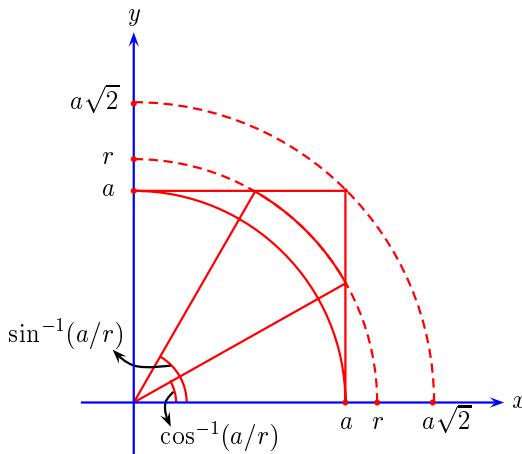
Fig. 8.10. Projections on the xy-plane of slivers of a cube by right coaxial cylinders

(ii) Let $a \in \mathbb{R}$ with $a > 0$, and consider the cube

$$D := \{(x, y, z) \in \mathbb{R}^3 : 0 \leq x, y, z \leq a\}$$

of side a. It lies inside the cylinder $x^2 + y^2 = (a\sqrt{2})^2 = 2a^2$. For a fixed $0 \leq r \leq a\sqrt{2}$, consider the sliver

$$\left\{(x, y, z) \in \mathbb{R}^3 : 0 \leq x, y, z \leq a \text{ and } x^2 + y^2 = r^2\right\}$$

of D obtained by intersecting it with the cylinder given by $x^2 + y^2 = r^2$. The projections of these slivers on the xy-plane are depicted in Figure 8.10. It is clear that if $0 \leq r \leq a$, then the sliver is given by

$$\left\{(x, y, z) \in \mathbb{R}^3 : x \geq 0, \ y \geq 0, \ x^2 + y^2 = r^2 \text{ and } 0 \leq z \leq a\right\},$$

and its parameter domain

$$E_r := \left\{(\theta, z) \in \mathbb{R}^2 : \theta \in [0, \pi/2] \text{ and } 0 \leq z \leq a\right\}$$

has area $B(r) = (\pi/2)a = a\pi/2$. On the other hand, if $a < r \leq a\sqrt{2}$, then the sliver is given by

$$\left\{(r\cos\theta, r\sin\theta, z) \in \mathbb{R}^3 : \cos^{-1}\frac{a}{r} \leq \theta \leq \sin^{-1}\frac{a}{r} \text{ and } 0 \leq z \leq a\right\},$$

and its parameter domain

$$E_r = \left\{(\theta, z) \in \mathbb{R}^2 : \cos^{-1}\frac{a}{r} \leq \theta \leq \sin^{-1}\frac{a}{r} \text{ and } 0 \leq z \leq a\right\}$$

has area $B(r) = \left[\sin^{-1}(a/r) - \cos^{-1}(a/r)\right]a$. Thus the volume of D is equal to

$$\int_0^{a\sqrt{2}} rB(r)dr = \int_0^a r\frac{a\pi}{2}dr + \int_a^{a\sqrt{2}} r\left(\sin^{-1}\frac{a}{r} - \cos^{-1}\frac{a}{r}\right)a\,dr.$$

Substituting $r = a\csc\theta$, and then integrating by parts, we obtain

$$\int_a^{a\sqrt{2}} r\sin^{-1}\frac{a}{r}dr = a^2\int_{\pi/4}^{\pi/2}\theta\csc^2\theta\cot\theta d\theta$$

$$= -\frac{a^2}{2}\left(\theta\cot^2\theta\Big|_{\pi/4}^{\pi/2} - \int_{\pi/4}^{\pi/2}\cot^2\theta d\theta\right) = \frac{a^2}{2},$$

while substituting $r = a\sec\theta$, and then integrating by parts, we obtain

$$\int_a^{a\sqrt{2}} r\cos^{-1}\frac{a}{r}dr = a^2\int_0^{\pi/4}\theta\sec^2\theta\tan\theta d\theta$$

$$= \frac{a^2}{2}\left(\theta\tan^2\theta\Big|_0^{\pi/4} - \int_0^{\pi/4}\tan^2\theta d\theta\right) = \frac{a^2}{2}\left(\frac{\pi}{2} - 1\right).$$

Hence we can conclude that the volume of D is equal to

$$a\frac{\pi}{2}\cdot\frac{a^2}{2} + a\cdot\frac{a^2}{2} - a\cdot\frac{a^2}{2}\left(\frac{\pi}{2} - 1\right) = a^3,$$

as expected.   ◇

Solids of Revolution

A subset of $\mathbb{R}^3$ that can be generated by revolving a planar region about an axis is known as a **solid of revolution**. For example, the spherical ball $\{(x, y, z) : x^2 + y^2 + z^2 \leq a^2\}$ of radius a can be generated by revolving the semidisk $\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq a^2 \text{ and } y \geq 0\}$ about the x-axis, or by revolving the semidisk $\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq a^2 \text{ and } x \geq 0\}$ about the y-axis. Likewise, the cylindrical solid $\{(x, y, z) \in \mathbb{R}^3 : y^2 + z^2 \leq a^2 \text{ and } 0 \leq x \leq h\}$ can be generated by revolving the rectangle $[0, h] \times [0, a]$ about the x-axis.

 If the planar region being revolved is bounded and the axis of revolution is one of the coordinate axes, then the volume of the corresponding solid of revolution can be found using one of the definitions of volume given earlier in this section. It may be remarked that the case in which a general plane domain is revolved about an arbitrary line in its plane can be treated in a course on multivariate calculus with the help of triple integrals.

 First let us consider slices of a solid of revolution by planes perpendicular to the axis of revolution. In general, each such slice is a circular 'washer'. If the region touches the axis of revolution at the point of slicing, then the slice is simply a disk. [See Figure 8.11.] For this reason, this method of finding the volume of a solid of revolution is known as the **Washer Method** or the **Disk Method**.
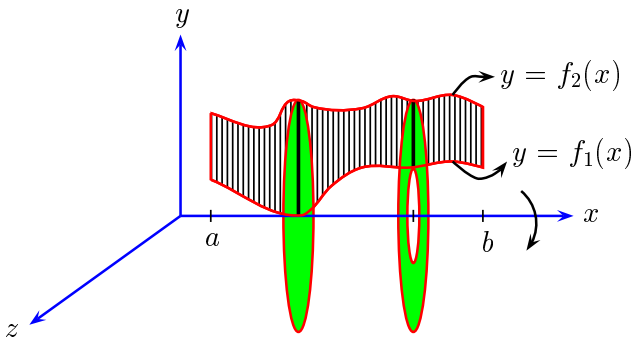
Fig. 8.11. Illustration of the Washer Method or the Disk Method

 For the sake of concreteness, let $f_1, f_2 : [a, b] \to \mathbb{R}$ be integrable functions such that $0 \leq f_1 \leq f_2$, and suppose that the region between the curves given by $y = f_1(x)$, $y = f_2(x)$ and between the lines given by $x = a$, $x = b$ is revolved about the x-axis. Let D denote the corresponding solid of revolution. Then for $s \in [a, b]$, the area $A(s)$ of the annular slice of D by the plane given by $x = s$ is equal to

$$\pi f_2(s)^2 - \pi f_1(s)^2 = \pi \left[f_2(s)^2 - f_1(s)^2 \right].$$

Hence the volume of D is equal to

$$\text{Vol}\,(D) = \pi \int_a^b \left[f_2(x)^2 - f_1(x)^2\right] dx.$$

Similarly, if $g_1, g_2 : [c, d] \to \mathbb{R}$ are integrable functions such that $0 \le g_1 \le g_2$, and the region between the curves given by $x = g_1(y)$, $x = g_2(y)$ and between the lines given by $y = c$, $y = d$ is revolved about the y-axis, then the volume of the solid D of revolution is equal to

$$\text{Vol}\,(D) = \pi \int_c^d \left[g_2(y)^2 - g_1(y)^2\right] dy.$$

Next, let us consider slivers of a solid of revolution by right circular cylinders whose axis is the same as the axis of revolution. In general, each such sliver is a cylindrical shell. For this reason, this method of finding the volume of a solid of revolution is known as the **shell method**. Note that if the radius of a cylindrical shell is r and its height is h, then the corresponding parameter domain is $[-\pi, \pi] \times [0, h]$. The latter has area $2\pi h$, and hence the surface area of the sliver is $r \cdot 2\pi h = 2\pi r h$.
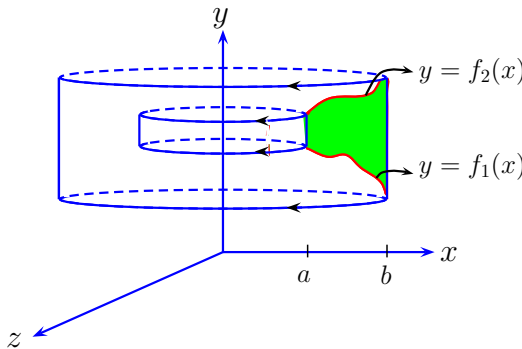
Fig. 8.12. Illustration of the Shell Method

For the sake of concreteness, let $f_1, f_2 : [a, b] \to \mathbb{R}$ be integrable functions such that $f_1 \le f_2$ and assume that $a \ge 0$. Suppose the region between the curves given by $y = f_1(x)$, $y = f_2(x)$ and between the lines given by $x = a$, $x = b$ is revolved about the y-axis to generate a solid D. Consider $s \in [a, b]$ and the sliver $\{(x, y, z) \in D : z^2 + x^2 = s^2\}$ of D by the cylinder given by $z^2 + x^2 = s^2$. Its parameter domain is $E_s := [-\pi, \pi] \times [f_1(s), f_2(s)]$. Since the area of E_s is equal to $B(s) := 2\pi[f_2(s) - f_1(s)]$, we see that the area of the sliver is equal to

$$sB(s) := 2\pi s[f_2(s) - f_1(s)].$$

Hence the volume of D is equal to

$$\text{Vol}\,(D) = 2\pi \int_a^b x[f_2(x) - f_1(x)]dx.$$

Similarly, if $g_1, g_2 : [c, d] \to \mathbb{R}$ are integrable functions such that $g_1 \leq g_2$ with $c \geq 0$, and if the region between the curves given by $x = g_1(y)$, $x = g_2(y)$ and between the lines given by $y = c$, $y = d$ is revolved about the x-axis, then the volume of the solid D of revolution is equal to

$$\text{Vol}\,(D) = 2\pi \int_c^d y[g_2(y) - g_1(y)]dy.$$
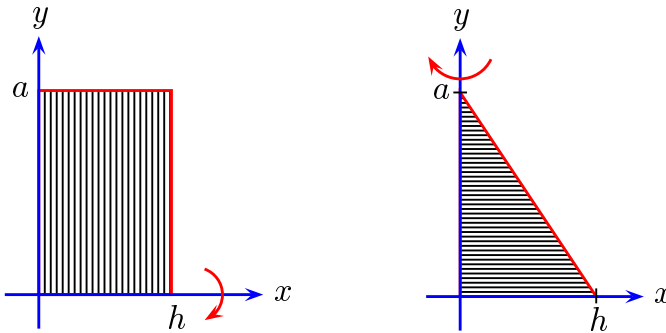
Fig. 8.13. Rectangular and triangular regions in Example 8.7 (i) and (ii)

Examples 8.7. (i) Let a and h be positive real numbers. A right circular cylindrical solid D of radius a and height h is obtained by revolving the rectangular region bounded by the lines given by $f_2(x) = a$, $f_1(x) = 0$, $x = 0$, and $x = h$ about the x-axis. [See Figure 8.13.] By the disk method, the volume of D is equal to

$$\text{Vol}\,(D) = \pi \int_0^h a^2 dx = \pi a^2 h,$$

whereas by the shell method, we also have

$$\text{Vol}\,(D) = 2\pi \int_0^a yh\, dy = \pi a^2 h.$$

(ii) Let a and h be positive real numbers. A right circular conical solid D of radius a and height h is obtained by revolving the triangular region bounded by the lines given by $x = 0$, $y = 0$, and $(x/a) + (y/h) = 1$ about

the y-axis. [See Figure 8.13.] By the disk method, the volume of D is equal
to

$$\text{Vol}\,(D) = \pi \int_0^h a^2 \left(1 - \frac{y}{h}\right)^2 dy = \pi a^2 \int_0^1 h u^2\, du = \frac{1}{3}\pi a^2 h,$$

whereas by the shell method, we also have

$$\text{Vol}\,(D) = 2\pi \int_0^a xh\left(1 - \frac{x}{a}\right) dx = 2\pi h\left(\frac{a^2}{2} - \frac{1}{a}\frac{a^3}{3}\right) = \frac{1}{3}\pi a^2 h.$$
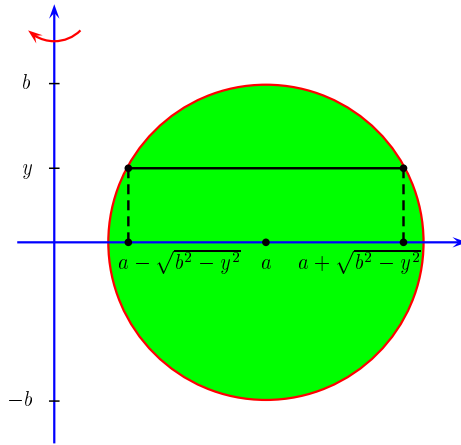
Fig. 8.14. The disk in Example 8.7 (iii) being revolved about the y-axis

(iii) Let $a, b \in \mathbb{R}$ with $0 < b < a$. If the disk $\left\{(x, y) \in \mathbb{R}^2 : (x - a)^2 + y^2 \le b^2\right\}$
is revolved about the y-axis, we obtain a solid torus D. [See Figure 8.14.]
By the Washer Method, the volume of D is equal to

$$\text{Vol}\,(D) = \pi \int_{-b}^b \left[\left(a + \sqrt{b^2 - y^2}\right)^2 - \left(a - \sqrt{b^2 - y^2}\right)^2\right] dy$$

$$= \pi \int_{-b}^b 4a\sqrt{b^2 - y^2}\, dy = 8\pi ab^2 \int_0^1 \sqrt{1 - u^2}\, du$$

$$= 8\pi ab^2 \frac{\sin^{-1} 1}{2} = 2\pi^2 ab^2.$$

(See Revision Exercise 46 (iii) given at the end of Chapter 7.)

(iv) Let R denote the region in the first quadrant between the parabolas given
by $y = x^2$ and $y = 2 - x^2$. [See Figure 8.15.] Consider the solid D_1
generated by revolving the region R about the x-axis. By the washer
method, the volume of D_1 is equal to

$$\text{Vol}(D_1) = \pi \int_0^1 \left[(2-x^2)^2 - (x^2)^2\right] dx = \pi \int_0^1 \left(4 - 4x^2\right) = 4\pi\left(1 - \frac{1}{3}\right) = \frac{8\pi}{3},$$

whereas by the shell method, we also have

$$\text{Vol}(D_1) = 2\pi \int_0^1 y\sqrt{y}\, dy + 2\pi \int_1^2 y\sqrt{2-y}\, dy$$

$$= 2\pi \frac{2}{5} + 2\pi \int_0^1 (2-u)\sqrt{u}\, du = 2\pi\left(\frac{2}{5} + \frac{4}{3} - \frac{2}{5}\right) = \frac{8\pi}{3}.$$

Consider next the solid D_2 generated by revolving the region R about the y-axis. By the disk method, the volume of D_2 is equal to

$$\text{Vol}(D_2) = \pi \int_0^1 \left(\sqrt{y}\right)^2 dy + \pi \int_1^2 \left(\sqrt{2-y}\right)^2 dy = \pi\left(\frac{1}{2} + \frac{1}{2}\right) = \pi,$$

whereas by the shell method, we also have

$$\text{Vol}(D_2) = 2\pi \int_0^1 x\left[(2-x^2) - x^2\right] dx = 4\pi \int_0^1 \left(x - x^3\right) dx = 4\pi\left(\frac{1}{2} - \frac{1}{4}\right) = \pi.$$
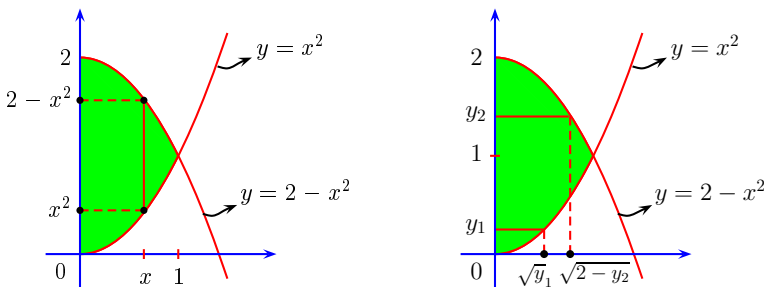
Fig. 8.15. Region in the first quadrant bounded by the parabolas $y = x^2$, $y = 2 - x^2$

It may be observed that, depending on the shape of a region relative to the axis of revolution, we may decide whether the washer method or the shell method turns out to be easier than the other. In any case, since both methods must give the same answer, one of them can be used as a check on the calculations for the other. ◇

We conclude this section by mentioning again that the definitions of volumes of various kinds of solids discussed here can all be unified in a course in multivariate calculus with the help of triple integrals. This would show that the volume of a solid calculated by using two different definitions given in this section must turn out to be the same!

8.3 Arc Length of a Curve

In this section, we shall discuss how to measure the distance covered while going along a curve, that is, how to calculate the 'length' of a curve. We shall base our discussion only on the *assumption* that the (Euclidean) distance between two points (x_1, y_1) and (x_2, y_2) in $\mathbb{R}^2$ is equal to $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$, which is in accordance with the Pythagorean Theorem of elementary geometry. Since our treatment here is in the form of an application of Riemann integration, we shall consider only those curves whose 'length' can be determined using Riemann integrals. The more general notion of a 'rectifiable curve' is treated in Exercise 70.

Let us first consider a special situation. Suppose $x^\circ, y^\circ, a_1, a_2$ are real numbers, and a curve is given by $(\phi_1(t), \phi_2(t))$, $t \in [\alpha, \beta]$, where

$$\phi_1(t) := x^\circ + a_1 t \quad \text{and} \quad \phi_2(t) := y^\circ + a_2 t \quad \text{for } t \in [\alpha, \beta].$$

The image of this curve is the line segment from the point $(x^\circ + a_1 \alpha, \ y^\circ + a_2 \alpha)$ to the point $(x^\circ + a_1 \beta, \ y^\circ + a_2 \beta)$ and its length is equal to

$$\sqrt{\left[(x^\circ + a_1 \beta) - (x^\circ + a_1 \alpha)\right]^2 + \left[(y^\circ + a_2 \beta) - (y^\circ + a_2 \alpha)\right]^2} = \sqrt{a_1^2 + a_2^2}\,(\beta - \alpha).$$

Note that $a_1 = \phi_1'(t)$ and $a_2 = \phi_2'(t)$ for all $t \in [\alpha, \beta]$. This observation is crucial in developing the notion of the length of a curve, because any 'nice' curve can be approximated locally by a line segment. To explain this, let t_0 be an interior point of an interval $[\alpha, \beta]$ and consider a curve C given by $(x(t), y(t))$, $t \in [\alpha, \beta]$, where the functions x and y are differentiable at t_0. Let

$$\phi_1(t) := x(t_0) + x'(t_0)(t - t_0) \quad \text{and} \quad \phi_2(t) := y(t_0) + y'(t_0)(t - t_0) \quad \text{for } t \in [\alpha, \beta].$$

Then by Proposition 5.11, we see that

$$x(t) - \phi_1(t) \to 0 \quad \text{and} \quad y(t) - \phi_2(t) \to 0 \quad \text{as } t \to t_0.$$

Thus the line segment given by $(\phi_1(t), \phi_2(t))$, $t \in [\alpha, \beta]$ approximates the curve C around t_0. It is therefore reasonable to expect that if $[\alpha, \beta]$ is a small interval about the point t_0, then the 'length' of the curve C should be approximately equal to the length of this line segment, which is equal to

$$\sqrt{\phi_1'(t_0)^2 + \phi_2'(t_0)^2}\,(\beta - \alpha) = \sqrt{x'(t_0)^2 + y'(t_0)^2}\,(\beta - \alpha).$$

We observe that this line segment is tangent to the curve C at $(x(t_0), y(t_0))$.

Keeping the above motivation in mind, we proceed as follows. A parametrically defined curve C in $\mathbb{R}^2$ given by $(x(t), y(t))$, $t \in [\alpha, \beta]$, is said to be **smooth** if the functions x and y are differentiable and their derivatives are continuous on $[\alpha, \beta]$. In this case, the **arc length** of C is defined to be
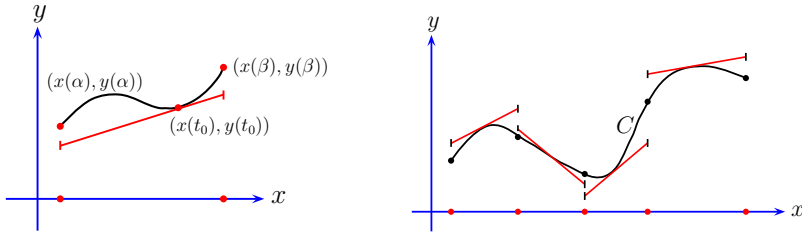
Fig. 8.16. Finding the arc length by considering the tangents to a curve

$$\ell(C) := \int_\alpha^\beta \sqrt{x'(t)^2 + y'(t)^2}\, dt.$$

Note that the arc length of C is well defined because by parts (i), (iii), and (v) of Proposition 3.3, the function $\sqrt{(x')^2 + (y')^2}$ is continuous, and hence by part (ii) of Proposition 6.9, it is integrable.

We emphasize that the arc length of a curve C is defined in terms of its given specific parametrization. The curve C should not be confused with its image $\{(x(t), y(t)) \in \mathbb{R}^2 : t \in [\alpha, \beta]\}$. For example, the curve C_1 given by $(\cos t, \sin t)$, $t \in [-\pi, \pi]$, and the curve C_2 given by $(\cos 2t, \sin 2t)$, $t \in [-\pi, \pi]$, have the same domain $[-\pi, \pi]$ and the same image $\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$, but they are obviously different curves, since C_1 winds around the origin $(0, 0)$ once, while C_2 winds around the origin $(0, 0)$ twice! We now show that the arc length of a curve does not change under certain 'reparametrizations'.

Proposition 8.8. *Let C be a smooth curve given by $(x(t), y(t))$, $t \in [\alpha, \beta]$. Suppose $\phi : [\gamma, \delta] \to \mathbb{R}$ is a differentiable function such that ϕ' is integrable, $\phi([\gamma, \delta]) = [\alpha, \beta]$, and $\phi'(u) \neq 0$ for every $u \in [\gamma, \delta]$. Let $\widetilde{C}$ denote the parametrically defined curve given by $(\tilde{x}(u), \tilde{y}(u))$, $u \in [\gamma, \delta]$, where the functions $\tilde{x}, \tilde{y} : [\gamma, \delta] \to \mathbb{R}$ are given by $\tilde{x} := x \circ \phi$, $\tilde{y} := y \circ \phi$. Then $\widetilde{C}$ is a smooth curve and*
$$\ell(\widetilde{C}) = \ell(C).$$

Proof. Consider the function $t \longmapsto \sqrt{x'(t)^2 + y'(t)^2}$ from $[\alpha, \beta]$ to $\mathbb{R}$. Since x' and y' are continuous on $[\alpha, \beta]$, it follows from parts (i), (iii), and (v) of Proposition 3.3 that this function is continuous (and hence integrable) on $[\alpha, \beta]$. Now part (ii) of Proposition 6.26 shows that

$$
\begin{aligned}
\ell(C) &= \int_\alpha^\beta \sqrt{x'(t)^2 + y'(t)^2}\, dt \\
&= \int_\gamma^\delta \sqrt{x'(\phi(u))^2 + y'(\phi(u))^2}\, |\phi'(u)|\, du \\
&= \int_\gamma^\delta \sqrt{\tilde{x}'(u)^2 + \tilde{y}'(u)^2}\, du,
\end{aligned}
$$

since $\tilde{x}'(u) = x'(\phi(u))\phi'(u)$ and $\tilde{y}'(u) = y'(\phi(u))\phi'(u)$ for all $u \in [\gamma, \delta]$ by the Chain Rule (Proposition 4.9). Thus $\ell(\widetilde{C}) = \ell(C)$. $\qquad\qquad\qquad\qquad\qquad\square$

Let us consider some important special cases of parametrically defined curves, namely curves defined by a Cartesian equation of the form $y = f(x)$ or of the form $x = g(y)$, and curves defined by a polar equation of the form $r = p(\theta)$ or of the form $\theta = \alpha(r)$.

1. Let $a, b \in \mathbb{R}$ with $a < b$, $f : [a, b] \to \mathbb{R}$, and a smooth curve C be given by $y = f(x)$, $x \in [a, b]$. Then the arc length of C is equal to

$$\ell(C) = \int_a^b \sqrt{1 + f'(x)^2} \, dx.$$

This follows by considering the Cartesian coordinate x as a parameter with $[a, b]$ as the parameter interval. Thus in this case $x' = 1$ and $y'(x) = f'(x)$ for $x \in [a, b]$.

Similarly, if $c, d \in \mathbb{R}$ with $c < d$, $g : [c, d] \to \mathbb{R}$ and a smooth curve C is given by $x = g(y)$, $y \in [c, d]$, then the arc length of C is equal to

$$\ell(C) = \int_c^d \sqrt{1 + g'(y)^2} \, dy.$$

2. Let $\alpha, \beta \in \mathbb{R}$, $p : [\alpha, \beta] \to [0, \infty)$ and a smooth curve C be given by $r = p(\theta)$, $\theta \in [\alpha, \beta]$. Then the arc length of C is equal to

$$\ell(C) = \int_\alpha^\beta \sqrt{p(\theta)^2 + p'(\theta)^2} \, d\theta.$$

This follows by considering the polar coordinate θ as a parameter with $[\alpha, \beta]$ as the parameter interval, so that C is given by the parametric equations

$$x(\theta) = p(\theta) \cos \theta \quad \text{and} \quad y(\theta) = p(\theta) \sin \theta,$$

which show that for all $\theta \in [\alpha, \beta]$,

$$x'(\theta)^2 + y'(\theta)^2 = [p'(\theta) \cos \theta - p(\theta) \sin \theta]^2 + [p'(\theta) \sin \theta + p(\theta) \cos \theta]^2$$
$$= p(\theta)^2 + p'(\theta)^2.$$

Arc length of a curve given by a polar equations of the form $\theta = \alpha(r)$ is treated in Exercises 30 and 31.

Proposition 8.9. (i) *For $\varphi \in [0, \pi]$, the length of the arc of a circle given by $x := a \cos t$, $y := a \sin t$, $0 \le t \le \varphi$ (which subtends an angle φ at the center), is equal to $a\varphi$.*

(ii) *The perimeter of the circle given by $x^2 + y^2 = a^2$ is equal to $2\pi a$. In other words, if C denotes a circle, then*

$$\pi = \frac{1}{2} \frac{\text{Perimeter of } C}{\text{Radius of } C}.$$

Proof. (i) The circular arc is given by the polar equation $r = p(\theta)$, where $p(\theta) := a$ for all $\theta \in [0, \varphi]$. Hence the length of the arc is equal to

$$\int_0^\varphi \sqrt{a^2 + 0^2}\, d\theta = a\varphi.$$

(ii) The perimeter of a circle is twice the arc length of its semicircle. Letting $\varphi = \pi$ in (i) above, we see that the perimeter of the circle is equal to $2\pi a$. The desired formula for π is then immediate. $\square$

The formula for π given in part (ii) of the above proposition makes it plain that the ratio of the perimeter of a circle to its diameter is independent of the radius. This fact is usually taken for granted when π is introduced in high-school geometry.

Part (i) of the above proposition says that the length of an arc of a semicircle is equal to the radius of the circle times the angle between 0 and π (in radian measure) that the arc subtends at the center. This explains the dictionary meaning of the word 'radian', namely an angle subtended at the center by an arc whose length is equal to the radius. Thus if the radius of a circle is 1, then the length of an arc of its semicircle is equal to the angle the arc subtends at the center. This also explains the use of the name 'arctangent' of the function whose inverse is the function $\tan : (-\pi/2, \pi/2) \to \mathbb{R}$. Indeed, for $x \in (0, \infty)$, we have $y = \arctan x$ if y is the length of an arc of the unit circle subtending an angle at the center whose tangent is x. For example, $\pi/3 = \arctan \sqrt{3}$ means that an arc of length $\pi/3$ of the unit circle subtends an angle θ at the center such that $\tan \theta = \sqrt{3}$.

Before considering some illustrative examples, we remark that the notion of the length of a smooth curve can be extended to slightly more general curves as follows. A parametrically defined curve C in $\mathbb{R}^2$ given by $(x(t), y(t))$, $t \in [\alpha, \beta]$, is said to be **piecewise smooth** if the functions x and y are continuous on $[\alpha, \beta]$ and if there is a finite number of points $\gamma_0 < \gamma_1 < \cdots < \gamma_n$ in $[\alpha, \beta]$, where $\gamma_0 = \alpha$ and $\gamma_n = \beta$, such that for each $i = 1, \ldots, n$, the curve given by $(x(t), y(t))$, $t \in [\gamma_{i-1}, \gamma_i]$, is smooth. If the curve C is piecewise smooth, then the **length** of C is defined to be

$$\ell(C) := \sum_{i=1}^n \int_{\gamma_{i-1}}^{\gamma_i} \sqrt{x'(t)^2 + y'(t)^2}\, dt.$$

In view of Propositions 6.7 and 6.12, we may write

$$\ell(C) := \int_\alpha^\beta \sqrt{x'(t)^2 + y'(t)^2}\, dt \quad \text{if } C \text{ is piecewise smooth.}$$

For example, if $x(t) := t$ and $y(t) := |t|$ for $t \in [-1, 1]$, then the curve given by $(x(t), y(t))$, $t \in [-1, 1]$, is piecewise smooth.

For a parametrically defined curve C in $\mathbb{R}^3$ given by $(x(t), y(t), z(t))$, $t \in [\alpha, \beta]$, we may define the concepts of 'smoothness' and 'piecewise smoothness'

analogously, and if C is a piecewise smooth curve, the **arc length** of C is defined to be

$$\ell(C) := \int_\alpha^\beta \sqrt{x'(t)^2 + y'(t)^2 + z'(t)^2}\, dt.$$

Examples 8.10. (i) Let $m, c \in \mathbb{R}$ and consider the line segment given by $y = mx + c$, $x \in [0, 1]$, from the point $(0, c)$ to the point $(1, m + c)$. Its length is equal to

$$\int_0^1 \sqrt{1 + m^2}\, dx = \sqrt{1 + m^2},$$

which is equal to the distance between the points $(0, c)$ and $(1, m + c)$.

(ii) Let $a \in \mathbb{R}$ and consider the parabolic curve given by $y = ax^2$, $x \in [0, 1]$. Its arc length is equal to

$$\int_0^1 \sqrt{1 + (2ax)^2}\, dx = \frac{1}{2a} \int_0^{2a} \sqrt{1 + u^2}\, du$$

$$= \frac{1}{2}\sqrt{1 + 4a^2} + \frac{1}{4a} \ln\left(2a + \sqrt{1 + 4a^2}\right).$$

(See Revision Exercise 46 (ii) given at the end of Chapter 7.)

(iii) Consider the curve given by $y = (2x^6 + 1)/8x^2$, $x \in [1, 2]$. Its arc length is equal to

$$\int_1^2 \sqrt{1 + \left(x^3 - \frac{1}{4x^3}\right)^2}\, dx = \int_1^2 \left(x^3 + \frac{1}{4x^3}\right) dx = \frac{123}{32}.$$

(iv) Let $a \in \mathbb{R}$ with $a > 0$, and consider the upper half of the cardioid given by $r = a(1 + \cos\theta)$, $\theta \in [0, \pi]$. Its arc length is equal to

$$\int_0^\pi \sqrt{a^2(1 + \cos\theta)^2 + a^2(-\sin\theta)^2}\, d\theta = \int_0^\pi \sqrt{2a^2(1 + \cos\theta)}\, d\theta$$

$$= 2a \int_0^\pi \cos(\theta/2)\, d\theta = 4a.$$

(v) Consider a **helix** in $\mathbb{R}^3$ given by the parametric equations

$$x(t) = a\cos t, \quad y(t) = a\sin t, \quad \text{and} \quad z(t) = bt + c, \quad t \in \mathbb{R},$$

where $a, b, c \in \mathbb{R}$ with $a > 0$ and $b \neq 0$. It lies on the cylinder given by $x^2 + y^2 = a^2$. [See Figure 8.17.] For $\alpha, \beta \in \mathbb{R}$ with $\alpha < \beta$, let C denote a part of the helix given by $(x(t), y(t), z(t))$, $t \in [\alpha, \beta]$. Then

$$\ell(C) = \int_\alpha^\beta \sqrt{(-a\sin t)^2 + (a\cos t)^2 + b^2}\, dt = (\beta - \alpha)\sqrt{a^2 + b^2}.$$

Consider points $P_1 := (x_1, y_1, z_1)$ and $P_2 := (x_2, y_2, z_2)$ on the cylinder given by $x^2 + y^2 = a^2$. Then $x_1^2 + y_1^2 = a^2 = x_2^2 + y_2^2$. Let us first assume
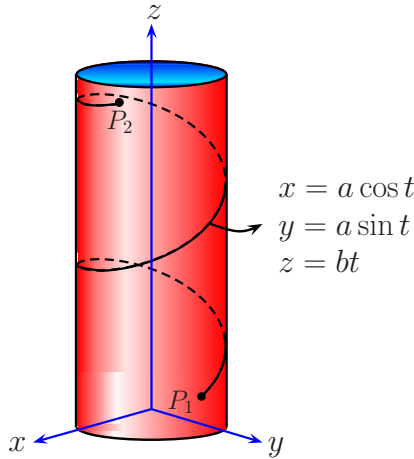
Fig. 8.17. A helix lying on the cylinder $x^2 + y^2 = a^2$

that $(x_1, y_1) \neq (x_2, y_2)$, that is, P_1 does not lie vertically above or below P_2, and also that $z_1 \neq z_2$, that is, P_1 and P_2 do not lie in a plane parallel to the xy-plane. If (a, θ_1) and (a, θ_2) denote the polar coordinates of (x_1, y_1) and (x_2, y_2) respectively, then $\theta_1 \neq \theta_2$ since $(x_1, y_1) \neq (x_2, y_2)$, and the helix given by the equations

$$x(t) = a \cos t, \ y(t) = a \sin t, \ z(t) = \frac{z_2 - z_1}{\theta_2 - \theta_1}(t - \theta_1) + z_1, \quad t \in \mathbb{R},$$

lies on the cylinder and passes through the points P_1 and P_2. We may assume that $\theta_1 < \theta_2$ without loss of generality. Letting $\alpha = \theta_1$ and $\beta = \theta_2$, it follows from what we have seen above that the arc length of the part of this helix from P_1 to P_2 is equal to

$$(\theta_2 - \theta_1)\sqrt{a^2 + \frac{(z_2 - z_1)^2}{(\theta_2 - \theta_1)^2}} = \sqrt{a^2(\theta_2 - \theta_1)^2 + (z_2 - z_1)^2}.$$

If we slit the cylinder vertically along a straight line parallel to the z-axis and open it up, then the points on the cylinder may be represented by $S :=$ $\{(s, z) \in \mathbb{R}^2 : -a\pi < s \leq a\pi\}$. In fact, a point $P = (x, y, z)$ on the cylinder corresponds to the point $Q := (a\theta, z)$ in S, where (a, θ) are the polar coordinates of (x, y). Let points $P_1 = (x_1, y_1, z_1)$ and $P_2 = (x_2, y_2, z_2)$ on the cylinder correspond to points $Q_1 := (a\theta_1, z_1)$ and $Q_2 := (a\theta_2, z_2)$ in S respectively. Then the part of the above-mentioned helix from P_1 to P_2 corresponds to the line segment from Q_1 to Q_2 in S. We note that the (Euclidean) distance between Q_1 and Q_2 is the same as the arc length of the part of this helix from P_1 to P_2. This is expressed by saying that if points P_1 and P_2 on a cylinder are neither one above the other nor at the

same height, then the **geodesic**, that is, the shortest path, on the cylinder from P_1 to P_2 is a helix.
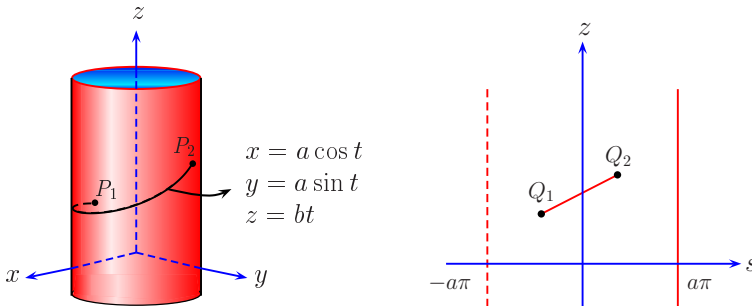
Fig. 8.18. Helix as a geodesic on the cylinder and the corresponding line segment when the cylinder is slit vertically and opened up

It is clear that if $P_1 = (x_1, y_1, z_1)$ and $P_2 = (x_2, y_2, z_2)$ lie one above the other, that is, if $(x_1, y_1) = (x_2, y_2)$, then the geodesic on the cylinder from P_1 to P_2 is the line segment given by $x(t) = x_1$, $y(t) = y_1$, $z(t) = (z_2 - z_1)t + z_1$, $t \in [0, 1]$. Also, it can be argued that if P_1 and P_2 are at the same height, that is, if $z_1 = z_2$, then the geodesic on the cylinder from P_1 to P_2 is the circular arc given by $x(t) = a \cos t$, $y(t) = a \sin t$, $z(t) = z_1$, $t \in [\theta_1, \theta_2]$. Likewise, if P_1 and P_2 are points on a sphere, then the geodesic on the sphere from P_1 to P_2 is an arc of the great circle passing through them. (A **great circle** is the intersection of the sphere with a plane passing through the center of the sphere.) To see this, one may rotate the sphere, if necessary, and assume that the great circle passing through P_1 and P_2 is in fact the equator of the sphere. ◇

Remark 8.11. Let us recall how we found the area of a circular disk of radius a in Section 8.1. We first found the area enclosed by the ellipse given by the equation $(x^2/a^2) + (y^2/b^2) = 1$, and then considered the special case $b = a$, which corresponds to a circle of radius a. Let us try to adopt a similar procedure and find the arc length of an ellipse. Consider an ellipse C parametrically given by $x(t) = a \cos t$, $y(t) = b \sin t$, $t \in [-\pi, \pi]$. Then we have

$$\ell(C) = \int_{-\pi}^{\pi} \sqrt{(-a \sin t)^2 + (b \cos t)^2}\, dt = 2 \int_0^{\pi} \sqrt{a^2 \sin^2 t + b^2 \cos^2 t}\, dt.$$

If $b = a$, then we easily get $\ell(C) = 2a\pi$, which gives the perimeter of a circle of radius a. On the other hand, if $b \neq a$, then the above integral cannot be evaluated in terms of known functions.

A similar situation occurs if we attempt to calculate the arc length of a lemniscate C given by the Cartesian equation $(x^2 + y^2)^2 = x^2 - y^2$, or by

the polar equation $r^2 = \cos 2\theta$. Considering its parametric equations $x(t) := \left(\cos t \sqrt{1 + \cos^2 t}\right)/\sqrt{2}$, $y(t) := (\cos t \sin t)/\sqrt{2}$, $t \in [-\pi, \pi]$, we obtain

$$\ell(C) = 2 \int_0^\pi \frac{1}{\sqrt{1 + \cos^2 t}} dt := 2\varpi, \text{ say.}$$

A special case of a formula of Gauss says that π/ϖ is equal to the arithmetic-geometric mean of $\sqrt{2}$ and 1. (See Exercise 12 of Chapter 2 for the definition. A simple proof of Gauss's formula is given in [50].) In the study of lemniscates, the number ϖ plays a role very similar to the role played by the number π in the study of circles. Notice, for example, that just as the length of a circle given by $x^2 + y^2 = 1$ is 2π, the length of the lemniscate C is 2ϖ. $\diamond$

8.4 Area of a Surface of Revolution

A surface of revolution is generated when a curve is revolved about a line. In this section, we shall define the area of such a surface and calculate it in several special cases. It may be remarked that the concept of the area of a general surface is usually developed in a course on multivariate calculus with the help of double integrals. It can be shown that the surfaces of revolution form a special case of the general development.

Let C be a parametrically defined curve in $\mathbb{R}^2$ given by $(x(t), y(t))$, $t \in [\alpha, \beta]$, and let L be a line in $\mathbb{R}^2$ given by the equation $ax + by + c = 0$, where $a, b, c \in \mathbb{R}$ and not both a and b are equal to zero. Let us begin by considering the curve C to be a line segment $P_1 P_2$ with endpoints $P_1 := (x_1, y_1)$ and $P_2 := (x_2, y_2)$. Thus C is parametrically given by $x(t) := (x_2 - x_1)t + x_1$, $y(t) := (y_2 - y_1)t + y_1$, $t \in [0, 1]$. Let us assume that the line segment $P_1 P_2$ does not cross the line L. Further, let d_1 and d_2 denote the distances of P_1 and P_2 from L, and let λ denote the length of the line segment $P_1 P_2$. We show that the 'area' of the surface generated by revolving $P_1 P_2$ about L is equal to

$$\pi(d_1 + d_2)\lambda.$$

To this end, first note that if $P_1 P_2 \perp L$, then $\lambda = |d_1 - d_2|$ and the surface of revolution is a circular washer with radii d_1 and d_2. [See the picture on the left in Figure 8.19.] Thus its area is equal to

$$|\pi d_1^2 - \pi d_2^2| = \pi(d_1 + d_2)|d_1 - d_2| = \pi(d_1 + d_2)\lambda.$$

Next, if $P_1 P_2 \parallel L$, then $d_1 = d_2 = d$ say, and the surface of revolution is a right circular cylinder with radius d and length λ. [See the picture on the right in Figure 8.19.] If we slit open this cylinder along a straight line parallel to its axis, we obtain a rectangle of sides $2\pi d$ and λ. Hence its area is equal to
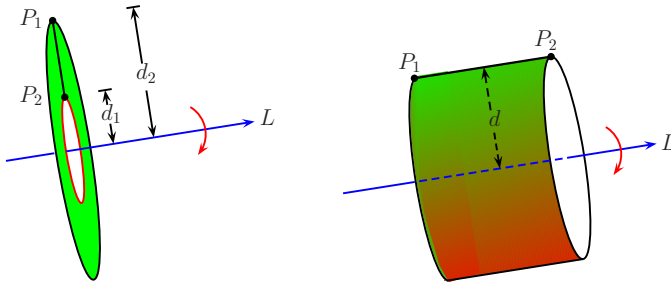
Fig. 8.19. Revolving the line segment P_1P_2 about L when $P_1P_2 \perp L$ and when $P_1P_2 \parallel L$

$$2\pi d\lambda = \pi(d_1 + d_2)\lambda.$$

Assume now that $P_1P_2 \not\perp L$ and $P_1P_2 \not\parallel L$. Then the surface of revolution is a frustum (that is, a piece) of a right circular cone with base radii d_1 and d_2, and slant height λ. In order to find the area of this frustum, let us first find the area of a right circular cone with base radius a and slant height ℓ.

If we slit open the cone along a straight line from its vertex to a point in its base, we obtain a sector of a circle of radius ℓ such that the length of its arc is equal to $2\pi a$. [See the picture on the left in Figure 8.20.] Note that $2\pi a = \ell\,\theta$, where $\theta = 2\pi a/\ell$. By part (iii) of Proposition 8.2, the area of this sector is equal to

$$\frac{1}{2}\ell^2\theta = \pi\ell a,$$

which is therefore the surface area of a right circular cone with base radius a and slant height ℓ.
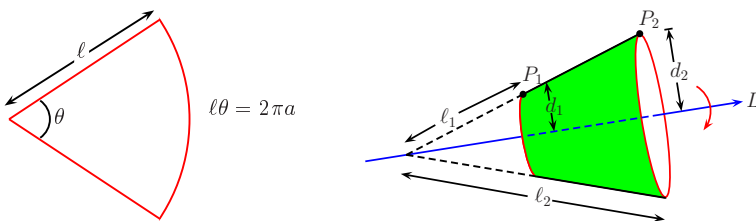
Fig. 8.20. Sector of a circle and the frustum of a right circular cone

To find the surface area of the frustum of the right circular cone with base radii d_1 and d_2, and slant height λ, we may assume without loss of generality that $d_1 < d_2$. Then this frustum is obtained by removing from a cone of radius d_2 and slant height ℓ_2 a smaller cone of radius d_1 and slant height ℓ_1, where $\ell_1, \ell_2 \in \mathbb{R}$ satisfy $\ell_2 > \ell_1 > 0$ and $\ell_2 - \ell_1 = \lambda$. [See the picture on the right

in Figure 8.20.] Using similarity of triangles, we have $d_1\ell_2 = d_2\ell_1$. Hence the area of the surface of the frustum is equal to

$$\pi d_2\ell_2 - \pi d_1\ell_1 = \pi(d_2\ell_2 - d_2\ell_1 + d_1\ell_2 - d_1\ell_1) = \pi(d_1 + d_2)(\ell_2 - \ell_1) = \pi(d_1 + d_2)\lambda,$$

as desired.

Consider now the general case in which C is a parametrically defined curve given by $(x(t), y(t))$, $t \in [\alpha, \beta]$, and let $\{t_0, t_1, \ldots, t_n\}$ be a partition of $[\alpha, \beta]$. Let us replace the piece $(x(t), y(t))$, $t \in [t_{i-1}, t_i]$, of the curve C by the line segment $P_{i-1}P_i$ for $i = 0, 1, \ldots, n$, where $P_i := (x(t_i), y(t_i))$. [See Figure 8.21.] Then the sum of the areas of the frustums of the cones generated by these line segments is equal to
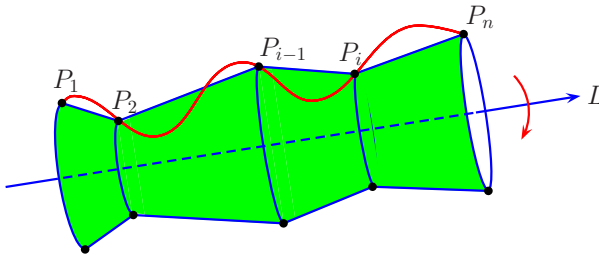
Fig. 8.21. A surface of revolution approximated by frustums of right circular cones

$$\sum_{i=1}^{n} \pi(d_{i-1} + d_i)\lambda_i,$$

where d_i is the distance of P_i from the line L for $i = 0, 1, \ldots, n$ and λ_i is the length of the line segment $P_{i-1}P_i$ for $i = 1, \ldots, n$. This sum can be considered as an approximation of the conceived area of the surface obtained by revolving the curve C about the line L. Now for each i, we have

$$d_i = \frac{|ax(t_i) + by(t_i) + c|}{\sqrt{a^2 + b^2}} \quad \text{and} \quad \lambda_i = \sqrt{[x(t_i) - x(t_{i-1})]^2 + [y(t_i) - y(t_{i-1})]^2}.$$

If the functions x and y are continuous on $[t_{i-1}, t_i]$ and differentiable on (t_{i-1}, t_i), then by the MVT, there are $s_i, u_i \in (t_{i-1}, t_i)$ such that

$$\lambda_i = \sqrt{x'(s_i)^2 + y'(u_i)^2}\,(t_i - t_{i-1}), \quad i = 1, \ldots, n,$$

and hence the sums $\sum_{i=1}^{n} d_{i-1}\lambda_i$ and $\sum_{i=1}^{n} d_i\lambda_i$ may be considered as approximations of the integral

$$\int_{\alpha}^{\beta} \frac{|ax(t) + by(t) + c|}{\sqrt{a^2 + b^2}} \sqrt{x'(t)^2 + y'(t)^2}\, dt.$$

These considerations lead to the following definition of the area of a surface of revolution.

Let C be a piecewise smooth curve given by $(x(t), y(t))$, $t \in [\alpha, \beta]$. Consider a line L given by $ax + by + c = 0$, where not both a and b are equal to zero. Assume that the line L does not cross the curve C. For $t \in [\alpha, \beta]$, let $\rho(t)$ denote the distance of the point $(x(t), y(t))$ on C from the line L, that is,

$$\rho(t) = \frac{|ax(t) + by(t) + c|}{\sqrt{a^2 + b^2}}.$$

Then the area of the surface S of revolution obtained by revolving the curve C about the line L is defined to be

$$\text{Area } (S) := 2\pi \int_{\alpha}^{\beta} \rho(t) \sqrt{x'(t)^2 + y'(t)^2} \, dt.$$

We note that since the line L does not cross the curve C, we have either $ax(t) + by(t) + c \geq 0$ for all $t \in [\alpha, \beta]$, or $ax(t) + by(t) + c \leq 0$ for all $t \in [\alpha, \beta]$. We consider some important special cases.

1. If a piecewise smooth curve C given by $y = f(x)$, $x \in [a, b]$, where $f(x) \geq 0$ for all $x \in [a, b]$ or $f(x) \leq 0$ for all $x \in [a, b]$, is revolved about the x-axis, then the area of the surface S of revolution so generated is equal to

$$\text{Area } (S) = 2\pi \int_{a}^{b} |f(x)| \sqrt{1 + f'(x)^2} \, dx.$$

Similarly, if a piecewise smooth curve C given by $x = g(y)$, $y \in [c, d]$, where $g(y) \geq 0$ for all $y \in [c, d]$ or $g(y) \leq 0$ for all $y \in [c, d]$, is revolved about the y-axis, then the area of the surface S of revolution so generated is equal to

$$\text{Area } (S) = 2\pi \int_{c}^{d} |g(y)| \sqrt{1 + g'(y)^2} \, dy.$$

2. Let a piecewise smooth curve C be given by $r = p(\theta)$, $\theta \in [\alpha, \beta]$, where $p(\theta) \geq 0$ for all $\theta \in [\alpha, \beta]$. If L denotes a line through the origin containing a ray given by $\theta = \gamma$, and not crossing the curve C, then the area of the surface S obtained by revolving the curve C about the line L is equal to

$$\text{Area } (S) = 2\pi \int_{\alpha}^{\beta} p(\theta) |\sin(\theta - \gamma)| \sqrt{p(\theta)^2 + p'(\theta)^2} \, d\theta.$$

This follows by considering the polar coordinate θ as a parameter and noting, as in Section 8.3, that $x'(\theta)^2 + y'(\theta)^2 = p(\theta)^2 + p'(\theta)^2$, $\theta \in [\alpha, \beta]$, and also that the distance of the point $(p(\theta) \cos \theta, p(\theta) \sin \theta)$ from the line L is equal to $p(\theta)|\sin(\theta - \gamma)|$ for all $\theta \in [\alpha, \beta]$. [See Figure 8.22.]

Area of a surface generated by revolving a curve given by a polar equations of the form $\theta = \alpha(r)$ is treated in Exercises 40 and 41.
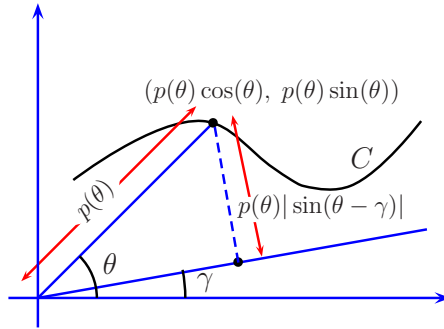
Fig. 8.22. Revolving a polar curve C about the line making an angle γ with the positive x-axis

Proposition 8.12. (i) *Let $\varphi \in [0, \pi]$ and C denote the arc of a circle given by $x := a \cos t$, $y := a \sin t$, $0 \le t \le \varphi$ (which subtends an angle φ at the center). Then the surface area of the spherical cap generated by revolving C about the x-axis is equal to $2\pi a^2 (1 - \cos \varphi)$.*

(ii) *The surface area of a sphere given by $x^2 + y^2 + z^2 = a^2$ is equal to $4\pi a^2$. In other words, if S denotes this sphere, then*

$$\pi = \frac{1}{4} \frac{Surface\ Area\ of\ S}{(Radius\ of\ S)^2}.$$

Proof. (i) The arc C is given by the polar equation $r = p(\theta)$, where $p(\theta) := a$ for $\theta \in [0, \varphi]$. If it is revolved about the x-axis, then the area of the spherical cap so generated is equal to

$$2\pi \int_0^\varphi a |\sin \theta| \sqrt{a^2 + 0^2}\, d\theta = 2\pi a^2 \int_0^\varphi \sin \theta\, d\theta = 2\pi a^2 (1 - \cos \varphi).$$

(ii) A sphere of radius a is obtained by revolving a semicircle of radius a about the line containing its diameter. Letting $\varphi = \pi$ in the formula obtained in (i) above, we see that the surface area of a sphere of radius a is equal to $2\pi a^2 (1 - (-1)) = 4\pi a^2$. $\square$

The formula for π given in part (ii) of the above proposition makes it plain that the ratio of the surface area of a sphere to the square of its radius is independent of the radius.

Remark 8.13. Let S be a surface lying on a sphere of radius a. Then S is said to subtend a **solid angle** Θ at the center of the sphere, where Θ is equal to the 'surface area' of S divided by a^2. For example, let $\varphi \in [0, \pi]$ and consider the arc of the circle of radius a given by $x := a \cos t$, $y := a \sin t$, $0 \le t \le \varphi$, which subtends an angle φ at the center of the circle. Then by part (i) of the above proposition, the spherical cap generated by revolving this arc about the x-axis

subtends a solid angle $\Theta_\varphi = 2\pi(1 - \cos\varphi)$ at the center of the sphere. By part (iii) of Proposition 8.5, the volume of the corresponding (solid) spherical cone is equal to $2\pi a^3(1 - \cos\varphi)/3 = a^3\Theta_\varphi/3$.

In particular, letting $\varphi = \pi/2$ and $\varphi = \pi$ in part (i) of the above proposition, we see that a hemisphere subtends a solid angle 2π at the center and the entire sphere subtends a solid angle 4π at the center. (Note that the volume of the solid enclosed by the entire sphere of radius a, that is, of the the ball of radius a, is equal to $(a^3 \times 4\pi)/3$, in conformity with part (ii) of Proposition 8.5.) The standard unit of a solid angle is known as **steradian**. The largest solid angle, therefore, is of 4π steradians, that is, of approximately 12.566 steradians. $\diamond$

Examples 8.14. (i) Consider the line segment given by $(x/a) + (y/h) = 1$, $x \in [0, a]$, where $a, h > 0$. The surface area of the cone S of radius a and height h generated by revolving this line segment about the y-axis is equal to

$$\text{Area}\,(S) = 2\pi \int_0^h a\left(1 - \frac{y}{h}\right)\sqrt{1 + \left(\frac{a}{h}\right)^2}\,dy = 2\pi a\frac{\sqrt{a^2 + h^2}}{h}\left(h - \frac{h}{2}\right)$$
$$= \pi a\sqrt{a^2 + h^2},$$

as expected.

(ii) Consider the spheroid S generated by revolving the ellipse $(x^2/a^2) + (y^2/b^2) = 1$, where $a, b > 0$, $a \neq b$, about the x-axis. It is given by the Cartesian equation $(x^2/a^2) + (y^2/b^2) + (z^2/b^2) = 1$. To find its surface area, it suffices to consider the curve C given by $y = (b/a)\sqrt{a^2 - x^2}$, $x \in [-a, a]$. We obtain

$$\text{Area}\,(S) = 2\pi \int_{-a}^a \frac{b}{a}\sqrt{a^2 - x^2}\sqrt{1 + \frac{b^2 x^2}{a^2(a^2 - x^2)}}\,dx$$
$$= \frac{2\pi b}{a} \int_{-a}^a \sqrt{a^2 + \frac{(b^2 - a^2)x^2}{a^2}}\,dx.$$

First consider the case $a < b$. If $c := \sqrt{b^2 - a^2}/a$, then we have $c > 0$ and

$$\text{Area}\,(S) = \frac{2\pi b}{a} \cdot 2a \int_0^a \sqrt{1 + \left(\frac{cx}{a}\right)^2}\,dx = \frac{4\pi ab}{c} \int_0^c \sqrt{1 + t^2}\,dt$$
$$= \frac{2\pi ab}{c}\left[c\sqrt{1 + c^2} + \ln\left(c + \sqrt{1 + c^2}\right)\right].$$

(See Revision Exercise 46 (ii) given at the end of Chapter 7.) Next, consider the case $a > b$. If $c := \sqrt{a^2 - b^2}/a$, then we have $0 < c < 1$ and

$$\text{Area}\,(S) = \frac{2\pi b}{a} \cdot 2a \int_0^a \sqrt{1 - \left(\frac{cx}{a}\right)^2}\,dx = \frac{4\pi ab}{c} \int_0^c \sqrt{1 - t^2}\,dt$$
$$= \frac{2\pi ab}{c}\left(c\sqrt{1 - c^2} + \sin^{-1} c\right).$$

(See Revision Exercise 46 (iii) given at the end of Chapter 7.) In both cases, if $b \to a$, then L'Hôpital's Rule for $\frac{0}{0}$ indeterminate forms (Proposition 4.37) shows that the surface area tends to $4\pi a^2$, which is the surface area of a sphere of radius a, as seen in part (ii) of Proposition 8.12.

We remark that the surface of the ellipsoid given by $(x^2/a^2) + (y^2/b^2) + (z^2/c^2) = 1$, where a, b, c are distinct positive numbers, is not a surface of revolution. In fact, the calculation of the surface area of such an ellipsoid involves the so-called elliptic integrals.

(iii) Consider the torus S obtained by revolving the circle given by $(x - a)^2 + y^2 = b^2$, where $0 < b < a$, about the y-axis. To find its area, we use the parametric equations $x(t) := a + b\cos t$, $y(t) := b\sin t, t \in [-\pi, \pi]$. Hence

$$\text{Area } (S) = 2\pi \int_{-\pi}^{\pi} (a + b\cos t)\sqrt{(-b\sin t)^2 + (b\cos t)^2} \, dt$$

$$= 2\pi b \int_{-\pi}^{\pi} (a + b\cos t)dt = 4\pi^2 ab$$

is the required area.   ◇

8.5 Centroids

Before introducing the concept of a centroid of a geometrical object, let us consider the notion of the average of a function. For this purpose, we recall that given $n \in \mathbb{N}$ and a function $f : \{1, \ldots, n\} \to \mathbb{R}$, the average of the values of f at the points $1, \ldots, n$, that is, the **average** of f, is given by

$$\text{Av}(f) := \frac{f(1) + \cdots + f(n)}{n}.$$

In general, there can be repetition among the values $f(1), \ldots, f(n)$ of f. If $y_1, \ldots, y_k$ are the distinct values of f, and if for each $j = 1, \ldots, k$, the function f assumes the value y_j at a total number of n_j points, that is, the set $\{i \in \mathbb{N} : 1 \le i \le n \text{ and } f(i) = y_j\}$ has n_j elements, then $n_1 + \cdots + n_k = n$ and we can also write

$$\text{Av}(f) = \frac{n_1 y_1 + \cdots + n_k y_k}{n_1 + \cdots + n_k}.$$

Simple examples show that $\text{Av}(f)$ need not be any of the values of f.

Next, consider a function $f : \mathbb{N} \to \mathbb{R}$. How can we define the average of f? One possibility is

$$\text{Av}(f) := \lim_{n \to \infty} \frac{f(1) + \cdots + f(n)}{n},$$

if the limit exists. Part (i) of Exercise 21 of Chapter 2 shows that if the sequence $(f(n))$ is convergent, then this limit exists and is equal to $\lim_{n \to \infty} f(n)$, but $\text{Av}(f)$ may exist even if the sequence $(f(n))$ is divergent.

Let us now consider a closed interval $[a, b]$ in $\mathbb{R}$ and a function $f : [a, b] \to \mathbb{R}$. How should we define the average of f? Suppose $P = \{x_0, x_1, \ldots, x_n\}$ is a partition of $[a, b]$ and we choose $s_i \in (x_{i-1}, x_i)$ for $i = 1, \ldots, n$. If f were to assume the value $f(s_i)$ at every point of the subinterval (x_{i-1}, x_i), then we may define

$$\mathrm{Av}(f) := \frac{f(s_1)(x_1 - x_0) + \cdots + f(s_n)(x_n - x_{n-1})}{(x_1 - x_0) + \cdots + (x_n - x_{n-1})} = \frac{1}{b - a} \sum_{i=1}^{n} f(s_i)(x_i - x_{i-1})$$

in analogy with the discrete case considered earlier.

Now assume that f is integrable on $[a, b]$. In view of the result of Darboux about Riemann sums (Proposition 6.31), we define the **average** of f by

$$\mathrm{Av}(f) := \frac{1}{b - a} \int_a^b f(x) dx.$$

As in the case of a function defined on $\{1, \ldots, n\}$, the average of f need not be any of the values of f. For example, consider $f : [-1, 1] \to \mathbb{R}$ defined by $f(x) := 1$ if $x \geq 0$ and $f(x) := -1$ if $x < 0$. Then

$$\mathrm{Av}(f) = \frac{1}{1 - (-1)} \int_{-1}^{1} f(x) dx = \frac{1}{2} \left(\int_{-1}^{0} f(x) dx + \int_0^1 f(x) dx \right)$$
$$= \frac{1}{2}(-1 + 1) = 0,$$

but f does not assume the value 0 at any point. If, however, f is continuous, then $\mathrm{Av}(f)$ is always one of the values of f. (See Exercise 72 of this chapter as well as Exercise 50 of Chapter 6.)

'Weighted' averages arise when we wish to give either more or less importance to some of the values of a function. Given $n \in \mathbb{N}$ and $f : \{1, \ldots, n\} \to \mathbb{R}$, let $w(1), \ldots, w(n)$ be nonnegative numbers such that $w(1) + \cdots + w(n) \neq 0$. If we decide to assign weights $w(1), \ldots, w(n)$ to the values $f(1), \ldots, f(n)$ respectively, then the **weighted average** of f with respect to these weights is given by

$$\mathrm{Av}(f; w) := \frac{w(1)f(1) + \cdots + w(n)f(n)}{w(1) + \cdots + w(n)}.$$

With this mind, we make the following definitions. An integrable function $w : [a, b] \to \mathbb{R}$ is called a **weight function** if w is nonnegative and $W := \int_a^b w(x) dx \neq 0$. Let $f : [a, b] \to \mathbb{R}$ be an integrable function and $w : [a, b] \to \mathbb{R}$ be a weight function. Then the **weighted average** of f with respect to w is defined by

$$\mathrm{Av}(f; w) := \frac{1}{W} \int_a^b f(x) w(x) dx.$$

Note that the product fw of the functions f and w is integrable by part (iii) of Proposition 6.15. Observe that if $w(x) = 1$ for all $x \in [a, b]$, then $\mathrm{Av}(f; w) = \mathrm{Av}(f)$.

Roughly speaking, the **centroid** of a set is a point whose coordinates are the averages (or the weighted averages) of the corresponding coordinate functions defined on the set. Thus the x-coordinate (or the first coordinate) of the centroid of a subset D of $\mathbb{R}^3$ is the average value of the function $f : D \to \mathbb{R}$ given by $f(x, y, z) = x$. Similar comments hold for the other coordinates. The difficulty in defining a centroid at this stage lies in the fact that so far we have only defined the average of a function defined on an *interval* $[a, b]$. To be able to deal with centroids of more general sets, we would have to extend the notion of an average to functions defined on them. This is usually done in a course on multivariate calculus. At present, we shall see how centroids of a limited variety of sets can still be defined using Riemann integrals. These definitions turn out to be special cases of the general treatment given in a course on multivariate calculus.

Curves and Surfaces

To begin with, let us consider a line segment $P_1 P_2$ in $\mathbb{R}^2$ with endpoints $P_1 := (x_1, y_1)$ and $P_2 := (x_2, y_2)$. Let $x, y : [0, 1] \to \mathbb{R}$ be the functions defined by

$$x(t) := (x_2 - x_1)t + x_1 \quad \text{and} \quad y(t) := (y_2 - y_1)t + y_1.$$

As t runs over the parameter interval $[0, 1]$, we obtain all the points $(x(t), y(t))$ on the line segment $P_1 P_2$. The **centroid** of $P_1 P_2$ is defined to be the point $(\overline{x}, \overline{y})$ given by

$$\overline{x} = \frac{1}{1 - 0} \int_0^1 x(t)dt = \frac{x_1 + x_2}{2} \quad \text{and} \quad \overline{y} = \frac{1}{1 - 0} \int_0^1 x(t)dt = \frac{y_1 + y_2}{2}.$$

Thus the centroid of $P_1 P_2$ is its midpoint.

More generally, consider a piecewise smooth curve C given by $(x(t), y(t))$, $t \in [\alpha, \beta]$. While defining the centroid of C, we shall use weighted averages with the weight function $w : [\alpha, \beta] \to \mathbb{R}$ given by $w(t) := \sqrt{x'(t)^2 + y'(t)^2}$. This is reasonable since the length of the curve C is given by

$$\ell(C) = \int_\alpha^\beta \sqrt{x'(t)^2 + y'(t)^2}\, dt.$$

If $\ell(C) \neq 0$, the **centroid** $(\overline{x}, \overline{y})$ of C is defined by

$$\overline{x} = \frac{1}{\ell(C)} \int_\alpha^\beta x(t)\sqrt{x'(t)^2 + y'(t)^2}\, dt$$

and

$$\overline{y} = \frac{1}{\ell(C)} \int_\alpha^\beta y(t)\sqrt{x'(t)^2 + y'(t)^2}\, dt.$$

Note that $\overline{x} = \mathrm{Av}(x; w)$ and $\overline{y} = \mathrm{Av}(y; w)$.

Let us now consider a surface S generated by revolving the above-mentioned curve C about a line L given by $ax + by + c = 0$ that does not cross C. As in the previous section, let $\rho(t) := |ax(t) + by(t) + c|/\sqrt{a^2 + b^2}$ denote the distance of the point $(x(t), y(t))$, $t \in [\alpha, \beta]$, on the curve C from this line. Assume that

$$\text{Area } (S) := 2\pi \int_\alpha^\beta \rho(t)\sqrt{x'(t)^2 + y'(t)^2}\, dt$$

is not equal to zero. If the line L is the x-axis, then we define the **centroid** $(\overline{x}, \overline{y}, \overline{z})$ of S by $\overline{y} := 0$, $\overline{z} := 0$ (because of symmetry), and

$$\overline{x} := \frac{2\pi}{\text{Area } (S)} \int_\alpha^\beta x(t)|y(t)|\sqrt{x'(t)^2 + y'(t)^2}\, dt.$$

On the other hand, if the line L is the y-axis, then we define the **centroid** $(\overline{x}, \overline{y}, \overline{z})$ of S by $\overline{x} := 0$, $\overline{z} := 0$ (because of symmetry), and

$$\overline{y} := \frac{2\pi}{\text{Area } (S)} \int_\alpha^\beta y(t)|x(t)|\sqrt{x'(t)^2 + y'(t)^2}\, dt.$$

Note that $\overline{x} = \text{Av}(x; w)$, $\overline{y} = \text{Av}(y; w)$, and $\overline{z} = \text{Av}(z; w)$, where the weight function $w : [\alpha, \beta] \to \mathbb{R}$ is given by $w(t) := \rho(t)\sqrt{x'(t)^2 + y'(t)^2}$.

We remark that the centroid of a surface of revolution about an arbitrary line L, and more generally the centroid of a surface in $\mathbb{R}^3$, can be defined with the help of double integrals. This is usually done in a course on multivariate calculus and it can be shown that the above formulas are particular cases of the general definition.

Examples 8.15. (i) Let $a \in \mathbb{R}$ with $a > 0$. Consider a semicircle C of radius a given by $x(t) := a \cos t$, $y(t) := a \sin t$, $t \in [0, \pi]$. Then

$$\ell(C) = \int_0^\pi \sqrt{(-a \sin t)^2 + (a \cos t)^2}\, dt = a\pi.$$

Hence

$$\overline{x} = \frac{1}{a\pi} \int_0^\pi a \cos t \cdot a\, dt = 0 \quad \text{and} \quad \overline{y} = \frac{1}{a\pi} \int_0^\pi a \sin t \cdot a\, dt = \frac{2a}{\pi}.$$

We note that the centroid $(0, 2a/\pi)$ of the semicircle C does not lie on C. Also, we could have directly obtained the x-coordinate $\overline{x}$ of the centroid to be equal to 0 by the symmetry of the semicircle about the y-axis.

(ii) Consider a cycloid C given by $x(t) := t - \sin t$ and $y(t) := 1 - \cos t$, $t \in [0, 2\pi]$. Then

$$\ell(C) = \int_0^{2\pi} \sqrt{(1 - \cos t)^2 + (\sin t)^2}\, dt = \int_0^{2\pi} \sqrt{2 - 2\cos t}\, dt$$

$$= \int_0^{2\pi} 2 \sin \frac{t}{2}\, dt = 8.$$

Hence

$$\overline{x} = \frac{1}{8} \int_0^{2\pi} (t - \sin t) \cdot 2 \sin \frac{t}{2} \, dt = \pi \quad \text{and} \quad \overline{y} = \frac{1}{8} \int_0^{2\pi} (1 - \cos t) \cdot 2 \sin \frac{t}{2} \, dt = \frac{4}{3}.$$

Thus the centroid of C is at $\left(\pi, \frac{4}{3}\right)$.

(iii) Let $a \in \mathbb{R}$ with $a > 0$. Consider the surface S generated by revolving an arc of the circle given by $x(t) := a \cos t$, $y(t) := a \sin t$, $t \in [0, \pi/2]$, about the line $y = -a$. Since the distance of $(x(t), y(t))$ from this line is equal to $a + a \sin t$ for $t \in [0, \pi]$, the area of S is equal to

$$2\pi \int_0^{\pi/2} a(1 + \sin t) \sqrt{(-a \sin t)^2 + (a \cos t)^2} \, dt = a^2 \pi (\pi + 2).$$

It can be seen that $\overline{y} = -a$ and $\overline{z} = 0$ by symmetry, whereas

$$\overline{x} = \frac{1}{a^2 \pi (\pi + 2)} \cdot 2\pi \int_0^{\pi/2} (a \cos t) a (1 + \sin t) a \, dt = \frac{3\pi a^3}{a^2 \pi (\pi + 2)} = \frac{3a}{\pi + 2}.$$

Thus the centroid of S is at $(3a/(\pi + 2), -a, 0)$.

(iv) Let a and h be positive real numbers. The surface S of a right circular cone of base radius a and height h is generated by revolving the line segment given by $(x/a) + (y/h) = 1$, $x \in [0, a]$, about the y-axis. As we have seen in Example 8.14 (i), $A(S) = \pi a \sqrt{a^2 + h^2}$. Then $\overline{x} = 0 = \overline{z}$ and

$$\overline{y} = \frac{2\pi}{\pi a \sqrt{a^2 + h^2}} \int_0^h y a \left(1 - \frac{y}{h}\right) \sqrt{1 + \frac{a^2}{h^2}} \, dy = \frac{h}{3},$$

as one may expect. $\diamond$

Planar Regions and Solid Bodies

Let us first consider the centroids of certain planar regions that we dealt with in Section 8.1. Let $f_1, f_2 : [a, b] \to \mathbb{R}$ be integrable functions such that $f_1 \leq f_2$, and consider the region

$$R := \{(x, y) \in \mathbb{R}^2 : a \leq x \leq b \text{ and } f_1(x) \leq y \leq f_2(x)\}.$$

Recall that the area of R is defined to be

$$\text{Area } (R) := \int_a^b [f_2(x) - f_1(x)] dx.$$

Let us assume that Area $(R) \neq 0$. In view of our comments in the introduction of this section, we define the x-coordinate of the **centroid** of R by

$$\overline{x} := \frac{1}{\text{Area } (R)} \int_a^b x[f_2(x) - f_1(x)] dx.$$

In order to define the y-coordinate of the centroid of R, we observe that for $x \in [a, b]$, the vertical slice of the region R at x has length $f_2(x) - f_1(x)$ and its centroid is its midpoint $(x, [f_1(x) + f_2(x)]/2)$. Since $[f_2(x) - f_1(x)] \cdot [f_1(x) + f_2(x)]/2 = [f_2(x)^2 - f_1(x)^2]/2$ for all $x \in [a, b]$, it is reasonable to define the y-coordinate of the **centroid** of R by

$$\overline{y} := \frac{1}{2\text{Area}(R)} \int_a^b [f_2(x)^2 - f_1(x)^2] dx.$$

Note that $\overline{x} = \text{Av}(f; w)$ and $\overline{y} = \text{Av}(g; w)$, where the functions $f, g, w : [a, b] \to \mathbb{R}$ are given by $f(x) := x$, $g(x) := [f_1(x) + f_2(x)]/2$, $w(x) := [f_2(x) - f_1(x)]$.

Similarly, if $g_1, g_2 : [c, d] \to \mathbb{R}$ are integrable functions such that $g_1 \leq g_2$, then the **centroid** $(\overline{x}, \overline{y})$ of the region

$$R := \{(x, y) \in \mathbb{R}^2 : c \leq y \leq d \text{ and } g_1(y) \leq x \leq g_2(y)\}$$

is defined by

$$\overline{x} = \frac{1}{2\text{Area}(R)} \int_c^d [g_2(y)^2 - g_1(y)^2] dy$$

and

$$\overline{y} = \frac{1}{\text{Area}(R)} \int_c^d y[g_2(y) - g_1(y)] dy,$$

provided the area

$$\text{Area}(R) := \int_d^c [g_2(y) - g_1(y)] dy$$

of R is not zero.

Finally, we shall consider the centroids of certain solid bodies that we dealt with in Section 8.2. First suppose that a bounded solid D lies between the planes given by $x = a$ and $x = b$, where $a, b \in \mathbb{R}$ with $a < b$, and for $x \in [a, b]$, let $A(x)$ denote the area of the slice of D at x by a plane perpendicular to the x-axis. Assume that

$$\text{Vol}(D) := \int_a^b A(x) dx$$

is not equal to zero. For each $x \in [a, b]$, the x-coordinate of the centroid of the slice of D at x is x itself. In view of this, we define the x-coordinate of the **centroid** of D by

$$\overline{x} := \frac{1}{\text{Vol}(D)} \int_a^b x A(x) dx.$$

Further, if for each $x \in [a, b]$, the y-coordinate and the z-coordinate of the centroid of the slice of D at x are given by $\tilde{y}(x)$ and $\tilde{z}(x)$, then we define the y-coordinate and the z-coordinate of the **centroid** of D by

$$\overline{y} := \frac{1}{\text{Vol}\,(D)} \int_a^b \tilde{y}(x)A(x)dx \quad \text{and} \quad \overline{z} := \frac{1}{\text{Vol}\,(D)} \int_a^b \tilde{z}(x)A(x)dx.$$

Note that $\overline{x} = \text{Av}(f; A)$, $\overline{y} = \text{Av}(g; A)$ and $\overline{z} = \text{Av}(h; A)$, where the functions $f, g, h : [a, b] \to \mathbb{R}$ are given by $f(x) := x$, $g(x) := \tilde{y}(x)$, $h(x) := \tilde{z}(x)$.

In particular, let the solid D be generated by revolving the region

$$\{(x, y) \in \mathbb{R}^2 : a \le x \le b \text{ and } f_1(x) \le y \le f_2(x)\}$$

about the x-axis, where $f_1, f_2 : [a, b] \to \mathbb{R}$ are integrable functions such that $0 \le f_1 \le f_2$. Recall that by the Washer Method, $A(x) = \pi[f_2(x)^2 - f_1(x)^2]$ for all $x \in [a, b]$ and

$$\text{Vol}\,(D) = \pi \int_b^a [f_2(x)^2 - f_1(x)^2]dx.$$

Then we have

$$\overline{x} = \frac{\pi}{\text{Vol}\,(D)} \int_a^b x[f_2(x)^2 - f_1(x)^2]dx,$$

whereas $\overline{y} = 0 = \overline{z}$, since $\tilde{y}(x) = 0 = \tilde{z}(x)$ for all $x \in [a, b]$ by symmetry.

Similar considerations hold for a solid whose volume is given by

$$\int_c^d A(y)dy \quad \text{or} \quad \int_p^q A(z)dz,$$

as described in Section 8.2.

Next, suppose that a bounded solid D lies between the cylinders given by $x^2 + y^2 = p^2$ and $x^2 + y^2 = q^2$, where $p, q \in \mathbb{R}$ with $0 \le p < q$. For $r \in [p, q]$, consider the sliver $\{(x, y, z) \in D : x^2 + y^2 = r^2\}$ of D by the cylinder given by $x^2 + y^2 = r^2$; let $E_r := \{(\theta, z) \in [-\pi, \pi] \times \mathbb{R} : (r\cos\theta, r\sin\theta, z) \in D\}$ denote the parameter domain for this sliver and let $B(r)$ denote the area of the planar region E_r. Assume that

$$\text{Vol}\,(D) := \int_p^q rB(r)dr$$

is not equal to zero. For each $r \in [p, q]$, if $B(r) \ne 0$ and $(\tilde{x}(r), \tilde{y}(r), \tilde{z}(r))$ denotes the centroid of the sliver of D at r for $r \in [p, q]$, then the **centroid** $(\overline{x}, \overline{y}, \overline{z})$ of the solid D is defined by

$$\overline{x} := \frac{\displaystyle\int_p^q \tilde{x}(r)rB(r)dr}{\text{Vol}\,(D)}, \quad \overline{y} := \frac{\displaystyle\int_p^q \tilde{y}(r)rB(r)dr}{\text{Vol}\,(D)}, \quad \text{and} \quad \overline{z} := \frac{\displaystyle\int_p^q \tilde{z}(r)rB(r)dr}{\text{Vol}\,(D)}.$$

Note that $\overline{x} = \text{Av}(f; A)$, $\overline{y} = \text{Av}(g; A)$, and $\overline{z} = \text{Av}(h; A)$, where the functions $A, f, g, h : [p, q] \to \mathbb{R}$ are given by $A(r) := rB(r)$, $f(r) := \tilde{x}(r)$, $g(r) := \tilde{y}(r)$, $h(r) := \tilde{z}(r)$.

Similar considerations hold for a solid whose volume is given by

$$\int_a^b xB(x)dx \quad \text{or} \quad \int_c^d yB(y)dy,$$

as described in Section 8.2.

In particular, let the solid D be generated by revolving the region

$$R := \{(x,y) \in \mathbb{R}^2 : a \leq x \leq b \text{ and } f_1(x) \leq y \leq f_2(x)\}$$

about the y-axis, where $0 \leq a < b$ and $f_1, f_2 : [a,b] \to \mathbb{R}$ are integrable functions such that $f_1 \leq f_2$. Recall that by the Shell Method, the volume of D is given by

$$\text{Vol}(D) = 2\pi \int_a^b x[f_2(x) - f_1(x)]dx.$$

As observed before, for each $x \in [a,b]$, the centroid of the vertical cut of the region R at $x \in [a,b]$ is at $(x, [f_1(x) + f_2(x)]/2)$, and hence the y-coordinate of the centroid of the sliver of D at x is given by $[f_1(x) + f_2(x)]/2$. Thus we have

$$\overline{y} := \frac{2\pi}{\text{Vol}(D)} \int_a^b \frac{[f_1(x) + f_2(x)]}{2} x[f_2(x) - f_1(x)]dx$$

$$= \frac{\pi}{\text{Vol}(D)} \int_a^b x[f_2(x)^2 - f_1(x)^2]dx,$$

whereas $\overline{x} = 0 = \overline{z}$ by symmetry.

Examples 8.16. (i) Let b and h be positive real numbers. Consider the planar region enclosed by the right-angled triangle whose vertices are at $(0,0)$, $(b,0)$, and $(0,h)$. The area of this triangular region is equal to $bh/2$. Hence the coordinates of its centroid are given by

$$\overline{x} = \frac{2}{bh} \int_0^b x\left[\frac{h}{b}(b-x) - 0\right]dx = \frac{2}{b^2} \int_0^b (bx - x^2)dx = \frac{2}{b^2}\left(\frac{b^3}{2} - \frac{b^3}{3}\right) = \frac{b}{3}$$

and

$$\overline{y} = \frac{1}{2} \cdot \frac{2}{bh} \int_0^b \left[\left(\frac{h}{b}(b-x)\right)^2 - 0^2\right]dx = \frac{h}{b^3} \int_0^b (b-x)^2 dx = \frac{h}{b^3} \cdot \frac{b^3}{3} = \frac{h}{3}.$$

(ii) Let b and h be positive real numbers. Consider the planar region enclosed by the rectangle whose vertices are at $(0,0)$, $(b,0)$, $(0,h)$ and (b,h). The area of this rectangular region is equal to bh. Hence the coordinates of its centroid are given by

$$\overline{x} = \frac{1}{bh} \int_0^b xh\, dx = \frac{b}{2} \quad \text{and} \quad \overline{y} = \frac{1}{2} \cdot \frac{1}{bh} \int_0^b (h^2 - 0^2)dx = \frac{h}{2}.$$

Thus the centroid of the rectangular region is at $(b/2, h/2)$.

(iii) Let $a \in \mathbb{R}$ with $a > 0$. Consider the semicircular region

$$\{(x, y) \in \mathbb{R}^2 : y \geq 0 \text{ and } x^2 + y^2 \leq a^2\}.$$

Its area is equal to $\pi a^2 / 2$ as we have seen in part (iii) of Proposition 8.2. Then $\overline{x} = 0$ by symmetry and

$$\overline{y} = \frac{2}{\pi a^2} \cdot \frac{1}{2} \int_{-a}^{a} \left[\left(\sqrt{a^2 - x^2} \right)^2 - 0^2 \right] dx = \frac{4a}{3\pi}.$$

Thus the centroid of the semicircular region is at $(0, 4a/3\pi)$.

(iv) Consider the region bounded by the curves given by $x = 2y - y^2$ and $x = 0$. Since $2y - y^2 = x = 0$ implies that $y = 0$ or $y = 2$, and since $2y - y^2 \geq 0$ for all $y \in [0, 2]$, the region is given by

$$\{(x, y) \in \mathbb{R}^2 : 0 \leq y \leq 2 \text{ and } 0 \leq x \leq 2y - y^2\}.$$

The area of this region is equal to $\int_0^2 (2y - y^2) dy = 4/3$. The curve given by $x = 2y - y^2$ is in fact the parabola given by $(y - 1)^2 = 1 - x$. Thus $\overline{y} = 1$ by symmetry and

$$\overline{x} = \frac{1}{2(4/3)} \int_0^2 \left[(2y - y^2)^2 - 0^2 \right] dy = \frac{1}{2(4/3)} \frac{16}{15} = \frac{2}{5}.$$

Thus the centroid of the region is at $(\frac{2}{5}, 1)$.

(v) Let b and h be positive real numbers. A right circular conical solid of base radius a and height h is generated by revolving the triangular region bounded by the lines given by $x = 0$, $y = 0$, and $(x/a) + (y/h) = 1$ about the y-axis. The volume of this solid cone is equal to $\pi a^2 h / 3$ as we have seen in Example 8.7 (ii). Hence

$$\overline{y} = \frac{\pi}{\pi a^2 h/3} \int_0^h y a^2 \left(1 - \frac{y}{h} \right)^2 dy = \frac{\pi}{(\pi a^2 h/3)} \frac{a^2 h^2}{12} = \frac{h}{4}$$

and $\overline{x} = 0 = \overline{z}$ by symmetry. Thus the centroid of the conical solid is at $(0, h/4, 0)$.

(vi) Let R denote the planar region in the first quadrant between the parabolas given by $y = x^2$ and $y = 2 - x^2$. Consider the solid D_1 obtained by revolving the region R about the x-axis. Its volume is equal to $8\pi/3$ as we have seen in Example 8.7 (iv). Hence

$$\overline{x} = \frac{\pi}{8\pi/3} \int_0^1 x \left[\left(2 - x^2 \right)^2 - \left(x^2 \right)^2 \right] dx = \frac{\pi}{(8\pi/3)} \cdot 1 = \frac{3}{8}$$

and $\overline{y} = 0 = \overline{z}$ by symmetry. Thus the centroid of the solid E_1 is at $(\frac{3}{8}, 0, 0)$. Next, consider the solid D_2 obtained by revolving the region R about the y-axis. Its volume is equal to π as we have seen in Example 8.7 (iv). Hence

$$\overline{y} = \frac{2\pi}{\pi} \int_0^1 x \left[\frac{x^2 + (2 - x^2)}{2} \right] (2 - x^2 - x^2) dx = \frac{2\pi}{\pi} \frac{1}{2} = 1$$

and $\overline{x} = 0 = \overline{z}$ by symmetry. Thus the centroid of the solid D_2 is at $(0, 1, 0)$. We could also have obtained this by symmetry. ◇

Theorems of Pappus

The following result relates the centroid of a curve with the area of the surface of revolution generated by it.

Proposition 8.17 (Theorem of Pappus for Surfaces of Revolution).
Let C be a piecewise smooth curve in $\mathbb{R}^2$ and L be a line in $\mathbb{R}^2$ that does not cross C. If C is revolved about L, then the area of the surface so generated is equal to the product of the arc length of C and the distance traveled by the centroid of C. Symbolically, we have

Area of Surface of Revolution = Arc length × Distance Traveled by Centroid.

Proof. Let the curve C be given by $(x(t), y(t))$, $t \in [\alpha, \beta]$, and the line L be given by $ax + by + c = 0$, where $a^2 + b^2 \neq 0$. Recall that the arc length of C is equal to

$$\ell(C) := \int_\alpha^\beta \sqrt{x'(t)^2 + y'(t)^2}\, dt.$$

Also, the area of the surface S generated by revolving C about L is equal to

$$A(S) := 2\pi \int_\alpha^\beta \rho(t) \sqrt{x'(t)^2 + y'(t)^2}\, dt,$$

where $\rho(t)$ is the distance of the point $(x(t), y(t))$, $t \in [\alpha, \beta]$, from the line L. On the other hand, the centroid $(\overline{x}, \overline{y})$ of C is given by

$$\overline{x} := \frac{\int_\alpha^\beta x(t) \sqrt{x'(t)^2 + y'(t)^2}\, dt}{\ell(C)} \quad \text{and} \quad \overline{y} := \frac{\int_\alpha^\beta y(t) \sqrt{x'(t)^2 + y'(t)^2}\, dt}{\ell(C)}.$$

Further, the distance d traveled by $(\overline{x}, \overline{y})$ about the line L is equal to 2π times its distance from the line L, which is equal to

$$2\pi \frac{|a\overline{x} + b\overline{y} + c|}{\sqrt{a^2 + b^2}} = \frac{2\pi}{\ell(C)\sqrt{a^2 + b^2}} \left| \int_\alpha^\beta [ax(t) + by(t) + c] \sqrt{x'(t)^2 + y'(t)^2}\, dt \right|$$

$$= \frac{2\pi}{\ell(C)} \int_\alpha^\beta \rho(t) \sqrt{x'(t)^2 + y'(t)^2}\, dt,$$

because $ax(t) + by(t) + c \geq 0$ for all $t \in [\alpha, \beta]$ or $ax(t) + by(t) + c \leq 0$ for all $t \in [\alpha, \beta]$. Thus

$$\text{Area}\,(S) = \ell(C) \times d.$$

This proves the theorem. □

The next result relates the centroid of a planar region lying between two curves with the volume of the solid obtained by revolving the region about the x-axis or the y-axis.

Proposition 8.18 (Theorem of Pappus for Solids of Revolution). *Let R be a planar region given by*

$$\{(x, y) \in \mathbb{R}^2 : a \le x \le b, \ f_1(x) \le y \le f_2(x)\},$$

where $f_1, f_2 : [a, b] \to \mathbb{R}$ are integrable functions such that $f_1 \le f_2$. If either $f_1(x) \ge 0$ and the region R is revolved about the x-axis, or $a \ge 0$ and the region R is revolved about the y-axis, then the volume of the solid so generated is equal to the product of the area of R and the distance traveled by the centroid of R. Symbolically, we have

$$\text{Volume of Solid of Revolution} = \text{Area} \times \text{Distance Traveled by Centroid.}$$

Proof. We note that the area of the region R is equal to

$$\text{Area } (R) := \int_a^b [f_2(x) - f_1(x)]dx.$$

Let $(\overline{x}, \overline{y})$ denote the centroid of R.

First assume that $f_1(x) \ge 0$ for all $x \in [a, b]$ and the region R is revolved about the x-axis. Then by the Washer Method, the volume of the solid D so generated is equal to

$$\text{Vol } (D) := \pi \int_a^b [f_2(x)^2 - f_1(x)^2]dx.$$

On the other hand, we have

$$\overline{y} = \frac{1}{\text{Area } (R)} \int_a^b \frac{[f_1(x) + f_2(x)]}{2} [f_2(x) - f_1(x)]dx$$

$$= \frac{1}{2\text{Area } (R)} \int_a^b [f_2(x)^2 - f_1(x)^2]dx.$$

Further, the distance d traveled by $(\overline{x}, \overline{y})$ about the x-axis is equal to 2π times its distance from the x-axis, that is,

$$d = 2\pi\overline{y}.$$

Thus we have

$$\text{Vol } (D) = \text{Area } (R) \times d.$$

This proves the desired result in the case that the planar region R is revolved about the x-axis.

Next, assume that $a \geq 0$ and the region R is revolved about the y-axis. Then by the Shell Method, the volume of the solid so generated is equal to

$$\text{Vol } (D) := 2\pi \int_a^b x[f_2(x) - f_1(x)]dx.$$

On the other hand, we have

$$\overline{x} = \frac{1}{\text{Area } (R)} \int_a^b x[f_2(x) - f_1(x)]dx.$$

Further, the distance d traveled by $(\overline{x}, \overline{y})$ about the y-axis is equal to 2π times its distance from the x-axis, that is,

$$d = 2\pi\overline{x}.$$

Thus, again, we have

$$\text{Vol } (D) = \text{Area } (R) \times d.$$

This proves the desired result in the case that the planar region R is revolved about the y-axis. $\qquad \square$

If we know any two of the three quantities (i) length of a planar curve, (ii) the distance of its centroid from a line in its plane, and (iii) the area of the surface obtained by revolving the curve about the line, then the result of Pappus allows us to find the remaining quantity easily. In case the curve is symmetric in some way, we can in fact determine its centroid. This also holds for the area of a planar region, the distance of its centroid from a line in its plane, and the volume of the solid obtained by revolving the region about the line. The following examples illustrate these comments.

Examples 8.19. We verify the conclusions of the Theorems of Pappus in several special cases. We present them in a tabular form for easy verification of the results of Pappus.

(i) Let $\ell(C)$ denote the length of a piecewise smooth curve C. If $(\overline{x}, \overline{y})$ denotes the centroid of C, then its distance from the y-axis is equal to $\overline{x}$. Let S denote the surface obtained by revolving C about the y-axis. Then by Proposition 8.17, we have Area $(S) = \ell(C) \times 2\pi\overline{x}$.

Curve C	Surface S	$\ell(C)$	$\overline{x}$	Area (S)
1. Line segment $(x/a) + (y/h) = 1,\ x \geq 0,\ y \geq 0$	Cone	$\sqrt{a^2 + h^2}$	$\dfrac{a}{2}$	$\pi a\sqrt{a^2 + h^2}$
2. Line segment $x = a,\ 0 \leq y \leq h$	Cylinder	h	a	$2\pi ah$
3. Semicircle $x^2 + y^2 = a^2,\ x \geq 0$	Sphere	πa	$\dfrac{2a}{\pi}$	$4\pi a^2$
4. Circle $(x - a)^2 + y^2 = b^2,\ 0 < b < a$	Torus	$2\pi b$	a	$4\pi^2 ab$

(ii) Let R be a planar region. If $(\overline{x}, \overline{y})$ denotes the centroid of R, then its distance from the y-axis is equal to $\overline{x}$. Let D denote the surface obtained by revolving R about the y-axis. Then by Proposition 8.18, we have Vol $(D) = $ Area $(R) \times 2\pi\overline{x}$.

Region R	Solid D	Area (R)	$\overline{x}$	Vol (D)
1. Triangle enclosed by the lines $x = 0,\ y = 0,\ (x/a) + (y/h) = 1$	Cone	$\dfrac{ah}{2}$	$\dfrac{a}{3}$	$\dfrac{\pi a^2 h}{3}$
2. Rectangle enclosed by the lines $x = 0,\ y = 0,\ x = a,\ y = h$	Cylinder	ah	$\dfrac{a}{2}$	$\pi a^2 h$
3. Semidisk $x^2 + y^2 \leq a^2,\ x \geq 0$	Ball	$\dfrac{\pi a^2}{2}$	$\dfrac{4a}{3\pi}$	$\dfrac{4\pi a^3}{3}$
4. Disk $(x - a)^2 + y^2 = b^2,\ 0 < b < a$	Torus	πb^2	a	$2\pi^2 ab^2$

In various examples given in this chapter, we have independently calculated all the quantities mentioned in the above tables. $\diamond$

8.6 Quadrature Rules

In Chapter 6, we have given various criteria for deciding the integrability of a bounded function $f : [a, b] \to \mathbb{R}$. The actual evaluation of the Riemann integral, however, can pose serious difficulties. If f is integrable and has an antiderivative F, then the Fundamental Theorem of Calculus tells us that $\int_a^b f(x)dx = F(b) - F(a)$. But an integrable function f need not have an antiderivative, and even if it has one, it may not be useful in evaluating $\int_a^b f(x)dx$ in terms of known functions. For example, if $f(x) := 1/x$ for $x \in [1, 2]$, then the function f has an antiderivative, namely, $F(x) := \int_1^x (1/t)dt,\ x \in [1, 2]$. But it is hardly useful in evaluating $\int_1^2 f(x)dx$. A similar comment holds for the function given by $f(x) := 1/(1 + x^2)$ for $x \in [0, 1]$. Sometimes integration by parts and integration by substitution are helpful in evaluating Riemann integrals, but the scope of such techniques is very limited. In fact, it is not possible to evaluate Riemann integrals of many of the functions that occur in practice. As we have mentioned in Section 6.4, the Riemann sums can then be employed to find approximate values of a Riemann integral. In the present section, we shall describe a number of efficient procedures for evaluating a Riemann integral approximately. These are known as **quadrature rules**. A quadrature rule for $[a, b]$ associates to an integrable function $f : [a, b] \to \mathbb{R}$ a real number

$$Q(f) := \sum_{i=1}^{n} w_i f(s_i),$$

where $n \in \mathbb{N}$, $w_i \in \mathbb{R}$, and $s_i \in [a, b]$ for $i = 1, \ldots, n$. The real numbers $w_1, \ldots, w_n$ are known as the **weights** and the points $s_1, \ldots, s_n$ are known as the **nodes** of the quadrature rule Q. For example, if $P = \{x_0, x_1, \ldots, x_n\}$ is a partition of $[a, b]$ and $s_i \in [x_{i-1}, x_i]$ for $i = 1, \ldots, n$, then the Riemann sum

$$S(P, f) := \sum_{i=1}^{n} f(s_i)(x_i - x_{i-1})$$

is an example of a quadrature rule whose nodes are $s_1, \ldots, s_n$ and whose weights are $x_1 - x_0, \ldots, x_n - x_{n-1}$. We shall now construct some simple quadrature rules by replacing the function f by a polynomial function of degree 0, 1, or 2, and by considering the 'signed area' under the curve given by the polynomial function.

1. Let us fix $s \in [a, b]$, and replace the function f by the polynomial function p_0 of degree 0 that is equal to the value $f(s)$ of f at s. The 'signed area' under the curve given by $y = p_0(x)$, whose graph is a horizontal line segment, is equal to the 'area' of the rectangle with base $[a, b]$ and 'height' $f(s)$. This gives the **Rectangular Rule**, which associates to f the number

$$R(f) := (b - a)f(s).$$

In particular, if s is the midpoint $(a + b)/2$ of $[a, b]$, then we obtain the **Midpoint Rule**, which associates to f the number

$$M(f) := (b - a)f\left(\frac{a + b}{2}\right).$$

2. Let us replace the function f by a polynomial function p_1 of degree 1 whose values at a and b are equal to $f(a)$ and $f(b)$. The 'signed area' under the curve given by $y = p_1(x)$, whose graph is an inclined line segment, is equal to the 'area' of the trapezoid with base $[a, b]$ and the 'lengths' of the two parallel sides equal to $f(a)$ and $f(b)$ respectively. This gives the **Trapezoidal Rule**, which associates to f the number

$$T(f) := \frac{(b - a)}{2}[f(a) + f(b)].$$

(See part (i) of Proposition 6.28.)

3. Let us replace the function f by a polynomial function p_2 of degree 2 whose values at a, $(a + b)/2$, and b are equal to $f(a)$, $f((a + b)/2)$, and $f(b)$ respectively. The 'signed area' under the curve given by $y = p_2(x)$, whose graph is, in general, a parabola, gives **Simpson's Rule**, which associates to f the number

$$S(f) := \frac{(b-a)}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right].$$

(See part (ii) of Proposition 6.28.)

The simple quadrature rules given above can be expected to yield only rough approximations of a Riemann integral of a function on $[a, b]$. To obtain more precise approximations, we may partition the interval $[a, b]$ into smaller intervals and apply the above quadrature rules to the function f restricted to each subinterval and then sum up the 'signed areas' so obtained. It is often convenient and also efficient to consider partitions of $[a, b]$ into equal parts.

For $n \in \mathbb{N}$, let $\mathsf{P_n} := \{x_{0,n}, x_{1,n}, \ldots, x_{n,n}\}$ denote the partition of $[a, b]$ into n equal parts. For the sake of simplicity of notation, we denote $x_{i,n}$ by x_i for $i = 0, 1, \ldots, n$. Let

$$h_n := \frac{b-a}{n} \quad \text{and} \quad y_i = f(x_i) \quad \text{for } i = 0, 1, \ldots, n.$$

Note that $x_i - x_{i-1} = h_n$ for $i = 1, \ldots, n$.

1. For $i = 1, \ldots, n$, let s_i be a point in the ith subinterval $[x_{i-1}, x_i]$ of P_n and let us replace the curve given by $y = f(x)$ on the ith subinterval $[x_{i-1}, x_i]$ by a horizontal line segment passing through the point $(s_i, f(s_i))$. Since the 'signed area' of the rectangle with base $x_i - x_{i-1}$ and 'height' $f(s_i)$ is $h_n f(s_i)$, we obtain the **Compound Rectangular Rule**, which associates to f the number

$$R_n(f) := h_n \sum_{i=1}^{n} f(s_i).$$

Since $R_n(f)$ is a Riemann sum for f corresponding to P_n and $\mu(P_n) = h_n \to 0$ as $n \to \infty$, it follows that $R_n(f) \to \int_a^b f(x)dx$ as $n \to \infty$. (See Remark 6.32.)
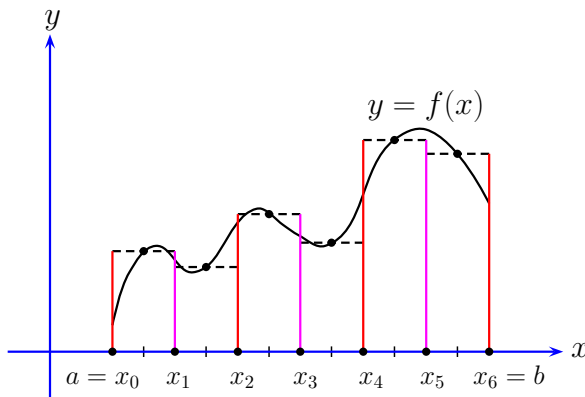
Fig. 8.23. Illustration of the Compound Midpoint Rule

In particular, if s_i is the midpoint $\overline{x}_i := (x_{i-1} + x_i)/2$ of the ith subinterval, we obtain the **Compound Midpoint Rule**, which associates to f the number

$$M_n(f) := h_n \sum_{i=1}^{n} f(\overline{x}_i).$$

[See Figure 8.23.]

2. For $i = 1, \ldots, n$, let us replace the curve given by $y = f(x)$ on the ith subinterval $[x_{i-1}, x_i]$ by a line segment joining the points (x_{i-1}, y_{i-1}) and (x_i, y_i). Since the 'signed area' of the trapezoid with base $x_i - x_{i-1}$ and parallel sides of 'lengths' y_{i-1} and y_i is equal to

$$\frac{h_n}{2}(y_{i-1} + y_i) = \frac{h_n}{2}[f(x_{i-1}) + f(x_i)],$$

we obtain the **Compound Trapezoidal Rule** which associates to f, the number

$$T_n(f) := \frac{h_n}{2} \sum_{i=1}^{n} (y_{i-1} + y_i) = \frac{h_n}{2}(y_0 + 2y_1 + \cdots + 2y_{n-1} + y_n)$$

$$= \frac{h_n}{2} \left[f(x_0) + 2 \sum_{i=1}^{n-1} f(x_i) + f(x_n) \right].$$

[See Figure 8.24.] We observe that

$$T_n(f) = \frac{1}{2} \left[R_n^{\ell}(f) + R_n^{r}(f) \right],$$

where

$$R_n^{\ell}(f) := \sum_{i=1}^{n} f(x_{i-1})(x_i - x_{i-1}) \quad \text{and} \quad R_n^{r}(f) := \sum_{i=1}^{n} f(x_i)(x_i - x_{i-1}).$$

Since $R_n^{\ell}(f) \to \int_a^b f(x)dx$ and $R_n^{r}(f) \to \int_a^b f(x)dx$ as $n \to \infty$, we have

$$T_n(f) \to \frac{1}{2} \left(\int_a^b f(x)dx + \int_a^b f(x)dx \right) = \int_a^b f(x)dx.$$

3. Assume that n is even. For $i = 1, 3, \ldots, n-1$, let us replace the curve given by $y = f(x)$ on the subinterval $[x_{i-1}, x_{i+1}]$ by a parabola passing through the points (x_{i-1}, y_{i-1}), (x_i, y_i), and (x_{i+1}, y_{i+1}). Since the 'signed area' under this quadratic curve is equal to

$$\frac{2h_n}{6}(y_{i-1} + 4y_i + y_{i+1}) = \frac{h_n}{3}[f(x_{i-1}) + 4f(x_i) + f(x_{i+1})],$$

we obtain **Compound Simpson's Rule**, which associates to f the number
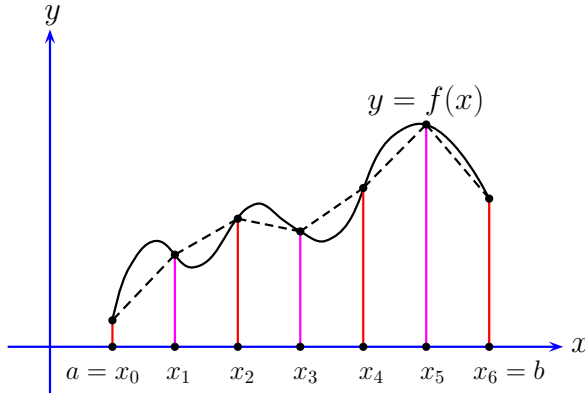
Fig. 8.24. Illustration of the Compound Trapezoidal Rule

$$S_n(f) := \frac{h_n}{3} \sum_{i=1,\, i\,\text{odd}}^{n-1} (y_{i-1} + 4y_i + y_{i+1})$$

$$= \frac{h_n}{3} [y_0 + 4(y_1 + y_3 + \cdots + y_{n-1}) + 2(y_2 + y_4 + \cdots + y_{n-2}) + y_n]$$

$$= \frac{h_n}{3} \left[f(x_0) + 4 \sum_{i=1}^{n/2} f(x_{2i-1}) + 2 \sum_{i=1}^{(n/2)-1} f(x_{2i}) + f(x_n) \right].$$

[See Figure 8.25.] We note that if $k := n/2$, then $h_k = 2h_n$ and
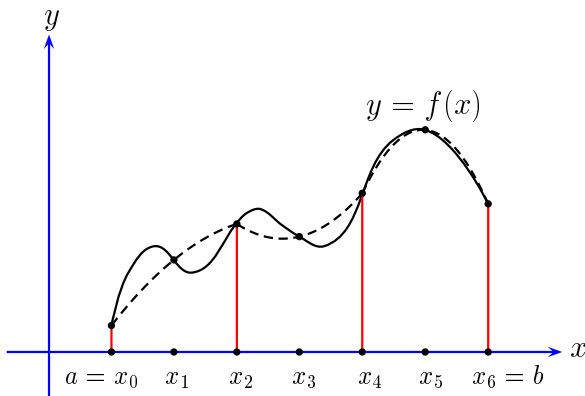
Fig. 8.25. Illustration of Compound Simpson's Rule

$$S_n(f) = \frac{h_k}{6} \sum_{j=1}^{k} \left[f(x_{2j-2}) + 4f(x_{2j-1}) + f(x_{2j}) \right].$$

Observe that each of the three sums $h_k \sum_{j=1}^{k} f(x_{2j-2})$, $h_k \sum_{j=1}^{k} f(x_{2j-1})$, and $h_k \sum_{j=1}^{k} f(x_{2j})$ is a Riemann sum for f corresponding to the partition $Q_k := \{x_0, x_2, \ldots, x_{2k-2}, x_{2k}\}$. Hence, as $n \to \infty$, we have

$$S_n(f) \to \frac{1}{6} \left(\int_a^b f(x)dx + 4 \int_a^b f(x)dx + \int_a^b f(x)dx \right) = \int_a^b f(x)dx.$$

If the function $f : [a, b] \to \mathbb{R}$ is sufficiently smooth, it is possible to obtain error estimates for the approximations $R_n(f)$, $M_n(f)$, $T_n(f)$, and $S_n(f)$ of $\int_a^b f(x)dx$. We shall show that such an error is $O(1/n)$ for $R_n(f)$ in general, while it is $O(1/n^2)$ for $M_n(f)$ and $T_n(f)$, and $O(1/n^4)$ for $S_n(f)$.

We first consider the Compound Rectangular Rule and the Compound Midpoint Rule. In the following result, we shall estimate the difference between the 'signed area' under the curve given by $y = f(x)$, $a \le x \le b$, and the 'signed area' obtained by replacing the function f by a constant function on the entire interval $[a, b]$. The key idea is to use Taylor's Theorem for the functions F, $G : [a, b] \to \mathbb{R}$ given by $F(x) = \int_a^x f(t)dt$ and $G(x) = \int_x^b f(t)dt$.

Lemma 8.20. *Consider a function $f : [a, b] \to \mathbb{R}$.*

(i) *Let f be continuous on $[a, b]$ and f' exist on (a, b). Given any $c \in [a, b]$, there are $\xi, \eta \in (a, b)$ such that*

$$\int_a^b f(x)dx = (b - a)f(c) + \frac{1}{2} \left[(b - c)^2 f'(\xi) - (a - c)^2 f'(\eta) \right].$$

(ii) *Let f' exist and be continuous on $[a, b]$, and f'' exist on (a, b). Then there is $\zeta \in (a, b)$ such that*

$$\int_a^b f(x)dx = (b - a)f\left(\frac{a + b}{2}\right) + \frac{(b - a)^3}{24} f''(\zeta).$$

Proof. (i) Consider the functions F, $G : [a, b] \to \mathbb{R}$ defined by

$$F(x) = \int_a^x f(t)dt \quad \text{and} \quad G(x) = \int_x^b f(t)dt.$$

Then by domain additivity (Proposition 6.7),

$$F(x) + G(x) = \int_a^b f(t)dt \quad \text{and hence} \quad F'(x) = -G'(x) \quad \text{for all } x \in [a, b].$$

Thus by part (ii) of the FTC (Proposition 6.21), we have

$$F'(x) = f(x) = -G'(x) \quad \text{for all } x \in [a, b].$$

Further, F'' and G'' exist on (a, b) and

$$F''(x) = f'(x) = -G''(x) \quad \text{for all } x \in (a, b).$$

Let $c \in [a, b]$ be given. By Taylor's Theorem (Proposition 4.23) for the function F on the interval $[c, b]$ and $n = 1$, there is $\xi \in (a, b)$ such that

$$F(b) = F(c) + (b - c)F'(c) + \frac{(b - c)^2}{2} F''(\xi),$$

that is,

$$\int_a^b f(t)dt = \int_a^c f(t)dt + (b - c)f(c) + \frac{(b - c)^2}{2} f'(\xi).$$

Also, by the version of Taylor's Theorem for right (hand) endpoint (Remark 4.24), there is $\eta \in (a, c)$ such that

$$G(a) = G(c) + (a - c)G'(c) + \frac{(a - c)^2}{2} G''(\eta),$$

that is,

$$\int_a^b f(t)dt = \int_c^b f(t)dt + (c - a)f(c) - \frac{(c - a)^2}{2} f'(\eta).$$

Adding the two equations for $\int_a^b f(t)dt$ given above, we obtain by domain additivity,

$$2 \int_a^b f(t)dt = \int_a^b f(t)dt + (b - a)f(c) + \frac{(b - c)^2}{2} f'(\xi) - \frac{(c - a)^2}{2} f'(\eta),$$

that is,

$$\int_a^b f(t)dt = (b - a)f(c) + \frac{1}{2} \left[(b - c)^2 f'(\xi) - (c - a)^2 f'(\eta) \right],$$

as desired.

(ii) Let F, G, and c be as in part (i) above. Since f'' exists on (a, b), we have

$$F'''(x) = f''(x) = -G'''(x) \quad \text{for all } x \in (a, b).$$

By Taylor's Theorem for $n = 2$, there are $\xi \in (c, b)$ and $\eta \in (a, c)$ such that

$$F(b) = F(c) + (b - c)F'(c) + \frac{(b - c)^2}{2} F''(c) + \frac{(b - c)^3}{6} F'''(\xi),$$

$$G(a) = G(c) + (a - c)G'(c) + \frac{(a - c)^2}{2} G''(c) + \frac{(a - c)^3}{6} G'''(\eta),$$

that is,

$$\int_a^b f(t)dt = \int_a^c f(t)dt + (b-c)f(c) + \frac{(b-c)^2}{2}f'(c) + \frac{(b-c)^3}{6}f''(\xi),$$

$$\int_a^b f(t)dt = \int_c^b f(t)dt + (c-a)f(c) - \frac{(c-a)^2}{2}f'(c) + \frac{(c-a)^3}{6}f''(\eta).$$

Adding the above two equations and letting $c = (a+b)/2$, we obtain

$$2\int_a^b f(t)dt = \int_a^b f(t)dt + (b-a)f\left(\frac{a+b}{2}\right) + \frac{(b-a)^3}{48}[f''(\xi) + f''(\eta)].$$

By the Intermediate Value Property of f'' (Proposition 4.14), we see that there is ζ between ξ and η such that $[f''(\xi) + f''(\eta)]/2 = f''(\zeta)$. Hence we have

$$\int_a^b f(t)dt = (b-a)f\left(\frac{a+b}{2}\right) + \frac{(b-a)^3}{24}f''(\zeta),$$

as desired. □

The above proof shows how the factor $(b-a)^3$ (in place of the factor $(b-a)^2$) arises in the remainder term when c is the midpoint $(a+b)/2$ of the interval $[a, b]$ and the given function f is twice differentiable. This is not possible for any other point c in $[a, b]$.

To obtain error estimates for $R_n(f)$ and $M_n(f)$, we apply the results of the above lemma with the interval $[a, b]$ replaced by the subintervals arising out of the partition $\mathsf{P_n} := \{x_0, x_1, \ldots, x_n\}$ of $[a, b]$ into n equal parts, and then sum up.

Proposition 8.21. *Consider a function $f : [a, b] \to \mathbb{R}$ and $n \in \mathbb{N}$.*

(i) *If f is continuous on $[a, b]$, f' exists on (a, b), and there is $\alpha \in \mathbb{R}$ such that $|f'(x)| \leq \alpha$ for all $x \in (a, b)$, then*

$$\left|\int_a^b f(x)dx - R_n(f)\right| \leq \frac{(b-a)^2\alpha}{2n}.$$

(ii) *If f' exists and is continuous on $[a, b]$, f'' exists on (a, b), and there is $\beta \in \mathbb{R}$ such that $|f''(x)| \leq \beta$ for all $x \in (a, b)$, then*

$$\left|\int_a^b f(x)dx - M_n(f)\right| \leq \frac{(b-a)^3\beta}{24n^2}.$$

Proof. Let $\mathsf{P_n} := \{x_0, x_1, \ldots, x_n\}$ denote the partition of $[a, b]$ into n equal parts, so that $x_i - x_{i-1} = (b-a)/n$ for $i = 1, \ldots, n$.

(i) Let $s_i \in [x_{i-1}, x_i]$ for $i = 1, \ldots, n$ and

$$R_n(f) = \sum_{i=1}^{n} f(s_i)(x_i - x_{i-1}).$$

By the domain additivity (Proposition 6.7), we have

$$\int_a^b f(x)dx - R_n(f) = \sum_{i=1}^{n} \left[\int_{x_{i-1}}^{x_i} f(x)dx - (x_i - x_{i-1})f(s_i) \right].$$

By Lemma 8.20 applied to the function f on the interval $[x_{i-1}, x_i]$ and with $c = s_i$, we see that the ith summand on the right (hand) side of the above equation is equal to

$$\frac{1}{2} \left[(x_i - s_i)^2 f'(\xi_i) - (x_{i-1} - s_i)^2 f'(\eta_i) \right]$$

for some $\xi_i, \eta_i \in (x_{i-1}, x_i)$. Since

$$(x_i - s_i)^2 + (x_{i-1} - s_i)^2 \le [(x_i - s_i) + (s_i - x_{i-1})]^2 = (x_i - x_{i-1})^2 = \frac{(b-a)^2}{n^2}$$

and since $|f'(\xi_i)|, |f'(\eta_i)| \le \alpha$ for $i = 1, \ldots, n$, we obtain

$$\left| \int_a^b f(x)dx - R_n(f) \right| \le \frac{1}{2} \frac{(b-a)^2 \alpha}{n^2} \cdot n = \frac{(b-a)^2 \alpha}{2n},$$

as desired.

(ii) Again by domain additivity we have

$$\int_a^b f(x)dx - M_n(f) = \sum_{i=1}^{n} \left[\int_{x_{i-1}}^{x_i} f(x)dx - (x_i - x_{i-1})f\left(\frac{x_{i-1} + x_i}{2}\right) \right].$$

By part (ii) of Lemma 8.20 applied to the function f on the interval $[x_{i-1}, x_i]$ for $i = 1, \ldots, n$, we see that the ith summand on the right (hand) side of the above equation is equal to

$$\frac{(x_i - x_{i-1})^3}{24} f''(\zeta_i)$$

for some ζ_i in (x_{i-1}, x_i). Since $x_i - x_{i-1} = (b-a)/n$ and $|f''(\zeta_i)| \le \beta$ for $i = 1, \ldots, n$, we obtain

$$\left| \int_a^b f(x)dx - M_n(f) \right| \le \frac{(b-a)^3 \beta}{24n^3} \cdot n = \frac{(b-a)^3 \beta}{24n^2},$$

as desired. □

We proceed to derive error estimates for the Compound Trapezoidal Rule and Compound Simpson's Rule. As before we shall estimate the difference between the 'signed area' under the curve given by $y = f(x)$, $a \leq x \leq b$, and the 'signed area' obtained by replacing the function f by a polynomial function of degree at most one (for the Trapezoidal Rule) as well as by a polynomial function of degree at most two (for Simpson's Rule) on the entire interval $[a, b]$. Then we apply the results with the interval $[a, b]$ replaced by the subintervals arising out of the partition $P_n := \{x_0, x_1, \ldots, x_n\}$ of $[a, b]$ into n equal parts, and sum up.

Lemma 8.22. *Consider a function* $f : [a, b] \to \mathbb{R}$.

(i) *Let* f' *exist and be continuous on* $[a, b]$, *and* f'' *exist on* (a, b). *Then there is* $\xi \in (a, b)$ *such that*

$$\int_a^b f(x)dx = \frac{(b-a)}{2}[f(a) + f(b)] - \frac{(b-a)^3}{12}f''(\xi).$$

(ii) *Let* f', f'', f''' *exist and be continuous on* $[a, b]$, *and* $f^{(4)}$ *exist on* (a, b). *Then there is* $\eta \in (a, b)$ *such that*

$$\int_a^b f(x)dx = \frac{(b-a)}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right] - \frac{(b-a)^5}{2880}f^{(4)}(\eta).$$

Proof. (i) Consider the function $F : [a, b] \to$ defined by

$$F(x) = \int_a^x f(t)dt - \frac{x-a}{2}[f(a) + f(x)] \quad \text{for } x \in [a, b].$$

Then $F(a) = 0$ and

$$F(b) = \int_a^b f(t)dt - \frac{b-a}{2}[f(a) + f(b)].$$

In order to express $F(b)$ as $-(b-a)^3 f''(\xi)/12$ for some $\xi \in (a, b)$, we consider the function $G : [a, b] \to \mathbb{R}$ defined by

$$G(x) = F(x) - \frac{(x-a)^3}{(b-a)^3}F(b).$$

Then $G(a) = 0 = G(b)$, and by part (ii) of the FTC (Proposition 6.21), we have

$$G'(x) = f(x) - \frac{1}{2}[f(a) + f(x)] - \frac{x-a}{2}f'(x) - \frac{3(x-a)^2}{(b-a)^3}F(b)$$

$$= \frac{f(x) - f(a)}{2} - \frac{x-a}{2}f'(x) - \frac{3(x-a)^2}{(b-a)^3}F(b) \quad \text{for } x \in [a, b].$$

Hence $G'(a) = 0$ and also

$$G''(x) = \frac{f'(x)}{2} - \frac{f'(x)}{2} - \frac{x-a}{2}f''(x) - \frac{6(x-a)}{(b-a)^3}F(b)$$

$$= -\frac{x-a}{2}\left[f''(x) + \frac{12}{(b-a)^3}F(b)\right] \quad \text{for } x \in (a,b).$$

By Taylor's Theorem (Proposition 4.23) for G and $n = 1$, there is $\xi \in (a,b)$ such that

$$G(b) = G(a) + G'(a)(b-a) + G''(\xi)\frac{(b-a)^2}{2},$$

that is, $G''(\xi) = 0$. Since $\xi \neq a$, it follows that

$$F(b) = -\frac{(b-a)^3}{12}f''(\xi),$$

as desired.

(ii) Consider the function $F : [a,b] \to \mathbb{R}$ defined by

$$F(x) = \int_{a+(b-x)/2}^{(b+x)/2} f(t)dt - \frac{x-a}{6}\left[f\left(a + \frac{b-x}{2}\right) + 4f\left(\frac{a+b}{2}\right) + f\left(\frac{b+x}{2}\right)\right].$$

Then $F(a) = 0$ and

$$F(b) = \int_a^b f(t)dt - \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right].$$

In order to express $F(b)$ as $-(b-a)^5 f^{(4)}(\eta)/2880$ for some $\eta \in (a,b)$, consider the function $G : [a,b] \to \mathbb{R}$ defined by

$$G(x) = F(x) - \frac{(x-a)^5}{(b-a)^5}F(b).$$

Then $G(a) = 0 = G(b)$, and by part (ii) of the FTC as well as the Chain Rule (Proposition 4.9), we see that for all $x \in (a,b)$,

$$G'(x) = \frac{1}{2}f\left(\frac{b+x}{2}\right) + \frac{1}{2}f\left(a + \frac{b-x}{2}\right)$$
$$- \frac{1}{6}\left[f\left(a + \frac{b-x}{2}\right) + 4f\left(\frac{a+b}{2}\right) + f\left(\frac{b+x}{2}\right)\right]$$
$$- \frac{x-a}{6}\left[-\frac{1}{2}f'\left(a + \frac{b-x}{2}\right) + \frac{1}{2}f'\left(\frac{b+x}{2}\right)\right] - \frac{5(x-a)^4}{(b-a)^5}F(b).$$

Hence $G'(a) = 0$. It can be easily verified that $G''(a) = 0$ and for $x \in (a,b)$,

$$G'''(x) = -\frac{x-a}{48}\left[f'''\left(\frac{b+x}{2}\right) - f'''\left(a + \frac{b-x}{2}\right)\right] - \frac{60(x-a)^2}{(b-a)^5}F(b).$$

By Taylor's Theorem for G and $n = 2$, there is $\xi \in (a, b)$ such that

$$G(b) = G(a) + G'(a)(b - a) + G''(a)\frac{(b-a)^2}{2} + G'''(\xi)\frac{(b-a)^3}{6},$$

that is, $G'''(\xi) = 0$. Since $\xi \neq a$, it follows that

$$f'''\left(\frac{b+\xi}{2}\right) - f'''\left(a + \frac{b-\xi}{2}\right) = -\frac{2880(\xi - a)}{(b-a)^5}F(b).$$

Now by the MVT (Proposition 4.18) for the function f''' on the interval $[a + (b - \xi)/2, (b + \xi)/2]$, there is $\eta \in (a + (b - \xi)/2, (b + \xi)/2) \subseteq (a, b)$ such that

$$f'''\left(\frac{b+\xi}{2}\right) - f'''\left(a + \frac{b-\xi}{2}\right) = \left(\frac{b+\xi}{2} - a - \frac{b-\xi}{2}\right)f^{(4)}(\eta) = (\xi - a)f^{(4)}(\eta).$$

Again, since $\xi \neq a$, it follows that

$$F(b) = -\frac{(b-a)^5}{2880}f^{(4)}(\eta),$$

as desired. $\qquad\square$

Proposition 8.23. *Consider a function $f : [a, b] \to \mathbb{R}$ and $n \in \mathbb{N}$.*

(i) *If f' exists and is continuous on $[a, b]$, f'' exists on (a, b), and there is $\beta \in \mathbb{R}$ such that $|f''(x)| \leq \beta$ for all $x \in (a, b)$, then*

$$\left|\int_a^b f(x)dx - T_n(f)\right| \leq \frac{(b-a)^3\beta}{12n^2}.$$

(ii) *If f', f'', f''' exist and are continuous on $[a, b]$, $f^{(4)}$ exists on (a, b), and there is $\gamma \in \mathbb{R}$ such that $|f^{(4)}(x)| \leq \gamma$ for all $x \in (a, b)$, and if $n \in \mathbb{N}$ is even, then*

$$\left|\int_a^b f(x)dx - S_n(f)\right| \leq \frac{(b-a)^4\gamma}{180n^4}.$$

Proof. Let $\mathsf{P_n} := \{x_0, x_1, \ldots, x_n\}$ denote the partition of $[a, b]$ into n equal parts, so that $x_i - x_{i-1} = (b - a)/n$ for $i = 1, \ldots, n$.

(i) By the domain additivity, we have

$$\int_a^b f(x)dx - T_n(f) = \sum_{i=1}^n \left[\int_{x_{i-1}}^{x_i} f(x)dx - \frac{(x_i - x_{i-1})}{2}[f(x_{i-1}) + f(x_i)]\right].$$

By part (i) of Lemma 8.22 applied to the function f on the interval $[x_{i-1}, x_i]$ for $i = 1, \ldots, n$, we see that the ith summand on the right side of the above equation is equal to

$$-\frac{(x_i - x_{i-1})^3}{12}f''(\xi_i)$$

for some ξ_i in (x_{i-1}, x_i). Since $x_i - x_{i-1} = (b-a)/n$ and $|f''(\xi_i)| \leq \beta$ for $i = 1, \ldots, n$, we obtain

$$\left| \int_a^b f(x)dx - T_n(f) \right| \leq \frac{(b-a)^3\beta}{12n^3} \cdot n = \frac{(b-a)^3\beta}{12n^2},$$

as desired.

(ii) Consider the partition $Q_n := \{x_0, x_2, \ldots, x_{n-2}, x_n\}$ of $[a, b]$. Then

$$\int_a^b f(x)dx - S_n(f)$$

is equal to the sum of the terms

$$\int_{x_{i-1}}^{x_{i+1}} f(x)dx - \frac{(x_{i+1} - x_{i-1})}{6}\left[f(x_{i-1}) + 4f\left(\frac{x_{i-1} + x_{i+1}}{2}\right) + f(x_{i+1}) \right]$$

for $i = 1, 3, \ldots, n-1$. By part (ii) of Lemma 8.22 applied to the function f on the interval $[x_{i-1}, x_{i+1}]$ for $i = 1, 3, \ldots, n-1$, we see that the ith term given above is equal to

$$-\frac{(x_{i+1} - x_{i-1})^5}{2880}f^{(4)}(\eta_i)$$

for some $\eta_i \in (x_{i-1}, x_{i+1})$. Since $x_{i+1} - x_{i-1} = 2(b-a)/n$ and $|f^{(4)}(\eta_i)| \leq \gamma$ for $i = 1, 3, \ldots, n-1$, we obtain

$$\left| \int_a^b f(x)dx - S_n(f) \right| \leq \frac{2^5(b-a)^5\gamma}{2880n^5} \cdot \frac{n}{2} = \frac{(b-a)^5\gamma}{180n^4},$$

as desired. $\square$

If we wish to approximate $\int_a^b f(x)dx$ by $R_n(f)$, $M_n(f)$, $T_n(f)$, or $S_n(f)$ with an error less than or equal to a given small positive number (like 10^{-3}, 10^{-4}, etc.), then Propositions 8.21 and 8.23 can be used to find how large n must be taken, provided we know an upper bound for $|f'|$, $|f''|$, or $|f^{(4)}|$ as the case may be.

The important point to be noted here is that the approximations $R_n(f)$, $M_n(f)$, $T_n(f)$, and $S_n(f)$ are available for use if we know the values of the function f only at certain n equally spaced points in $[a, b]$.

Example 8.24. Consider the function $f : [1, 2] \to \mathbb{R}$ defined by $f(x) := 1/x$. We know from Chapter 7 that $\int_1^2 f(x)dx = \ln 2$, and to estimate this value, we may apply the quadrature rules discussed earlier in this section. For $n \in \mathbb{N}$, let $\mathsf{P}_n := \{x_0, x_1, \ldots, x_n\}$ denote the partition of the interval $[1, 2]$ into n equal parts. Then

$$x_i = 1 + \frac{i}{n}, \quad i = 0, \ldots, n,$$

and so $h_n = 1/n$. If we use right (hand) endpoints of the subintervals for calculating $R_n(f)$, then

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+(i/n)} = \sum_{i=1}^{n}\frac{1}{n+i},$$

$$M_n(f) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{(x_{i-1}+x_i)/2} = \frac{1}{n}\sum_{i=1}^{n}\frac{2}{2+[(2i-1)/n]} = 2\sum_{i=1}^{n}\frac{1}{2(n+i)-1},$$

$$T_n(f) = \frac{1}{2n}\left[\frac{1}{1} + 2\left(\frac{1}{1+(1/n)} + \frac{1}{1+(2/n)} + \cdots + \frac{1}{2-(1/n)}\right) + \frac{1}{2}\right]$$

$$= \frac{3}{4n} + \frac{1}{n+1} + \frac{1}{n+2} + \cdots + \frac{1}{2n-1},$$

and if n is even, then

$$S_n(f) = \frac{1}{3n}\left[\frac{1}{1} + 4\left(\frac{1}{1+(1/n)} + \frac{1}{1+(3/n)} + \cdots + \frac{1}{2-(1/n)}\right)\right.$$

$$\left. + 2\left(\frac{1}{1+(2/n)} + \frac{1}{1+(4/n)} + \cdots + \frac{1}{2-(2/n)}\right) + \frac{1}{2}\right]$$

$$= \frac{1}{2n} + \frac{4}{3}\left(\frac{1}{n+1} + \frac{1}{n+3} + \cdots + \frac{1}{2n-1}\right)$$

$$+ \frac{2}{3}\left(\frac{1}{n+2} + \frac{1}{n+4} + \cdots + \frac{1}{2n-2}\right).$$

For all $x \in (1,2)$, we have

$$|f'(x)| = \left|\frac{-1}{x^2}\right| \leq 1, \quad |f''(x)| = \left|\frac{2}{x^3}\right| \leq 2 \text{ and } \left|f^{(4)}(x)\right| = \left|\frac{24}{x^4}\right| \leq 24.$$

Hence Proposition 8.21 shows that for each $n \in \mathbb{N}$,

$$\left|\int_1^2 \frac{1}{x}dx - R_n(f)\right| \leq \frac{1}{2n} \quad \text{and} \quad \left|\int_1^2 \frac{1}{x}dx - M_n(f)\right| \leq \frac{1}{12n^2},$$

while Proposition 8.23 shows that for each $n \in \mathbb{N}$,

$$\left|\int_1^2 \frac{1}{x}dx - T_n(f)\right| \leq \frac{1}{6n^2} \quad \text{and if } n \text{ is even, then } \left|\int_1^2 \frac{1}{x}dx - S_n(f)\right| \leq \frac{2}{15n^4}.$$

To approximate the Riemann integral $\int_1^2(1/x)dx$ with an error less than 10^{-3}, we must choose

(i) $n \geq 501$ if we use $R_n(f)$, so as to have $\dfrac{1}{2n} < 10^{-3}$,

(ii) $n \geq 10$ if we use $M_n(f)$, so as to have $\dfrac{1}{12n^2} < 10^{-3}$,

(iii) $n \geq 13$ if we use $T_n(f)$, so as to have $\dfrac{1}{6n^2} < 10^{-3}$, and

(iv) $n \geq 4$ if we use $S_n(f)$, so as to have $\dfrac{2}{15n^4} < 10^{-3}$. $\diamond$

Notes and Comments

We have given in this chapter a systematic development of the notion of an area of a region between two curves given by Cartesian equations of the form $y = f(x)$ or $x = g(y)$, or by polar equations of the form $r = p(\theta)$ or $\theta = \alpha(r)$. Two methods of finding the volume of a solid body are described in this chapter: (i) by considering the slices of the solid body by planes perpendicular to a given line and (ii) by considering the slivers of the solid body by cylinders having a common axis. For solids obtained by revolving planar regions about a line, these two methods specialize to the Washer Method and the Shell method.

We have motivated the definition of the length of a smooth curve by considering the tangent line approximations of such a curve. The following alternative motivation is often given. If a curve C is given by $(x(t), y(t))$, $t \in [\alpha, \beta]$, consider a partition $\{t_0, t_1, \ldots, t_n\}$ of the interval $[\alpha, \beta]$. The sum

$$\sum_{i=1}^{n} \sqrt{[x(t_i) - x(t_{i-1})]^2 + [y(t_i) - y(t_{i-1})]^2}$$

of the lengths of the line segments joining the points $(x(t_{i-1}), y(t_{i-1}))$ and $(x(t_i), y(t_i))$ for $i = 1, \ldots, n$ can be considered as an approximation of the 'length' of the curve C. If the functions x and y are continuous on $[\alpha, \beta]$ and are differentiable on (α, β), then by the MVT this sum can be written as

$$\sum_{i=1}^{n} \sqrt{x'(s_i)^2 + y'(u_i)^2}\,(t_i - t_{i-1}),$$

where $s_i, u_i \in (t_{i-1}, t_i)$ for $i = 1, \ldots, n$. We are then naturally led to the definition of the length of C given in the text. We have opted for a motivation based on tangent lines because this consideration extends analogously to a motivation for the definition of the 'area of a smooth surface' given in a course on multivariate calculus. On the contrary, the analogue of the limit of the sums of the lengths of chords, namely, the limit of the areas of inscribed polyhedra formed of triangles, may not exist even for a simple-looking surface such as a cylinder. See, for example, Appendix A.4 of Chapter 4 in Volume II of the book by Courant and John [19].

We have extended the notion of length to a piecewise smooth curve using domain additivity. In fact, the length of any 'rectifiable' curve can be defined. However, we have relegated this to an exercise since the present chapter deals with applications of integration, and rectifiability is defined without any reference to integration.

The basic idea behind the definition of the area of a surface generated by revolving a curve about a line is to approximate the curve by a piecewise linear curve and to consider the areas of the frustums of cones generated by revolving the line segments that approximate the curve.

For $\varphi \in [0, \pi]$, the sector $\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq a^2 \text{ and } 0 \leq \theta(x, y) \leq \varphi\}$ of a disk of radius a subtends an angle φ at the center. We have shown that the area of this sector is $a^2\varphi/2$, and if this sector is revolved about the x-axis, then it generates a (solid) spherical cone whose volume is $2a^3(1 - \cos\varphi)/3$, while the surface area of the spherical cap so generated is $2\pi a^2(1 - \cos\varphi)$. Letting $\varphi = \pi$, we may obtain the area of a disk of radius a, the volume of a ball of radius a, and the surface area of a sphere of radius a.

We have calculated the area enclosed by an ellipse and the volume enclosed by an ellipsoid. However, it is not possible to calculate the arc length of an ellipse or the surface area of an ellipsoid in terms of algebraic functions and elementary transcendental functions. The same holds for the arc length of a lemniscate. To find these, one is led to the so-called 'elliptic integrals' or 'lemniscate integrals'. Inverting functions defined by elliptic integrals gives rise to a new class of functions known as 'elliptic functions', just as inverting the function arctan defined by

$$\arctan x = \int_0^x \frac{1}{1 + t^2} dt \quad \text{for } x \in \mathbb{R}$$

led us to the tangent function in Section 7.2. The study of elliptic functions, initiated by Abel, Jacobi, and Gauss, is a rich and fascinating topic, which connects many branches of mathematics. For a relatively accessible introduction, see the book of Silverman and Tate [54].

The results given in this chapter show that the real number π introduced in Section 7.2 is equal to each of the following:

$$\frac{Area(D)}{Radius(D)^2}, \quad \frac{3}{4}\frac{Volume(B)}{Radius(B)^3}, \quad \frac{1}{2}\frac{Perimeter(C)}{Radius(C)}, \quad \frac{1}{4}\frac{Surface\ Area(S)}{Radius(S)^2},$$

where D, B, C, and S denote a disk, a ball, a circle, and a sphere respectively. These formulas are often used in high-school geometry without any proofs.

The results of Pappus regarding the centroids of surfaces of revolution and of solids of revolution are truly remarkable, especially since they were conceived as early as the fourth century A.D. They reduce the calculations of areas of surfaces of revolution and volumes of solids of revolution to the calculations of arc lengths and planar areas respectively.

As we have remarked in this chapter, one needs the notions of multiple integrals to introduce the general concepts of area and volume. This is usually done in a course on multivariate calculus and one can show that the definitions of area and volume given in this chapter are indeed special cases of the general treatment. Then one would be sure, for example, that the volume of a solid body calculated by the Washer Method and by the Shell Method must come out to be the same!

In the section on quadrature rules, our proofs of error estimates do not use divided differences; they are based only on the Fundamental Theorem of Calculus and Taylor's Theorem. These proofs are inspired by the treatment on pages 328–330 of Hardy's book [31]. Admittedly, these proof are quite involved. But they display the power of Taylor's Theorem. If f is an infinitely differentiable function and the 'Taylor series' of f converges, then these error estimates can be obtained more easily, as indicated in Exercise 61 of Chapter 9.

Exercises

Part A

1. Find the average of the function $f : [1, 2] \to \mathbb{R}$ defined by $f(x) := 1/x$.
2. Given a circle of radius a and a diameter AB of the circle, chords are drawn perpendicular to AB intercepting equal segments at each point of AB. Find the average length of these chords.
3. Given a circle of radius a and a diameter AB of the circle, for each $n \in \mathbb{N}$, n chords are drawn perpendicular to AB so as to intercept equal arcs along the circumference of the circle. Find the limit of the average length of these n chords as $n \to \infty$.
4. Let $f : [a, b] \to \mathbb{R}$ be differentiable such that f' is integrable on $[a, b]$. Show that the average of f' is equal to the average rate of change of f on $[a, b]$, namely $[f(b) - f(a)]/(b - a)$.
5. Let a, b be positive real numbers. If $f(x) := (b/a)\sqrt{a^2 - x^2}$ and $w(x) := x$ for $0 \le x \le a$, find the average of
   (i) f^2 with respect to w,   (ii) f with respect to w^2.
6. If $f, g : [a, b] \to \mathbb{R}$ are integrable functions, then show that $\mathrm{Av}(f + g) = \mathrm{Av}(f) + \mathrm{Av}(g)$, but $\mathrm{Av}(fg)$ may not be equal to $\mathrm{Av}(f)\mathrm{Av}(g)$.
7. Let $f : [0, 1] \to \mathbb{R}$ be defined by $f(x) := x$. Find $\mathrm{Av}(f, w)$ and $\mathrm{Av}(w, f)$ if $w : [0, 1] \to \mathbb{R}$ is defined by
   (i) $w(x) := x$, (ii) $w(x) := x^2$, (iii) $w(x) := 1 - x$, (iv) $w(x) := x(1 - x)$.
8. Find the area of the region bounded by the given curves in each of the following cases:
   (i) $y = 0$, $y = 2x + 3$, $x = 0$ and $x = 1$,   (ii) $y = 4 - x^2$ and $y = 0$,
   (iii) $\sqrt{x} + \sqrt{y} = 1$, $x = 0$ and $y = 0$,   (iv) $y = x^4 - 2x^2$ and $y = 2x^2$,
   (v) $y = 3x^5 - x^3$, $x = -1$ and $x = 1$,   (vi) $x = y^3$ and $x = y^2$,
   (vii) $y = 2 - (x - 2)^2$ and $y = x$,   (viii) $x = 3y - y^2$ and $x + y = 3$.
9. Find the area of the region bounded on the right by the line given by $x + y = 2$, on the left by the parabola given by $y = x^2$, and below by the x-axis.
10. Let $a \in \mathbb{R}$. Define $f(x) := x - x^2$ and $g(x) := ax$ for $x \in \mathbb{R}$. Determine a so that the region above the graph of g and below the graph of f has area equal to $\frac{9}{2}$.

11. Show that the area of the elliptical region given by $ax^2 + 2bxy + cy^2 \le 1$, where $a, b, c \in \mathbb{R}$, $c > 0$, and $ac - b^2 > 0$, is equal to $\pi/\sqrt{ac - b^2}$.

12. Let $\alpha, \beta \in \mathbb{R}$. Show that the areas $A_0, A_1, A_2, \ldots$ of the regions bounded by the x-axis and the half-waves of the curve $y = e^{\alpha x} \sin \beta x$, $x \ge 0$, form a geometric progression with the common ratio $e^{\alpha \pi/\beta}$.

13. Let $a \in \mathbb{R}$ with $a > 0$. Find the area enclosed by the lemniscate given by the polar equation $r^2 = 2a^2 \cos 2\theta$.

14. Let $a \in \mathbb{R}$ with $a > 0$. Find the area of the region inside the circle given by $r = 6a \cos \theta$ and outside the cardioid given by $r = 2a(1 + \cos \theta)$.

15. Let $a \in \mathbb{R}$ with $a > 0$. Find the area of the region enclosed by the loop of the **folium of Descartes** given by $x^3 + y^3 = 3axy$.

16. Let $p, q \in \mathbb{R}$ satisfy $0 \le p < q$ and let $\alpha_1, \alpha_2 : [p, q] \to \mathbb{R}$ be integrable functions such that $-\pi \le \alpha_1 \le \alpha_2 \le \pi$. Let $R := \{(r \cos \theta, r \sin \theta) \in \mathbb{R}^2 : p \le r \le q \text{ and } \alpha_1(r) \le \theta \le \alpha_2(r)\}$ denote the region between the curves given by $\theta = \alpha_1(r)$, $\theta = \alpha_2(r)$ and between the circles given by $r = p$, $r = q$. Define
$$\text{Area } (R) := \int_p^q r[\alpha_2(r) - \alpha_1(r)]dr.$$

Give a motivation for the above definition along the lines of the motivation given in the text for the definition of the area of the region between curves given by polar equations of the form $r = p(\theta)$.

17. (i) Let $p, q \in \mathbb{R}$ be such that $0 \le p < q$ and $\varphi \in [0, \pi]$. Using the formula given in Exercise 16, show that the area of the circular strip $\{(r \cos \theta, r \sin \theta) \in \mathbb{R}^2 : p \le r \le q \text{ and } 0 \le \theta \le \varphi\}$ is $(q^2 - p^2)\varphi/2$.

   (ii) Let $\alpha : [1, 2] \to \mathbb{R}$ be given by $\alpha(r) := 4\pi(r - 1)(2 - r)$, and let $R := \{(r \cos \theta, r \sin \theta) \in \mathbb{R}^2 : 1 \le r \le 2 \text{ and } 0 \le \theta \le \alpha(r)\}$ Show that the area of R is equal to π.

   (iii) Let $R := \{(r \cos \theta, r \sin \theta) \in \mathbb{R}^2 : 1 \le r \le 2 \text{ and } r \le \theta \le r\sqrt{r}\}$. Find Area (R).

18. Let $a \in \mathbb{R}$ with $a > 0$. The base of a certain solid body is the disk given by $x^2 + y^2 \le a^2$. Each of its slices by a plane perpendicular to the x-axis is an isosceles right-angled triangular region with one of the two equal sides in the base of the solid body. Find the volume of the solid body.

19. A solid body lies between the planes given by $y = -2$ and $y = 2$. Each of its slices by a plane perpendicular to the y-axis is a disk with a diameter extending between the curves given by $x = y^2$ and $x = 8 - y^2$. Find the volume of the solid body.

20. A twisted solid is generated as follows. A fixed line L in 3-space and a square of side s in a plane perpendicular to L are given. One vertex of the square is on L. As this vertex moves a distance h along L, the square turns through a full revolution with L as the axis. Find the volume of the solid generated by this motion. What would the volume be if the square had turned through two full revolutions in moving the same distance along the line L?

21. Let $a, b \in \mathbb{R}$ with $0 \le a < b$. Suppose that a planar region R lies between the lines given by $x = a$ and $x = b$, and for each $s \in [a, b]$, the line given by $x = s$ intersects R in a finite number of line segments whose total length is $\ell(s)$. If the function $\ell : [a, b] \to \mathbb{R}$ is integrable, then show that the volume of the solid body obtained by revolving the region R about the y-axis is equal to

$$2\pi \int_a^b x\,\ell(x)dx.$$

22. Find the volume of the solid of revolution obtained by revolving the region bounded by the curves given by $y = 3 - x^2$ and $y = -1$ about the line given by $y = -1$ by both the Washer Method and the Shell Method.

23. The disk given by $x^2 + (y - b)^2 \le a^2$, where $0 < a < b$, is revolved about the x-axis to generate a solid torus. Find the volume of this solid torus by both the Washer Method and the Shell Method.

24. A round hole of radius $\sqrt{3}$ cm. is bored through the center of a solid ball of radius 2 cm. Find the volume cut out.

25. Find the volume of the solid generated by revolving the region in the first quadrant bounded by the curves given by $y = x^3$ and $y = 4x$ about the x-axis by both the Washer Method and the Shell Method.

26. Let $f : [0, \infty) \to [0, \infty)$ be a continuous function. If for each $a > 0$, the volume of the solid obtained by revolving the region under the curve $y = f(x)$, $0 \le x \le a$, about the x-axis is equal to $a^2 + a$, determine f.

27. Find the volume of the solid generated by revolving the region bounded by the curves given by $y = \sqrt{x}$, $y = 2$, and $x = 0$ about the x-axis by both the Washer Method and the Shell Method. If the region is revolved about the line given by $x = 4$, what is the volume of the solid so generated?

28. If the region bounded by the curves given by $y = \tan x$, $y = 0$, and $x = \pi/3$ is revolved about the x-axis, find the volume of the solid so generated.

29. Find the arc length of each of the curves mentioned below.
    (i) the cuspidal cubic given by $y^2 = x^3$ between the points $(0, 0)$ and $(4, 8)$,
    (ii) the cycloid given by $x = t - \sin t$, $y = 1 - \cos t$, $-\pi \le t \le \pi$,
    (iii) the curve given by $(y + 1)^2 = 4x^3$, $0 \le x \le 1$,
    (iv) the curve given by $y = \int_0^x \sqrt{\cos 2t}\ dt$, $0 \le x \le \pi/4$.

30. Let $p, q \in \mathbb{R}$ with $0 \le p < q$ and $\alpha : [p, q] \to \mathbb{R}$. Suppose a piecewise smooth curve C is given by $\theta = \alpha(r)$, $r \in [p, q]$. Show that the arc length of C is equal to

$$\ell(C) = \int_p^q \sqrt{1 + r^2\alpha'(r)^2}\ dr.$$

    (Hint: If $x(r) := r\cos\alpha(r)$ and $y(r) := r\sin\alpha(r)$ for $r \in [p, q]$, then $x'(r)^2 + y'(r)^2 = 1 + r^2\alpha'(r)^2$.)

31. Show that the arc length of the spiral given by $\theta = r$, $r \in [0, \pi]$, is equal to

$$\frac{1}{2}\pi\sqrt{1+\pi^2} + \frac{1}{2}\ln\left(\pi + \sqrt{1+\pi^2}\right).$$

(Hint: Revision Exercise 46 (ii) given at the end of Chapter 7.)

32. For each of the following curves, find the arc length as well as the area of the surface generated by revolving the curve about the x-axis.
    (i) the asteroid given by $x = a\cos^3\theta$, $y = a\sin^3\theta$, $-\pi \le \theta \le \pi$,
    (ii) the loop of the curve given by $9x^2 = y(3-y)^2$, $0 \le y \le 3$.

33. For each of the following curves, find the arc length as well as the area of the surface generated by revolving the curve about the line given by $y = -1$.
    (i) $y = \dfrac{x^3}{3} + \dfrac{1}{4x}$, $1 \le x \le 3$,   (ii) $x = \dfrac{3}{5}y^{5/3} - \dfrac{3}{4}y^{1/3}$, $1 \le y \le 8$.

34. Find the arc length of the curve given by

$$y = \frac{2}{3}x^{3/2} - \frac{1}{2}x^{1/2}, \quad 1 \le x \le 4,$$

    and find the area of the surface generated by revolving the curve about the y-axis.

35. Show that the surface area of the torus obtained by revolving the circle given by $x^2 + (y-b)^2 = a^2$, where $0 < a < b$, about the x-axis is equal to $4\pi^2 ab$. (Compare Example 8.14 (iii).)

36. For each of the following curves, find the area of the surface generated by revolving the curve about the y-axis.
    (i) $y = (x^2 + 1)/2$, $0 \le x \le 1$,
    (ii) $x = t + 1$, $y = (t^2/2) + t$, $0 \le t \le 1$.

37. Let $a \in \mathbb{R}$ with $a > 0$. An arc of the catenary given by $y = a\cosh(x/a)$ whose endpoints have abscissas 0 and a is revolved about the x-axis. Show that the surface area A and the volume V of the solid thus generated are related by the formula $A = 2V/a$.

38. How accurately should we measure the radius of a ball in order to calculate its surface area within 3 percent of its exact value?

39. Given a right circular cone of base radius a and height h, find the radius and the height of the right circular cylinder having the largest lateral surface area that can be inscribed in the cone.

40. Let $p, q \in \mathbb{R}$ with $0 \le p < q$ and $\alpha : [p, q] \to \mathbb{R}$. Suppose a piecewise smooth curve given by $\theta = \alpha(r)$, $r \in [p, q]$, is revolved about a line through the origin containing a ray given by $\theta = \gamma$, and not crossing the curve. If S denotes the surface so generated, then show that

$$\text{Area }(S) = 2\pi \int_p^q r|\sin(\alpha(r) - \gamma)|\sqrt{1 + r^2\alpha'(r)^2}\, dr.$$

    (Hint: Compare Exercise 30 and note that for $r \in [p, q]$, the distance of the point $(r\cos\alpha(r), r\sin\alpha(r))$ from the line L is equal to $r|\sin(\alpha(r) - \gamma)|$.)

41. Let $\ell, \phi \in \mathbb{R}$ with $\ell > 0$. Consider the line segment given by $\theta = \alpha(r)$, where $\alpha(r) := \varphi$ for $r \in [0, \ell]$. If this line segment is revolved about the x-axis, show that the area of the cone S so generated is equal to $\pi\ell^2|\sin\varphi|$. [Note: Since the right circular cone S has slant height ℓ and base radius $\ell|\sin\varphi|$, the result matches with the earlier calculation of the surface area of a right circular cone done by splitting it open.]

42. If a piecewise smooth curve C is given by $y = f(x)$, $x \in [a, b]$, and $\ell(C) = \int_a^b \sqrt{1 + f'(x)^2}\, dx \neq 0$, then show that the centroid $(\overline{x}, \overline{y})$ of C is given by

$$\overline{x} = \frac{1}{\ell(C)} \int_a^b x\sqrt{1 + f'(x)^2}\, dx \quad \text{and} \quad \overline{y} = \frac{1}{\ell(C)} \int_a^b f(x)\sqrt{1 + f'(x)^2}\, dx.$$

43. If a piecewise smooth curve C is given by $r = p(\theta)$, $\theta \in [\alpha, \beta]$, and $\ell(C) = \int_\alpha^\beta \sqrt{p(\theta)^2 + p'(\theta)^2}\, d\theta \neq 0$, then show that the centroid $(\overline{x}, \overline{y})$ of C is given by

$$\overline{x} = \frac{1}{\ell(C)} \int_\alpha^\beta p(\theta)\cos\theta\sqrt{p(\theta)^2 + p'(\theta)^2}\, d\theta$$

and

$$\overline{y} = \frac{1}{\ell(C)} \int_\alpha^\beta p(\theta)\sin\theta\sqrt{p(\theta)^2 + p'(\theta)^2}\, d\theta.$$

44. Let $a > 0$ and $\varphi \in [0, \pi]$. Find the centroid of the arc of the circle given by the polar equation $r = a$, $0 \leq \theta \leq \varphi$.

45. By choosing a suitable coordinate system, find the centroids of (i) a hemisphere of radius a and (ii) a cylinder of radius a and height h.

46. Let $a \in \mathbb{R}$ with $a > 0$. Find the centroid of the region bounded by the curves given by $y = -a$, $x = a$, $x = -a$, and $y = \sqrt{a^2 - x^2}$.

47. Find the centroid of the region enclosed by the curves given by $y^2 = 8x$ and $y = x^2$.

48. Find the centroid of the region in the first quadrant bounded by the curves given by $4y = x^2$, $x = 0$, and $y = 4$.

49. Find the centroid of the region in the first quadrant bounded by the curves given by $4x^2 + 9y^2 = 36$ and $x^2 + y^2 = 9$.

50. Let $a \in \mathbb{R}$ with $a > 0$. Show that the centroid of the ball $\{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 \leq a^2\}$ is $(0, 0, 0)$.

51. Let $a \in \mathbb{R}$ with $a > 0$. Find the centroid of the hemispherical solid body generated by revolving the region under the curve given by $y = \sqrt{a^2 - x^2}$, $0 \leq x \leq a$.

52. Find the centroid of the region bounded by the curves given by $x = y^2 - y$ and $x = y$. If this region is revolved about the x-axis, find the centroid of the solid body so generated.

53. The region bounded by the curves given by $y = 0$, $x = 3$, and $y = x^2$ is revolved about the x-axis. Find the centroid of the solid body so generated.

54. Let $a > 0$. Use a result of Pappus to find the centroid of the region bounded by the curves given by $y = \sqrt{a^2 - x^2}$, $y = 0$, and $x = 0$. (Hint: Revolve

the given region about the x-axis or the y-axis to generate a hemispherical solid.)

55. Let $a > 0$. Use a result of Pappus to find the centroid of the semicircular region bounded by the curves given by $y = \sqrt{a^2 - x^2}$ and $y = 0$. If this region is revolved about the line given $y = -a$, find the volume of the solid so generated.

56. Let $a > 0$. Use a result of Pappus to find the centroid of the semicircular arc $y = \sqrt{a^2 - x^2}$. If this arc is revolved about the line given by $y = a$, find the surface area so generated.

57. Let a and b be positive real numbers such that $a < b$. Find the y-coordinate of the centroid of the region bounded by curves given by $y = \sqrt{a^2 - x^2}$, $y = \sqrt{b^2 - x^2}$, and $y = 0$.

58. Use a result of Pappus to find (i) the volume of a cylinder with height h and radius a (ii) the volume of a cone with height h and base radius a.

59. Use a result of Pappus to show that the lateral surface area of a cone of base radius a and slant height ℓ is $\pi \ell a$.

60. Let $f : [a, b] \to \mathbb{R}$ be a function, $n \in \mathbb{N}$, and $P_n := \{x_0, x_1, \ldots, x_n\}$ be any partition of $[a, b]$. Define

$$R(P_n, f) := \sum_{i=1}^{n} f(x_{i-1})(x_i - x_{i-1}),$$

$$M(P_n, f) := \sum_{i=1}^{n} f\left(\frac{x_{i-1} + x_i}{2}\right)(x_i - x_{i-1}),$$

$$T(P_n, f) := \frac{1}{2} \sum_{i=1}^{n} [f(x_{i-1}) + f(x_i)](x_i - x_{i-1}),$$

and

$$S(P_n, f) := \frac{1}{6} \sum_{i=1}^{n} \left[f(x_{i-1}) + 4f\left(\frac{x_{i-1} + x_i}{2}\right) + f(x_i) \right] (x_i - x_{i-1}).$$

If f is a polynomial function of degree at most 1, then show that

$$R(P_n, f) = M(P_n, f) = T(P_n, f) = \int_a^b f(x)dx,$$

and if f is a polynomial function of degree at most 2, then show that

$$S(P_n, f) = \int_a^b f(x)dx.$$

61. If $f : [a, b] \to \mathbb{R}$ is a polynomial function of degree at most 3, then show that for every $n \in \mathbb{N}$,

$$S_n(f) = \int_a^b f(x)dx.$$

(Compare part (ii) of Proposition 8.23.)

62. If $f : [a, b] \to \mathbb{R}$ is a convex function, then show that for every $n \in \mathbb{N}$, the error

$$\int_a^b f(x)dx - T_n(f)$$

in using $T_n(f)$ as an approximation of $\int_a^b f(x)dx$ is nonpositive, and if f is a concave function, then it is nonnegative.

63. Let $f : [a, b] \to \mathbb{R}$ be any function. Let $n \in \mathbb{N}$ be even and $\mathsf{P_n} := \{x_0, x_1, \ldots, x_n\}$ be the partition of $[a, b]$ into n equal parts. If $k := n/2$ and $\mathsf{Q_k} := \{x_0, x_2, \ldots, x_{2k-2}, x_{2k}\}$, show that

$$S_n(f) = \frac{1}{3}\left[T_k(f) + 2M_k(f)\right],$$

where $S_n(f)$ is defined with respect to $\mathsf{P_n}$ and $T_k(f)$, $M_k(f)$ are defined with respect to $\mathsf{Q_k}$. Deduce that if f is integrable, then

$$S_n(f) \to \int_a^b f(x)dx \quad \text{as} \quad n \to \infty.$$

64. If f is continuous on $[a, b]$, f' exists on (a, b), and there is $\alpha \in \mathbb{R}$ such that $|f'(x)| \leq \alpha$ for all $x \in (a, b)$, then show that

$$\left|\int_a^b f(x)dx - M_n(f)\right| \leq \frac{(b-a)^2\alpha}{4n}.$$

(Compare parts (i) and (ii) of Proposition 8.21.)

65. Consider the function $f : [0, 1] \to \mathbb{R}$ defined by $f(x) := 1/(1 + x^2)$. Find $R_n(f)$, $M_n(f)$, and $T_n(f)$ for $n \in \mathbb{N}$, and $S_n(f)$ for even $n \in \mathbb{N}$. Prove that

$$\left|\int_0^1 f(x)dx - R_n(f)\right| \leq \frac{1}{n}, \quad \left|\int_0^1 f(x)dx - M_n(f)\right| \leq \frac{1}{6n^2},$$

while

$$\left|\int_0^1 f(x)dx - T_n(f)\right| \leq \frac{1}{3n^2} \quad \text{and} \quad \left|\int_0^1 f(x)dx - S_n(f)\right| \leq \frac{2}{15n^4} \ (n \text{ even}).$$

Find how large n must be taken if we wish to approximate $\int_0^1 f(x)dx$ with an error less than 10^{-4} using $R_n(f)$, $M_n(f)$, $T_n(f)$, or $S_n(f)$.

66. Let $f : [0, 1] \to \mathbb{R}$ be defined by $f(x) := (1 - x^2)^{3/2}$. Find $R_n(f)$, $M_n(f)$, $T_n(f)$, and $S_n(f)$ for $n = 4$ and $n = 6$. Also, find the corresponding error estimates.

67. Consider the **error function** $\mathrm{erf} : \mathbb{R} \to \mathbb{R}$ defined by

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt.$$

Use Compound Simpson's Rule with $n = 4$ to find an approximation α to $\mathrm{erf}(1)$ in terms of π and e. Show that $|\mathrm{erf}(1) - \alpha| \leq 19/5760$.

68. Consider the function $f : [0,1] \to \mathbb{R}$ defined by $f(x) = xe^{-x^2}$. Find $T_n(f)$ and $S_n(f)$ with $n = 2$ and $n = 4$. Obtain the corresponding error estimates, and compare them with the actual errors

$$\int_0^1 f(x)dx - T_n(f) \quad \text{and} \quad \int_0^1 f(x)dx - S_n(f).$$

Part B

69. Let $h > 0$. For each $x \in [0, h]$, the area of the slice at x of a solid body by a plane perpendicular to the x-axis is given by $A(x) := ax^2 + bx + c$. If $B_1 := A(0) = c$, $M := A(h/2) = (ah^2 + 2bh + 4c)/4$, and $B_2 := A(h) = ah^2 + bh + c$, then show that the volume of the solid body is equal to $(B_1 + 4M + B_2)/6$.
    [Note: This formula is known as the **Prismoidal Formula**.]
70. Let a curve C in $\mathbb{R}^2$ be given by $(x(t), y(t))$, $t \in [\alpha, \beta]$. For a partition $\{t_0, t_1, \ldots, t_n\}$ of $[\alpha, \beta]$, let

$$\ell(C, P) := \sum_{i=1}^n \sqrt{[x(t_i) - x(t_{i-1})]^2 + [y(t_i) - y(t_{i-1})]^2}.$$

    If the set $\{\ell(C, P) : P \text{ is a partition of } [\alpha, \beta]\}$ is bounded above, then the curve C is said to be **rectifiable**, and the **length** of C is defined to be

$$\ell(C) := \sup\{\ell(C, P) : P \text{ is a partition of } [\alpha, \beta]\}.$$

    [Analogous definitions hold for a curve in $\mathbb{R}^3$.]
    (i) If $\gamma \in (\alpha, \beta)$, and the curves C_1 and C_2 are given by $(x(t), y(t))$, $t \in [\alpha, \gamma]$ and by $(x(t), y(t))$, $t \in [\gamma, \beta]$ respectively, then show that C is rectifiable if and only if C_1 and C_2 are rectifiable.
    (ii) Suppose that the functions x and y are differentiable on $[\alpha, \beta]$, and one of the derivatives x' and y' is continuous on $[\alpha, \beta]$, while the other is integrable on $[\alpha, \beta]$. Show that the curve C is rectifiable and

$$\ell(C) = \int_\alpha^\beta \sqrt{x'(t)^2 + y'(t)^2}\, dt.$$

    (Hint: Propositions 4.18, 6.31, and 3.17 and Exercise 43 of Chapter 6.) (Compare Exercise 48 of Chapter 6.)
    (iii) Show that the conclusion in (ii) above holds if the functions x and y are continuous on $[\alpha, \beta]$ and if there are a finite number of points $\gamma_0 < \gamma_1 < \cdots < \gamma_n$ in $[\alpha, \beta]$, where $\gamma_0 = \alpha$ and $\gamma_n = \beta$, such that the assumptions made in (ii) above about the functions x and y hold on each of the subintervals $[\gamma_{i-1}, \gamma_i]$ for $i = 1, \ldots, n$.

[Note: The result in (iii) above shows that the definition of the length of a piecewise smooth curve given in Section 8.3 is consistent with the definition of the length of a rectifiable curve given above.]

71. Let $f : [0, 1] \to \mathbb{R}$ be defined by $f(0) = 0$ and $f(x) = x^2 \sin(\pi/x^2)$ for $x \in (0, 1]$. Given any $n \in \mathbb{N}$, consider the partition

$$P_n := \left\{ 0, n^{-1/2}, \left(n - \frac{1}{2} \right)^{-1/2}, (n-1)^{-1/2}, \dots, (3/2)^{-1/2}, 1 \right\}$$

of $[0, 1]$ and write $P_n := \{x_0, x_1, \dots, x_{2n-2}\}$. Show that

$$\sum_{i=1}^{2n-2} \sqrt{[x_i - x_{i-1}]^2 + [f(x_i) - f(x_{i-1})]^2} \geq \left(\frac{1}{3} + \frac{1}{5} + \dots + \frac{1}{2n-1} \right).$$

Deduce that the curve $y = f(x)$, $0 \leq x \leq 1$, is not rectifiable even though the function f is differentiable. (Hint: Exercise 10 of Chapter 2.)

72. Let $f : [a, b] \to \mathbb{R}$ be a bounded function that is continuous on (a, b), and $w : [a, b] \to \mathbb{R}$ be a weight function that is continuous and positive on (a, b). Show that there is $c \in (a, b)$ such that $\mathrm{Av}(f; w) = f(c)$. (Hint: Apply Cauchy's Mean Value Theorem (Proposition 4.36) to the functions F, $G : [a, b] \to \mathbb{R}$ defined by $F(x) := \int_a^x f(t)w(t)dt$ and $G(x) := \int_a^x w(t)dt$.)

73. Let $f : [a, b] \to \mathbb{R}$ be a bounded function that is continuous on (a, b). If the range of f is contained in (α, β) and $\phi : [\alpha, \beta] \to \mathbb{R}$ is a convex function that is continuous at α and β, then show that $\mathrm{Av}(f) \in (\alpha, \beta)$, the function $\phi \circ f : [a, b] \to \mathbb{R}$ is integrable, and $\phi(\mathrm{Av}(f)) \leq \mathrm{Av}(\phi \circ f)$. (Hint: Considering partitions of $[a, b]$ into equal parts, use Exercise 72 of this chapter, Exercise 42 of Chapter 6, Exercise 47 of Chapter 3, and Proposition 6.31.)

9

Infinite Series and Improper Integrals

If $a_1, \ldots, a_n$ are any real numbers, then we can add them together and form their sum $a_1 + \cdots + a_n$. In this chapter, we shall investigate whether we can 'add' infinitely many real numbers. In other words, if (a_k) is a sequence of real numbers, then we ask whether we can give a meaning to a symbol such as '$a_1 + a_2 + \cdots$' or '$\sum_{k=1}^{\infty} a_k$'. This leads us to consider what is known as an infinite series, or simply a series, of real numbers. The study of infinite series is taken up in the first three sections of this chapter, and this may be viewed as a sequel to the theory of sequences developed in Chapter 2. In Section 9.1 below, we define the notion of convergence of a series and thus give a precise meaning to the idea of forming the sum of infinitely many real numbers. A number of useful tests for the convergence of series are given in Section 9.2. In Section 9.3, we study a special kind of series, known as power series. We also discuss here the Taylor series, which is a natural analogue of the notion studied in Chapter 4 of the Taylor polynomial of a function.

In the last three sections of this chapter, we develop the theory of improper integrals, which are a continuous analogue of infinite series and which extend the theory of integration developed in Chapter 6. The notion of convergence of improper integrals and some basic properties are discussed in Section 9.4. A number of useful tests for the convergence of improper integrals are given in Section 9.5. In Section 9.6, we discuss some 'integrals' that are related to improper integrals of the kind studied in the previous sections. We also discuss here the beta function and the gamma function, which are quite important and useful in analysis.

9.1 Convergence of Series

An **infinite series**, or, for short, a **series** of real numbers is an ordered pair $\big((a_k), (A_n)\big)$ of sequences of real numbers such that

$$A_n = a_1 + \cdots + a_n \quad \text{for all } n \in \mathbb{N}.$$

Equivalently, it is an ordered pair $\big((a_k), (A_n)\big)$ of sequences such that

$$a_k = A_k - A_{k-1} \quad \text{for all } k \in \mathbb{N}, \quad \text{where} \quad A_0 := 0, \text{ by convention.}$$

The first sequence (a_k) is called the **sequence of terms** and the second sequence (A_n) is called the **sequence of partial sums** of the (infinite) series $\big((a_k), (A_n)\big)$. For simplicity and brevity, we shall use an informal but suggestive notation $\sum_{k=1}^{\infty} a_k$ for the infinite series $\big((a_k), (A_n)\big)$. In this notation, prominence is given to the first sequence (a_k), but the second sequence (A_n) is just as important. At any rate, the two sequences (a_k) and (A_n) determine each other uniquely.

In some cases, it is convenient to consider the sequence (a_k) of terms indexed as $a_0, a_1, a_2, \ldots$, or more generally, as $a_m, a_{m+1}, \ldots$ for some $m \in \mathbb{Z}$. In such cases, the sequence (A_n) of partial sums will be indexed as $A_0, A_1, A_2, \ldots$ or more generally, as $A_m, A_{m+1}, \ldots$ for some $m \in \mathbb{Z}$. Accordingly, the convention $A_0 := 0$ is replaced by $A_{-1} := 0$ or more generally, $A_{m-1} := 0$. In general, the indexing of (a_k) will be clear from the context, and we may simply use the notation $\sum_k a_k$ in place of the more elaborate $\sum_{k=1}^{\infty} a_k$, or $\sum_{k=0}^{\infty} a_k$, or $\sum_{k=m}^{\infty} a_k$.

We say that a series $\sum_{k=1}^{\infty} a_k$ is **convergent** if

$$\lim_{n \to \infty} A_n = \lim_{n \to \infty} \sum_{k=1}^{n} a_k$$

exists, that is, if the sequence (A_n) of its partial sums is convergent. If (A_n) converges to A, then by part (i) of Proposition 2.2, the real number A is unique, and it is called the **sum** of the series $\sum_{k=1}^{\infty} a_k$. If a series $\sum_{k=1}^{\infty} a_k$ is convergent, we may denote its sum by the same symbol $\sum_{k=1}^{\infty} a_k$ used to denote the series. Thus, when we write

$$\sum_{k=1}^{\infty} a_k = A,$$

we mean that A is a real number, the series $\sum_{k=1}^{\infty} a_k$ is convergent, and its sum is equal to A. In this case we may also say that $\sum_{k=1}^{\infty} a_k$ **converges** to A. An infinite series that is not convergent is said to be **divergent**. In particular, we say that the series **diverges** to ∞ or to $-\infty$ according as its sequence of partial sums tends to ∞ or to $-\infty$. It is useful to keep in mind that the convergence of a series is not affected by changing a finite number of its terms, although its sum may change by doing so. (See Exercise 2.)

In Chapter 2 we have considered many sequences that are in fact sequences of partial sums of some important series. We list them here for convenience.

Examples 9.1. (i) **(Geometric Series)** Let $a \in \mathbb{R}$. Define $a_0 := 0$ and $a_k := a^k$ for $k \in \mathbb{N}$. If $a \neq 1$, then for $n = 0, 1, 2, \ldots$, we have

$$A_n := a_0 + a_1 + \cdots + a_n = 1 + a + \cdots + a^n = \frac{1 - a^{n+1}}{1 - a}.$$

Suppose $|a| < 1$. We have seen in Example 2.7 (i) that $A_n \to 1/(1-a)$. Thus $\sum_{k=0}^{\infty} a_k$ is convergent and its sum is equal to $1/(1-a)$, that is,

$$1 + \sum_{k=1}^{\infty} a^k = \frac{1}{1-a} \quad \text{for } a \in \mathbb{R} \text{ with } |a| < 1.$$

This is perhaps the most important example of a convergent series. Its special feature is that we are able to give a simple closed-form formula for each of its partial sums as well as its sum. On the other hand, if $|a| \geq 1$, then $\sum_{k=0}^{\infty} a_k$ is divergent and this can be seen as follows. If $a \geq 1$, then $A_n \geq n + 1$ for $n = 0, 1, 2, \ldots$, and so $A_n \to \infty$. Thus, in this case $\sum_{k=0}^{\infty} a_k$ diverges to ∞. Next, if $a = -1$, then $A_{2n} = 1$ and $A_{2n+1} = 0$ for all $n = 0, 1, 2, \ldots$, and so $\sum_{k=0}^{\infty} a_k$ is divergent. Finally, if $a < -1$, then $A_{2n} \to \infty$, whereas $A_{2n+1} \to -\infty$, and hence $\sum_{k=0}^{\infty} a_k$ is divergent.

(ii) **(Exponential Series)** For $k = 0, 1, 2, \ldots$, define $a_k := 1/k!$. Then for $n = 0, 1, 2, \ldots$, we have

$$A_n := a_0 + a_1 + \cdots + a_n = 1 + \frac{1}{1!} + \cdots + \frac{1}{n!}.$$

We have seen in Example 2.10 (i) that (A_n) is convergent. Moreover, Example 2.10 (ii) and Corollary 7.6 show that $A_n \to e$. Thus $\sum_{k=0}^{\infty} a_k$ is convergent and its sum is equal to e. More generally, given any $x \in \mathbb{R}$, if we define $a_0 := 0$ and $a_k := x^k/k!$ for $k \in \mathbb{N}$, then we shall see in Example 9.31 that $\sum_{k=0}^{\infty} a_k$ is convergent and its sum is equal to e^x, that is

$$1 + \sum_{k=1}^{\infty} \frac{x^k}{k!} = e^x \quad \text{for } x \in \mathbb{R}.$$

(iii) **(Harmonic Series and its variants)** As seen in Example 2.10 (iii),

$$\sum_{k=1}^{\infty} \frac{1}{k} \text{ diverges to } \infty, \text{ but } \sum_{k=1}^{\infty} (-1)^{k-1} \frac{1}{k} \text{ converges.}$$

The divergent series $\sum_{k=1}^{\infty} (1/k)$ is called the **harmonic series**. Replacing k by its powers, we obtain important and useful variants of this series. Let p be a rational number. Example 2.10 (v) shows that

$$\sum_{k=1}^{\infty} \frac{1}{k^p} \text{ diverges to } \infty \text{ if } p \leq 1, \text{ but converges if } p > 1.$$

Further, we shall see in Examples 9.7 (i) and 9.23 (i) that

$$\sum_{k=1}^{\infty} (-1)^{k-1} \frac{1}{k^p} \text{ diverges if } p \leq 0, \text{ but converges if } p > 0. \qquad \diamond$$

Since the convergence of a series is defined in terms of the convergence of a particular sequence, namely the sequence of its partial sums, many results about the convergence of series follow from the corresponding results given in Chapter 2 for the convergence of sequences. We mention them below without giving detailed proofs.

1. The sequence of partial sums of a convergent series is bounded. (Compare part (ii) of Proposition 2.2.)
2. Let $\sum_k a_k = A$ and $\sum_k b_k = B$. Then

$$\sum_k (a_k + b_k) = A + B \quad \text{and} \quad \sum_k (ra_k) = rA \text{ for any } r \in \mathbb{R}.$$

Further, if $a_k \leq b_k$ for all k, then $A \leq B$. (Compare parts (i) and (ii) of Proposition 2.3 and part (i) of Proposition 2.4.) For products, see Exercises 1 and 51.
3. **(Sandwich Theorem)** If (a_k), (b_k), and (c_k) are sequences of real numbers such that $a_k \leq c_k \leq b_k$ for each k, and further, $\sum_k a_k = A$ and $\sum_k b_k = A$, then $\sum_k c_k = A$. (Compare Proposition 2.5.)
4. **(Cauchy Criterion)** A series $\sum_k a_k$ is convergent if and only if for every $\epsilon > 0$, there is $n_0 \in \mathbb{N}$ such that

$$\left| \sum_{k=n+1}^{m} a_k \right| < \epsilon \quad \text{for all } m \geq n \geq n_0.$$

This follows from Proposition 2.19 by noting that the sequence (A_n) of partial sums satisfies $A_m - A_n = \sum_{k=n+1}^{m} a_k$ for all $m \geq n$.

Remark 9.2. As a simple application of the geometric series and the second property above, we can strengthen Exercise 29 of Chapter 2. Indeed, if the decimal expansion of $y \in [0, 1)$ is finite or recurring and $b_1, b_2, \ldots$ denote the digits of y, then there are $i, j \in \mathbb{N}$ with $i < j$ such that

$$y = \sum_{k=0}^{i-1} \frac{b_k}{10^k} + \left(\frac{b_i}{10^i} + \frac{b_{i+1}}{10^{i+1}} + \cdots + \frac{b_{j-1}}{10^{j-1}} \right) A,$$

where

$$A = \sum_{k=0}^{\infty} \frac{1}{10^{(j-i)k}} = \frac{10^{(j-i)}}{10^{(j-i)} - 1}.$$

Consequently, y is a rational number. Thus, we can conclude that $y \in [0, 1)$ is a rational number if and only if its decimal expansion is finite or recurring. A similar result holds for any real number. $\diamond$

Telescoping Series and Series with Nonnegative Terms

If (b_k) is a sequence of real numbers, the series $\sum_{k=1}^{\infty}(b_k - b_{k+1})$ is known as a **telescoping series**. We have the following result regarding its convergence.

Proposition 9.3. *A telescoping series $\sum_{k=1}^{\infty}(b_k - b_{k+1})$ is convergent if and only if the sequence (b_k) is convergent, and in this case*

$$\sum_{k=1}^{\infty}(b_k - b_{k+1}) = b_1 - \lim_{k \to \infty} b_k.$$

Proof. For every $n \in \mathbb{N}$, we have

$$\sum_{k=1}^{n}(b_k - b_{k+1}) = b_1 - b_{n+1}.$$

This yields the desired result. $\square$

It may be noted that every series $\sum_{k=1}^{\infty} a_k$ can be written as a telescoping series. In fact, if A_n is the nth partial sum of the series $\sum_{k=1}^{\infty} a_k$, then letting $b_1 := 0$ and $b_k := -A_{k-1}$ for $k \geq 2$, we obtain $a_k = b_k - b_{k+1}$ for all $k \in \mathbb{N}$. But then determining whether the sequence (b_k) is convergent is the same as determining the convergence of the given series $\sum_{k=1}^{\infty} a_k$. In some special cases, however, it is possible write $a_k = b_k - b_{k+1}$ for all $k \in \mathbb{N}$ without considering the partial sums A_n. In these cases, we can determine the convergence of the series and find its sum using Proposition 9.3. For example, consider the series $\sum_{k=1}^{\infty} 1/k(k+1)$. We have

$$a_k = \frac{1}{k(k+1)} = \frac{1}{k} - \frac{1}{k+1} = b_k - b_{k+1} \quad \text{for all } k \in \mathbb{N},$$

where $b_k := 1/k$ for $k \in \mathbb{N}$. Since $b_k \to 0$, we see that $\sum_{k=1}^{\infty} 1/k(k+1)$ is convergent and

$$\sum_{k=1}^{\infty} \frac{1}{k(k+1)} = \sum_{k=1}^{\infty}(b_k - b_{k+1}) = b_1 - \lim_{k \to \infty} b_k = 1 - 0 = 1.$$

Our next result is a characterization of the convergence of a series with non-negative terms. An interesting application of this result, known as **Cauchy's Condensation Test**, is given in Exercise 7.

Proposition 9.4. *Let (a_k) be a sequence such that $a_k \geq 0$ for all $k \in \mathbb{N}$. Then $\sum_{k=1}^{\infty} a_k$ is convergent if and only if the sequence (A_n) of its partial sums is bounded above, and in this case*

$$\sum_{k=1}^{\infty} a_k = \sup\{A_n : n \in \mathbb{N}\}.$$

If (A_n) is not bounded above, then $\sum_{k=1}^{\infty} a_k$ diverges to ∞.

Proof. Since $a_k \geq 0$ for all $k \in \mathbb{N}$, we see that $A_{n+1} = A_n + a_{n+1} \geq A_n$ for all $n \in \mathbb{N}$, that is, the sequence (A_n) of the partial sums of $\sum_{k=1}^{\infty} a_k$ is monotonically increasing. By part (ii) of Proposition 2.2 and part (i) of Proposition 2.8, we see that the sequence (A_n) is convergent if and only if it is bounded above, and in this case

$$\sum_{k=1}^{\infty} a_k = \lim_{n \to \infty} A_n = \sup\{A_n : n \in \mathbb{N}\}.$$

Also, as we have seen in Remark 2.12, if (A_n) is not bounded above, then $A_n \to \infty$, that is, $\sum_{k=1}^{\infty} a_k$ diverges to ∞. $\square$

As an easy application of Proposition 9.4, we can extend the result in Example 9.1 (iii) from rational powers to real powers (See Exercise 5.) A result similar to Proposition 9.4 holds if $a_k \leq 0$ for all $k \in \mathbb{N}$. (See Exercise 6.) More generally, if a_k has the same sign **for all large** k, that is, if there is $k_0 \in \mathbb{N}$ such that a_k has the same sign for all $k \geq k_0$, then $\sum_k a_k$ is convergent if and only if (A_n) is bounded. However, if there is no $k_0 \in \mathbb{N}$ such that a_k is of the same sign for all $k \geq k_0$, then the series $\sum_k a_k$ may diverge even though its sequence of partial sums is bounded. This is illustrated by the series $\sum_{k=1}^{\infty} (-1)^{k-1}$, for which the sequence (A_n) of partial sums is given by $A_{2n-1} := 1$ and $A_{2n} := 0$ for all $n \in \mathbb{N}$.

If each term a_k of a series $\sum_k a_k$ is either equal to 0 or has the same sign, then clearly, the series $\sum_k a_k$ is convergent if and only if the series $\sum_k |a_k|$ is convergent. This may not hold if the terms a_k are of mixed signs. Thus we are led to the following concept. A series $\sum_k a_k$ is said to be **absolutely convergent** if the series $\sum_k |a_k|$ is convergent. We now give an important result about absolutely convergent series of real numbers.

Proposition 9.5. *An absolutely convergent series is convergent.*

Proof. Let $\sum_k a_k$ be an absolutely convergent series. For each k, define

$$a_k^+ := \frac{|a_k| + a_k}{2} \quad \text{and} \quad a_k^- := \frac{|a_k| - a_k}{2}.$$

Let (A_n), (A_n^+), (A_n^-), and (B_n) denote the sequences of partial sums of the series $\sum_k a_k$, $\sum_k a_k^+$, $\sum_k a_k^-$, and $\sum_k |a_k|$, respectively. Since $\sum_k |a_k|$ is convergent, the sequence (B_n) is bounded. Also, since

$$0 \leq A_n^+ \leq B_n \quad \text{and} \quad 0 \leq A_n^- \leq B_n \quad \text{for all } n,$$

we see that the sequences (A_n^+) and (A_n^-) are bounded. Further, since $a_k^+ \geq 0$ and $a_k^- \geq 0$ for all k, it follows from Proposition 9.4 that the series $\sum_k a_k^+$ and $\sum_k a_k^-$ are convergent. But $a_k = a_k^+ - a_k^-$ for all k. Hence we can conclude that the series $\sum_k a_k$ is convergent. $\square$

The converse of the above result does not hold, as can be seen by considering the series $\sum_{k=1}^{\infty} (-1)^{k-1}/k$, which is convergent but not absolutely convergent. A convergent series that is not absolutely convergent is said to be **conditionally convergent**.

9.2 Convergence Tests for Series

In this section we shall consider several practical tests that enable us to test the convergence or the divergence of a wide variety of series. We begin with a simple result on which most of the tests for the divergence of a series are based.

Proposition 9.6 (kth Term Test). *If $\sum_k a_k$ is convergent, then $a_k \to 0$ as $k \to \infty$. In other words, if $a_k \nrightarrow 0$, then $\sum_k a_k$ is divergent.*

Proof. Let $\sum_k a_k$ be a convergent series. If (A_n) is its sequence of partial sums and A is its sum, then we have $a_k = A_k - A_{k-1} \to A - A = 0$. $\qquad\square$

Examples 9.7. (i) If $p \in \mathbb{R}$ with $p \le 0$, then $|(-1)^{k-1}k^{-p}| \ge 1$ for all $k \in \mathbb{N}$. Hence by the kth Term Test (Proposition 9.6),

$$\sum_{k=1}^{\infty}(-1)^{k-1}\frac{1}{k^p} \text{ is divergent if } p \le 0.$$

(ii) The converse of the kth Term Test (Proposition 9.6) does not hold, as can be seen by considering the harmonic series $\sum_{k=1}^{\infty} 1/k$. $\qquad\diamond$

Remark 9.8. A variant of the kth Term Test (Proposition 9.6), known as **Abel's kth Term Test**, is given in Exercise 8. This variant can also be useful in establishing the divergence of a series. $\qquad\diamond$

Tests for Absolute Convergence

We shall now give a variety of tests to determine the absolute convergence (and hence, the convergence) of a series.

Proposition 9.9 (Comparison Test). *Let $a_k, b_k \in \mathbb{R}$ be such that $|a_k| \le b_k$ for all $k \in \mathbb{N}$. If $\sum_k b_k$ is convergent, then $\sum_k a_k$ is absolutely convergent and*

$$\left|\sum_k a_k\right| \le \sum_k b_k.$$

Proof. Let (A_n), (B_n), and (C_n) denote the sequences of partial sums of the series $\sum_k a_k$, $\sum_k b_k$, and $\sum_k |a_k|$ respectively. Suppose $\sum_k b_k$ is convergent. Then (B_n) is a bounded sequence. Since $|a_k| \le b_k$ for all k, we see that $0 \le C_n \le B_n$, and hence (C_n) is also a bounded sequence. Further, since $|a_k| \ge 0$ for all k, it follows from Proposition 9.4 that $\sum_k |a_k|$ is convergent, that is, $\sum_k a_k$ is absolutely convergent. By Proposition 9.5, $\sum_k a_k$ is convergent. Further, since $-b_k \le a_k \le b_k$ for all k, we have $-B_n \le A_n \le B_n$ for all n. Taking the limit as $n \to \infty$, we get $-\sum_k b_k \le \sum_k a_k \le \sum_k b_k$, that is, $\left|\sum_k a_k\right| \le \sum_k b_k$. $\qquad\square$

It follows from the above result that if $a_k = O(b_k)$ and $b_k \geq 0$ for all k, then the convergence of $\sum_k b_k$ implies the absolute convergence of $\sum_k a_k$. The above result can also be stated as follows. If $|a_k| \leq b_k$ for all k and $\sum_k |a_k|$ diverges to ∞, then $\sum_k b_k$ also diverges to ∞.

As seen below, the geometric series and the series $\sum_{k=1}^{\infty} 1/k^p$, where $p \in \mathbb{R}$, are often useful in employing the comparison tests.

Examples 9.10. (i) For $k = 0, 1, 2, \ldots$, let $a_k := (2^k + k)/(3^k + k)$. If we let $b_k := (2/3)^k$, then $\sum_{k=0}^{\infty} b_k$ is convergent and

$$|a_k| = \frac{2^k + k}{3^k + k} \leq \frac{2^k + 2^k}{3^k} = 2 \left(\frac{2}{3}\right)^k = 2b_k \quad \text{for all } k \geq 0.$$

Hence by the Comparison Test, $\sum_{k=0}^{\infty} a_k$ is convergent.
(ii) Let $a_k := 1/(1+k^2+k^4)^{1/3}$ for $k \in \mathbb{N}$. If we let $b_k := 1/k^{4/3}$, then $\sum_{k=0}^{\infty} b_k$ is convergent and

$$|a_k| = \frac{1}{(1 + k^2 + k^4)^{1/3}} \leq \frac{1}{k^{4/3}} = b_k \quad \text{for all } k \in \mathbb{N}.$$

Hence by the Comparison Test, $\sum_{k=0}^{\infty} a_k$ is convergent. $\diamond$

Given a series $\sum_k a_k$, it may be difficult to look for a convergent series $\sum_k b_k$ such that $|a_k| \leq b_k$ for *each* k. It is often easier to find a convergent series $\sum_k b_k$ of nonzero terms such that the ratio a_k/b_k approaches a limit as $k \to \infty$. In these cases, the following result is useful.

Proposition 9.11. *Let (a_k) and (b_k) be sequences such that $b_k \neq 0$ for all large k. Assume that $a_k/b_k \to \ell$ as $k \to \infty$, where $\ell \in \mathbb{R}$ or $\ell = \infty$ or $\ell = -\infty$.*

(i) *If $b_k > 0$ for all large k, $\sum_k b_k$ is convergent, and $\ell \in \mathbb{R}$, then $\sum_k a_k$ is absolutely convergent.*
(ii) *If $a_k > 0$ for all large k, $\sum_k a_k$ is convergent, and $\ell \neq 0$, then $\sum_k b_k$ is absolutely convergent.*

Proof. (i) Suppose $b_k > 0$ for all large k, and $\ell \in \mathbb{R}$. Since $a_k/b_k \to \ell$ as $k \to \infty$, there is $k_0 \in \mathbb{N}$ such that

$$-1 < \frac{a_k}{b_k} - \ell < 1, \quad \text{that is,} \quad (\ell - 1)b_k < a_k < (\ell + 1)b_k \quad \text{for all } k \geq k_0.$$

If $\alpha := \max\{|\ell + 1|, |\ell - 1|\}$, then $|a_k| < \alpha b_k$ for all $k \geq k_0$. So by Proposition 9.9, the convergence of $\sum_k b_k$ implies the absolute convergence of $\sum_k a_k$.

(ii) Suppose $a_k > 0$ for all large k and $\ell \neq 0$. If $\ell \in \mathbb{R}$, then we have $\lim_{k \to \infty}(b_k/a_k) = 1/\ell$, and if $\ell = \infty$ or $\ell = -\infty$, then $\lim_{k \to \infty}(b_k/a_k) = 0$. So the desired result follows from (i) above by interchanging a_k and b_k. $\square$

In practice, the following simpler version of the above proposition turns out to be particularly useful.

Corollary 9.12 (Limit Comparison Test). *Let (a_k) and (b_k) be sequences such that $a_k > 0$ and $b_k > 0$ for all large k. Assume that*

$$\lim_{k \to \infty} \frac{a_k}{b_k} = \ell, \quad where \quad \ell \in \mathbb{R} \text{ with } \ell \neq 0.$$

Then

$$\sum_k a_k \text{ is convergent} \iff \sum_k b_k \text{ is convergent.}$$

Proof. The implication '$\Longrightarrow$' follows from part (ii) of Proposition 9.11, while '$\Longleftarrow$' follows from part (i) of Proposition 9.11. $\qquad\square$

A version of the Comparison Test known as the **Ratio Comparison Test** is given in Exercise 10.

Examples 9.13. (i) Let $a_k := (2^k + k)/(3^k - k)$ for $k \in \mathbb{N}$. If we let $b_k := (2/3)^k$, then $a_k > 0$ and $b_k > 0$ for all $k \in \mathbb{N}$. Moreover,

$$\frac{a_k}{b_k} = \frac{1 + (k/2^k)}{1 - (k/3^k)} \to 1 \text{ as } k \to \infty.$$

Since $\sum_{k=0}^{\infty} b_k$ is convergent, by the Limit Comparison Test, we see that $\sum_{k=0}^{\infty} a_k$ is convergent.
(ii) Let $a_k := \sin(1/k)$ for all $k \in \mathbb{N}$. If we let $b_k := 1/k$, then $a_k > 0$ and $b_k > 0$ for all $k \in \mathbb{N}$. Moreover,

$$\frac{a_k}{b_k} = \frac{\sin(1/k)}{(1/k)} \to 1 \text{ as } k \to \infty.$$

Since $\sum_{k=1}^{\infty} b_k$ is divergent, by the Limit Comparison Test, we see that $\sum_{k=1}^{\infty} a_k$ is divergent. $\diamond$

Sometimes, it is better to apply the stronger version of the Limit Comparison Test given in Proposition 9.11.

Examples 9.14. (i) Let $p \in \mathbb{R}$ and $a_k := (\ln k)/k^p$ for $k \in \mathbb{N}$. First assume that $p > 1$ and let $q := (p+1)/2$. Then $1 < q < p$. If we let $b_k := 1/k^q$ for $k \in \mathbb{N}$, then $b_k > 0$ for all $k \in \mathbb{N}$ and by L'Hôpital's Rule,

$$\frac{a_k}{b_k} = \frac{(\ln k)/k^p}{1/k^q} = \frac{\ln k}{k^{p-q}} \to 0 \text{ as } k \to \infty.$$

Since $\sum_{k=1}^{\infty} b_k$ is convergent, by part (i) of Proposition 9.11, we see that $\sum_{k=1}^{\infty} a_k$ is convergent. On the other hand, if $p \leq 1$ and we let $b_k := 1/k^p$ for $k \in \mathbb{N}$, then $b_k > 0$ for all $k \in \mathbb{N}$ and

$$\frac{a_k}{b_k} = \frac{(\ln k)/k^p}{1/k^p} = \ln k \to \infty \text{ as } k \to \infty.$$

Since $a_k > 0$ for $k \geq 2$ and $\sum_{k=1}^{\infty} b_k$ is divergent, by part (ii) of Proposition 9.11, we see that $\sum_{k=1}^{\infty} a_k$ is divergent. Thus we can conclude that

$$\sum_{k=1}^{\infty} \frac{\ln k}{k^p} \quad \text{converges if } p > 1 \text{ and diverges if } p \leq 1.$$

(ii) Let $p > 0$ and $a_k := 1/(\ln k)^p$ for $k \in \mathbb{N}$ with $k \geq 2$. If we let $b_k := 1/k$ for $k = 2, 3, \ldots$, then $b_k > 0$ for $k \geq 2$ and by L'Hôpital's Rule,

$$\frac{a_k}{b_k} = \frac{1/(\ln k)^p}{1/k} = \frac{k}{(\ln k)^p} \to \infty \quad \text{as} \quad k \to \infty.$$

Since $a_k > 0$ for $k \geq 2$ and $\sum_{k=2}^{\infty} b_k$ is divergent, by part (ii) of Proposition 9.11, we see that $\sum_{k=2}^{\infty} a_k$ is divergent. $\diamond$

The following result, known as **Cauchy's Root Test**, or simply the **Root Test**, is one of the most basic tests to determine the convergence of a series.

Proposition 9.15 (Root Test). *Let (a_k) be a sequence of real numbers.*

(i) *If there is $\alpha \in \mathbb{R}$ with $\alpha < 1$ such that $|a_k|^{1/k} \leq \alpha$ for all large k, then $\sum_{k=1}^{\infty} a_k$ is absolutely convergent.*
(ii) *If $|a_k|^{1/k} \geq 1$ for infinitely many $k \in \mathbb{N}$, then $\sum_{k=1}^{\infty} a_k$ is divergent.*

In particular, if

$$|a_k|^{1/k} \to \ell \quad \text{as} \quad k \to \infty, \quad \text{where } \ell \in \mathbb{R} \text{ or } \ell = \infty,$$

then

$$\sum_{k=1}^{\infty} a_k \text{ is absolutely convergent when } \ell < 1, \text{ and it is divergent when } \ell > 1.$$

Proof. (i) Suppose $\alpha < 1$ and $k_0 \in \mathbb{N}$ is such that $|a_k|^{1/k} \leq \alpha$ for all $k \geq k_0$. If we let $b_k := \alpha^k$ for $k \in \mathbb{N}$, then $|a_k| \leq b_k$ for all $k \geq k_0$. Moreover, since $\sum_{k=1}^{\infty} b_k$ is convergent, the Comparison Test (Proposition 9.9) shows that $\sum_{k=1}^{\infty} a_k$ is absolutely convergent.

(ii) If $|a_k|^{1/k} \geq 1$ for infinitely many $k \in \mathbb{N}$, then $|a_k| \geq 1$ for infinitely many $k \in \mathbb{N}$ and therefore $a_k \nrightarrow 0$ as $k \to \infty$. Hence the kth Term Test (Proposition 9.6) shows that $\sum_{k=1}^{\infty} a_k$ is divergent.

Now assume that $|a_k|^{1/k} \to \ell$ as $k \to \infty$. Suppose $\ell \in \mathbb{R}$ with $\ell < 1$. If $\alpha := (1+\ell)/2$, then $\ell < \alpha < 1$ and there is $k_0 \in \mathbb{N}$ such that $|a_k|^{1/k} < \alpha$ for all $k \geq k_0$. Hence by (i) above, $\sum_{k=1}^{\infty} a_k$ is absolutely convergent. Next, suppose $\ell = \infty$ or $\ell \in \mathbb{R}$ with $\ell > 1$. Then there is $k_1 \in \mathbb{N}$ such that $|a_k|^{1/k} > 1$ for all $k \geq k_1$. Hence by (ii) above, $\sum_{k=1}^{\infty} a_k$ is divergent. $\square$

The following consequence of the Root Test, known as **D'Alembert's Ratio Test**, or simply the **Ratio Test**, is one of the most widely employed tests to determine the convergence of a series.

Proposition 9.16 (Ratio Test). *Let (a_k) be a sequence of real numbers.*

(i) *If there is $\alpha \in \mathbb{R}$ with $\alpha < 1$ such that $|a_{k+1}| \leq \alpha |a_k|$ for all large k, then $\sum_{k=1}^{\infty} a_k$ is absolutely convergent.*

(ii) *If $|a_{k+1}| \geq |a_k| > 0$ for all large k, then $\sum_{k=1}^{\infty} a_k$ is divergent.*

In particular, if $a_k \neq 0$ for all large k and

$$\frac{|a_{k+1}|}{|a_k|} \to \ell \quad as \quad k \to \infty, \qquad where \; \ell \in \mathbb{R} \; or \; \ell = \infty,$$

then

$$\sum_{k=1}^{\infty} a_k \; is \; absolutely \; convergent \; if \; \ell < 1, \; and \; it \; is \; divergent \; if \; \ell > 1.$$

Proof. (i) Suppose $\alpha < 1$ and $k_0 \in \mathbb{N}$ is such that $|a_{k+1}| \leq \alpha |a_k|$ for all $k \geq k_0$. Clearly, $\alpha \geq 0$. If $\alpha = 0$, there is nothing to prove. If $\alpha > 0$, then

$$|a_k| \leq \alpha |a_{k-1}| \leq \alpha^2 |a_{k-2}| \leq \cdots \leq \alpha^{k-k_0} |a_{k_0}| = (|a_{k_0}| \alpha^{-k_0}) \alpha^k \quad \text{for all } k \geq k_0.$$

Also, since $0 < \alpha < 1$, the series $\sum_{k=1}^{\infty} \alpha^k$ is convergent, and thus, the Comparison Test (Proposition 9.9) shows that $\sum_{k=1}^{\infty} a_k$ is absolutely convergent.

(ii) Let $k_1 \in \mathbb{N}$ be such that $|a_{k+1}| \geq |a_k| > 0$ for all $k \geq k_1$. Since $|a_k| \geq |a_{k-1}| \geq \cdots \geq |a_{k_1}| > 0$ for all $k \geq k_1$, it follows that $a_k \not\to 0$ as $k \to \infty$. Hence by the kth Term Test (Proposition 9.6), $\sum_{k=1}^{\infty} a_k$ is divergent.

Next, assume that $a_k \neq 0$ for all large k and $|a_{k+1}|/|a_k| \to \ell$ as $k \to \infty$. Suppose $\ell \in \mathbb{R}$ with $\ell < 1$. If $\alpha := (1+\ell)/2$, then $\ell < \alpha < 1$ and there is $k_0 \in \mathbb{N}$ such that $(|a_{k+1}|/|a_k|) < \alpha$, that is, $|a_{k+1}| < \alpha |a_k|$ for all $k \geq k_0$. Hence by (i) above, $\sum_{k=1}^{\infty} a_k$ is absolutely convergent. On the other hand, suppose $\ell \in \mathbb{R}$ with $\ell > 1$ or $\ell = \infty$. Then there is $k_1 \in \mathbb{N}$ such that $|a_{k+1}|/|a_k| > 1$, that is, $|a_{k+1}| > |a_k|$ for all $k \geq k_1$. Hence by (ii) above, $\sum_{k=1}^{\infty} a_k$ is divergent. $\qquad \square$

Remarks 9.17. (i) Both the Root Test and the Ratio Test deduce absolute convergence of a series by comparing it with the geometric series. The Ratio Test is often simpler to use than the Root Test because it is easier to calculate ratios than roots. But the Root Test has a wider applicability than the Ratio Test in the following sense. Whenever the Ratio Test gives (absolute) convergence of a series, so does the Root Test (Exercise 55), and moreover, the Root Test can yield (absolute) convergence of a series for which the Ratio Test is inconclusive (Example 9.18 (iv)).

(ii) Both the Root Test and the Ratio Test deduce divergence of a series by appealing to the kth Term Test (Proposition 9.6). It may be observed that for deducing the divergence of a series $\sum_{k=1}^{\infty} a_k$, the Root Test requires $|a_k| \geq 1$ for infinitely many $k \in \mathbb{N}$, while the Ratio Test requires $|a_{k+1}| \geq |a_k|$ for all $k \geq k_1$, where $k_1 \in \mathbb{N}$ and $a_{k_1} \neq 0$. The series $\sum_k a_k$ may not diverge if we only have $|a_{k+1}| \geq |a_k|$ for infinitely many $k \in \mathbb{N}$. For example, consider

$$a_1 := 1, \quad a_{2k} := \frac{1}{(k+1)^2} \quad \text{and} \quad a_{2k+1} := \frac{1}{k^2} \quad \text{for all } k \in \mathbb{N}.$$

Then $|a_{2k+1}| \geq |a_{2k}|$ for all $k \in \mathbb{N}$, but $\sum_{k=1}^{\infty} a_k$ is convergent because it has positive terms and for $k \in \mathbb{N}$,

$$a_3 + a_5 + \cdots + a_{2k+1} = \sum_{j=1}^{k} \frac{1}{j^2} \quad \text{and} \quad a_1 + a_2 + a_4 + \cdots + a_{2k} = \sum_{j=1}^{k+1} \frac{1}{j^2}.$$

Hence the sequence of partial sums of the series $\sum_{k=1}^{\infty} a_k$ is bounded.

(iii) If $a_k \neq 0$ for all large k and $|a_{k+1}/a_k| \to 1$ as $k \to \infty$, then the Ratio Test is inconclusive in deducing the convergence or divergence of $\sum_k a_k$. In this case, we have $|a_k|^{1/k} \to 1$ as well (Exercise 56). Hence the Root Test is also inconclusive. In this event, $\sum_{k=1}^{\infty} a_k$ may be divergent or convergent, as the examples $\sum_{k=1}^{\infty} (1/k)$ and $\sum_{k=1}^{\infty} (1/k^2)$ show.

Using the Ratio Comparison Test (Exercise 10) in conjunction with the series $\sum_{k=1}^{\infty} 1/k^p$, where $p > 0$, one can obtain a result known as **Raabe's Test**, which is useful when $|a_{k+1}|/|a_k| \to 1$. See Exercises 13, 14, 15.

Another test for the convergence for a series of nonnegative terms, known as the Integral Test and based on 'improper integrals', will be given in Proposition 9.39. Examples 9.40 (i) and (ii) illustrate the use of this test.   ◇

Examples 9.18. (i) Let $a_k := k^2/2^k$ for $k \in \mathbb{N}$. Then for each $k \in \mathbb{N}$, we have

$$\frac{|a_{k+1}|}{|a_k|} = \frac{(k+1)^2}{2^{k+1}} \frac{2^k}{k^2} = \frac{1}{2}\left(1 + \frac{1}{k}\right)^2.$$

Hence $|a_{k+1}|/|a_k| \to 1/2$ as $k \to \infty$. So by the Ratio Test, $\sum_{k=1}^{\infty} a_k$ is (absolutely) convergent. Alternatively, we may the Root Test. We have $|a_k|^{1/k} = (k^{1/k})^2/2$ for all $k \in \mathbb{N}$. Since

$$\left(\frac{4}{3}\right)^k = \left(1 + \frac{1}{3}\right)^k \geq \frac{k}{3} + \frac{k(k-1)}{2 \cdot 3^2} = k\left(\frac{1}{3} + \frac{k-1}{18}\right) \geq k \quad \text{for all } k \geq 13,$$

we see that $k^{1/k} \leq \frac{4}{3}$ and hence $|a_k|^{1/k} \leq \frac{8}{9}$ for all $k \geq 13$. Hence $\sum_{k=1}^{\infty} a_k$ is (absolutely) convergent. (In fact, Exercise 7 of Chapter 2 shows that $|a_k|^{1/k} \to \frac{1}{2}$ as $k \to \infty$.)

(ii) Let $a_k := k!/2^k$ for $k \in \mathbb{N}$. Then for each $k \in \mathbb{N}$, we have

$$\frac{|a_{k+1}|}{|a_k|} = \frac{(k+1)!}{2^{k+1}} \frac{2^k}{k!} = \frac{k+1}{2}.$$

Hence $|a_{k+1}|/|a_k| \to \infty$ as $k \to \infty$. So by the Ratio Test, $\sum_{k=1}^{\infty} a_k$ is divergent. Alternatively, we may the Root Test. We have $|a_k|^{1/k} = (k!)^{1/k}/2$ for all $k \in \mathbb{N}$. Since $k! \geq 2^k$ for all $k \geq 4$, we see that $|a_k|^{1/k} \geq 1$ for all $k \geq 4$. Hence $\sum_{k=1}^{\infty} a_k$ is divergent. (In fact, Exercise 11 of Chapter 2 shows that $|a_k|^{1/k} \to \infty$ as $k \to \infty$.)

(iii) Let $a_k := k!/k^k$ for $k \in \mathbb{N}$. Then for each $k \in \mathbb{N}$, we have

$$\frac{|a_{k+1}|}{|a_k|} = \frac{(k+1)!}{(k+1)^{k+1}} \frac{k^k}{k!} = \frac{k+1}{k+1} \left(\frac{k}{k+1}\right)^k = \frac{1}{(1+1/k)^k}.$$

Hence $|a_{k+1}|/|a_k| \to 1/e$ by Corollary 7.6. So by the Ratio Test, $\sum_{k=1}^{\infty} a_k$ is (absolutely) convergent.

(iv) For $k \in \mathbb{N}$, let $a_{2k-1} := 1/4^k$ and $a_{2k} := 1/9^k$. Since

$$\frac{|a_{2k}|}{|a_{2k-1}|} = \left(\frac{4}{9}\right)^k \leq \frac{4}{9} \quad \text{and} \quad \frac{|a_{2k+1}|}{|a_{2k}|} = \frac{1}{4}\left(\frac{9}{4}\right)^k \geq 1 \quad \text{for all } k \geq 1,$$

the Ratio Test is inconclusive. On the other hand,

$$|a_{2k-1}|^{1/(2k-1)} = \frac{1}{2}\left(\frac{1}{2}\right)^{1/(2k-1)} \quad \text{and} \quad |a_{2k}|^{1/2k} = \frac{1}{3} \quad \text{for all } k \geq 1,$$

and hence $|a_k|^{1/k} \leq \frac{1}{2}$ for all $k \geq 1$. Consequently, by the Root Test, $\sum_{k=1}^{\infty} a_k$ is (absolutely) convergent. $\diamond$

Tests for Conditional Convergence

The tests considered so far give either the absolute convergence or the divergence of a series. We now consider some tests which give conditional convergence. They are based on the following simple result, which may be compared with Exercise 14 of Chapter 1.

Proposition 9.19 (Partial Summation Formula). *If $a_k, b_k \in \mathbb{R}$ for $k \in \mathbb{N}$ and $B_n := \sum_{k=1}^{n} b_k$ for $n \in \mathbb{N}$, then*

$$\sum_{k=1}^{n} a_k b_k = \sum_{k=1}^{n-1} (a_k - a_{k+1})B_k + a_n B_n \quad \text{for all } n \geq 2.$$

Proof. Given any $n \in \mathbb{N}$ with $n \geq 2$, we have

$$\sum_{k=1}^{n} a_k b_k = a_1 B_1 + a_2(B_2 - B_1) + \cdots + a_n(B_n - B_{n-1})$$

$$= (a_1 - a_2)B_1 + (a_2 - a_3)B_2 + \cdots + (a_{n-1} - a_n)B_{n-1} + a_n B_n.$$

This yields the desired formula. $\qquad\square$

Proposition 9.20 (Dirichlet's Test). *Let (a_k) and (b_k) be sequences such that (a_k) is monotonic, $a_k \to 0$ as $k \to \infty$, and the sequence (B_n) defined by $B_n := \sum_{k=1}^{n} b_k$ for $n \in \mathbb{N}$ is bounded. Then the series $\sum_{k=1}^{\infty} a_k b_k$ is convergent.*

Proof. Since (B_n) is bounded, there is $\beta \in \mathbb{R}$ such that $|B_n| \leq \beta$ for all $n \in \mathbb{N}$. Also, since (a_k) is monotonic, we have for all $n \geq 2$,

$$\sum_{k=1}^{n-1} |(a_k - a_{k+1})B_k| \leq \beta \sum_{k=1}^{n-1} |a_k - a_{k+1}| = \beta \left| \sum_{k=1}^{n-1} (a_k - a_{k+1}) \right| = \beta |a_1 - a_n|.$$

Now, the sequence (a_n) is convergent and hence bounded. Consequently, the series $\sum_{k=1}^{\infty} (a_k - a_{k-1})B_k$ is absolutely convergent, and so it is convergent by Proposition 9.5. Let C denote its sum. Using the Partial Summation Formula and the fact that $a_n \to 0$, we obtain

$$\sum_{k=1}^{n} a_k b_k = \sum_{k=1}^{n-1} (a_k - a_{k+1})B_k + a_n B_n \to C + 0 = C \quad \text{as} \ \ n \to \infty.$$

Thus $\sum_{k=1}^{\infty} a_k b_k$ is convergent. $\square$

A similar result, known as **Abel's Test**, is given in Exercise 19. Generalizations, due to Dedekind, of the tests of Dirichlet and Abel are given in Exercise 17.

Corollary 9.21 (Leibniz Test). *Let (a_k) be a monotonic sequence such that $a_k \to 0$. Then $\sum_{k=1}^{\infty} (-1)^{k-1} a_k$ is convergent.*

Proof. Define $b_k := (-1)^{k-1}$ for $k \in \mathbb{N}$ and $B_n := \sum_{k=1}^{n} b_k$ for $n \in \mathbb{N}$. Then $B_n = 1$ if n is odd and $B_n = 0$ if n is even. Thus the sequence (B_n) is bounded. Hence Dirichlet's Test shows that $\sum_{k=1}^{\infty} (-1)^{k-1} a_k$ is convergent. $\square$

Corollary 9.22 (Convergence Test for Trigonometric Series). *Let (a_k) be a monotonic sequence such that $a_k \to 0$. Then*

(i) $\sum_{k=1}^{\infty} a_k \sin k\theta$ *is convergent for each $\theta \in \mathbb{R}$.*
(ii) $\sum_{k=1}^{\infty} a_k \cos k\theta$ *is convergent for each $\theta \in \mathbb{R}$ with $\theta \neq 2m\pi$ for any $m \in \mathbb{Z}$.*

Proof. (i) Let $\theta \in \mathbb{R}$. Define $b_k := \sin k\theta$ for $k \in \mathbb{N}$ and $B_n := \sum_{k=1}^{n} b_k$ for $n \in \mathbb{N}$. Now, $2 \sin k\theta \sin(\theta/2) = \cos [k\theta - (\theta/2)] - \cos [k\theta + (\theta/2)]$ for each $k \in \mathbb{N}$, and hence

$$2B_n \sin \frac{\theta}{2} = \sum_{k=1}^{n} \left[\cos \frac{(2k-1)\theta}{2} - \cos \frac{(2k+1)\theta}{2} \right] = \cos \frac{\theta}{2} - \cos \frac{(2n+1)\theta}{2}.$$

If $\sin(\theta/2) = 0$, that is, if $\theta = 2m\pi$ for some $m \in \mathbb{Z}$, then $b_k = 0$ for each $k \in \mathbb{N}$ and so $B_n = 0$ for each $n \in \mathbb{N}$. If $\sin(\theta/2) \neq 0$, then for each $n \in \mathbb{N}$,

$$|2B_n \sin (\theta/2)| \leq 2 \quad \text{and hence} \quad |B_n| \leq \frac{1}{|\sin(\theta/2)|}.$$

Thus the sequence (B_n) is bounded in all cases. Hence the desired result follows from Dirichlet's Test (Proposition 9.20).

(ii) Let $\theta \in \mathbb{R}$ with $\theta \neq 2m\pi$ for any $m \in \mathbb{Z}$. Define $b_k := \cos k\theta$ for $k \in \mathbb{N}$ and $B_n := \sum_{k=1}^{n} b_k$ for $n \in \mathbb{N}$. Now $2\cos k\theta \sin(\theta/2) = \sin[k\theta + (\theta/2)] - \sin[k\theta - (\theta/2)]$ for each $k \in \mathbb{N}$, and hence

$$2B_n \sin \frac{\theta}{2} = \sum_{k=1}^{n} \left[\sin \frac{(2k+1)\theta}{2} - \sin \frac{(2k-1)\theta}{2} \right] = \sin \frac{(2n+1)\theta}{2} - \sin \frac{\theta}{2}.$$

Since $\sin(\theta/2) \neq 0$, the desired result follows as in (i) above. $\qquad\square$

It may be observed that Corollary 9.21 is a special case of part (ii) of Corollary 9.22 with $\theta = \pi$.

Examples 9.23. (i) Let $p > 0$ and $a_k := 1/k^p$ for $k \in \mathbb{N}$. Then (a_k) is monotonic and $a_k \to 0$. Hence by the Leibniz Test, the series

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

is convergent. If $p > 1$, then it is absolutely convergent and if $p \leq 1$, then it is conditionally convergent (Example 9.1 (iii)).

(ii) Let $p > 0$ and $a_k := 1/(\ln k)^p$ for $k \in \mathbb{N}$ with $k \geq 2$. Then (a_k) is monotonic and $a_k \to 0$. Hence by the Leibniz Test, the series

$$\sum_{k=2}^{\infty} \frac{(-1)^{k-1}}{(\ln k)^p}$$

is convergent. In fact, it is conditionally convergent (Example 9.14(ii)).

(iii) Even if the signs of the terms of a sequence (a_k) alternate and $a_k \to 0$, the series $\sum_{k=1}^{\infty} a_k$ may not converge, that is, the monotonicity assumption in the Leibniz Test (Corollary 9.21) cannot be omitted. For example, consider the series

$$\frac{1}{2} - \frac{1}{3} + \frac{1}{2^2} - \frac{1}{4} + \frac{1}{2^3} - \frac{1}{5} + \frac{1}{2^4} - \frac{1}{6} + \cdots .$$

Since the sequence of partial sums of the series $\sum_{k=1}^{\infty} 1/2^k$ is bounded and the sequence of partial sums of the series $\sum_{k=3}^{\infty} 1/k$ is unbounded, it follows that the sequence of partial sums of the series displayed above is unbounded. Hence it is divergent, although the signs of its terms alternate and the kth term tends to zero as $k \to \infty$.

(iv) Let $p > 0$ and $\theta \in \mathbb{R}$. Then the series $\sum_{k=1}^{\infty} (\sin k\theta)/k^p$ is convergent. This follows by letting $a_k := 1/k^p$ for $k \in \mathbb{N}$ in the Convergence Test for trigonometric series (part (i) of Corollary 9.22). Similarly, if $\theta \neq 2m\pi$ for any $m \in \mathbb{Z}$, then the series $\sum_{k=1}^{\infty} (\cos k\theta)/k^p$ is convergent. On the other hand, if $\theta = 2m\pi$ for some $m \in \mathbb{Z}$, then the series

$$\sum_{k=1}^{\infty} \frac{\cos k(2m\pi)}{k^p} = \sum_{k=1}^{\infty} \frac{1}{k^p}$$

is divergent if $p \leq 1$ and convergent if $p > 1$. $\qquad\diamond$

9.3 Power Series

For $k = 0, 1, 2, \ldots$, let $c_k \in \mathbb{R}$. The series

$$\sum_{k=0}^{\infty} c_k x^k := c_0 + \sum_{k=1}^{\infty} c_k x^k, \quad \text{where} \quad x \in \mathbb{R},$$

is called a **power series** and the numbers $c_0, c_1, c_2, \ldots$ are called its **coefficients**. It is clear that if $x = 0$, then for any choice of $c_0, c_1, c_2, \ldots$, the power series $\sum_{k=0}^{\infty} c_k x^k$ is convergent and its sum is equal to c_0. On the other hand, if there is $k_0 \in \mathbb{N}$ such that $c_k = 0$ for all $k > k_0$, then for every $x \in \mathbb{R}$, the power series $\sum_{k=0}^{\infty} c_k x^k$ is convergent and its sum is equal to $c_0 + c_1 x + \cdots + c_{k_0} x^{k_0}$. More generally, if $a \in \mathbb{R}$, then the series $\sum_{k=0}^{\infty} c_k (x - a)^k$, where $x \in \mathbb{R}$, is called a **power series** around a. The treatment of such a series can be reduced to a power series around 0 by letting $\tilde{x} := x - a$. Here are some simple but nontrivial examples of power series.

Examples 9.24. (i) Let $c_0 := 0$ and $c_k := k^k$ for $k \in \mathbb{N}$. Given any $x \in \mathbb{R}$ with $x \neq 0$, we have $|c_k x^k| > 1$, for all $k \in \mathbb{N}$ satisfying $k > 1/|x|$. Hence $c_k x^k \nrightarrow 0$ as $k \to \infty$. So by the kth Term Test, $\sum_{k=0}^{\infty} c_k x^k$ is divergent for every nonzero $x \in \mathbb{R}$.

(ii) For $k = 0, 1, 2, \ldots$, let $c_k := 1/k!$. Given any $x \in \mathbb{R}$, we have

$$\frac{|c_{k+1} x^{k+1}|}{|c_k x^k|} = \frac{|x|}{k+1} \to 0 \quad \text{as} \quad k \to \infty.$$

So by the Ratio Test, $\sum_{k=0}^{\infty} c_k x^k$ is absolutely convergent for every $x \in \mathbb{R}$.

(iii) For all $k = 0, 1, 2, \ldots$, let $c_k := 1$. Then $\sum_{k=0}^{\infty} c_k x^k$ is the geometric series $1 + x + x^2 + \cdots$, and we have seen in Example 9.1 (i) that it is convergent if $|x| < 1$ and its sum is $1/(1 - x)$, while it is divergent if $|x| \geq 1$.   ◇

The above examples are typical as far as the convergence of a power series $\sum_{k=0}^{\infty} c_k x^k$ for various values of x is concerned. The general phenomenon is described by the following basic result.

Lemma 9.25 (Abel's Lemma). *Let x_0 and $c_0, c_1, c_2, \ldots$ be real numbers. If the set $\{c_k x_0^k : k \in \mathbb{N}\}$ is bounded, then the power series $\sum_{k=0}^{\infty} c_k x^k$ is absolutely convergent for every $x \in \mathbb{R}$ with $|x| < |x_0|$. In particular, if $\sum_{k=0}^{\infty} c_k x_0^k$ is convergent, then the power series $\sum_{k=0}^{\infty} c_k x^k$ is absolutely convergent for every $x \in \mathbb{R}$ with $|x| < |x_0|$.*

Proof. If $x_0 = 0$, then there is nothing to prove. Suppose $x_0 \neq 0$. Let $\alpha \in \mathbb{R}$ be such that $|c_k x_0^k| \leq \alpha$ for all $k \in \mathbb{N}$. Given any $x \in \mathbb{R}$ with $|x| < |x_0|$, let $\beta := |x|/|x_0|$. Then $|c_k x^k| = |c_k x_0^k|\beta^k \leq \alpha\beta^k$ for all $k \in \mathbb{N}$. Since $|\beta| < 1$, the geometric series $\sum_k \beta^k$ is convergent. So by the Comparison Test, it follows that $\sum_{k=0}^{\infty} c_k x^k$ is absolutely convergent. In case $\sum_{k=0}^{\infty} c_k x_0^k$ is convergent, then by the kth Term Test, $c_k x_0^k \to 0$, and hence the sequence $(c_k x_0^k)$ is bounded, that is, the set $\{c_k x_0^k : k \in \mathbb{N}\}$ is bounded.   □

Proposition 9.26. *A power series $\sum_{k=0}^{\infty} c_k x^k$ is either absolutely convergent for all $x \in \mathbb{R}$, or there is a unique nonnegative real number r such that the series is absolutely convergent for all $x \in \mathbb{R}$ with $|x| < r$ and is divergent for all $x \in \mathbb{R}$ with $|x| > r$.*

Proof. Let $E := \{|x| : x \in \mathbb{R}$ and $\sum_{k=0}^{\infty} c_k x^k$ is convergent$\}$. Then $0 \in E$. If E is not bounded above, then given $x \in \mathbb{R}$, we may find $x_0 \in E$ such that $|x| < |x_0|$, and then $\sum_{k=0}^{\infty} c_k x^k$ is absolutely convergent by Lemma 9.25. Next, suppose E is bounded above and let $r := \sup E$. If $x \in \mathbb{R}$ and $|x| < r$, then by the definition of a supremum, we may find $x_0 \in E$ such that $|x| < |x_0|$, and so by Lemma 9.25, we see that $\sum_{k=0}^{\infty} c_k x^k$ is absolutely convergent. If $x \in \mathbb{R}$ and $|x| > r$, then by the definition of the set E, the power series $\sum_{k=0}^{\infty} c_k x^k$ is divergent. This proves the existence of the nonnegative real number r with the desired properties. The uniqueness of r is obvious. $\qquad\square$

We say that the **radius of convergence** of a power series is ∞ if the power series is absolutely convergent for all $x \in \mathbb{R}$; otherwise, it is defined to be the unique nonnegative real number r such that the power series is absolutely convergent for all $x \in \mathbb{R}$ with $|x| < r$ and divergent for all $x \in \mathbb{R}$ with $|x| > r$. If r is the radius of convergence of a power series, then the open interval $\{x \in \mathbb{R} : |x| < r\}$ is called the **interval of convergence** of that power series; note that the interval of convergence is the empty set if $r = 0$ and is $\mathbb{R}$ if $r = \infty$. Given a power series $\sum_{k=0}^{\infty} c_k x^k$, the set

$$ S := \left\{ x \in \mathbb{R} : \sum_{k=0}^{\infty} c_k x^k \text{ is convergent} \right\} $$

may be distinct from the interval of convergence. In fact, S is always nonempty and it can equal $\{0\}$ or $\mathbb{R}$ or an interval of the form $(-r, r)$, $[-r, r]$, $[-r, r)$, $(-r, r]$ for some $r > 0$. In following table we illustrate various possibilities for the set S.

Power series	Radius of convergence	S
(i) $\sum_{k=1}^{\infty} k^k x^k$	0	$\{0\}$
(ii) $\sum_{k=1}^{\infty} x^k / k!$	∞	$(-\infty, \infty)$
(iii) $\sum_{k=0}^{\infty} x^k$	1	$(-1, 1)$
(iv) $\sum_{k=0}^{\infty} x^k / k^2$	1	$[-1, 1]$
(v) $\sum_{k=0}^{\infty} x^k / k$	1	$[-1, 1)$
(vi) $\sum_{k=0}^{\infty} (-1)^k x^k / k$	1	$(-1, 1]$

The entries in (i), (ii), and (iii) above follow from Examples 9.24 (i), (ii), and (iii) respectively. For the power series in (iv) above, we note that

$$\frac{|x^{k+1}|}{(k+1)^2} \frac{k^2}{|x^k|} = \left(\frac{k}{k+1}\right)^2 |x| \to |x| \ \text{ as } \ k \to \infty.$$

So by the Ratio Test, the series is absolutely convergent if $|x| < 1$ and it is divergent if $|x| > 1$. Letting $p = 2$ in Example 9.1 (iii), we see that it is convergent if $x = 1$. Further, its convergence for $x = -1$ follows from Proposition 9.5. Similarly, the entries in (v) and (vi) above follow from the Ratio Test and Example 9.1 (iii) with $p = 1$.

The following result characterizes the radius of convergence of a power series $\sum_{k=0}^{\infty} c_k x^k$ in terms of the sequence $(|c_k|^{1/k})$.

Proposition 9.27. *Let $\sum_{k=0}^{\infty} c_k x^k$ be a power series and r be its radius of convergence.*

(i) *If $(|c_k|^{1/k})$ is unbounded, then $r = 0$.*
(ii) *If $(|c_k|^{1/k})$ is bounded, and we let*

$$M_k := \sup\{|c_j|^{1/j} : j \in \mathbb{N} \text{ and } j \geq k\} \quad \text{for } k \in \mathbb{N},$$

then the sequence (M_k) is convergent. Further, if $\ell := \lim_{k \to \infty} M_k$, then

$$r = \infty \text{ when } \ell = 0 \quad \text{and} \quad r = \frac{1}{\ell} \text{ when } \ell \neq 0.$$

Proof. (i) Suppose $(|c_k|^{1/k})$ is unbounded. Consider $x \in \mathbb{R}$ with $x \neq 0$. There are infinitely many $k \in \mathbb{N}$ such that $|c_k|^{1/k} \geq 1/|x|$, that is, $|c_k x^k| \geq 1$. Hence by the kth Term Test, $\sum_{k=0}^{\infty} c_k x^k$ is divergent. This shows that $r = 0$.

(ii) Suppose $(|c_k|^{1/k})$ is bounded. For each $k \in \mathbb{N}$, consider the set $D_k := \{|c_j|^{1/j} : j \in \mathbb{N} \text{ and } j \geq k\}$. Then D_k is bounded, $D_{k+1} \subseteq D_k$, and hence $M_{k+1} \leq M_k$ for all $k \in \mathbb{N}$. Thus (M_k) is a monotonically decreasing sequence that is bounded below by 0. By part (ii) of Proposition 2.8, $\lim_{k \to \infty} M_k$ exists and is equal to $\ell := \inf\{M_k : k \in \mathbb{N}\}$.

Consider $x \in \mathbb{R}$ such that $x \neq 0$ and $\ell < 1/|x|$. Choose $s \in \mathbb{R}$ such that $\ell < s < (1/|x|)$ and define $\alpha := s|x|$. Since $M_k \to \ell$ and $\ell < s$, there is $k_0 \in \mathbb{N}$ such that $M_{k_0} < s = \alpha/|x|$. By the definition of M_{k_0}, we have $|c_k|^{1/k} < \alpha/|x|$, that is, $|c_k x^k| < \alpha^k$ for all $k \geq k_0$. Since $0 \leq \alpha < 1$, by the Comparison Test, we see that the series $\sum_{k=0}^{\infty} c_k x^k$ is absolutely convergent. Since this holds for all nonzero $x \in \mathbb{R}$ with $\ell < 1/|x|$, it follows that $r = \infty$ when $\ell = 0$ and $r \geq 1/\ell$ when $\ell > 0$.

Finally, suppose $\ell > 0$ and consider $x \in \mathbb{R}$ such that $|x| > 1/\ell$, that is, $1/|x| < \ell$. Since $\ell := \inf\{M_k : k \in \mathbb{N}\}$, it follows that for every $k \in \mathbb{N}$, we have $(1/|x|) < M_k$ and by the definition of M_k, there is $j_k \geq k$ with

$$\frac{1}{|x|} < |c_{j_k}|^{1/j_k}, \text{ that is, } 1 < |c_{j_k} x^{j_k}|.$$

So by the kth Term Test, $\sum_{k=0}^{\infty} c_k x^k$ is divergent. Since this holds for all $x \in \mathbb{R}$ with $|x| > 1/\ell$, we obtain $r \leq 1/\ell$. Thus $r = 1/\ell$ when $\ell \neq 0$. $\qquad \square$

We note that the real number ℓ defined in part (ii) of the above proposition is in fact equal to $\limsup_{k \to \infty} |c_k|^{1/k}$ as defined in Exercise 36 of Chapter 2. A result similar to Proposition 9.27 involving the ratios $|c_{k+1}|/|c_k|$, $k \in \mathbb{N}$, in place of the roots $|c_k|^{1/k}$, $k \in \mathbb{N}$, is given in Exercise 59.

The following result is useful in calculating the radius of convergence of a power series.

Proposition 9.28. *Let $\sum_{k=0}^{\infty} c_k x^k$ be a power series and r be its radius of convergence.*

(i) *If $|c_k|^{1/k} \to \ell$, then*
$$r = \begin{cases} 0 & \text{if } \ell = \infty, \\ \infty & \text{if } \ell = 0, \\ 1/\ell & \text{if } 0 < \ell < \infty. \end{cases}$$

(ii) *If $c_k \neq 0$ for all large k and $|c_{k+1}|/|c_k| \to \ell$, then the conclusion of (i) above holds.*

Proof. (i) Suppose $|c_k|^{1/k} \to \ell$. The desired result can be deduced from Proposition 9.27. However, we give an independent proof. Let $x \in \mathbb{R}$ and let us define $a_0(x) := c_0$ and $a_k(x) := c_k x^k$ for $k \in \mathbb{N}$. Then $|a_k(x)|^{1/k} = |c_k|^{1/k} |x| \to l|x|$ as $k \to \infty$. So by the Root Test, the series $\sum_{k=0}^{\infty} a_k(x)$ is convergent if $l|x| < 1$, and it is divergent if $l|x| > 1$. Hence the desired result follows from the very definition of the radius of convergence of a power series.

(ii) Suppose $c_k \neq 0$ for all large k and $|c_{k+1}|/|c_k| \to \ell$. A proof similar to the one above holds if we use the Ratio Test in place of the Root Test. $\square$

Examples 9.29. (i) For $k = 1, 2, \ldots$, let
$$c_{2k-1} = \frac{1}{4^k} \quad \text{and} \quad c_{2k} = \frac{1}{9^k}.$$

Then $|c_j|^{1/j}$ for $j = 1, 2, \ldots$ are given by
$$\frac{1}{4}, \frac{1}{3}, \frac{1}{4^{2/3}}, \frac{1}{3}, \frac{1}{4^{3/5}}, \frac{1}{3}, \frac{1}{4^{4/7}}, \cdots$$

and $M_k := \sup\{|c_j|^{1/j} : j \in \mathbb{N} \text{ and } j \geq k\} = \frac{1}{2}$ for all $k \in \mathbb{N}$. Hence the radius of convergence of the power series $\sum_{k=0}^{\infty} c_k x^k$ is 2.

(ii) For $k = 0, 1, 2, \ldots$, let $c_k := k^3/3^k$. Since
$$\frac{|c_{k+1}|}{|c_k|} = \frac{(k+1)^3}{3^{k+1}} \frac{3^k}{k^3} = \frac{1}{3} \left(1 + \frac{1}{k} \right)^3 \to \frac{1}{3} \quad \text{as } k \to \infty,$$

the radius of convergence of $\sum_{k=0}^{\infty} c_k x^k$ is 3.

(iii) For $k = 0, 1, 2, \ldots$, let $c_k := 2^k/k!$. Since
$$\frac{|c_{k+1}|}{|c_k|} = \frac{2^{k+1}}{(k+1)!} \frac{k!}{2^k} = \frac{2}{k+1} \to 0 \quad \text{as } n \to \infty,$$

the radius of convergence of $\sum_{k=0}^{\infty} c_k x^k$ is ∞. $\diamond$

Taylor Series

If r is the radius of convergence of a power series $\sum_{k=0}^{\infty} c_k x^k$ and $x_0 \in \mathbb{R}$ is such that $|x_0| < r$, then the series $\sum_{k=0}^{\infty} c_k x_0^k$ is convergent, but it may not be easy to find the sum of this series. Essentially the only power series whose sum we have found so far is the geometric series:

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x} \quad \text{for } x \in (-1, 1).$$

If we consider the function $f : (-1, 1) \to \mathbb{R}$ given by $f(x) = 1/(1-x)$, then we observe that $f(0) = 1$, f is infinitely differentiable and $f^{(k)}(0) = k!$ for each $k \in \mathbb{N}$. Hence for each $n \in \mathbb{N}$, the nth Taylor polynomial of f around 0 is given by $1 + x + \cdots + x^n$. This is also the nth partial sum of the power series $\sum_{k=0}^{\infty} x^k$. These considerations lead us to a special kind of series.

Let $a \in \mathbb{R}$, I be an interval containing a, and $f : I \to \mathbb{R}$ be an infinitely differentiable function. For $n = 0, 1, \ldots$, let $P_n(x)$ denote the nth Taylor polynomial of f around a, that is,

$$P_n(x) := \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x - a)^k.$$

The power series around the point a given by

$$\sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!}(x - a)^k$$

is called the **Taylor series** of f around a. Let $R_n(x) := f(x) - P_n(x)$ for $x \in I$. If $R_n(x) \to 0$ as $n \to \infty$ for some $x \in I$, then the nth partial sum $P_n(x)$ of the Taylor series of f around a converges to $f(x)$, that is, $f(x)$ is the sum of this series. In the special case $a = 0$, the Taylor series of f around a is sometimes called the **Maclaurin series** of f.

For all $n = 0, 1, \ldots$, we have $P_n(a) = f(a)$, that is, $R_n(a) = 0$. In particular, the Taylor series of f around a converges to $f(a)$ at $x = a$. However, at some other point $x \in I$, this series may be divergent, or even when it is convergent, it may not converge to $f(x)$. For example, consider $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) := |x|$. Then the Taylor series of f around 1 is easily seen to be $1 + (x - 1)$. But this is not equal to $f(x)$ when $x < 0$. As an extreme case, consider $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) := \begin{cases} e^{-1/x^2} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

Using Revision Exercise 29 given at the end of Chapter 7, we see that f is infinitely differentiable on $\mathbb{R}$ with $f^{(k)}(0) = 0$ for all $k = 0, 1, 2, \ldots$. Thus the

Taylor series of f around 0 is identically zero and it does not converge to $f(x)$ at any $x \neq 0$.

Taylor's Theorem says that for $x \in I$ and each $n \in \mathbb{N}$,

$$R_n(x) = \frac{f^{(n+1)}(c_{x,n})}{(n+1)!}(x-a)^{n+1} \quad \text{for some } c_{x,n} \text{ between } a \text{ and } x.$$

As we have mentioned in Chapter 4, the above expression for $R_n(x)$ is known as the Lagrange form of remainder in Taylor formula. Other forms of remainder are given in Exercise 49 of Chapter 4 and Exercise 46 of Chapter 6. The following simple result is useful in dealing with the sum of a Taylor series of a function.

Proposition 9.30. *Let $a \in \mathbb{R}$, I be an interval containing a and $f : I \to \mathbb{R}$ be an infinitely differentiable function. If there is $\alpha > 0$ such that*

$$\left| f^{(n)}(x) \right| \leq \alpha^n \quad \text{for all } n \in \mathbb{N} \text{ and } x \in I,$$

then the Taylor series of f converges to $f(x)$ for each $x \in I$.

Proof. Using the Lagrange form of remainder in the Taylor formula, we obtain

$$|R_n(x)| \leq \frac{|\alpha(x-a)|^{n+1}}{(n+1)!} \quad \text{for each } x \in I.$$

Consequently, by Example 2.7 (ii) of Chapter 2, we see that for each $x \in I$, $R_n(x) \to 0$ as $n \to \infty$. $\qquad \square$

We shall use the above result tacitly in the examples below, in which we determine the Taylor series of some classical functions.

Examples 9.31. (i) Let $a := 0$, $I := \mathbb{R}$, and $f : I \to \mathbb{R}$ be the sine function given by $f(x) := \sin x$. Then $|f^{(n)}(x)| \leq 1$ for all $n \in \mathbb{N}$ and $x \in I$. Using the Taylor polynomials of f around 0 found in Section 7.2, we see that the Taylor series of f is convergent for $x \in I$ and

$$\sin x = x - \frac{x^3}{3!} + \frac{x^s}{5!} - \cdots = \sum_{k=1}^{\infty}(-1)^{k-1}\frac{x^{2k-1}}{(2k-1)!} \quad \text{for } x \in \mathbb{R}.$$

For the Taylor series of the cosine function, see Exercise 24.

(ii) Let $a := 0$, $\beta > 0$, $I := (-\beta, \beta)$, and $f : I \to \mathbb{R}$ be the exponential function given by $f(x) := e^x$. Then $|f^{(n)}(x)| = e^x \leq e^\beta$ for all $n \in \mathbb{N}$ and $x \in I$. Using the Taylor polynomials for f found in Section 7.1, we see that the Taylor series of f is convergent for $x \in I$ and

$$e^x = 1 + x + \frac{x^2}{2!} + \cdots = \sum_{k=0}^{\infty}\frac{x^k}{k!} \quad \text{for } x \in (-\beta, \beta).$$

Since $\beta > 0$ is arbitrary, we see that

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} \quad \text{for all } x \in \mathbb{R}.$$

(iii) Let $a := 0$, $I := (-1, 1)$, and $f : I \to \mathbb{R}$ be given by $f(x) := \ln(1 + x)$. Then $f(0) = 0$ and for each $k \in \mathbb{N}$, we have

$$f^{(k)}(x) = (-1)^{k-1} \frac{(k-1)!}{(1+x)^k} \quad \text{for } x \in (-1, 1).$$

In particular, $f^{(k)}(0) = (-1)^{k-1}(k-1)!$ for each $k \in \mathbb{N}$. Hence the Taylor series of f around 0 is given by

$$\sum_{k=1}^{\infty} (-1)^{k-1} \frac{x^k}{k}, \quad x \in \mathbb{R}.$$

Since it is not easy to show that $R_n(x) \to 0$ as $n \to \infty$ for $x \in (-1, 1)$ using Lagrange's Form of Remainder, we proceed as follows. By the definition of the logarithmic function, we have

$$f(x) = \int_1^{1+x} \frac{dt}{t} = \int_0^x \frac{ds}{1+s} \quad \text{for } x \in (-1, 1).$$

Now for each $s \in \mathbb{R}$ with $s \neq -1$ and each $n \in \mathbb{N}$, we note that

$$\frac{1}{1+s} = \frac{1}{1-(-s)} = 1 - s + s^2 + \cdots + (-1)^{n-1} s^{n-1} + (-1)^n \frac{s^n}{1+s}.$$

Given any $x \in (-1, 1)$, we integrate both sides from 0 to x and obtain

$$f(x) = \sum_{k=1}^{n} (-1)^{k-1} \frac{x^k}{k} + (-1)^n \int_0^x \frac{s^n}{1+s} ds.$$

If $x = 0$, this gives $f(0) = 1$. Suppose $x \in (0, 1)$. Then

$$\left| \int_0^x \frac{s^n}{1+s} ds \right| \leq \int_0^x s^n ds = \frac{x^{n+1}}{n+1},$$

which tends to 0 as $n \to \infty$. Next, suppose $x \in (-1, 0)$. Substituting $u = -s$, we obtain

$$\left| \int_0^x \frac{s^n}{1+s} ds \right| = \left| \int_0^{-x} \frac{(-1)^{n+1} u^n}{1-u} du \right| \leq \int_0^{-x} \frac{u^n}{1+x} du = \frac{1}{1+x} \left(\frac{(-x)^{n+1}}{n+1} \right),$$

which tends to 0 as $n \to \infty$. Hence we see that the Taylor series of f is convergent for $x \in (-1, 1)$ and

$$\ln(1+x) = \sum_{k=0}^{\infty} (-1)^{k-1} \frac{x^k}{k} \quad \text{for } x \in (-1, 1).$$

(iv) Let $a := 0$, $I = (-1, 1)$, $r \in \mathbb{R}$, and $f : I \to \mathbb{R}$ be given by $f(x) := (1+x)^r$. Then $f(0) = 1$ and $f^{(k)}(0) = r(r-1) \cdots (r - k + 1)$ for $k \in \mathbb{N}$. If r is a nonnegative integer, then $f^{(k)}(0) = 0$ for all $k > r + 1$. Hence

$$f(x) = 1 + \sum_{k=1}^{r} \binom{r}{k} x^k \quad \text{for } x \in I.$$

(Note that we obtain the same result by the Binomial Theorem.) Suppose now that r is not a nonnegative integer. Then $f^{(k)}(0) \neq 0$ for each $k \geq 0$. Thus for each $n \in \mathbb{N}$, the nth Taylor polynomial of f around 0 is given by

$$P_n(x) := 1 + \sum_{k=1}^{n} \frac{r(r-1) \cdots (r - k + 1)}{k!} x^k.$$

Using Cauchy's Form of Remainder (Exercise 49 of Chapter 4), it can be shown that $f(x) - P_n(x) \to 0$ as $n \to \infty$ for each $x \in (-1, 1)$. (See Exercise 60.) Hence we see that the Taylor series of f is convergent for $x \in I$ and

$$(1 + x)^r = 1 + \sum_{k=1}^{\infty} \frac{r(r-1) \cdots (r - k + 1)}{k!} x^k \quad \text{for } x \in (-1, 1).$$

This series is known as the **binomial series**. In particular, if $r = -1$, we have

$$\frac{1}{1 + x} = 1 + \sum_{k=1}^{\infty} \frac{(-1)(-2) \cdots (-k)}{k!} x^k = \sum_{k=0}^{\infty} (-1)^k x^k \quad \text{for } x \in (-1, 1).$$

Replacing x by $-x$, we thus recover the geometric series

$$\frac{1}{1 - x} = \sum_{k=0}^{\infty} x^k \quad \text{for } x \in (-1, 1),$$

with which we started our discussion of Taylor series.   ◇

Remark 9.32. Let I be an open interval in $\mathbb{R}$ and $f : I \to \mathbb{R}$ be infinitely differentiable at every point of I. If for every $a \in I$, there is $r > 0$ such that

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!} (x - a)^k \quad \text{for every } x \in I \text{ with } |x - a| < r,$$

then f is said to be a **real analytic function**. In other words, f is real analytic on I if the Taylor series of f around each point of I converges to $f(x)$ for every $x \in I$. Clearly, polynomial functions are real analytic functions on $\mathbb{R}$. Also, as seen above, the exponential function as well as the sine function are real analytic functions on $\mathbb{R}$. It is easy to see that sums of real analytic functions are real analytic. Also, it can be shown that products of real analytic functions and reciprocals of nowhere-vanishing real analytic functions are real analytic. On the other hand, there do exist infinitely differentiable functions on $\mathbb{R}$ that are not real analytic. Indeed, as noted earlier, it suffices to consider $f : \mathbb{R} \to \mathbb{R}$ defined by $f(0) := 0$ and $f(x) := e^{-1/x^2}$ for $x \neq 0$.   ◇

9.4 Convergence of Improper Integrals

In Chapter 6, we considered the Riemann integral of a bounded function defined on a closed and bounded interval. In this and the next two sections, we shall extend the process of integration to functions defined on a semi-infinite interval or a doubly infinite interval, and also to unbounded functions defined on bounded or unbounded intervals.

We begin by considering bounded functions defined on a semi-infinite interval of the form $[a, \infty)$, where $a \in \mathbb{R}$. Our treatment will run parallel to that of infinite series given in Sections 9.1 and 9.2. In analogy with an infinite series, we shall first give a formal (and pedantic) definition of an improper integral and then adopt suitable conventions in order to simplify our treatment.

Let $a \in \mathbb{R}$. An **improper integral** on $[a, \infty)$ is an ordered pair (f, F) of real-valued functions f and F defined on $[a, \infty)$ such that f is integrable on $[a, x]$ for every $x \geq a$ and

$$F(x) = \int_a^x f(t)dt \quad \text{for all } x \in [a, \infty).$$

Note that in view of the Fundamental Theorem of Calculus (Proposition 6.21), we have the following. If (f, F) is an improper integral on $[a, \infty)$, then F is continuous with $F(a) = 0$, and moreover, if f is continuous, then F is differentiable with $f = F'$. Conversely, if $f, F : [a, \infty) \to \mathbb{R}$ are such that f is continuous and F is differentiable with $F(a) = 0$ and $f = F'$, then (f, F) is an improper integral on $[a, \infty)$. For simplicity and brevity, we shall use the informal but suggestive notation $\int_a^\infty f(t)dt$ for the improper integral (f, F) on $[a, \infty)$. In this notation, prominence is given to the first function f, but the second function F is just as important. At any rate, F is uniquely determined by f, and if f is continuous, then f is uniquely determined by F.

We say that an improper integral $\int_a^\infty f(t)dt$ is **convergent** if the limit

$$\lim_{x \to \infty} F(x) = \lim_{x \to \infty} \int_a^x f(t)dt$$

exists. It is clear that if this limit exists, then it is unique, and we may denote it by the same symbol $\int_a^\infty f(t)dt$ used to denote the improper integral. Usually, when we write

$$\int_a^\infty f(t)dt = I,$$

we mean that I is a real number and the improper integral $\int_a^\infty f(t)dt$ is convergent with I as its limiting value. In this case we may also say that $\int_a^\infty f(t)dt$ **converges** to I. An improper integral that is not convergent is said to be **divergent**. In particular, if $\int_a^x f(t)dt \to \infty$ or $\int_a^x f(t)dt \to -\infty$ as $x \to \infty$, then we say that the improper integral **diverges** to ∞ or to $-\infty$, as the case may be.

It is useful to keep in mind that the convergence of an improper integral $\int_a^\infty f(t)dt$ is not affected by changing the initial point a of the interval $[a, \infty)$, although the limiting value $\lim_{x\to\infty} F(x)$ may change by doing so. Indeed, if $a' \geq a$, then

$$\int_a^\infty f(t)dt \text{ is convergent} \iff \int_{a'}^\infty f(t)dt \text{ is convergent,}$$

and if this holds, then the limiting values are related by the equation

$$\int_a^\infty f(t)dt = \int_a^{a'} f(t)dt + \int_{a'}^\infty f(t)dt.$$

Examples 9.33. (i) Let $a > 0$ and consider the improper integral $\int_1^\infty a^t dt$. Given any $x \in [1, \infty)$, we have

$$\int_1^x a^t dt = \begin{cases} (a^x - a)/\ln a & \text{if } a \neq 1, \\ x - 1 & \text{if } a = 1. \end{cases}$$

Thus, in view of part (iii) of Proposition 7.7, it follows that if $a < 1$, then $\int_1^\infty a^t dt = -(a/\ln a)$, while if $a \geq 1$, then $\int_a^\infty f(t)dt$ diverges to ∞.

(ii) The improper integral $\int_0^\infty te^{-t^2} dt$ converges to $\frac{1}{2}$, since

$$\int_0^x te^{-t^2} dt = \frac{1}{2}\int_0^{x^2} e^{-s} ds = \frac{1}{2}\left(1 - e^{-x^2}\right) \to \frac{1}{2} \quad \text{as } x \to \infty.$$

(iii) Let $p \in \mathbb{R}$ and consider the improper integral $\int_1^\infty (1/t^p)dt$. Given any $x \in [1, \infty)$, we have

$$\int_1^x \frac{1}{t^p} dt = \begin{cases} (x^{1-p} - 1)/(1 - p) & \text{if } p \neq 1, \\ \ln x & \text{if } p = 1. \end{cases}$$

It follows that if $p > 1$, then $\int_1^\infty (1/t^p)dt$ converges to $1/(p-1)$, while if $p \leq 1$, then it diverges to ∞.

(iv) The improper integral $\int_0^\infty \left(1/(1+t^2)\right) dt$ converges to $\pi/2$, since

$$\int_0^x \frac{dt}{1+t^2} = \arctan x \to \frac{\pi}{2} \quad \text{as } x \to \infty,$$

(v) The improper integral $\int_0^\infty \cos t\, dt$ is divergent, since $\int_0^x \cos t\, dt = \sin x$ for all $x \in \mathbb{R}$ and $\lim_{x\to\infty} \sin x$ does not exist. $\diamond$

It may be observed that there is a remarkable analogy between the definition of an infinite series $\sum_{k=1}^\infty a_k$ and the definition of an infinite integral $\int_a^\infty f(t)dt$. The sequence of terms (a_k) corresponds to the function $f : [a, \infty) \to \mathbb{R}$, and a partial sum $A_n := \sum_{k=1}^n a_k$, where $n \in \mathbb{N}$, corresponds to a 'partial integral' $F(x) = \int_a^x f(t)dt$, where $x \in [a, \infty)$. The convention that

A_0, being the empty sum, is 0 corresponds to the initial condition $F(a) = 0$. Further, the difference quotient

$$a_k = A_k - A_{k-1} = \frac{A_k - A_{k-1}}{k - (k-1)}, \quad \text{where } k \in \mathbb{N},$$

corresponds to the derivative

$$f(t) = F'(t) = \lim_{s \to t} \frac{f(s) - f(t)}{s - t}, \quad \text{where } t \in [a, \infty).$$

This analogy will become more and more apparent as we develop the theory of improper integrals. At the same time, we will point out an instance in which the analogy breaks down.

The following results follow from the corresponding results for limits of functions of a real variable as the variable tends to infinity, just as similar results in the case of infinite series followed from the corresponding results for limits of sequences. In what follows, we have let $a \in \mathbb{R}$ and f, g, h denote real-valued functions on $[a, \infty)$.

1. If $\int_a^\infty f(t)dt$ is convergent, then the set $\{\int_a^x f(t)dt : x \in [a, \infty)\}$ of 'partial integrals' is bounded. (To see this, let $F(x) := \int_a^x f(t)dt$ for $x \in [a, \infty)$ and note that since $\int_a^\infty f(t)dt$ is convergent, there is $x_0 \geq a$ such that $|F(x)| \leq 1 + |\int_a^\infty f(t)dt|$ for all $x \geq x_0$, and moreover, since F is continuous on the closed and bounded interval $[a, x_0]$, it is bounded on $[a, x_0]$.)

2. Let $\int_a^\infty f(t)dt = I$ and $\int_a^\infty g(t)dt = J$. Then

$$\int_a^\infty (f(t) + g(t))dt = I + J \quad \text{and} \quad \int_a^\infty (rf)(t)dt = rI \quad \text{for any } r \in \mathbb{R}.$$

Further, if $f(t) \leq g(t)$ for all $t \in [a, \infty)$, then $I \leq J$.

3. **(Sandwich Theorem)** If for each $t \in [a, \infty)$, we have $f(t) \leq h(t) \leq g(t)$, and $\int_a^\infty f(t)dt = I = \int_a^\infty g(t)dt$, then $\int_a^\infty h(t)dt = I$.

4. **(Cauchy Criterion)** An improper integral $\int_a^\infty f(t)dt$ is convergent if and only if for every $\epsilon > 0$, there is $x_0 \in [a, \infty)$ such that

$$\left| \int_x^y f(t)dt \right| < \epsilon \quad \text{for all } y \geq x \geq x_0.$$

(To see this, let F denote the partial integral, so that $F(y) - F(x) = \int_x^y f(t)dt$ for all $y \geq x \geq x_0$, and use the analogue of Proposition 3.28 for the case $x \to \infty$.)

Integrals of Derivatives and of Nonnegative Functions

We now consider two results about improper integrals of functions of a special kind. The following result is an analogue of the result about the convergence of a telescoping series (Proposition 9.3).

Proposition 9.34. *Let $g : [a, \infty) \to \mathbb{R}$ be a differentiable function such that its derivative g' is integrable on $[a, x]$ for every $x \geq a$. Then $\int_a^\infty g'(t)dt$ is convergent if and only if $\lim_{x \to \infty} g(x)$ exists, and in this case,*

$$\int_a^\infty g'(t)dt = \lim_{x \to \infty} g(x) - g(a).$$

Proof. By part (i) of the FTC (Proposition 6.21), we have

$$\int_a^x g'(t)dt = g(x) - g(a) \quad \text{for all } x \in [a, \infty).$$

This implies the desired result. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

It may be noted that if a function $f : [a, \infty) \to \mathbb{R}$ is continuous and if we define $g : [a, \infty) \to \mathbb{R}$ by

$$g(x) := \int_a^x f(t)dt,$$

that is, if g is the 'partial integral' of $\int_a^\infty f(t)dt$, then by part (i) of the FTC, the improper integral $\int_a^\infty f(t)dt$ can be written as $\int_a^\infty g'(t)dt$. But then determining whether $\lim_{x \to \infty} g(x)$ exists is the same as determining whether the given improper integral $\int_a^\infty f(t)dt$ is convergent. In some special cases, however, it is possible to find an antiderivative g of the function f without considering any 'partial integral' of f. In these cases, we can determine the convergence of the improper integral $\int_a^\infty f(t)dt$ using Proposition 9.34. In fact, Examples 9.33 (i)–(v) illustrate this technique.

Our next result is regarding the convergence of an improper integral of a nonnegative function. It is an analogue of the result about the convergence of a series of nonnegative terms (Proposition 9.4).

Proposition 9.35. *Let $f : [a, \infty) \to \mathbb{R}$ be a nonnegative function. Then $\int_a^\infty f(t)dt$ is convergent if and only if the function $F : [a, \infty) \to \mathbb{R}$ defined by $F(x) = \int_a^x f(t)dt$ is bounded above, and in this case,*

$$\int_a^\infty f(t)dt = \sup\{F(x) : x \in [a, \infty)\}.$$

If the function F is not bounded above, then $\int_a^\infty f(t)dt$ diverges to ∞.

Proof. Since $f(t) \geq 0$ for all $t \in [a, \infty)$, using the domain additivity of Riemann integrals (Proposition 6.7) we see that

$$F(x) = \int_a^y f(t)dt + \int_y^x f(t)dt \geq \int_a^y f(t)dt = F(y) \quad \text{for all } x \geq y \geq a.$$

Hence the function F is monotonically increasing. By part (i) of Proposition 3.35 with $b = \infty$, we see that $\lim_{x \to \infty} F(x)$ exists if and only if F is bounded above, and in this case

$$\int_a^\infty f(t)dt = \lim_{x\to\infty} F(x) = \sup\{F(x) : x \in [a, \infty)\}.$$

Also, by part (i) of Proposition 3.35 with $b = \infty$, if F is not bounded above, then $F(x) \to \infty$ as $x \to \infty$, that is, $\int_a^\infty f(t)dt$ diverges to ∞. □

A result similar to the one above holds if $f(t) \leq 0$ for all $t \in [a, \infty)$. (See Exercise 27.) More generally, if $f(t)$ has the same sign **for all large** t, that is, if there is $t_0 \in [a, \infty)$ such that $f(t)$ has the same sign for all $t \geq t_0$, then $\int_a^\infty f(t)dt$ is convergent if and only if F is bounded. However, if there is no $t_0 \in [a, \infty)$ such that $f(t)$ is of the same sign for all $t \geq t_0$, then the improper integral $\int_a^\infty f(t)dt$ may diverge even though F is bounded. This is illustrated by the improper integral $\int_1^\infty f(t)dt$, where $f : [1, \infty) \to \mathbb{R}$ is defined by $f(x) := (-1)^{[x]}$. Here, the partial integral $F : [1, \infty) \to \mathbb{R}$ is given by $F(x) = -1 + x - [x]$ if $[x]$ is even and $F(x) = -x + [x]$ if $[x]$ is odd. Clearly, F is bounded on $[1, \infty)$, but since $F(2n - 1) = 0$ and $F(2n) = -1$ for all $n \in \mathbb{N}$, the limit of $F(x)$ as $x \to \infty$ does not exist, that is, $\int_1^\infty f(t)dt$ is divergent.

Example 9.36. Consider $f : [0, \infty) \to \mathbb{R}$ defined by $f(t) := (1+\sin t)/(1+t^2)$. Then $f(t) \geq 0$ for all $t \in [0, \infty)$ and

$$F(x) = \int_0^x \frac{1 + \sin t}{1 + t^2}dt \leq \int_0^x \frac{2}{1 + t^2}dt = 2\arctan x \leq \pi \quad \text{for all } x \in [0, \infty).$$

Hence $\int_0^\infty f(t)dt$ is convergent. On the other hand, consider $g : [0, \infty) \to \mathbb{R}$ defined by $g(t) := (2 + \cos t)/t$. We have $g(t) \geq 0$ for all $t \in [1, \infty)$ and

$$G(x) = \int_1^x \frac{2 + \cos t}{t}dt \geq \int_1^x \frac{1}{t}dt = \ln x \quad \text{for all } x \in [1, \infty).$$

Since $\ln x \to \infty$ as $x \to \infty$, it follows that $\int_1^\infty g(t)dt$ diverges to ∞. ◇

An improper integral $\int_a^\infty f(t)dt$ is said to be **absolutely convergent** if the improper integral $\int_a^\infty |f(t)|dt$ is convergent. The following result is an analogue of Proposition 9.5.

Proposition 9.37. *An absolutely convergent improper integral is convergent.*

Proof. Let $a \in \mathbb{R}$ and $\int_a^\infty f(t)dt$ be an absolutely convergent improper integral on $[a, \infty)$. Consider the positive part $f^+ : [a, \infty) \to \mathbb{R}$ and the negative part $f^- : [a, \infty) \to \mathbb{R}$ of f defined by

$$f^+(t) := \frac{|f(t)| + f(t)}{2} \quad \text{and} \quad f^-(t) = \frac{|f(t)| - f(t)}{2} \quad \text{for } t \in [a, \infty).$$

For $x \in [a, \infty)$, let $F(x), F^+(x), F^-(x)$, and $G(x)$ denote the 'partial integrals' of the improper integrals $\int_a^\infty f(t)dt$, $\int_a^\infty f^+(t)dt$, $\int_a^\infty f^-(t)dt$, and $\int_a^\infty |f(t)|dt$ respectively. Since $\int_a^\infty |f(t)|dt$ is convergent, the function G is bounded. Also,

$$0 \le F^+(x) \le G(x) \quad \text{and} \quad 0 \le F^-(x) \le G(x) \quad \text{for all } x \in [a, \infty).$$

So by Proposition 9.35, both $\int_a^\infty f^+(t)dt$ and $\int_a^\infty f^-(t)dt$ are convergent. But $f(t) = f^+(t) - f^-(t)$ for all $t \in [a, \infty)$. Hence $\int_a^\infty f(t)dt$ is convergent. □

Example 9.38 below shows that the converse of the above result does not hold. A convergent improper integral that is not absolutely convergent is said to be **conditionally convergent**. Another example of a conditionally convergent improper integral (which is modeled on the conditionally convergent infinite series $\sum_{k=1}^\infty (-1)^{k-1}/k$) is given in Exercise 42.

Example 9.38. Consider the improper integral $\int_1^\infty (\cos t/t)dt$. Integrating by parts, we have

$$\int_1^x \frac{\cos t}{t}dt = \frac{\sin x}{x} - \sin 1 + \int_1^x \frac{\sin t}{t^2}dt \quad \text{for all } x \ge 1.$$

Further, $(\sin x)/x \to 0$ as $x \to \infty$ and also

$$\int_1^x \left|\frac{\sin t}{t^2}\right| dt \le \int_1^x \frac{1}{t^2}dt = 1 - \frac{1}{x} \to 1 \quad \text{as } x \to \infty.$$

Hence by Proposition 9.35, the improper integral $\int_1^\infty |(\sin t)/t^2|dt$ is convergent, that is, the improper integral $\int_1^\infty (\sin t)/t^2 dt$ is absolutely convergent, and so by Proposition 9.37, it is convergent. Consequently,

$$\int_1^\infty \frac{\cos t}{t}dt = -\sin 1 + \int_1^\infty \frac{\sin t}{t^2}dt.$$

On the other hand, for each $n \in \mathbb{N}$ with $n \ge 2$, we have

$$\int_\pi^{n\pi} \left|\frac{\cos t}{t}\right| dt = \sum_{k=2}^n \int_{(k-1)\pi}^{k\pi} \frac{|\cos t|}{t}dt \ge \sum_{k=2}^n \int_{(k-1)\pi}^{k\pi} \frac{|\cos t|}{k\pi} dt = \sum_{k=2}^n \frac{2}{k\pi}.$$

Since the series $\sum_{k=2}^\infty 1/k$ diverges to ∞, it follows from Proposition 9.35 that the improper integral

$$\int_1^\infty \left|\frac{\cos t}{t}\right| dt$$

diverges to ∞. Thus $\int_1^\infty (\cos t/t)dt$ is conditionally convergent. ◇

Let us now discuss whether the convergence of an improper integral $\int_1^\infty f(t)dt$ is related to the convergence of the infinite series $\sum_{k=1}^\infty f(k)$. Consider $f : [1, \infty) \to \mathbb{R}$ given by $f(t) := 1$ if $t \in \mathbb{N}$ and $f(t) := 0$ if $t \notin \mathbb{N}$. Then it is easy to see that $\int_a^\infty f(t)dt$ is convergent but $\sum_{k=1}^\infty f(k)$ is divergent. On the other hand, if we let $g := f - 1$, then it is easily seen that $\int_1^\infty g(t)dt$ is divergent, but $\sum_{k=1}^\infty g(k)$ is convergent. Thus, in general, the convergence of $\int_1^\infty f(t)dt$ is independent of the convergence of $\sum_{k=1}^\infty f(k)$. In view of this, the following result is noteworthy.

Proposition 9.39 (Integral Test). *Let $f : [1, \infty) \to \mathbb{R}$ be a nonnegative monotonically decreasing function. Then the improper integral $\int_1^\infty f(t)dt$ is convergent if and only if the infinite series $\sum_{k=1}^\infty f(k)$ is convergent, and in this case we have*

$$\sum_{k=2}^\infty f(k) \leq \int_1^\infty f(t)dt \leq \sum_{k=1}^\infty f(k),$$

or, equivalently,

$$\int_1^\infty f(t)dt \leq \sum_{k=1}^\infty f(k) \leq f(1) + \int_1^\infty f(t)dt.$$

Also, the improper integral $\int_1^\infty f(t)dt$ diverges to ∞ if and only if the infinite series $\sum_{k=1}^\infty f(k)$ diverges to ∞.

Proof. First, note that since f is monotonic, by part (i) of Proposition 6.9, f is integrable on $[a, x]$ for every $x \in [1, \infty)$. Define $F : [1, \infty) \to \mathbb{R}$ by $F(x) := \int_1^x f(t)dt$. Since the function f is nonnegative, the function F is monotonically increasing. Hence Proposition 9.35 implies that $\int_1^\infty f(t)dt$ is convergent if and only if the set $\{F(n) : n \in N\}$ is bounded above, and in this case

$$\int_1^\infty f(t)dt = \sup\{F(n) : n \in \mathbb{N}\} = \lim_{n \to \infty} F(n).$$

Also, using Proposition 9.35 and the fact that f is monotonically increasing, we see that

$$\int_1^\infty f(t)dt \text{ diverges to } \infty \iff F(n) \to \infty \text{ as } n \to \infty.$$

Define

$$a_k := \int_k^{k+1} f(t)dt \quad \text{for } k \in \mathbb{N} \quad \text{and} \quad A_n := \sum_{k=1}^n a_k \quad \text{for } n \in \mathbb{N}.$$

Then $A_n = F(n + 1)$ for all $n \in \mathbb{N}$. Further, since $a_k \geq 0$ for all $k \in \mathbb{N}$, it follows from Proposition 9.4 that the series $\sum_{k=1}^\infty a_k$ is convergent if and only if the sequence $(F(n))$ is bounded above, that is, the improper integral $\int_1^\infty f(t)dt$ is convergent. Also, $\int_1^\infty f(t)dt$ diverges to ∞ if and only if $\sum_{k=1}^\infty a_k$ diverges to ∞.

Now since f is monotonically decreasing, we have

$$f(k + 1) \leq a_k \leq f(k) \quad \text{for all } k \in \mathbb{N}.$$

The Comparison Test for series (Proposition 9.9) shows that $\sum_{k=1}^\infty a_k$ is convergent if and only if $\sum_{k=1}^\infty f(k)$ is convergent, and $\sum_{k=1}^\infty a_k$ diverges to ∞ if and only if $\sum_{k=1}^\infty f(k)$ diverges to ∞. Finally, since

$$\sum_{k=2}^{n+1} f(k) = \sum_{k=1}^{n} f(k+1) \leq A_n \leq \sum_{k=1}^{n} f(k) \quad \text{for all } n \in \mathbb{N},$$

we see that

$$\sum_{k=2}^{\infty} f(k) \leq \lim_{n\to\infty} A_n = \int_1^{\infty} f(t)dt \leq \sum_{k=1}^{\infty} f(k)$$

whenever $\int_1^{\infty} f(t)dt$ is convergent. $\qquad\square$

The above result can be extremely useful in deducing the convergence or the divergence of infinite series. Further, it can be used to obtain lower bounds and upper bounds for the partial sums of a series. These yield, in case the series converges, a lower bound and an upper bound for the sum of the series. These remarks are illustrated by Examples 9.40 below.

If in the Integral Test, the hypothesis that $f : [1, \infty) \to \mathbb{R}$ is a nonnegative monotonically decreasing function is not satisfied, but f is a differentiable function such that f' is integrable on $[1, x]$ for every $x \geq 1$, then the convergence of $\int_1^{\infty} f(t)dt$ and $\sum_{k=1}^{\infty} f(k)$ can be related by what is known as **Euler's Summation Formula**. We refer the interested reader to pages 74–75 of [59].

Examples 9.40. (i) Let $p > 0$ and $f(t) := 1/t^p$ for $t \in [1, \infty)$. Then f is a nonnegative monotonically decreasing function. We have seen in Example 9.1 (iv) that $\int_1^{\infty} f(t)dt$ is convergent if $p > 1$ and it diverges to ∞ if $p \leq 1$. Hence by Proposition 9.39 we see that $\sum_{k=1}^{\infty}(1/k^p)$ is convergent if $p > 1$ and it diverges to ∞ if $p \leq 1$. We thus obtain an alternative proof the result given in Example 9.1 (iii). Further, if $p > 1$, we can estimate the sum using Proposition 9.39 as follows:

$$\frac{1}{p-1} = \int_1^{\infty} \frac{1}{t^p}dt \leq \sum_{k=1}^{\infty} \frac{1}{k^p} \leq 1 + \int_1^{\infty} \frac{1}{t^p}dt = \frac{p}{p-1}.$$

(ii) Let $p > 0$ and consider the infinite series

$$\sum_{k=2}^{\infty} \frac{1}{k(\ln k)^p} = \sum_{k=1}^{\infty} \frac{1}{(k+1)(\ln(k+1))^p}.$$

If $f : [1, \infty) \to \mathbb{R}$ is defined by

$$f(t) := \frac{1}{(t+1)(\ln(t+1))^p} \quad \text{for } t \in [1, \infty),$$

then f is a nonnegative function and since

$$f'(t) = -\frac{(\ln(t+1))^p + p(\ln(t+1))^{p-1}}{(t+1)^2 (\ln(t+1))^{2p}} < 0 \quad \text{for all } t \in [1, \infty),$$

we see that f is monotonically decreasing. Now for $x \in [1, \infty)$, we have

$$\int_1^x f(t)dt = \int_{\ln 2}^{\ln(x+1)} \frac{1}{s^p}\, ds = \begin{cases} \dfrac{(\ln(x+1))^{1-p} - (\ln 2)^{1-p}}{1-p} & \text{if } p \neq 1, \\[3mm] \ln\left(\ln(x+1)\right) - \ln(\ln 2) & \text{if } p = 1. \end{cases}$$

Letting $x \to \infty$, we see that

$$\int_1^\infty f(t)dt = \frac{1}{(p-1)(\ln 2)^{p-1}} \text{ if } p > 1 \quad \text{and} \quad \int_1^\infty f(t)dt \text{ is divergent if } p \leq 1.$$

This shows that the infinite series $\sum_{k=2}^\infty 1/[k(\ln k)^p]$ is convergent if $p > 1$ and it diverges to ∞ if $p \leq 1$. Further, if $p > 1$, we have

$$\frac{1}{(p-1)(\ln 2)^{p-1}} \leq \sum_{k=2}^\infty \frac{1}{k(\ln k)^p} \leq \frac{1}{(\ln 2)^{p-1}}\left[\frac{1}{2\ln 2} + \frac{1}{p-1}\right].$$

The upper and lower bounds on the sums of the series in (i) and (ii) above are noteworthy.   ◇

9.5 Convergence Tests for Improper Integrals

In this section we shall consider several tests that enable us to conclude the convergence or divergence of improper integrals. We begin by pointing out that even if an improper integral $\int_a^\infty f(t)dt$ is convergent, $f(t)$ may not tend to 0 as $t \to \infty$. In other words, the kth Term Test for series (Proposition 9.6) does not have an analogue for improper integrals.
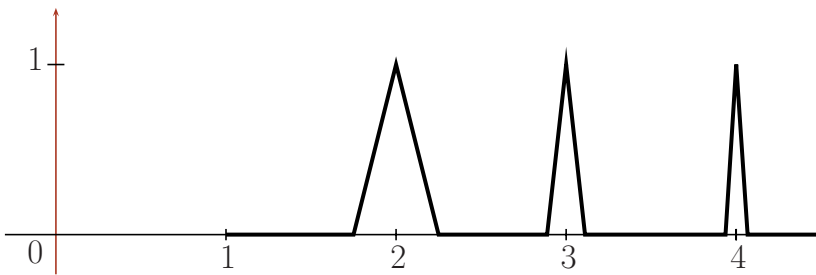
Fig. 9.1. Graph of the piecewise linear function in Example 9.41

Example 9.41. Let $f : [1, \infty) \to \mathbb{R}$ be the piecewise linear function whose graph is as in Figure 9.1. Formally, f is defined as follows. Given any $k \in \mathbb{N}$ with $k \geq 2$, let

$$f(t) := \begin{cases} k^2t - k^3 + 1 & \text{if } k - \dfrac{1}{k^2} \leq t \leq k, \\ -k^2t + k^3 + 1 & \text{if } k < t \leq k + \dfrac{1}{k^2} \end{cases}$$

Moreover, let $f(t) := 0$ if $1 \leq t < 2 - (1/2)^2$ and also

$$f(t) := 0 \quad \text{if } k + \frac{1}{k^2} < t < (k+1) - \frac{1}{(k+1)^2} \text{ for any } k \in \mathbb{N} \text{ with } k \geq 2.$$

Note that the function f is continuous. Let $F(x) := \int_1^x f(t)dt$ for $x \in [1, \infty)$. Since the area of a triangle having height 1 and base $2/k^2$ is equal to $1/k^2$, we see that for each $n \in \mathbb{N}$ with $n \geq 2$,

$$F(x) = \sum_{k=2}^{n} \frac{1}{k^2} \quad \text{if } n + \frac{1}{n^2} \leq x < (n+1) - \frac{1}{(n+1)^2}.$$

Also, since the series $\sum_{k=2}^{\infty} 1/k^2$ is convergent, it follows that the function F is bounded. So by Proposition 9.35, $\int_1^\infty f(t)dt$ is convergent. However, since $f(k) = 1$ for each $k \in \mathbb{N}$ with $k \geq 2$, we see that $f(t) \nrightarrow 0$ as $t \to \infty$. $\diamond$

By modifying the function in the above example, one can obtain a continuous function $g : [1, \infty) \to \infty$ such that $\int_1^\infty g(t)dt$ is convergent, but $(g(k))$ is in fact an unbounded sequence. (See Exercise 26.) For more examples of this type, see Exercise 28. On the other hand, suppose $f : [a, \infty) \to \mathbb{R}$ is differentiable. If $\int_a^\infty f(t)dt$ and $\int_a^\infty f'(t)dt$ are convergent, then it can be shown that $f(t) \to 0$ as $t \to \infty$. (See Exercise 29.)

Tests for Absolute Convergence of Improper Integrals

We shall now consider, wherever possible, analogues of the tests for absolute convergence of infinite series in the case of improper integrals.

Proposition 9.42 (Comparison Test for Improper Integrals). *Suppose $a \in \mathbb{R}$ and $f, g : [a, \infty) \to \mathbb{R}$ are such that both f and g are integrable on $[a, x]$ for every $x \geq a$ and $|f| \leq g$. If $\int_a^\infty g(t)dt$ is convergent, then $\int_a^\infty f(t)dt$ is absolutely convergent and*

$$\left| \int_a^\infty f(t)dt \right| \leq \int_a^\infty g(t)dt.$$

Proof. For $x \in [1, \infty)$, let

$$F(x) := \int_a^x f(t)dt, \quad G(x) := \int_a^x g(t)dt, \quad \text{and} \quad H(x) = \int_a^x |f(t)|dt.$$

Assume that $\int_a^\infty g(t)dt$ is convergent. Then the function G is bounded above. Since $|f| \leq g$, we see that $H \leq G$, and hence the function H is bounded

above. Also, since $|f| \geq 0$, it follows from Proposition 9.35 that $\int_a^\infty |f(t)|dt$ is convergent, that is, $\int_a^\infty f(t)dt$ is absolutely convergent. Further, since $-f \leq |f| \leq g$ and $f \leq |f| \leq g$, we have $-F(x) \leq H(x) \leq G(x)$ and $F(x) \leq H(x) \leq G(x)$ for all $x \geq a$. Letting $x \to \infty$, we obtain

$$-\int_a^\infty f(t)dt \leq \int_a^\infty g(t)dt \quad \text{and} \quad \int_a^\infty f(t)dt \leq \int_a^\infty g(t)dt,$$

that is, $\left|\int_a^\infty f(t)dt\right| \leq \int_a^\infty g(t)dt$, as desired.   □

Examples 9.43. The assertions in (i) and (ii) below follow from the Comparison Test for improper integrals (Proposition 9.42) in exactly the same manner as the assertions in Examples 9.13 followed from Proposition 9.9.

(i) $\displaystyle\int_0^\infty \frac{2^t + t}{3^t + t}\, dt$ is convergent.

(ii) $\displaystyle\int_1^\infty \frac{1}{(1 + t^2 + t^4)^{1/3}}\, dt$ is convergent.   ◇

Proposition 9.44. *Let $a \in \mathbb{R}$ and $f, g : [a, \infty) \to \mathbb{R}$ be such that both f and g are integrable on $[a, x]$ for every $x \geq a$ and $g(t) \neq 0$ for all large t. Assume that $f(t)/g(t) \to \ell$ as $t \to \infty$, where $\ell \in \mathbb{R}$ or $\ell = \infty$ or $\ell = -\infty$.*

(i) *If $g(t) > 0$ for all large t, $\int_a^\infty g(t)dt$ is convergent, and $\ell \in \mathbb{R}$, then $\int_a^\infty f(t)dt$ is absolutely convergent.*

(ii) *If $f(t) > 0$ for all large t, $\int_a^\infty f(t)dt$ is convergent, and $\ell \neq 0$, then $\int_a^\infty g(t)dt$ is absolutely convergent.*

Proof. The desired results can be deduced from Proposition 9.42 in a similar way as the results in Proposition 9.11 were deduced from Proposition 9.9.   □

Corollary 9.45 (Limit Comparison Test for Improper Integrals). *Let $a \in \mathbb{R}$ and $f, g : [a, \infty) \to \mathbb{R}$ be such that both f and g are integrable on $[a, x]$ for every $x \geq a$ with $f(t) > 0$ and $g(t) > 0$ for all large t. Assume that*

$$\lim_{t \to \infty} \frac{f(t)}{g(t)} = \ell, \quad \text{where} \quad \ell \in \mathbb{R} \text{ with } \ell \neq 0.$$

Then

$$\int_a^\infty f(t)dt \text{ is convergent} \iff \int_a^\infty g(t)dt \text{ is convergent.}$$

Proof. The implication '$\Longrightarrow$' follows from part (ii) of Proposition 9.44, while '$\Longleftarrow$' follows from part (i) of Proposition 9.44.   □

Examples 9.46. The assertions in (i) and (ii) below follow from Corollary 9.45 in the same manner as the assertions in Example 9.13 followed from Corollary 9.12.

(i) $\displaystyle\int_0^\infty \frac{2^t + t}{3^t - t}\, dt$ is convergent.

(ii) $\displaystyle\int_1^\infty \sin\frac{1}{t}\, dt$ is divergent.   ◇

Sometimes, it is better to apply the stronger version of the Limit Comparison Test for improper integrals given in Proposition 9.44.

Examples 9.47. The assertions in (i) and (ii) below follow from Proposition 9.44 in the same manner as the assertions in Example 9.14 followed from Proposition 9.11. In (iii) below we give an additional example.

(i) $\displaystyle\int_1^\infty \frac{\ln t}{t^p}\, dt$ is convergent if $p > 1$ and it is divergent if $p \leq 1$.

(ii) $\displaystyle\int_2^\infty \frac{1}{(\ln t)^p}\, dt$ is divergent if $p > 0$.

(iii) Let $q \in \mathbb{R}$ and $f : [1, \infty) \to \mathbb{R}$ be given by $f(t) := e^{-t}t^q$. Then $\int_1^\infty f(t)dt$ is convergent. To see this, choose $k \in \mathbb{N}$ such that $k > q + 1$, and define $g : [1, \infty) \to \mathbb{R}$ by $g(t) := t^{q-k}$. Then $g(t) \neq 0$ for all $t \in [1, \infty)$ and

$$\frac{f(t)}{g(t)} = \frac{t^k}{e^t} \to 0 \ \text{ as } \ t \to \infty.$$

Since $k - q > 1$, we see that $\int_1^\infty g(t)dt$ is convergent. Hence $\int_1^\infty f(t)dt$ is convergent by part (i) of Corollary 9.44.   ◇

The following result is an analogue of (Cauchy's) Root Test for infinite series (Proposition 9.15).

Proposition 9.48 (Root Test for Improper Integrals). *Let $a \in \mathbb{R}$ and $f : [a, \infty) \to \mathbb{R}$ be a function that is integrable on $[a, x]$ for every $x \geq a$.*

(i) *If there is $\alpha \in \mathbb{R}$ with $\alpha < 1$ such that $|f(t)|^{1/t} \leq \alpha$ for all large t, then $\int_a^\infty f(t)dt$ is absolutely convergent.*

(ii) *If there is $\delta \in \mathbb{R}$ with $\delta > 0$ such that $f(t) \geq \delta$ for all large t, then $\int_a^\infty f(t)dt$ diverges to ∞.*

In particular, if

$$|f(t)|^{1/t} \to \ell \ \text{ as } \ t \to \infty, \qquad \text{where } \ell \in \mathbb{R} \text{ or } \ell = \infty,$$

then $\int_a^\infty f(t)dt$ is absolutely convergent when $\ell < 1$, and it diverges to ∞ when f is nonnegative and $\ell > 1$.

Proof. The first part follows by letting $g(t) := \alpha^t$ for $t \in [a, \infty)$ and using the Comparison Test (Proposition 9.42). For the second part, let $\delta > 0$ and $t_1 \in [a, \infty)$ be such that $f(t) \geq \delta$ for all $t \geq t_1$. Then

$$\int_a^x f(t)dt \geq \int_a^{t_1} f(t)dt + \delta(x - t_1) \quad \text{ for all } x \in [t_1, \infty).$$

Hence $\int_a^x f(t)dt \to \infty$ as $x \to \infty$.

The last assertion can be proved using (i) and (ii) above and arguing as in the proof of the Root Test for infinite series (Proposition 9.15).   □

Examples 9.49. (i) Let $f(t) := t^2/2^t$ for $t \in [1, \infty)$. Then $|f(t)|^{1/t} = (t^{1/t})^2/2 \to 1/2$ as $t \to \infty$. Hence $\int_1^\infty f(t)dt$ is (absolutely) convergent. On the other hand, if $g(t) := 2^t/t^2$ for $t \in [1, \infty)$, then g is nonnegative and $|g(t)|^{1/t} \to 2$ as $t \to \infty$, and so $\int_1^\infty g(t)dt$ diverges to ∞.

(ii) If $f : [a, \infty) \to \mathbb{R}$ and $|f(t)|^{1/t} \to 1$ as $t \to \infty$, then $\int_a^\infty f(t)dt$ may be convergent or divergent. For example, if $f(t) := 1/t$ and $g(t) := 1/t^2$ for $t \in [1, \infty)$, then as we have seen in Remark 7.12, $|f(t)|^{1/t} \to 1$ and $|g(t)|^{1/t} = \left(|f(t)|^{1/t}\right)^2 \to 1$ as $t \to \infty$. However, $\int_a^\infty f(t)dt$ is divergent, whereas $\int_a^\infty g(t)dt$ is convergent.   ◇

Remark 9.50. The Ratio Test for infinite series does not have a meaningful analogue for improper integrals.   ◇

Tests for Conditional Convergence of Improper Integrals

We shall now consider some tests that give conditional convergence of an improper integral. They are based on the following formula for Integration by Parts (Proposition 6.25), which can be considered as an analogue of the Partial Summation Formula (Proposition 9.19).

Let $f, g : [a, b] \to \mathbb{R}$ be such that f is differentiable and g is continuous. If f' is integrable and $G(x) := \int_a^x g(t)dt$ for $x \in [a, b]$, then

$$\int_a^b f(t)g(t)dt = f(b)G(b) - \int_a^b f'(t)G(t)dt.$$

Proposition 9.51 (Dirichlet's Test for Improper Integrals). *Let $a \in \mathbb{R}$ and $f, g : [a, \infty) \to \mathbb{R}$ be such that f is monotonic, $f(x) \to 0$ as $x \to \infty$, f is differentiable, f' is integrable on $[a, x]$ for every $x \geq a$, g is continuous, and the function $G : [a, \infty) \to \mathbb{R}$ defined by $G(x) := \int_a^x g(t)dt$ is bounded. Then the improper integral $\int_a^\infty f(t)g(t)dt$ is convergent.*

Proof. Since G is bounded, there is $\beta > 0$ such that $|G(x)| \leq \beta$ for all $x \geq a$. Also, since $f(x) \to 0$ as $x \to \infty$, we obtain $f(x)G(x) \to 0$ as $x \to \infty$. Further, since f is monotonic, for each $x \geq a$ we have

$$\int_a^x |f'(t)G(t)|dt \leq \beta \int_a^x |f'(t)|dt = \beta \left| \int_a^x f'(t)dt \right| = \beta |f(x) - f(a)|.$$

Now since f is monotonic and $f(x) \to 0$ as $x \to \infty$, we see that f is bounded. Hence $\int_a^\infty f'(t)G(t)dt$ is absolutely convergent, and so it is convergent by Proposition 9.37. Using the formula for Integration by Parts quoted above, we obtain

$$\int_a^x f(t)g(t)dt = f(x)G(x) - \int_a^x f'(t)G(t)dt \rightarrow -\int_a^\infty f'(t)G(t)dt \quad \text{as } x \rightarrow \infty.$$

Thus $\int_a^\infty f(t)g(t)dt$ is convergent. □

A similar result, known as **Abel's Test for Improper Integrals**, is given in Exercise 34. Dedekind's generalizations of the tests of Dirichlet and Abel are given in Exercise 36. While the Leibniz Test (Corollary 9.21) has no straightforward analogue for improper integrals, Dirichlet's Test for trigonometric series (Corollary 9.22) admits the following analogue for what could be called Fourier sine integrals and Fourier cosine integrals.

Corollary 9.52 (Convergence Test for Fourier Integrals). *Let $a \in \mathbb{R}$ and $f : [a, \infty) \rightarrow \mathbb{R}$ be a monotonic and differentiable function such that $f(x) \rightarrow 0$ as $x \rightarrow \infty$ and f' is integrable on $[a, x]$ for every $x \geq a$. Then*

(i) *$\int_a^\infty f(t) \sin \theta t \, dt$ is convergent for each $\theta \in \mathbb{R}$.*
(ii) *$\int_a^\infty f(t) \cos \theta t \, dt$ is convergent for each $\theta \in \mathbb{R}$ with $\theta \neq 0$.*

Proof. (i) Let $\theta \in \mathbb{R}$. Define $g : [a, \infty) \rightarrow \mathbb{R}$ by $g(t) := \sin \theta t$. Then g is a continuous function. For $x \in [a, \infty)$, let $G(x) := \int_a^x g(t)dt$. If $\theta = 0$, then $g = 0$ and so $G = 0$. If $\theta \neq 0$, then by part (i) of the FTC (Proposition 6.21), we have

$$|G(x)| = \frac{|\cos \theta a - \cos \theta x|}{|\theta|} \leq \frac{2}{|\theta|} \quad \text{for all } x \geq a.$$

Thus, in any event, the function G is bounded. Hence the desired result follows from Proposition 9.51.

(ii) Let $\theta \in \mathbb{R}$ with $\theta \neq 0$. By part (i) of the FTC (Proposition 6.21), we have

$$\left| \int_a^x \cos \theta t \, dt \right| = \frac{|\sin \theta x - \sin \theta a|}{|\theta|} \leq \frac{2}{|\theta|} \quad \text{for all } x \geq a.$$

Hence the desired result follows as in (i) above. □

Example 9.53. Let $p \in (0, 1]$ and $\theta \in \mathbb{R}$. Then the improper integral

$$\int_1^\infty \frac{\sin \theta t}{t^p} dt$$

is convergent. This follows by applying Corollary 9.52 to $f : [1, \infty) \rightarrow \mathbb{R}$ defined by $f(t) := 1/t^p$. Similarly, if $\theta \in \mathbb{R}$ and $\theta \neq 0$, then the improper integral

$$\int_1^\infty \frac{\cos \theta t}{t^p} dt$$

is convergent. On the other hand, if $\theta = 0$, then the improper integral

$$\int_1^\infty \frac{\cos 0}{t^p} dt = \int_1^\infty \frac{1}{t^p} dt$$

is divergent. ◇

9.6 Related Integrals

In the previous two sections we have considered convergence of an improper integral $\int_a^\infty f(t)dt$, where $a \in \mathbb{R}$ and $f : [a, \infty) \to \mathbb{R}$ is integrable on $[a, x]$ for all $x \geq a$. We shall now show that this treatment can be used to discuss the convergence of other types of 'improper integrals'.

Suppose $b \in \mathbb{R}$ and $f : (-\infty, b] \to \mathbb{R}$ is integrable on $[x, b]$ for every $x \leq b$. Define $\tilde{f} : [-b, \infty) \to \mathbb{R}$ by $\tilde{f}(u) := f(-u)$. Then for every $x \leq b$, that is, for every $-x \geq -b$, we have

$$\int_x^b f(t)dt = \int_{-b}^{-x} \tilde{f}(u)du.$$

We say that $\int_{-\infty}^b f(t)dt$ is **convergent** if the improper integral $\int_{-b}^\infty \tilde{f}(u)du$ is convergent, that is, if the limit

$$\lim_{y \to \infty} \int_{-b}^y \tilde{f}(u)du = \lim_{x \to -\infty} \int_x^b f(t)dt$$

exists. In this case, this limit will be denoted by $\int_{-\infty}^b f(t)dt$ itself. Otherwise, we say that $\int_{-\infty}^b f(t)dt$ is **divergent**.

Next, let $f : \mathbb{R} \to \mathbb{R}$ be a function that is integrable on $[a, b]$ for all $a, b \in \mathbb{R}$ with $a \leq b$. We say that $\int_{-\infty}^\infty f(t)dt$ is **convergent** if both $\int_0^\infty f(t)dt$ and $\int_{-\infty}^0 f(t)dt$ are convergent, that is, if the limits

$$\lim_{x \to \infty} \int_0^x f(t)dt \quad \text{and} \quad \lim_{x \to -\infty} \int_x^0 f(t)dt$$

both exist. In this case, the sum of these two limits is denoted by $\int_{-\infty}^\infty f(t)dt$ itself. If any one of these limits does not exist, we say that $\int_{-\infty}^\infty f(t)dt$ is **divergent**.

If the limit

$$\lim_{x \to \infty} \int_{-x}^x f(t)dt$$

exists, then this limit is called the **Cauchy principal value** of the integral of f on $\mathbb{R}$. If $\int_{-\infty}^\infty f(t)dt$ is convergent, then since

$$\int_{-x}^x f(t)dt = \int_{-x}^0 f(t)dt + \int_0^x f(t)dt \quad \text{for all } x \geq 0,$$

the Cauchy principal value of the integral of f on $\mathbb{R}$ exists and is equal to $\int_{-\infty}^\infty f(t)dt$. But the Cauchy principal value of the integral of f on $\mathbb{R}$ may exist even when $\int_{-\infty}^\infty f(t)dt$ is divergent. For example, consider $f : \mathbb{R} \to \mathbb{R}$ defined by $f(t) := \sin t$. For every $x \geq 0$, we have

$$\int_0^x \sin t\, dt = 1 - \cos x = -\int_{-x}^0 \sin t\, dt \quad \text{and so} \quad \int_{-x}^x \sin t\, dt = 0.$$

Hence $\lim_{x\to\infty} \int_{-x}^x f(t)dt = 0$, but neither of the two limits $\lim_{x\to\infty} \int_0^x f(t)dt$ and $\lim_{x\to\infty} \int_{-x}^0 f(t)dt$ exists.

However, if $f : \mathbb{R} \to \mathbb{R}$ is a nonnegative function and the Cauchy principal value of the integral of f on $\mathbb{R}$ exists, then $\int_{-\infty}^\infty f(t)dt$ is convergent. This can be seen as follows. For $x \geq 0$, let $F_1(x) := \int_0^x f(t)dt$, $F_2(x) := \int_{-x}^0 f(t)dt$, and $\ell := \lim_{x\to\infty} \int_{-x}^x f(t)dt$. Then F_1 and F_2 are monotonically increasing functions with $F_1(x) \leq \ell$ and $F_2(x) \leq \ell$ for all $x \geq 0$. Hence by part (i) of Proposition 3.35 with $b = \infty$, both $\lim_{x\to\infty} F_1(x)$ and $\lim_{x\to\infty} F_2(x)$ exist, that is, $\int_{-\infty}^\infty f(t)dt$ is convergent.

Integrals of the type $\int_a^\infty f(t)dt$, $\int_{-\infty}^b f(t)dt$, and $\int_{-\infty}^\infty f(t)dt$ are sometimes known as **improper integrals of the first kind**, in contrast to those of the second kind, which we now describe.

Improper Integrals of the Second Kind

Let $a, b \in \mathbb{R}$ with $a < b$. An **improper integral of the second kind** on $(a, b]$ is an ordered pair (f, F) of functions $f : (a, b] \to \mathbb{R}$ and $F : (a, b] \to \mathbb{R}$ such that f is unbounded on $(a, b]$ but integrable on $[x, b]$ for each $x \in (a, b]$ and

$$F(x) = \int_x^b f(t)dt \quad \text{for } x \in (a, b].$$

For simplicity and brevity, we use the suggestive notation $\int_{a+}^b f(t)dt$ for the improper integral (f, F) on $(a, b]$. We say that $\int_{a+}^b f(t)dt$ is **convergent** if the right (hand) limit

$$\lim_{x\to a+} F(x) = \lim_{x\to a+} \int_x^b f(t)dt$$

exists. In this case, the right (hand) limit is denoted by the same symbol $\int_{a+}^b f(t)dt$. If $\int_{a+}^b f(t)dt$ is not convergent, then it is said to be **divergent**.

The study of improper integrals of the second kind can be reduced to the theory discussed in Sections 9.4 and 9.5 as follows. Given an improper integral of the second kind $\int_{a+}^b f(t)dt$, define $c \in \mathbb{R}$ and $\tilde{f} : [c, \infty) \to \mathbb{R}$ by

$$c := \frac{1}{b-a} \quad \text{and} \quad \tilde{f}(u) := \frac{1}{u^2} f\left(a + \frac{1}{u}\right) \quad \text{for } u \in [c, \infty).$$

Then for every $x \in (a, b]$, we have

$$\int_x^b f(t)dt = \int_c^v \tilde{f}(u)du, \quad \text{where} \quad v := \frac{1}{x-a}.$$

Moreover, $x \to a^+$ if and only if $v \to \infty$. Consequently,

$$\int_{a+}^{b} f(t)dt \text{ is convergent} \iff \int_{c}^{\infty} \tilde{f}(u)du \text{ is convergent,}$$

and in this case $\int_{a+}^{b} f(t)dt = \int_{c}^{\infty} \tilde{f}(u)du$.

A variant of the above is an improper integral $\int_{a}^{b^-} f(t)dt$ of an unbounded function $f : [a, b) \to \mathbb{R}$ that is integrable on $[a, y]$ for every $y \in [a, b)$. The convergence of such integrals is defined analogously. Moreover, if we define $c \in \mathbb{R}$ and $\tilde{f} : [c, \infty) \to \mathbb{R}$ by

$$c := \frac{1}{b-a} \quad \text{and} \quad \tilde{f}(u) := \frac{1}{u^2} f\left(b - \frac{1}{u}\right) \quad \text{for } u \in [c, \infty),$$

then for every $y \in [a, b)$, we have

$$\int_{a}^{y} f(t)dt = \int_{c}^{v} \tilde{f}(u)du, \quad \text{where} \quad v := \frac{1}{b-y}.$$

Moreover, $y \to b^-$ if and only if $v \to \infty$. Consequently,

$$\int_{a}^{b^-} f(t)dt \text{ is convergent} \iff \int_{c}^{\infty} \tilde{f}(u)du \text{ is convergent,}$$

and in this case $\int_{a}^{b^-} f(t)dt = \int_{c}^{\infty} \tilde{f}(u)du$.

Finally, consider an unbounded function $f : (a, b) \to \mathbb{R}$ that is integrable on $[x, y]$ for all $x, y \in (a, b)$ with $x \le y$. Let $c := (a + b)/2$. We say that $\int_{a+}^{b^-} f(t)dt$ is **convergent** if both $\int_{a+}^{c} f(t)dt$ and $\int_{c}^{b^-} f(t)$ are convergent, that is, if

$$\lim_{x \to a^+} \int_{x}^{c} f(t)dt \quad \text{and} \quad \lim_{x \to b^-} \int_{c}^{x} f(t)dt$$

both exist. In this case, the sum of these two limits is denoted by $\int_{a+}^{b^-} f(t)dt$ itself. If any one of these limits does not exist, we say that $\int_{a+}^{b^-} f(t)dt$ is **divergent**.

If the right (hand) limit

$$\lim_{\epsilon \to 0^+} \int_{a+\epsilon}^{b-\epsilon} f(t)dt$$

exists, then it is called the **Cauchy principal value** of the integral of f on (a, b). If $\int_{a+}^{b^-} f(t)$ is convergent, then for any $\epsilon > 0$ with $a + \epsilon \le c \le b - \epsilon$, we have

$$\int_{a+\epsilon}^{c} f(t)dt + \int_{c}^{b-\epsilon} f(t)dt = \int_{a+\epsilon}^{b-\epsilon} f(t)dt.$$

Hence the Cauchy principal value of the integral of f on (a, b) is equal to $\int_{a+}^{b-} f(t)dt$. But the Cauchy principal value of f on (a, b) may exist even when $\int_{a+}^{b-} f(t)dt$ is divergent. For example, let $f(t) := t/(t^2 - 1)$ for $t \in (-1, 1)$. Then

$$f(t) = \frac{1}{2}\left[\frac{1}{1+t} - \frac{1}{1-t}\right] \quad \text{for } t \in (-1, 1).$$

For each $\epsilon \in \mathbb{R}$ satisfying $0 < \epsilon < 1$, we have

$$\int_0^{1-\epsilon} f(t)dt = \frac{1}{2}[\ln(2-\epsilon) + \ln \epsilon] \quad \text{and} \quad \int_{-1+\epsilon}^0 f(t)dt = -\frac{1}{2}[\ln \epsilon + \ln(2-\epsilon)].$$

Thus we see that $\lim_{\epsilon \to 0+} \int_{-1+\epsilon}^{1-\epsilon} f(t)dt = 0$, but neither of the two limits $\lim_{\epsilon \to 0+} \int_0^{1-\epsilon} f(t)dt$ and $\lim_{\epsilon \to 0+} \int_{-1+\epsilon}^0 f(t)dt$ exists. However, it can be shown that if $f : (a, b) \to \mathbb{R}$ is a nonnegative function and the Cauchy principal value of the integral of f on (a, b) exists, then $\int_{a+}^{b-} f(t)dt$ is convergent. The proof is similar to the proof given earlier for the Cauchy principal value of the integral of a nonnegative function on $\mathbb{R}$.

Definitions of absolute and conditional convergence as well as tests for the convergence of $\int_{-\infty}^b f(t)dt$, $\int_{a+}^b f(t)dt$, and $\int_a^{b-} f(t)dt$ can be obtained by reducing these to improper integrals considered in Section 9.4 and 9.5. Alternatively, such tests can be developed independently along similar lines. To illustrate these two procedures, let us consider the comparison test for $\int_{a+}^b f(t)dt$.

Proposition 9.54 (Comparison Test for Improper Integrals of the Second Kind). *Let $a, b \in \mathbb{R}$ with $a < b$ and $f, g : (a, b] \to \mathbb{R}$ be such that both f and g are integrable on $[x, b]$ for every $x \geq a$ and $|f| \leq g$. If $\int_{a+}^b g(t)$ is convergent, then $\int_{a+}^b f(t)dt$ is convergent and*

$$\left|\int_{a+}^b f(t)dt\right| \leq \int_{a+}^b g(t)dt.$$

Proof. Let $c := 1/(b-a)$. Define $\tilde{f} : [c, \infty) \to \mathbb{R}$ and $\tilde{g} : [c, \infty) \to \mathbb{R}$ by

$$\tilde{f}(u) := \frac{1}{u^2} f\left(a + \frac{1}{u}\right) \quad \text{and} \quad \tilde{g}(u) := \frac{1}{u^2} g\left(a + \frac{1}{u}\right).$$

Now, $|\tilde{f}| \leq \tilde{g}$. Also, for every $y \geq c$, $\tilde{f}$ is integrable on $[c, y]$. Assume that $\int_{a+}^b g(t)$ is convergent, that is, the improper integral $\int_c^\infty \tilde{g}(u)du$ is convergent. Then by Proposition 9.42, it follows that $\int_c^\infty \tilde{f}(u)du$ is convergent, that is, $\int_{a+}^b f(t)dt$ is convergent. Further,

$$\left|\int_{a+}^b f(t)dt\right| = \left|\int_c^\infty \tilde{f}(u)du\right| \leq \int_c^\infty \tilde{g}(u)du = \int_{a+}^b g(t)dt,$$

as desired.

Alternatively, we can give a proof from first principles as follows. Consider $f^+, f^- : (a, b] \to \mathbb{R}$ defined by

$$f^+(t) := \frac{|f(t)| + f(t)}{2} \quad \text{and} \quad f^-(t) := \frac{|f(t)| - f(t)}{2} \quad \text{for } t \in (a, b].$$

Define $F^+, F^-, G : (a, b] \to \mathbb{R}$ by

$$F^+(x) := \int_x^b f^+(t)dt, \quad F^-(x) := \int_x^b f^-(t)dt, \quad \text{and} \quad G(x) := \int_x^b g(t)dt.$$

Since the functions f^+, f^-, and g are nonnegative, the functions F^+, F^-, and G are monotonically decreasing. Assume that $\int_{a+}^b g(t)dt$ is convergent, that is, $\lim_{x \to a+} G(x)$ exists. Then the function G is bounded above. Since $f^+ \le |f| \le g$ and $f^- \le |f| \le g$, we see that the functions F^+ and F^- are bounded above, and hence both the limits $\lim_{x \to a+} F^+(x)$ and $\lim_{x \to a+} F^-(x)$ exist. (Compare Exercise 32 (ii) of Chapter 3.) Since $f = f^+ - f^-$, we have $\int_x^b f(t)dt = F^+(x) - F^-(x)$ for all $x \in (a, b]$. Hence $\lim_{x \to a+} \int_x^b f(t)dt$ exists, that is, $\int_{a+}^b f(t)dt$ is convergent. Further, since

$$-\int_x^b f(t)dt \le G(x) \quad \text{and} \quad \int_x^b f(t)dt \le G(x) \quad \text{for all } x \in (a, b],$$

upon letting $x \to a+$, we obtain

$$-\int_{a+}^b f(t)dt \le \int_{a+}^b g(t)dt \quad \text{and} \quad \int_{a+}^b f(t)dt \le \int_{a+}^b g(t)dt,$$

that is, $\left| \int_{a+}^b f(t)dt \right| \le \int_{a+}^b g(t)dt$, as desired. $\square$

From the above result we can deduce the following analogue of Proposition 9.44 for improper integrals of the second kind.

Proposition 9.55. *Let $a, b \in \mathbb{R}$ with $a < b$ and $f, g : (a, b] \to \mathbb{R}$ be such that both f and g are integrable on $[x, b]$ for every $x \ge a$. Assume that there is $a_0 \in (a, b]$ such that $g(t) \ne 0$ for all $t \in (a, a_0]$ and that $f(t)/g(t) \to \ell$ as $t \to a^+$, where $\ell \in \mathbb{R}$ or $\ell = \infty$ or $\ell = -\infty$.*

(i) *If $g(t) > 0$ for all $t \in (a, a_0]$, $\int_{a+}^b g(t)dt$ is convergent, and $\ell \in \mathbb{R}$, then $\int_{a+}^b f(t)dt$ is absolutely convergent.*
(ii) *If $f(t) > 0$ for all $t \in (a, a_0]$, $\int_{a+}^b f(t)dt$ is convergent, and $\ell \in \mathbb{R}$, then $\int_{a+}^b g(t)dt$ is absolutely convergent.*

Proof. Both (i) and (ii) follow from Proposition 9.54 in a similar manner as the proof of Proposition 9.44 using Proposition 9.42. $\square$

In turn, Proposition 9.55 can be used to deduce a Limit Comparison Test for improper integrals of the second kind, analogous to Corollary 9.45. This time, we leave the formulation of the statement and a proof to the reader.

Examples 9.56. (i) Let $f : (-\infty, 0] \to \mathbb{R}$ be given by $f(t) := e^t$. Since

$$\int_x^0 f(t)dt = 1 - e^x \to 1 \quad \text{as } x \to -\infty,$$

we see that $\int_{-\infty}^0 f(t)dt$ is convergent.

(ii) Let $f : (-\infty, \infty) \to \mathbb{R}$ be given by $f(t) := e^{-t^2}$. For $x \geq 1$, we have

$$0 \leq \int_0^x f(t)dt = \int_0^1 e^{-t^2} dt + \int_1^x e^{-t^2} dt \leq \int_0^1 e^{-t^2} dt + \int_1^x e^{-t} dt$$

$$= \int_0^1 e^{-t^2} dt + e^{-1} - e^{-x} \leq \int_0^1 e^{-t^2} dt + e^{-1}.$$

Hence $\int_0^\infty f(t)dt$ is convergent. Also, if $\tilde{f} : [0, \infty) \to \mathbb{R}$ is defined by $\tilde{f}(u) := f(-u)$, then the improper integral

$$\int_0^\infty \tilde{f}(u)du = \int_0^\infty e^{-u^2} du$$

is convergent, that is, $\int_{-\infty}^0 f(t)dt$ is convergent. Thus $\int_{-\infty}^\infty f(t)dt$ is convergent.

(iii) Let $p \in \mathbb{R}$, $b \in (0, \infty)$, and $f(t) := 1/t^p$ for $t \in (0, b]$. If $c \in \mathbb{R}$ and $\tilde{f} : [c, \infty) \to \mathbb{R}$ is defined by

$$c := \frac{1}{b} \quad \text{and} \quad \tilde{f}(u) = \frac{1}{u^2} f\left(\frac{1}{u}\right),$$

then

$$\int_c^\infty \tilde{f}(u)du = \int_c^\infty \frac{1}{u^{2-p}} du.$$

As we have seen in Example 9.33 (iii), $\int_c^\infty u^{p-2} du$ is convergent if and only if $2 - p > 1$, that is, $p < 1$. Hence

$$\int_{0+}^b \frac{1}{t^p} dt \text{ is convergent if and only if } p < 1.$$

If $p < 0$, then the function f is bounded on $(0, b]$, and if we define $f(0) := 0$, then $f : [0, b] \to \mathbb{R}$ is in fact continuous, and therefore integrable, on $[0, b]$. Alternatively, for $x \in (0, b]$, we have

$$\int_x^1 \frac{1}{t^p} dt = \begin{cases} (1 - x^{1-p})/(1-p) & \text{if } p \neq 1, \\ -\ln x & \text{if } p = 1. \end{cases}$$

This shows that $\int_{0+}^1 (1/t^p)dt$ if convergent if and only if $p < 1$ and when $p < 1$, it converges to $1/(1-p)$.

(iv) Let $f(t) := \ln t$ for $t \in (0, 1]$. For $x \in (0, 1]$, we have

$$\int_x^1 f(t)dt = (t \ln t - t))\Big|_x^1 = x - 1 - x \ln x.$$

Since $x \ln x \to 0$ as $x \to 0^+$, we see that $\int_{0+}^1 \ln t \, dt$ is convergent and is equal to -1. Alternatively, define $g(t) := 1/\sqrt{t}$ for $t \in (0, 1]$. Then by (iii) above, $\int_{0+}^1 g(t)dt$ is convergent and by L'Hôpital's Rule for $\frac{\infty}{\infty}$ indeterminate forms (Proposition 4.40), we have

$$\lim_{t \to 0^+} \frac{f(t)}{g(t)} = \lim_{t \to 0^+} \frac{\ln t}{1/\sqrt{t}} = \lim_{t \to 0^+} \frac{1/t}{-1/2t^{3/2}} = \lim_{t \to 0^+} -2\sqrt{t} = 0.$$

By part (i) of Proposition 9.55, we see that $\int_{0+}^1 \ln t \, dt$ is convergent.   ◇

We can also consider a combination of improper integrals of the first kind and the second kind, that is, integrals of unbounded functions on unbounded intervals. The notion of convergence can be readily defined as follows using the theory we have developed earlier. Let $a \in \mathbb{R}$ and $f : (a, \infty) \to \mathbb{R}$ be an unbounded function that is integrable (and in particular bounded) on $[x, y]$ for all $x, y \in \mathbb{R}$ such that $a < x < y$. We say that $\int_{a+}^\infty f(t)dt$ is **convergent** if both $\int_{a+}^{a+1} f(t)dt$ and $\int_{a+1}^\infty f(t)dt$ are convergent. In this case, the sum $\int_{a+}^{a+1} f(t)dt +$ $\int_{a+1}^\infty f(t)dt$ is denoted by the same symbol $\int_{a+}^\infty f(t)dt$. Similarly, if $b \in \mathbb{R}$ and $f : (-\infty, b) \to \mathbb{R}$ is an unbounded function such that f is integrable on $[x, y]$ for all $x, y \in \mathbb{R}$ with $x < y < b$, then we say that $\int_{-\infty}^{b-} f(t)dt$ is **convergent** if both $\int_{-\infty}^{b-1} f(t)dt$ and $\int_{b-1}^{b-} f(t)dt$ are convergent. In this case, the sum $\int_{-\infty}^{b-1} f(t)dt + \int_{b-1}^{b-} f(t)dt$ is denoted by the same symbol $\int_{-\infty}^{b-} f(t)dt$. It is clear that the study of such integrals easily reduces to that of improper integrals of the first kind and of the second kind.

Examples 9.57. (i) Let $f(t) := 1/\sqrt{t}(t + 1)$ for $t \in (0, \infty)$. Define $g(t) :=$ $1/\sqrt{t}$ for $t \in (0, 1]$ and $h(t) := 1/t\sqrt{t}$ for $t \in [1, \infty)$. Since

$$\lim_{t \to 0^+} \frac{f(t)}{g(t)} = \lim_{t \to 0^+} \frac{1}{t + 1} = 1$$

and $\int_{0+}^1 g(t)dt$ is convergent, part (i) of Proposition 9.55 shows that $\int_{0+}^1 f(t)dt$ is convergent. Also, since

$$\lim_{t \to \infty} \frac{f(t)}{h(t)} = \lim_{t \to \infty} \frac{t}{t + 1} = 1$$

and $\int_1^\infty h(t)dt$ is convergent, the Limit Comparison Test (Corollary 9.45) shows that $\int_1^\infty f(t)dt$ is convergent. It follows that $\int_{0+}^\infty f(t)dt$ is convergent.

(ii) Let $f_1(t) := 1/t^2$ and $f_2(t) = 1/\sqrt{t}$ for $t \in (0, \infty)$. Since $\int_{0+}^{1} f_1(t)dt$ and $\int_{1}^{\infty} f_2(t)dt$ are divergent, it follows that both $\int_{0+}^{\infty} f_1(t)dt$ and $\int_{0+}^{\infty} f_2(t)dt$ are divergent. $\diamond$

The Beta and Gamma Functions

We shall now consider two general examples of improper integrals, which lead to certain important functions in analysis.

To begin with, let p, q be any real numbers and $f : (0, 1) \to \mathbb{R}$ be the function defined by $f(t) := t^{p-1}(1-t)^{q-1}$. Let us consider the improper integrals

$$\int_{0+}^{1/2} f(t)dt \quad \text{and} \quad \int_{1/2}^{1^-} f(t)dt.$$

If $p \geq 1$, then f is bounded on $\left(0, \frac{1}{2}\right]$ and if we define $f(0) := 0$, then it is continuous and, therefore, integrable on $\left(0, \frac{1}{2}\right]$. Suppose $p < 1$ and let $g(t) := 1/t^{1-p}$ for $t \in \left(0, \frac{1}{2}\right]$. Then

$$\lim_{t \to 0+} \frac{f(t)}{g(t)} = \lim_{t \to 0+} (1-t)^{q-1} = 1.$$

By Example 9.56 (i), $\int_{0+}^{1/2} g(t)dt$ is convergent if and only if $1 - p < 1$, that is, $p > 0$. So by part (i) of Proposition 9.55, the improper integral $\int_{0+}^{1/2} f(t)dt$ is convergent if and only if $p > 0$.

Next, suppose $q \geq 1$. Then f is bounded on $\left[\frac{1}{2}, 1\right)$ and if we define $f(1) := 0$, then f is continuous and therefore, integrable on $\left[\frac{1}{2}, 1\right]$. On the other hand, suppose $q < 1$. For $x \in \left[\frac{1}{2}, 1\right)$, if we let $y := 1 - x$, then $y \in \left(0, \frac{1}{2}\right]$ and we have

$$\int_{1/2}^{x} f(t)dt = \int_{y}^{1/2} u^{q-1}(1-u)^{p-1}du.$$

Hence using the result in the previous paragraph, we see that the improper integral $\int_{1/2}^{1^-} f(t)dt$ is convergent if and only if $q > 0$.

Thus, if $p \geq 1$ and $q \geq 1$ and if we set $f(0) := 0$ and $f(1) = 0$, then f is integrable on $[0, 1]$, and in all other cases, the improper integral

$$\int_{0+}^{1^-} f(t)dt$$

is convergent if and only if $p > 0$ and $q > 0$. With this in view, we obtain a well-defined function $\beta : (0, \infty) \times (0, \infty) \to \mathbb{R}$ given by

$$\beta(p, q) := \int_{0+}^{1^-} t^{p-1}(1-t)^{q-1}dt \quad \text{for } p > 0 \text{ and } q > 0.$$

This is known as the **beta function**.

We shall now proceed to define another important function, which is a neat generalization of the factorial function in the sense that it extends the real-valued function on $\mathbb{N}$ given by $n \mapsto (n-1)!$ to the set $(0, \infty)$ of all positive real numbers. To motivate its definition, let us consider the improper integral

$$I_n := \int_0^\infty e^{-t}t^{n-1}dt, \quad \text{where } n \in \mathbb{N}.$$

First we show that $I_n = (n-1)!$ for all $n \in \mathbb{N}$. For $n = 1$, we have

$$\int_0^x e^{-t}dt = 1 - e^{-x} \to 1 \quad \text{as } x \to \infty.$$

Thus $I_1 = 1$. Now assuming that $I_n = (n-1)!$ for $n \geq 1$, we deduce $I_{n+1} = n!$. Let $x \geq 0$. Integration by Parts gives us

$$\int_0^x e^{-t}t^n dt = -e^{-x}x^n + n \int_0^x e^{-t}t^{n-1}dt.$$

We have seen in Example 7.4 (ii) that $e^{-x}x^n \to 0$ as $x \to \infty$, and by our assumption

$$\lim_{x \to \infty} \int_0^x e^{-t}t^{n-1}dt = I_n = (n-1)!.$$

Hence

$$\lim_{x \to \infty} \int_0^x e^{-t}t^n dt = 0 + n \cdot (n-1)! = n!,$$

that is, $I_{n+1} = n!$, as desired.

Thus we see that $I_1 = 1$ and $I_{n+1} = nI_n$ for all $n \in \mathbb{N}$. In an attempt to generalize this relation, let us fix $s \in \mathbb{R}$ and consider $f : (0, \infty) \to \mathbb{R}$ given by $f(t) := e^{-t}t^{s-1}$. Let

$$J_1 := \int_{0+}^1 f(t)dt \quad \text{and} \quad J_2 := \int_1^\infty f(t)dt.$$

If $s \geq 1$, then f is a bounded function on $(0, 1]$ and if we define $f(0) := 1$, then it is continuous, and therefore integrable, on $[0, 1]$. Now suppose $s < 1$. Define $g : (0, 1] \to \mathbb{R}$ by $g(t) := 1/t^{1-s}$. Then

$$\lim_{t \to 0+} \frac{f(t)}{g(t)} = \lim_{t \to 0+} e^{-t} = 1.$$

By Example 9.56 (iii), $\int_{0+}^1 g(t)dt$ is convergent if and only if $1 - s < 1$, that is, $s > 0$. Hence by Proposition 9.55, J_1 is convergent if and only if $s > 0$.

To consider the convergence of J_2, choose $n \in \mathbb{N}$ such that $n > s$ and define $g : [1, \infty) \to \mathbb{R}$ by $g(t) := 1/t^{1-s+n}$. Then

$$\lim_{t \to \infty} \frac{f(t)}{g(t)} = \lim_{t \to \infty} e^{-t} t^n = 0.$$

Since $1 - s + n > 1$, Example 9.33 (iii) shows that $\int_1^\infty g(t)dt$ is convergent. By part (i) of Proposition 9.44, it follows that J_2 is convergent for every $s \in \mathbb{R}$.

Thus $\int_{0+}^\infty f(t)dt$ is convergent if and only if $s > 0$. With this in view, we obtain a well-defined function $\Gamma : (0, \infty) \to \mathbb{R}$ given by

$$\Gamma(s) := \int_{0+}^\infty e^{-t} t^{s-1} dt \quad \text{for } s > 0.$$

This is known as the **gamma function**.

Proposition 9.58 (Properties of Gamma Function).

(i) $\Gamma(s) > 0$ for all $s > 0$ and $\Gamma(s) \to \infty$ as $s \to 0^+$.
(ii) $\Gamma(s+1) = s\,\Gamma(s)$ for all $s > 0$.
(iii) $\Gamma(n) = (n-1)!$ for all $n \in \mathbb{N}$.

Proof. (i) For $s > 0$, we have

$$\Gamma(s) \geq \int_{0+}^1 e^{-t} t^{s-1} dt \geq e^{-1} \int_{0+}^1 t^{s-1} dt = \frac{e^{-1}}{s}.$$

This shows that $\Gamma(s) > 0$ for all $s > 0$ and $\Gamma(s) \to \infty$ as $s \to 0^+$.

(ii) Fix any $s > 0$. For $\epsilon > 0$ and $x \geq \epsilon$, we have

$$\int_\epsilon^x e^{-t} t^s dt = -e^{-t} t^s \Big|_\epsilon^x + s \int_\epsilon^x e^{-t} t^{s-1} dt$$

$$= (e^{-\epsilon} \epsilon^s - e^{-x} x^s) + s \int_\epsilon^x e^{-t} t^{s-1} dt.$$

Now, $e^{-\epsilon} \to 1$ as $\epsilon \to 0^+$, and by part (iii) of Proposition 7.9, $\epsilon^s \to 0$ as $\epsilon \to 0^+$. Thus $e^{-\epsilon} \epsilon^s \to 0$ as $\epsilon \to 0^+$. Also, if we choose $n \in \mathbb{N}$ such that $n > s$, then for all $x \geq 1$, we have $0 \leq e^{-x} x^s \leq e^{-x} x^n$. From Example 7.4 (ii) we know that $e^{-x} x^n \to 0$ as $x \to \infty$. Thus $e^{-x} x^s \to 0$ as $x \to \infty$. Consequently,

$$\int_{0+}^1 e^{-t} t^s dt = \lim_{\epsilon \to 0^+} \int_\epsilon^1 e^{-t} t^s dt = (0 - e^{-1}) + s \left(\lim_{\epsilon \to 0^+} \int_\epsilon^1 e^{-t} t^{s-1} dt \right)$$

and

$$\int_1^\infty e^{-t} t^s dt = \lim_{x \to \infty} \int_1^x e^{-t} t^s dt = (e^{-1} - 0) + s \left(\lim_{x \to \infty} \int_1^x e^{-t} t^{s-1} dt \right).$$

Combining these, we obtain $\Gamma(s+1) = s\,\Gamma(s)$, as desired.

(iii) We have already shown that $\Gamma(n) = (n-1)!$. Alternatively, it follows from (ii) above using induction on n and the identity $\Gamma(1) = \int_0^\infty e^{-t} dt = 1$, which has also been verified earlier. $\qquad \square$

Remark 9.59. While $\Gamma(s)$ is easy to calculate when s is a positive integer, determining the value at other positive real numbers s is not obvious. For example, it can be shown that $\Gamma(\frac{1}{2}) = 2\int_0^\infty e^{-u^2} du$. (See Exercise 49.) Further, using double integrals it can be shown that $\int_0^\infty e^{-u^2} du = \sqrt{\pi}/2$. Assuming these facts, we can use part (ii) of Proposition 9.58 to deduce that

$$\Gamma\left(n + \frac{1}{2}\right) = \frac{1 \cdot 3 \cdots (2n-3)(2n-1)}{2^n}\sqrt{\pi} \quad \text{for all } n \in \mathbb{N}.$$

Another crucial property of the gamma function is that it is a **log-convex** function, that is, the function $\Gamma_\ell : (0, \infty) \to \mathbb{R}$ defined by $\Gamma_\ell(s) := \ln\Gamma(s)$ is convex on $(0, \infty)$. A proof of this result is outlined in Exercise 64.

There is an interesting relationship between the beta function and the gamma function, namely,

$$\beta(p, q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} \quad \text{for all } p > 0 \text{ and } q > 0.$$

We refer the reader to (2.12) of [2] for a proof of this result. Using $\Gamma(1/2) = \sqrt{\pi}$, the above identity implies that $\beta\left(\frac{1}{2}, \frac{1}{2}\right) = \pi$. In general, if p, q are positive nonintegral rational numbers, then by a theorem of Schneider, $\beta(p, q)$ is a transcendental number. For a proof, we refer to Section 6.2 of [7]. $\diamond$

Notes and Comments

Logically, the theory of infinite series is a particular case of the theory of sequences. In fact, the two are equivalent. However, from a pedagogical and historical point of view, it seems preferable to treat infinite series separately and at a stage when tools from the theory of integration are at our disposal. Also, it appears natural to treat improper integrals alongside infinite series.

We have followed Apostol [3] to define an infinite series as a pair of sequences, the first comprising of the terms of the series and the second formed by the partial sums of the series. This might seem pedantic but it avoids 'defining' an infinite series as an expression or a symbol. Similar considerations apply to improper integrals. However, we quickly adopt the usual conventions for denoting infinite series and improper integrals.

The treatments of the infinite series in Sections 9.1 and 9.2, and of the improper integrals in Sections 9.3 and 9.5, run parallel. Our development brings home the fact that they are the discrete and the continuous representations of the same thing. For example, the partial sum $A_n := \sum_{k=1}^n a_k$ of an infinite series $\sum_{k=1}^\infty a_k$ is analogous to the 'partial' integral $F(x) := \int_a^x f(t)dt$ of the improper integral $\int_a^\infty f(t)dt$. Further, just as $A_0 = 0$ and $a_n = A_n - A_{n-1}$ for all $n \geq 1$, we have $F(a) = 0$ and $f(x) = F'(x)$, whenever $x \geq a$ and f is continuous at x. Tests of convergence for the two are based on the same

principles. However, there are a few exceptions. The 'kth Term Test' and the 'Ratio Test' for infinite series fail to have an analogue in the setting of improper integrals.

While the convergence of an infinite series can usually be determined using one of the several tests, finding the sum is often far more difficult. The only cases in which we have actually found the sum of an infinite series are those involving a geometric series, or in which a series can be written as a genuine telescopic series or the 'tail' of a Taylor series can be shown to tend to zero. In fact, essentially the only series whose partial sums have a 'closed form expression' is the geometric series. The situation for actually evaluating improper integrals is similar. But if a function $f : [a, \infty) \to \mathbb{R}$ is integrable on $[a, x]$ for all $x \geq a$ and equals the derivative of a known function g, then the 'partial integral' F of the improper integral $\int_a^\infty f(t)dt$ is given by $F(x) = g(x) - g(a)$ for $x \geq a$. As a result, to evaluate the improper integral $\int_a^\infty f(t)dt$, one only needs to find $\lim_{x \to \infty} g(x)$ (Proposition 9.34). Needless to say, this procedure can be carried through in only a limited number of cases.

The fact that an absolutely convergent infinite series of real numbers is convergent is intimately related to the Completeness Property of the real numbers via the Cauchy Criterion. In fact, the Cauchy Criterion for real numbers can be proved using only the fact that an absolutely convergent infinite series of real numbers is convergent (Exercise 54).

While we have given a number of tests for convergence of infinite series in the text and also in the exercises, the list is not meant to be comprehensive. A wealth of material, including a plethora of convergence tests, can be found in old classics on infinite series such as the books of Bromwich [12] and Knopp [42]. See also the more specialized books of Dienes [22] and Hardy [32].

Power series are an important class of infinite series whose terms depend on a parameter. Their peculiar convergence behavior is brought out in Lemma 9.25. This result allows us to introduce the concept of the radius of convergence of a power series without any reference to the terms of the series. Of course, the calculation of the radius of convergence of a given power series will be based on the Root Test or the Ratio Test, for which either the roots of the absolute values of the terms of the series or the ratios of the successive terms of the series are needed. Taylor series form a special class of power series. Their convergence can be proved by showing that the remainder after the nth term tends to zero. This process does not use the Root Test or the Ratio Test and, when successful, yields also the sum of the series. Many classical functions admit a Taylor series, which can be effectively used to understand and study these functions. Conversely, new functions can be introduced by means of convergent power series, just as we introduced the logarithmic and the arctangent function by means of integrals of rational functions. It may be interesting to note that any power series is the Taylor series of some function. See the article of Meyerson [48] for a proof. Apart from power series and Taylor series, another important class of infinite series whose terms depend on a parameter is that of Fourier series. These are series of the form

$a_0 + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx)$. *The study of Fourier series is a rich and fascinating topic in mathematics, and for more on this, we suggest the recent book of Stein and Shakarchi [60].*

In order to retain the parallelism between infinite series and improper integrals, we have restricted the definition of an improper integral to cover only the 'integrals' of the type $\int_a^{\infty} f(t)dt$, where $a \in \mathbb{R}$ and $f : [a, \infty) \to \mathbb{R}$ is a function that is integrable on $[a, x]$ for each $x \geq a$. In doing so, there is no real loss of generality since the treatment of improper integrals of other kinds can be reduced to improper integrals of the above type. As an application of improper integrals of the second kind, we have defined the beta function and the gamma function. We have restricted ourselves to discussing only the most rudimentary properties of these functions. The article of Davis [21] gives a very readable introduction to the gamma function. A lucid development of the gamma function and the beta function can be found in the book of Artin [2].

Exercises

Part A

1. Give examples to show that if $\sum_k a_k$ and $\sum_k b_k$ are convergent series of real numbers, then the series $\sum_k a_k b_k$ may not be convergent. Also show that if $\sum_k a_k = A$ and $\sum_k b_k = B$, then $\sum_k a_k b_k$ may be convergent, but its sum may not be equal to AB.

2. Consider a series $\sum_k a_k$ and for each $n = 0, 1, 2, \ldots$, let $b_{n,k} := a_{n+k}$. Show that the series $\sum_k a_k$ is convergent if and only if for some $n = 0, 2, \ldots$, the series $\sum_k b_{n,k}$ is convergent. In this case, prove that the series $\sum_k b_{n,k}$ is convergent for every $n = 0, 1, 2, \ldots$, and $\sum_k b_{n,k} = \sum_k a_k - A_n$, where A_n is the nth partial sum of the series $\sum_k a_k$.

3. Show that the series $\sum_{k=1}^{\infty} 2/(k+1)(2k+1)$ is convergent and its sum is less than or equal to 1. (Hint: Compare the given series with the series $\sum_{k=1}^{\infty} 1/k(k+1)$.)

4. Let $a \in \mathbb{R}$ with $a > 1$. Show that the series $\sum_{k=1}^{\infty} \left(1/a^{k!}\right)$ is convergent.

5. Let $p \in \mathbb{R}$. Use Example 9.1 (iii) together with Proposition 9.4 to show that $\sum_{k=1}^{\infty}(1/k^p)$ is convergent if $p > 1$ and divergent if $p \leq 1$.

6. Let $a_k \in \mathbb{R}$ with $a_k \leq 0$ for all $k \in \mathbb{N}$. Show that $\sum_{k=1}^{\infty} a_k$ is convergent if and only if the sequence (A_n) of its partial sums is bounded below, and in this case $\sum_{k=1}^{\infty} a_k = \inf\{A_n : n \in \mathbb{N}\}$. If (A_n) is not bounded below, then show that $\sum_{k=1}^{\infty} a_k$ diverges to $-\infty$.

7. **(Cauchy's Condensation Test)** Let (a_k) be a monotonically decreasing sequence of nonnegative real numbers. Show that the series $\sum_{k=1}^{\infty} a_k$ is convergent if and only if the series $\sum_{k=0}^{\infty} 2^k a_{2^k}$ is convergent. (Hint: Proposition 9.4.) Deduce the convergence and divergence of the series $\sum_{k=1}^{\infty} 1/k^p$ and $\sum_{k=2}^{\infty} 1/k(\ln k)^p$, where $p \in \mathbb{R}$. (Compare Example 9.40.)

8. **(Abel's kth Term Test)** Suppose (a_k) is a monotonically decreasing sequence of nonnegative real numbers. If the series $\sum_k a_k$ is convergent, then show that $ka_k \to 0$ as $k \to \infty$. (Hint: Exercise 7.) Also, show that the converse of this result does not hold.

9. A sequence (a_k) is said to be of **bounded variation** if $\sum_{k=1}^{\infty} |a_k - a_{k+1}|$ is convergent. Prove the following:

   (i) A sequence of bounded variation is convergent.

   (ii) Let (a_k) and (b_k) be of bounded variation and let $r \in \mathbb{R}$. Then $(a_k + b_k)$, (ra_k), and $(a_k b_k)$ are of bounded variation. If $a_k \neq 0$ for all $k \in \mathbb{N}$, is $(1/a_k)$ of bounded variation?

   (iii) Every bounded monotonically increasing sequence is of bounded variation. Further, if (b_k) and (c_k) are bounded monotonically increasing sequences and we define $a_k := b_k - c_k$ for $k \in \mathbb{N}$, then the sequence (a_k) is of bounded variation.

   (iv) If (a_k) is of bounded variation, then there are bounded monotonically increasing sequences (b_k) and (c_k) such that $a_k = b_k - c_k$ for $k \in \mathbb{N}$. (Hint: Let $a_0 := 0$ and $v_k := |a_1| + |a_1 - a_2| + \cdots + |a_{k-1} - a_k|$ for $k \in \mathbb{N}$. Define $b_k := (v_k + a_k)/2$ and $c_k := (v_k - a_k)/2$ for $k \in \mathbb{N}$.)

10. **(Ratio Comparison Test)** Let (a_k) and (b_k) be sequences and suppose $b_k > 0$ for all k. Prove the following:

    (i) If $|a_{k+1}|b_k \leq |a_k|b_{k+1}$ for all large k and $\sum_{k=1}^{\infty} b_k$ is convergent, then $\sum_{k=1}^{\infty} a_k$ is absolutely convergent.

    (ii) If $|a_{k+1}|b_k \geq |a_k|b_{k+1}$ for all large k and $\sum_{k=1}^{\infty} b_k$ is divergent, then $\sum_{k=1}^{\infty} a_k$ is not absolutely convergent.

11. Let $a, b \in \mathbb{R}$ be such that $0 < a < b$. For $k \in \mathbb{N}$, define

$$a_{2k-1} := a^{k-1}b^{k-1} \quad \text{and} \quad a_{2k} := a^k b^{k-1}.$$

   Consider the series $\sum_{k=1}^{\infty} a_k = 1 + a + ab + a^2 b + a^2 b^2 + a^3 b^2 + \cdots$.

    (i) Use the Ratio Test to show that $\sum_{k=1}^{\infty} a_k$ is convergent if $b < 1$, and it is divergent if $a \geq 1$.

    (ii) Use the Root Test to show that $\sum_{k=1}^{\infty} a_k$ is convergent if $ab < 1$, and it is divergent if $ab > 1$.

12. For $k \in \mathbb{N}$, let $a_{2k-1} := 4^{k-1}/9^{k-1}$ and $a_{2k} := 4^{k-1}/9^k$. Show that $|a_{2k}/a_{2k-1}| = \frac{1}{9}$ and $|a_{2k+1}/a_{2k}| = 4$ for all $k \in \mathbb{N}$ and so the Ratio Test for the convergence of $\sum_{k=1}^{\infty} a_k$ is inconclusive. Prove that $|a_k|^{1/k} \to \frac{2}{3}$ as $k \to \infty$ and use the Root Test to conclude that $\sum_{k=1}^{\infty} a_k$ is convergent.

13. **(Raabe's Test)** Let (a_k) be a sequence of real numbers. If there is $p > 1$ such that

$$|a_{k+1}| \leq \left(1 - \frac{p}{k}\right)|a_k| \quad \text{for all large } k,$$

   then show that $\sum_{k=1}^{\infty} a_k$ is absolutely convergent. On the other hand, if

$$|a_{k+1}| \geq \left(1 - \frac{1}{k}\right)|a_k| \quad \text{for all large } k,$$

then show that $\sum_{k=1}^{\infty} a_k$ is divergent. (Hint: If $p > 1$ and $x \in [0, 1]$, then $1 - px \le (1 - x)^p$. Use Exercise 10.)

14. (i) If $a_1 := 1$ and $a_{k+1} := (k-1)a_k/(k+1)$ for $k \ge 2$, then show that $\sum_{k=1}^{\infty} a_k$ is convergent.
   (ii) If $a_1 := 1$ and $a_{k+1} := (2k-1)a_k/2k$ for $k \in \mathbb{N}$, then show that $\sum_{k=1}^{\infty} a_k$ diverges to ∞.
   (Hint: Exercise 13.)

15. **(Hypergeometric Series)** Let α, β, γ be positive real numbers. If $a_0 := 1$ and

$$a_k := \frac{\alpha(\alpha+1)\cdots(\alpha+k-1)\beta(\beta+1)\cdots(\beta+k-1)}{\gamma(\gamma+1)\cdots(\gamma+k-1)k!} \quad \text{for } k \in \mathbb{N},$$

then show that $\sum_{k=0}^{\infty} a_k$ is convergent if and only if $\gamma > \alpha + \beta$. (Hint: Exercise 13.)

16. Suppose the partial sums of a series $\sum_{k=1}^{\infty} b_k$ are bounded. If $p > 0$ and $x \in (0, 1)$, then show that the series

$$\sum_{k=1}^{\infty} \frac{b_k}{k^p}, \quad \sum_{k=1}^{\infty} \frac{b_k}{(\ln k)^p} \quad \text{and} \quad \sum_{k=1}^{\infty} b_k x^k$$

are convergent. (Hint: Proposition 9.20.)

17. **(Abel's Test)** If (a_k) is a bounded monotonic sequence and $\sum_{k=1}^{\infty} b_k$ is a convergent series, then show that the series $\sum_{k=1}^{\infty} a_k b_k$ is convergent. (Hint: Partial Summation Formula.)

18. Let $\sum_{k=1}^{\infty} b_k$ be a convergent series. Show that the series

$$\sum_{k=1}^{\infty} k^{1/k} b_k \quad \text{and} \quad \sum_{k=1}^{\infty} \left(1 + \frac{1}{k}\right)^k b_k$$

are also convergent. (Hint: Exercise 17 and Exercises 7, 8 of Chapter 2.)

19. **(Dedekind's Tests)** Let (a_k) and (b_k) be sequences of real numbers.
   (i) If the series $\sum_{k=1}^{\infty} |a_k - a_{k+1}|$ is convergent, $a_k \to 0$ as $k \to \infty$, and the sequence of partial sums of $\sum_{k=1}^{\infty} b_k$ is bounded, then show that the series $\sum_{k=1}^{\infty} a_k b_k$ is convergent.
   (ii) If the series $\sum_{k=1}^{\infty} |a_k - a_{k+1}|$ and $\sum_{k=1}^{\infty} b_k$ are convergent, then show that the series $\sum_{k=1}^{\infty} a_k b_k$ is convergent.
   (Hint: Partial Summation Formula. Alternatively, use Exercise 9 (iv), Proposition 9.20, and Exercise 17.)

20. Let $p \in \mathbb{R}$ with $p > 1$. Show that

$$\frac{1}{(p-1)(\ln 2)^{p-1}} \le \sum_{k=2}^{\infty} \frac{1}{k(\ln k)^p} \le \frac{p-1+2\ln 2}{2(p-1)(\ln 2)^p}.$$

(Hint: Proposition 9.39 and Example 9.40 (ii).)

21. Test the series $\sum_{k=1}^{\infty} a_k$ for absolute/conditional convergence if for $k \in \mathbb{N}$, a_k is given as follows. In (vii)–(x) below, q, r, p, and θ are real numbers.

(i) $(-1)^k \dfrac{k}{3k-2}$, (ii) $\dfrac{k!}{2^k}$, (iii) ke^{-k}, (iv) $\dfrac{1}{\sqrt{1+k^3}}$,

(v) $(-1)^{k-1} \dfrac{k}{k^2+1}$, (vi) $(-1)^{k-1} \dfrac{1}{\ln(\ln k)}$, (vii) $\dfrac{k^q}{1+k^q}$,

(viii) $\dfrac{r^k}{1+r^{2k}}$, (ix) $(-1)^{k-1} \sin\left(\dfrac{1}{k^p}\right)$, (x) $\dfrac{\cos k\theta}{\sqrt{k}}$.

22. Find the radius of convergence of the power series $\sum_{k=0}^{\infty} c_k x^k$ whose coefficients are defined by $c_{2k-1} := 3^{-k}$ and $c_{2k} := 2^k 5^{-k}$ for $k \in \mathbb{N}$.

23. Find the radius of convergence of the power series $\sum_{k=0}^{\infty} c_k x^k$ if for $k \in \mathbb{N}$, the coefficient c_k is given as follows:

(i) $k!$, (ii) k^2, (iii) $\dfrac{k}{k^2+1}$, (iv) ke^{-k}, (v) c^{k^2}, where $c \in \mathbb{R}$,

(vi) $\dfrac{k^k}{k!}$, (vii) $\dfrac{2^k}{k^2}$, (viii) $\dbinom{k+m}{k}$, where $m \in \mathbb{N}$.

24. Let $f(x) := \cos x$ for $x \in \mathbb{R}$. Show that the Taylor series of f is convergent for $x \in \mathbb{R}$. Deduce that

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!} \quad \text{for } x \in \mathbb{R}.$$

25. Let $a \in \mathbb{R}$ and I be an open interval containing a. Show that if $\sum_{k=0}^{\infty} a_k (x-a)^k$ is the Taylor series of some $f : I \to \mathbb{R}$ around a, then there are infinitely many functions $g : I \to \mathbb{R}$ that have the same Taylor series around a.

26. Modify the function given in Example 9.41 to obtain a piecewise linear continuous function $g : [1, \infty) \to \mathbb{R}$ such that $g(1) = 0$ and for $k \geq 2$,

$$g(k) = \sqrt{k} \quad \text{and} \quad g\left(k - \frac{1}{k^2\sqrt{k}}\right) = 0 = g\left(k + \frac{1}{k^2\sqrt{k}}\right).$$

Show that $\int_1^{\infty} g(t)dt$ is convergent, but g is not bounded.

27. Let $a \in \mathbb{R}$ and $f : [a, \infty) \to \mathbb{R}$ be such that $f(t) \leq 0$ for all $t \geq a$ and f is integrable on $[a, x]$ for each $x \geq a$. Show that $\int_a^{\infty} f(t)$ is convergent if and only if its partial integral $F : [a, \infty) \to \mathbb{R}$ defined by $F(x) := \int_a^x f(t)dt$ is bounded below, and in this case $\int_a^{\infty} f(t) = \inf\{F(t) : t \in [a, \infty)\}$. If F is not bounded below, then show that $\int_a^{\infty} f(t)$ diverges to $-\infty$.

28. Let $f, g : [2, \infty) \to \mathbb{R}$ be defined by

$$f(t) := \begin{cases} 1 & \text{if } k \leq t < k + (1/k^2) \text{ for some } k \in \mathbb{N}, \\ 0 & \text{otherwise}, \end{cases}$$

and

$$g(t) := \begin{cases} k & \text{if } k \leq t < k + (1/k^3) \text{ for some } k \in \mathbb{N}, \\ 0 & \text{otherwise}. \end{cases}$$

Show that $\int_2^\infty f(t)dt$ and $\int_2^\infty g(t)dt$ are convergent, $f(k) = 1$ for each $k \in \mathbb{N}$ with $k \geq 2$, and $g(k) \to \infty$ as $k \to \infty$.

29. Let $a \in \mathbb{R}$ and $f : [a, \infty) \to \mathbb{R}$ be such that f is integrable on $[a, x]$ for all $x \geq a$. Prove the following:

   (i) If $\int_a^\infty f(t)dt$ is convergent and $f(x) \to \ell$ as $x \to \infty$, then $\ell = 0$.

   (ii) If f is differentiable and $\int_a^\infty f'(t)dt$ is convergent, then there is $\ell \in \mathbb{R}$ such that $f(x) \to \ell$ as $x \to \infty$. (Hint: Use part (i) of Proposition 6.21.)

   (iii) If f is differentiable and both $\int_a^\infty f(t)dt$ and $\int_a^\infty f'(t)dt$ are convergent, then $f(x) \to 0$ as $x \to \infty$.

30. Use Exercise 29 to conclude that the improper integral $\int_0^\infty t \sin t^2 \, dt$ is divergent.

31. Show that $\int_1^\infty (\cos t/t^p)dt$ and $\int_1^\infty (\sin t/t^p)dt$ are absolutely convergent if $p > 1$ and that they are conditionally convergent if $0 < p \leq 1$.

32. Let $f : [1, \infty) \to \mathbb{R}$ be such that f is integrable on $[1, x]$ for every $x \geq 1$. Prove the following using Proposition 9.42:

   (i) If there are $p > 1$ and $\ell \in \mathbb{R}$ such that $t^p f(t) \to \ell$ as $t \to \infty$, then $\int_1^\infty f(t)dt$ is absolutely convergent.

   (ii) Suppose $f(t) > 0$ for all $t \in [1, \infty)$. If there are $p \leq 1$ and $\ell \neq 0$ such that $t^p f(t) \to \ell$ as $t \to \infty$, then $\int_1^\infty f(t)dt$ is divergent.

33. Let $g : [1, \infty) \to \mathbb{R}$ be a continuous real-valued function such that the function $G : [a, \infty) \to \mathbb{R}$ defined by $G(x) := \int_a^x g(t)dt$ is bounded. If $p \in \mathbb{R}$ with $p > 0$ and $x \in (0, 1)$, then show that the improper integrals

$$\int_1^\infty \frac{g(t)}{t^p}dt, \quad \int_1^\infty \frac{g(t)}{(\ln t)^p}dt, \quad \text{and} \quad \int_1^\infty x^t g(t)dt$$

are convergent. (Hint: Proposition 9.51.)

34. **(Abel's Test for Improper Integrals)** Let $a \in \mathbb{R}$ and $f, g : [a, \infty) \to \mathbb{R}$ be such that f is bounded, monotonic, and differentiable, f' is integrable on $[a, x]$ for every $x \geq a$, g is continuous, and $\int_a^\infty g(t)dt$ is convergent. Show that $\int_a^\infty f(t)g(t)dt$ is convergent. (Hint: Use Integration by Parts.) [Note: Compare with Proposition 9.51.]

35. Let $\int_1^\infty g(t)dt$ be a convergent improper integral. Show that the improper integrals $\int_1^\infty t^{1/t} g(t)dt$ and $\int_1^\infty \left(1 + \frac{1}{t}\right)^t g(t)dt$ are also convergent. (Hint: Exercise 34, and Revision Exercise 15 given at the end of Chapter 7.)

36. **(Dedekind's Tests for Improper Integrals)** Let $a \in \mathbb{R}$ and let $f, g : [a, \infty) \to \mathbb{R}$ be any functions.

   (i) If f is differentiable, $\int_a^\infty |f'(t)|dt$ is convergent, $f(x) \to 0$ as $x \to \infty$, g is continuous, and the function $G : [a, \infty) \to \mathbb{R}$ defined by $G(x) := \int_a^x g(t)dt$ is bounded, then show that $\int_a^\infty f(t)g(t)dt$ is convergent.

   (ii) If f is differentiable, $\int_a^\infty |f'(t)|dt$ is convergent, g is continuous, and $\int_a^\infty g(t)dt$ is convergent, then show that $\int_a^\infty f(t)g(t)dt$ is convergent.

   (Hint: Use Integration by Parts.)

37. Show that the improper integrals $\int_1^\infty \sin t^2 \, dt$ and $\int_1^\infty \cos t^2 \, dt$ are convergent. (Hint: Substitute $s = t^2$ and use Corollary 9.52.)

38. Let $p \in \mathbb{R}$ with $p > 0$. Show that the improper integrals

$$\int_2^\infty \frac{\sin t}{(\ln t)^p}dt \quad \text{and} \quad \int_2^\infty \frac{\cos t}{(\ln t)^p}dt$$

are conditionally convergent. (Hint: Corollary 9.52 and Exercise 31.)

39. Let $f : [0, \infty) \to \mathbb{R}$ be such that $\int_0^\infty f(t)dt$ is absolutely convergent. Show that the improper integrals

$$\mathcal{L}(f)(u) := \int_0^\infty f(t)e^{-ut}dt, \quad \text{where } u \in \mathbb{R} \text{ and } u \geq 0,$$

$$\mathcal{F}_s(f)(u) := \frac{2}{\pi}\int_0^\infty f(t)\sin ut\, dt, \quad \text{where } u \in \mathbb{R},$$

$$\mathcal{F}_c(f)(u) := \frac{2}{\pi}\int_0^\infty f(t)\cos ut\, dt, \quad \text{where } u \in \mathbb{R},$$

are absolutely convergent.

[Note: $\mathcal{L}(f)$, $\mathcal{F}_s(f)$, and $\mathcal{F}_c(f)$ are called the **Laplace Transform**, the **Fourier Sine Transform**, and the **Fourier Cosine Transform** of f.]

40. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(t) := (1+t)/(1+t^2)$ for $t \in \mathbb{R}$. Show that $\int_{-\infty}^\infty f(t)dt$ is divergent, but the Cauchy principal value of the integral of f on $\mathbb{R}$ exists and is equal to π.

41. Let $a, b \in \mathbb{R}$ be such that $a < b$ and $f : [a, b] \to \mathbb{R}$ be an integrable function. If g denotes the restriction of f to (a, b), then show that $\int_{a+}^{b-} g(t)dt$ exists and is equal to $\int_a^b f(t)dt$.

42. Let $f : [1, \infty) \to \mathbb{R}$ be defined as follows. If $t \in [1, \infty)$ and $k \leq t < k+1$ with $k \in \mathbb{N}$, let $f(t) := (-1)^{k-1}/k$. Show that $\int_a^\infty f(t)dt$ is conditionally convergent. (Hint: $\sum_{k=1}^\infty (-1)^{k-1}/k$ is conditionally convergent.)

43. Show that the improper integral $\int_0^\infty e^{t^2}dt$ is divergent, but the improper integral $\int_0^\infty e^{-t^2}dt$ is convergent. (Hint: Comparison with $\int_0^\infty e^t dt$ and $\int_0^\infty e^{-t}dt$)

44. Let $p(t)$ and $q(t)$ be polynomials of degrees m and n respectively. Suppose $q(t) \neq 0$ for all $t \geq a$ and let $f : [a, \infty) \to \mathbb{R}$ be defined by $f(t) := p(t)/q(t)$. Show that $\int_a^\infty f(t)dt$ is absolutely convergent if $n \geq m+2$ and $\int_a^\infty f(t)dt$ is divergent if $n < m+2$.

45. Let $f : (a, b] \to \mathbb{R}$ be a nonnegative function that is integrable on $[x, b]$ for every $x \in (a, b]$. Prove the following:
   (i) If there is $p \in (0, 1)$ such that $(t - a)^p f(t) \to \ell$ for some $\ell \in \mathbb{R}$, then $\int_{a+}^b f(t)dt$ is convergent.
   (ii) If there is $p \geq 1$ such that $(t - a)^p f(t) \to \ell$ for some $\ell \neq 0$, then $\int_{a+}^b f(t)dt$ is divergent.
   (Hint: Corollary 9.44 with $g(t) := 1/(t - a)^p$ for $t \in (a, b]$.)

46. Show that $\int_{1+}^2 (\sqrt{t}/\ln t)dt$ is divergent. (Hint: Exercise 45.)

47. Let $p, q \in \mathbb{R}$ with $p > 0$ and $q > 0$. Show that

$$\beta(p, q) = 2 \int_{0+}^{(\pi/2)^-} (\sin u)^{2p-1} (\cos u)^{2q-1} du,$$

and in particular, $\beta\left(\frac{1}{2}, \frac{1}{2}\right) = \pi$. (Hint: Substitute $t := \sin^2 u$.)

48. Let $p, q > 0$. Show that

$$\beta(p, q) = \int_{0+}^{\infty} \frac{u^{p-1}}{(1+u)^{p+q}} du = \int_{0+}^{1} \frac{v^{p-1} + v^{q-1}}{(1+v)^{p+q}} dv.$$

(Hint: Substitute $t := u/(1+u)$ and then $v := 1/u$.)

49. Show that $\Gamma(s) = 2 \int_0^{\infty} e^{-u^2} u^{2s-1} du$ for all $s > 0$ and in particular, that $\Gamma\left(\frac{1}{2}\right) = 2 \int_0^{\infty} e^{-u^2} du$. (Hint: Substitute $t := u^2$.)

50. Test the following for absolute/conditional convergence:

(i) $\int_1^{\infty} \frac{1}{\sqrt{1+t^3}} dt$, (ii) $\int_1^{\infty} \frac{t^q}{1+t^q} dt$, where $q \in \mathbb{R}$, (iii) $\int_2^{\infty} \frac{1}{\ln t} dt$,

(iv) $\int_{0+}^{1} \sin\left(\frac{1}{t}\right) dt$, (v) $\int_{0+}^{1} e^{1/t} t^q dt$, where $q \in \mathbb{R}$, (vi) $\int_{0+}^{1-} \frac{1}{t \ln t} dt$.

Part B

51. **(Cauchy Product)** Suppose one of the series $\sum_{k=0}^{\infty} a_k$ and $\sum_{k=0}^{\infty} b_k$ is absolutely convergent and the other is convergent. Let A and B denote their respective sums. For each $k = 0, 1, \ldots$, let $c_k := \sum_{j=0}^{k} a_j b_{k-j}$. Show that the series $\sum_{k=0}^{\infty} c_k$ is convergent and its sum is equal to AB. Give an example to show that the result may not hold if both the series $\sum_{k=0}^{\infty} a_k$ and $\sum_{k=0}^{\infty} b_k$ are conditionally convergent.

52. **(Grouping of Terms)** Let $m_0 := 0$ and $m_1 < m_2 < \cdots$ be natural numbers. Given a series $\sum_{k=1}^{\infty} a_k$, define $b_k := a_{m_{k-1}+1} + \cdots + a_{m_k}$ for $k \in \mathbb{N}$. If the series $\sum_{k=1}^{\infty} a_k$ is convergent, then show that the series $\sum_{k=1}^{\infty} b_k$ is convergent and has the same sum. Give an example to show that $\sum_{k=1}^{\infty} b_k$ may be convergent although $\sum_{k=1}^{\infty} a_k$ is divergent.

53. **(Rearrangement of Terms)** Let $k \longmapsto j(k)$ be a bijection from $\mathbb{N}$ to $\mathbb{N}$. Given a series $\sum_{k=1}^{\infty} a_k$, consider the series $\sum_{k=1}^{\infty} b_k$, where $b_k := a_{j(k)}$. Then the series $\sum_{k=1}^{\infty} b_k$ is called a **rearrangement** of the series $\sum_{k=1}^{\infty} a_k$. Show that a series $\sum_{k=1}^{\infty} a_k$ is absolutely convergent if and only if every rearrangement of it is convergent. In this case, the sum of a rearrangement is unchanged.

54. Use the triangle inequality and the Cauchy Criterion (Propositions 1.8 and 2.19) to conclude that if a series of real numbers is absolutely convergent, then it is convergent. Conversely, assuming that every absolutely convergent series of real numbers is convergent, deduce the Cauchy Criterion. (Hint: Given a Cauchy sequence (A_n) of real numbers, inductively construct a subsequence (A_{n_k}) such that $|A_{n_{k+1}} - A_{n_k}| \leq 1/k^2$ for all $k \in \mathbb{N}$ and consider $a_k := A_{n_{k+1}} - A_{n_k}$.)

55. For $k \in \mathbb{N}$, let $a_k \in \mathbb{R}$ with $a_k > 0$. Show that

$$\liminf_{k \to \infty} \frac{a_{k+1}}{a_k} \leq \liminf_{k \to \infty} a_k^{1/k} \quad \text{and} \quad \limsup_{k \to \infty} a_k^{1/k} \leq \limsup_{k \to \infty} \frac{a_{k+1}}{a_k}.$$

56. For $k \in \mathbb{N}$, let $a_k \in \mathbb{R}$ with $a_k \neq 0$. If $|a_{k+1}|/|a_k| \to \ell$ as $k \to \infty$, then show that $|a_k|^{1/k} \to \ell$ as $k \to \infty$.

57. For $k \in \mathbb{N}$, let $a_k \in \mathbb{R}$ and define $\alpha := \limsup_{k \to \infty} |a_k|^{1/k}$. Show that if $\alpha < 1$, then $\sum_{k=1}^{\infty} a_k$ is absolutely convergent and if $\alpha > 1$, then $\sum_{k=1}^{\infty} a_k$ is divergent.

58. For $k \in \mathbb{N}$, let $a_k \in \mathbb{R}$ with $a_k \neq 0$. Define $\alpha := \limsup_{k \to \infty} |a_{k+1}|/|a_k|$ and $\beta := \liminf_{k \to \infty} |a_{k+1}|/|a_k|$. Show that if $\alpha < 1$, then $\sum_{k=1}^{\infty} a_k$ is absolutely convergent and if $\beta > 1$, then $\sum_{k=1}^{\infty} a_k$ is divergent.

59. Let $\sum_{k=0}^{\infty} c_k x^k$ be a power series with $c_k \neq 0$ for all $k \in \mathbb{N}$ and let r denote its radius of convergence Prove the following:
   (i) Suppose $|c_{k+1}|/|c_k| \to \infty$ as $k \to \infty$. Then $r = 0$.
   (ii) Suppose the sequence $(|c_{k+1}|/|c_k|)$ is bounded. For $k \in \mathbb{N}$, define $M_k := \sup\{|c_{j+1}|/|c_j| : j \in \mathbb{N} \text{ and } j \geq k\}$. Then (M_k) is a monotonically decreasing sequence of nonnegative real numbers. Let $L = \lim_{k \to \infty} M_k$. If $L = 0$, then $r = \infty$ and if $L > 0$, then $r \geq 1/L$.
   (iii) Suppose $|c_{k+1}|/|c_k| \not\to \infty$ as $k \to \infty$. For $k \in \mathbb{N}$, define $m_k := \inf\{|c_{j+1}|/|c_j| : j \in \mathbb{N} \text{ and } j \geq k\}$. Then (m_k) is a monotonically increasing sequence that is bounded above. Let $\ell := \lim_{k \to \infty} m_k$. If $\ell \neq 0$, then $r \leq 1/\ell$.
   [Note: $L = \limsup_{k \to \infty} |c_{k+1}|/|c_k|$ and $\ell = \liminf_{k \to \infty} |c_{k+1}|/|c_k|$.]

60. **(Binomial Series)** Let $r \in \mathbb{R}$ be such that $r \notin \{0, 1, 2, \ldots\}$, and define $f : (-1, 1) \to \mathbb{R}$ by $f(x) = (1 + x)^r$. Show that

$$f(x) = 1 + \sum_{k=1}^{\infty} \frac{r(r-1)\cdots(r-k+1)}{k!} x^k \quad \text{for } x \in (-1, 1).$$

   (Hint: If $x \in (-1, 1)$, $n \in \mathbb{N}$, and $R_n(x)$ denotes the Cauchy form of remainder as given in Exercise 49 of Chapter 4, then

$$|R_n(x)| \leq \left| r(r-1)\left(\frac{r}{2} - 1\right) \cdots \left(\frac{r}{n} - 1\right) \right| (1 + c)^{r-1} |x|^{n+1}$$

   for some c between 0 and x.)

61. Let $f : [a, b] \to \mathbb{R}$ be an infinitely differentiable function. Assume that the Taylor series of f around a converges to $f(x)$ at every $x \in [a, b]$, that is,

$$f(x) = f(a) + \sum_{n=1}^{\infty} \frac{f^{(n)}(a)}{n!} (x - a)^n, \quad x \in [a, b].$$

   Also, assume that the series obtained by integrating the above series term by term converges to $\int_a^b f(x)$, that is,

$$\int_a^b f(x)dx = f(a)(b-a) + \sum_{n=1}^{\infty} \frac{f^{(n)}(a)}{(n+1)!}(b-a)^{n+1}.$$

If $\mathsf{M}(f)$, $\mathsf{T}(f)$, and $\mathsf{S}(f)$ denote the Midpoint Rule, the Trapezoidal Rule, and Simpson's Rule for f, show that there are $\alpha_n, \beta_n, \gamma_n$ in $\mathbb{R}$ such that the series $\sum_{n=4}^{\infty} \alpha_n(b-a)^n, \sum_{n=4}^{\infty} \beta_n(b-a)^n$, and $\sum_{n=6}^{\infty} \gamma_n(b-a)^n$ converge and

(i) $\displaystyle\int_a^b f(x)dx - M(f) = \frac{f''(a)}{24}(b-a)^3 + \sum_{n=4}^{\infty} \alpha_n(b-a)^n,$

(ii) $\displaystyle\int_a^b f(x)dx - T(f) = -\frac{f''(a)}{12}(b-a)^3 + \sum_{n=4}^{\infty} \beta_n(b-a)^n,$

(iii) $\displaystyle\int_a^b f(x)dx - S(f) = -\frac{f^{(4)}(a)}{2880}(b-a)^5 \sum_{n=6}^{\infty} \beta_n(b-a)^n).$

(Compare Lemmas 8.20 and 8.22, and the subsequent error estimates.)

62. Let $f : [1, \infty) \to \mathbb{R}$ be a nonnegative monotonically decreasing function such that $\int_1^{\infty} f(t)dt$ is convergent. For $n \in \mathbb{N}$, let $B_n := \sum_{k=1}^{n} f(k)$ denote the nth partial sum of the convergent series $\sum_{k=1}^{\infty} f(k)$. Show that

$$B_n + \int_{n+1}^{\infty} f(t)dt \le \sum_{k=1}^{\infty} f(k) \le B_n + \int_{n+1}^{\infty} f(t)dt + f(n+1).$$

Use this result to show that

$$\sum_{k=1}^{n} \frac{1}{k^2} + \frac{1}{n+1} \le \sum_{k=1}^{\infty} \frac{1}{k^2} \le \sum_{k=1}^{n} \frac{1}{k^2} \le \sum_{k=1}^{n} \frac{1}{k^2} + \frac{1}{n+1} + \frac{1}{(n+1)^2}.$$

Further, show that if $n \ge 31$, then

$$\left| \sum_{k=n+1}^{\infty} \frac{1}{k^2} - \frac{1}{n+1} \right| < \frac{1}{1000}.$$

63. Let $f : [1, \infty) \to \mathbb{R}$ be a nonnegative monotonically decreasing function. For $n \in \mathbb{N}$, define $c_n := \sum_{k=1}^{n} f(k) - \int_1^n f(t)dt$. Show that $\lim_{n \to \infty} c_n$ exists and

$$0 \le f(1) - \int_1^2 f(t)dt \le \lim_{n \to \infty} c_n \le f(1).$$

Use this result to show that if

$$c_n := 1 + \frac{1}{2} + \cdots + \frac{1}{n} - \ln n,$$

then $c_n \to \gamma$, where γ satisfies $1 - \ln 2 < \gamma < 1$.

64. **(Log-Convexity of the Gamma Function)** Let $\Gamma_\ell : (0, \infty) \to \mathbb{R}$ be defined by $\Gamma_\ell(s) = \ln \Gamma(s)$. Show that Γ_ℓ is a convex function. (Hint: Use Exercise 55 (iv) of Chapter 7 to show that if $p, q \in (1, \infty)$ with $\frac{1}{p} + \frac{1}{q} = 1$, then $\Gamma((s/p) + (u/q)) \le \Gamma(s)^{1/p} \Gamma(u)^{1/q}$ for all $s, u \in (0, \infty)$.)

References

1. S. S. Abhyankar, *Algebraic Geometry for Scientists and Engineers*, American Mathematical Society, Providence, RI, 1990.
2. E. Artin, *Gamma Functions*, Holt, Rinehart and Winston, New York, 1964.
3. T. Apostol, *Mathematical Analysis*, first ed. and second ed., Addison-Wesley, Reading, MA, 1957 and 1974.
4. T. Apostol et al. (eds.), *A Century of Calculus*, vols. I and II, Mathematical Association of America, 1969 and 1992.
5. T. Apostol, *Calculus*, second ed., vols. 1 and 2, John Wiley, New York, 1967.
6. J. Arndt and C. Haenel, π *Unleashed*, Springer-Verlag, New York, 2001.
7. A. Baker, *Transcendental Number Theory*, Cambridge University Press, Cambridge, 1975.
8. R. G. Bartle and D. R. Sherbert, *Introduction to Real Analysis*, second ed., John Wiley, New York, 1992.
9. E. F. Beckenbach and R. Bellman, *Inequalities*, Springer-Verlag, New York, 1965.
10. L. Bers, On avoiding the mean value theorem, *American Mathematical Monthly* **74** (1967), p. 583. [Reprinted in [4]: Vol. I, p. 224.]
11. G. Birkhoff and S. Mac Lane, *A Survey of Modern Algebra*, third ed., Macmillan, New York, 1965.
12. T. J. Bromwich, *An Introduction to the Theory of Infinite Series*, third ed., Chelsea, New York, 1991.
13. C. B. Boyer, *The History of Calculus and Its Conceptual Development*, Dover, New York, 1959.
14. R. P. Boas, Who needs those mean-value theorems, anyway?, *College Mathematics Journal* **12** (1981), pp. 178–181. [Reprinted in [4]: Vol. II, pp. 182–186.]
15. P. S. Bullen, D. S. Mitrinovic, and P. M. Vasic, *Means and Their Inequalities*, D. Reidel, Dordrecht, 1988.
16. G. Chrystal, *Algebra: An Elementary Text-book for the Higher Classes of Secondary Schools and for Colleges*, sixth ed., parts I and II, Chelsea, New York, 1959.

17. L. Cohen, On being mean to the mean value theorem, *American Mathematical Monthly* **74** (1967), pp. 581–582.

18. T. Cohen and W. J. Knight, Convergence and divergence of $\sum_{n=1}^{\infty} 1/n^p$, *Mathematics Magazine* **52** (1979), p. 178. [Reprinted in [4]: Vol. II, p. 400.]

19. R. Courant and F. John, *Introduction to Calculus and Analysis*, vols. I and II, Springer-Verlag, New York, 1989.

20. D. A. Cox, The arithmetic-geometric mean of Gauss, *L'Enseignement Mathématique* **30** (1984), pp. 275–330.

21. P. J. Davis, Leonhard Euler's integral: A historical profile of the gamma function, *American Mathematical Monthly* **66** (1959), pp. 849–869.

22. P. Dienes, *The Taylor Series: An Introduction to the Theory of Functions of a Complex Variable*, Dover, New York, 1957.

23. C. H. Edwards Jr., *The Historical Development of the Calculus*, Springer-Verlag, New York, 1979.

24. H. B. Enderton, *Elements of Set Theory*, Academic Press, New York, 1977.

25. B. Fine and G. Rosenberger, *The Fundamental Theorem of Algebra*, Springer-Verlag, New York, 1997.

26. A. R. Forsyth, *A Treatise on Differential Equations*, Reprint of the sixth (1929) ed., Dover, New York, 1996.

27. C. Goffman, *Introduction to Real Analysis*, Harper & Row, New York-London, 1966.

28. R. Goldberg, *Methods of Real Analysis*, second ed., John Wiley, New York, 1976.

29. P. R. Halmos, *Naive Set Theory*, Springer-Verlag, New York, 1974.

30. R. W. Hamming, An elementary discussion of the transcendental nature of the elementary transcendental functions, *American Mathematical Monthly* **77** (1972), pp. 294–297. [Reprinted in [4]: Vol. II, pp. 80–83.]

31. G. H. Hardy, *A Course of Pure Mathematics*, Reprint of the (1952) tenth ed., Cambridge University Press, Cambridge, 1992.

32. G. H. Hardy, *Divergent Series*, second ed., Chelsea, New York, 1992.

33. G. H. Hardy, D. E. Littlewood and G. Polya, *Inequalities*, second ed., Cambridge University Press, Cambridge, 1952.

34. J. Havil, *Gamma: Exploring Euler's constant*, Princeton University Press, Princeton, NJ, 2003.

35. T. Hawkins, *Lebesgue's Theory of Integration: Its Origins and Development*, Reprint of the (1979) corrected second ed., AMS Chelsea Publishing, Providence, RI, 2001.

36. E. Hewitt and K. Stromberg, *Real and Abstract Analysis*, Springer-Verlag, New York, 1965.

37. E. W. Hobson, *The Theory of Functions of a Real Variable and the Theory of Fourier Series*, vol. I, third ed., Dover, New York, 1927.

38. K. D. Joshi, *Introduction to General Topology*, John Wiley, New York, 1983.

39. R. Kaplan, *The Nothing That Is: A Natural History of Zero*, Oxford University Press, New York, 2000.

40. G. Klambauer, *Aspects of Calculus*, Springer-Verlag, New York, 1986.

41. M. Kline, *Mathematical Thought from Ancient to Modern Times*, vols. 1, 2, and 3, second ed., Oxford University Press, New York, 1990.

42. K. Knopp, *Theory and Application of Infinite Series*, second ed., Dover, New York, 1990.

43. P. P. Korovkin, *Inequalities*, Little Mathematics Library No. 5, Mir Publishers, Moscow, 1975. [A reprint is distributed by Imported Publications, Chicago, 1986.]

44. E. Landau, *Foundations of Analysis: The Arithmetic of Whole, Rational, Irrational and Complex Numbers* (translated by F. Steinhardt), Chelsea, New York, 1951.

45. Jitan Lu, Is the composite function integrable?, *American Mathematical Monthly* **106** (1999), pp. 763–766.

46. R. Lyon and M. Ward, The Limit for *e*, *American Mathematical Monthly* **59** (1952), pp. 102–103 [Reprinted in [4]: Vol. I, pp. 432–433.]

47. E. Maor, *e: The Story of a Number*, Princeton University Press, Princeton, NJ, 1994.

48. M. D. Meyerson, Every power series is a Taylor series, *American Mathematical Monthly* **88** (1981), pp. 51–52.

49. D. J. Newman, *A Problem Seminar*, Springer-Verlag, New York, 1982.

50. D. J. Newman, A simplified version of the fast algorithms of Brent and Salamin, *Mathematics of Computation* **44** (1985), pp. 207–210.

51. K. A. Ross, *Elementary Analysis: The Theory of Calculus*, Springer-Verlag, New York, 1980.

52. H. L. Royden, *Real Analysis*, third ed., Prentice Hall, Englewood Cliffs, NJ, 1988.

53. W. Rudin, *Principles of Mathematical Analysis*, third ed., McGraw-Hill, New York-Auckland-Düsseldorf, 1976.

54. J. H. Silverman and J. Tate, *Rational Points on Elliptic Curves*, Springer-Verlag, New York, 1992.

55. G. F. Simmons, *Differential Equations, with Applications and Historical Notes*, McGraw-Hill, New York-Düsseldorf-Johannesburg, 1972.

56. R. S. Smith, Rolle over Lagrange – another shot at the mean value theorems, *College Mathematics Journal* **17** (1986), pp. 403–406.

57. M. Spivak, *Calculus*, first ed., Benjamin, New York, 1967; third ed., Publish or Perish Inc., Houston, 1994.

58. J. Stillwell, *Mathematics and Its History*, Springer-Verlag, New York, 1989.

59. O. E. Stanaitis, *An Introduction to Sequences, Series and Improper Integrals*, Holden-Day, San Francisco, 1967.

60. E. M. Stein and R. Shakarchi, *Fourier Analysis: An Introduction*, Princeton University Press, Princeton, NJ, 2003.

61. A. E. Taylor, L'Hospital's rule, *American Mathematical Monthly* **59** (1952), pp. 20–24.

62. G. B. Thomas and R. L. Finney, *Calculus and Analytic Geometry*, ninth ed., Addison-Wesley, Reading, MA, 1996.

63. J. van Tiel, *Convex Analysis: An Introductory Text*, John Wiley, New York, 1984.

64. J.-P. Tignol, *Galois' Theory of Algebraic Equations*, World Scientific Publishing Co., Inc., River Edge, NJ, 2001.

65. N. Y. Vilenkin, *Method of Successive Approximations*, Little Mathematics Library, Mir Publishers, Moscow, 1979.

List of Symbols and Abbreviations

	Definition/Description	Page
$x \in D$	x is an element of D	1
$\mathbb{N}$	the set of all positive integers	1
$\mathbb{Z}$	the set of all integers	1
$\mathbb{Q}$	the set of all rational numbers	1
$\mathbb{R}$	the set of all real numbers	2
$\sum$	sum	3
$\prod$	product	3
$A := B$	A is defined to be equal to B	3
$\mathbb{R}^+$	the set of all positive real numbers	4
$\emptyset$	the empty set	4
$\sup S$	the supremum of a subset S of $\mathbb{R}$	5
$\inf S$	the infimum of a subset S of $\mathbb{R}$	5
$\max S$	the maximum of a subset S of $\mathbb{R}$	5
$\min S$	the minimum of a subset S of $\mathbb{R}$	5
$[x]$	the integer part of a real number x	6
$\lfloor x \rfloor$	the integer part or the floor of a real number x	6
$\lceil x \rceil$	the ceiling of a real number x	6
$\sqrt[n]{a}$	the nth root of a nonnegative real number a	7
$\sqrt{a}$	the square root of a nonnegative real number a	7
$m \mid n$	m divides n	8, 18
$m \nmid n$	m does not divide n	8, 18
$\pm$	plus or minus	8
$C \subseteq D$	C is a subset of D	9
$\Longrightarrow$	implies	9
$D \setminus C$	the complement of C in D, namely, $\{x \in D : x \notin C\}$	9
(a, b)	the open interval $\{x \in \mathbb{R} : a < x < b\}$	9
$[a, b]$	the closed interval $\{x \in \mathbb{R} : a \le x \le b\}$	9
$[a, b)$	the semiopen interval $\{x \in \mathbb{R} : a \le x < b\}$	9
$(a, b]$	the semiopen interval $\{x \in \mathbb{R} : a < x \le b\}$	9
∞	the symbol ∞ or the fictional right endpoint of $\mathbb{R}$	9

	Definition/Description	**Page**
$\angle(OP_1, OP_2)$	the angle between OP_1 and OP_2	264
$L_1 \parallel L_2$	the lines L_1 and L_2 are parallel	266
$L_1 \nparallel L_2$	the lines L_1 and L_2 are not parallel	266
$\sphericalangle(L_1, L_2)$	the (acute) angle between L_1 and L_2	266
$L_1 \perp L_2$	the lines L_1 and L_2 are perpendicular	267
$L_1 \not\perp L_2$	the lines L_1 and L_2 are not perpendicular	267
$\sphericalangle(C_1, C_2; P)$	the angle at P between C_1 and C_2	268
Area (R)	the area of a region R	292
Vol (D)	the volume of a solid body D	299, 303
$\ell(C)$	the length of a curve C	311
Area (S)	the area of a surface S	321
$\mathrm{Av}(f)$	the average of a function f	325
$\mathrm{Av}(f; w)$	the weighted average of f with respect to w	325
$(\overline{x}, \overline{y})$	the centroid of a curve or a planar region	326, 329
$(\overline{x}, \overline{y}, \overline{z})$	the centroid of a surface or a solid body	327, 330
$Q(f)$	a Quadrature Rule for f	336
$R(f)$	Rectangular Rule for f	337
$M(f)$	Midpoint Rule for f	337
$T(f)$	Trapezoidal Rule for f	337
$S(f)$	Simpson's Rule for f	337
$R_n(f)$	Compound Rectangular Rule for f	338
$M_n(f)$	Compound Midpoint Rule for f	339
$T_n(f)$	Compound Trapezoidal Rule for f	339
$S_n(f)$	Compound Simpson's Rule for f	339
$\displaystyle\sum_{k=1}^{\infty} a_k$	the series whose sequence of terms is (a_k)	362
$\displaystyle\int_a^{\infty} f(t)dt$	the improper integral of f on $[a, \infty)$	384
$\displaystyle\int_{-\infty}^{b} f(t)dt$	the improper integral of f on $(-\infty, b]$	398
$\displaystyle\int_{-\infty}^{\infty} f(t)dt$	the improper integral of f on $(-\infty, \infty)$	398
$\displaystyle\int_{a+}^{b} f(t)dt$	the improper integral of f on $(a, b]$	399
$\displaystyle\int_{a}^{b^-} f(t)dt$	the improper integral of f on $[a, b)$	400
$\displaystyle\int_{a+}^{b^-} f(t)dt$	the improper integral of f on (a, b)	400
$\beta(p, q)$	the beta function for $p > 0$ and $q > 0$	405
$\gamma(s)$	the gamma function for $s > 0$	407

Index

Undergraduate Texts in Mathematics *(continued from p.ii)*