Kenneth R. Davidson, Allan P. Donsig

# Real Analysis and Applications: Theory in Practice

# Supplementary Material

Springer

# Contents

## 1.A The Language of Mathematics

The language of mathematics has to be precise, because mathematical statements must be interpreted with as little ambiguity as possible. Indeed, the rigour in mathematics is much greater than in law. There should be no doubts, reasonable or otherwise, when a theorem is proved. It is either completely correct, or it is wrong. Consequently, mathematicians have adopted a very precise language so that statements may not be misconstrued.

In complicated situations, it is easy to fool yourself. By being very precise and formal now, we can build up a set of tools that will help prevent mistakes later. The history of mathematics is full of stories in which mathematicians have fooled themselves with incorrect proofs. Clarity in mathematical language, like clarity in all other kinds of writing, is essential to communicating your ideas.

We begin with a brief discussion of the logical usage of certain innocuous words *if, then, only if, and, or* and *not*. Let $A, B, C$ represent statements that may or may not be true in a specific instance. For example, consider the statements

$A$. It is raining.
$B$. The sidewalk is wet.

The statement "If $A$, then $B$" means that whenever $A$ is true, it follows that $B$ must also be true. We also formulate this as "$A$ **implies** $B$." This statement does not claim either that the sidewalk is wet or that it is not. It tells you that if you look outside and see that it is raining, then without looking at the sidewalk, you will know that the sidewalk is wet as a result. As in the English language, "if $A$, then $B$" is a conditional statement meaning that only when the hypothesis $A$ is verified can you deduce that $B$ is valid. One also writes "Suppose $A$. Then $B$" with essentially the same meaning.

On the other hand, $A$ implies $B$ is quite different from $B$ implies $A$. For example, the sidewalk may be wet because

$C$. The lawn sprinkler is on.

The statement "if $B$, then $A$" is known as the **converse** of "if $A$, then $B$." This amounts to reversing the direction of the implication. As you can see from this example, one may be true but not the other.

We can also say "$A$ if $B$" to mean "if $B$, then $A$." The statement "$B$ only if $A$" means that in order that $B$ be true, it is necessary that $A$ be true. A bit of thought reveals that this is yet another reformulation of "if $A$, then $B$." For reasons of clarity, these two expressions are rarely used alone and are generally restricted to the combined statement "$A$ if and only if $B$." Parsing this sentence, we arrive at two statements "$A$ if $B$" and "$A$ only if $B$." The former means "$B$ implies $A$" and the latter means "$A$ implies $B$." Together they mean that either both statements are true or both are false. In this case, we say that statements $A$ and $B$ are **equivalent**.

The words *and*, *or*, and *not* are used with a precise mathematical meaning that does not always coincide with English usage. It is easy to be tripped up by these changes in meaning; be careful. "Not $A$" is the **negation** of the statement $A$. So "not $A$" is true if and only if $A$ is false. To say that "$A$ and $B$" is true, we mean that both

$A$ is true and $B$ is true. On the other hand, "$A$ or $B$" is true when at least one is true, but both being true is also possible. For example, the statement "if $A$ or $C$, then $B$' means that if either $A$ is true or $C$ is true, then $B$ is true.

Consider these statements about an integer $n$:

$D$. $n$ is even.
$E$. $n$ is a multiple of 4.
$F$. There is an integer $k$ so that $n = 4k+2$.

The statement "not $F$" is "there is no integer $k$ so that $n = 4k+2$." The statement "$D$ and not $F$" says that "$n$ is even, and there is no integer $k$ so that $n = 4k+2$." One can easily check that this is equivalent to statement $E$. Here are some valid statements:

(1) $D$ if and only if ($E$ or $F$).
(2) If ($D$ and not $E$), then $F$.
(3) If $F$, then $D$.

In the usual logical system of mathematics, a statement is either true or false, even if one cannot determine which is valid. A statement that is always true is a **tautology**. For example, "$A$ or not $A$" is a tautology. A more complicated tautology known as **modus ponens** is "If $A$ is true, and $A$ implies $B$, then $B$ is true." It is more common that a statement may be true or false depending on the situation (e.g., statement $D$ may be true or false depending on the value assigned to $n$).

The words *not* and *and* can be used together, but you must be careful to interpret statements accurately. The statement "not ($A$ and $B$)" is true if ($A$ and $B$) is false. If $A$ is false, then ($A$ and $B$) is false. Likewise if $B$ is false, then ($A$ and $B$) is false. While if both $A$ and $B$ are true, then ($A$ and $B$) is true. So "not ($A$ and $B$)" is true if either $A$ is false or $B$ is false. Equivalently, one of "not $A$" or "not $B$" is true. Thus "not ($A$ and $B$)" means the same thing as "(not $A$) or (not $B$)."

This kind of thinking may sound pedantic, but it is an important way of looking at a problem from another angle. The statement "$A$ implies $B$" means that $B$ is true whenever $A$ is true. Thus if $B$ is false, $A$ cannot be true, and thus $A$ is false. That is, "not $B$ implies not $A$." For example, if the sidewalk is not wet, then it is not raining. Conversely, if "not $B$ implies not $A$", then "$A$ implies $B$." Go through the same reasoning to see this through. You may have to use that "not (not $A$)" is equivalent to $A$. The statement "not $B$ implies not $A$" is called the **contrapositive** of "$A$ implies $B$." This discussion shows that the two statements are equivalent.

In addition to the converse and contrapositive of the statement "$A$ implies $B$," there is the negation, "not ($A$ implies $B$)." For "$A$ implies $B$" to be false, there must be *some instance* in which $A$ is true and $B$ is false. Such an instance is called a **counterexample** to the claim that "$A$ implies $B$." So the truth of $A$ has no direct implication on the truth of $B$. For example, "not ($C$ implies $B$)" means that it is possible for the lawn sprinkler to be on, yet the sidewalk remains dry. Perhaps the sprinkler is in the backyard, well out of reach of the sidewalk. It does not allow one to deduce any sensible conclusion about the relationship between $B$ and $C$ *except* that there are counterexamples to the statement "$C$ implies $B$."

$G$. If 2 divides 3, then 10 is prime.

*H*. If 2 divides $n$, then $n^2+1$ is prime.

One common point of confusion is the fact that false statements can imply anything. For example, statement $G$ is a tautology because the condition “2 divides 3” is never satisfied, so one never arrives at the false conclusion. One the other hand, $H$ is sometimes false (e.g., when $n=8$).

Another important use of precise language in mathematics is the phrases **for every** (or **for all**) and **there exists**, which are known as **quantifiers**. For example,

*I*. For every integer $n$, the integer $n^2-n$ is even.

This statement means that every substitution of an integer for $n$ in $n^2-n$ yields an even integer. This is correct because $n^2-n=n(n-1)$ is the product of the two integers $n$ and $n-1$, and one of them is even.

On the other hand, look at

*J*. For every integer $n \geq 0$, the integer $n^2+n+41$ is prime.

The first few terms 41, 43, 47, 53, 61, 71, 83, 97, 113, 131 are all prime. But to disprove this statement, it only takes a single instance where the statement fails. Indeed, $40^2+40+41=41^2$ is not prime. So this statement is false. We established this by demonstrating instead that

*K*. There is an integer $n$ so that $n^2+n+41$ is not prime.

This is the negation of statement $J$, and exactly one of them is true.

Things can get tricky when several quantifiers are used together. Consider

*L*. For every integer $m$, there is an integer $n$ so that 13 divides $m^2+n^2$.

To verify this, one needs to take each $m$ and prove that $n$ exists. This can be done by noting that $n=5m$ does the job since $m^2+(5m)^2=13(2m^2)$. On the other hand, consider

*M*. For every integer $m$, there is an integer $n$ so that 7 divides $m^2+n^2$.

To disprove this, one needs to find just one $m$ for which this statement is false. Take $m=1$. To show that this statement is false for $m=1$, it is necessary to check *every n* to make sure that $n^2+1$ is not a multiple of 7. This could take a rather long time by brute force. However observe that every integer may be written as $n=7k \pm j$ where $j$ is 0, 1, 2 or 3. Therefore

$$n^2+1=(7k \pm j)^2+1=7(7k^2 \pm 2j)+j^2+1.$$

Note that $j^2+1$ takes the values 1, 2, 5 and 10. None of these is a multiple of 7, and thus all of these possibilities are eliminated.

The order in which quantifiers is critical. Suppose the words in the statement $L$ are reordered as

*N*. There is an integer $n$ so that for every integer $m$, 13 divides $m^2+n^2$.

This has exactly the same words as statement $L$, but it claims the existence of an integer $n$ that works with *every* choice of $m$. We can dispose of this by showing that for every possible $n$, there is at least one value of $m$ for which the statement is false. Let us consider $m = 0$ and $m = 1$. If $N$ is true, then for the number $n$ satisfying the statement, we would have that both $n^2 + 1$ and $n^2 + 0$ are multiples of 13. But then 13 would divide the difference, which is 1. This contradiction shows that $n$ does not validate statement $N$. As $n$ was arbitrary, we conclude that $N$ is false.

## Exercises for Section 1.A

**A.** Which of the following are statements? That is, can they be true or false?

(a) Are all cats black?
(b) All integers are prime.
(c) $x + y$.
(d) $|x|$ is continuous.
(e) Don't divide by zero.

**B.** Which of the following statements implies which others?

(1) $X$ is a quadrilateral.
(2) $X$ is a square.
(3) $X$ is a parallelogram.
(4) $X$ is a trapezoid.
(5) $X$ is a rhombus.

**C.** Give the converse and contrapositive statements of the following:

(a) An equilateral triangle is isosceles.
(b) If the wind blows, the cradle will rock.
(c) If Jack Sprat could eat no fat and his wife could eat no lean, then together they can lick the platter clean.
(d) ($A$ and $B$) implies ($C$ or $D$).

**D.** Three young hoodlums accused of stealing CDs make the following statements:

(1) Ed: "Fred did it, and Ted is innocent."
(2) Fred: "If Ed is guilty, then so is Ted."
(3) Ted: "I'm innocent, but at least one of the others is guilty."

(a) If they are all innocent, who is lying?
(b) If all these statements are true, who is guilty?
(c) If the innocent told the truth and the guilty lied, who is guilty?
HINT: Remember that false statements imply anything.

**E.** Which of the following statements is true? For those that are false, write down the negation of the statement.

(a) For every $n \in \mathbb{N}$, there is an $m \in \mathbb{N}$ so that $m > n$.
(b) For every $m \in \mathbb{N}$, there is an $n \in \mathbb{N}$ so that $m > n$.
(c) There is an $m \in \mathbb{N}$ so that for every $n \in \mathbb{N}$, $m \geq n$.
(d) There is an $n \in \mathbb{N}$ so that for every $m \in \mathbb{N}$, $m \geq n$.

**F.** Let $A, B, C, D, E$ be statements. Make the following inferences.

(a) Suppose that $(A \text{ or } B)$ and $(A \text{ implies } B)$. Prove $B$.
(b) Suppose that $((\text{not } A) \text{ implies } B)$ and $(B \text{ implies } (\text{not } C))$ and $C$. Prove $A$.
(c) Suppose that $(A \text{ or } (\text{not } D))$, $((A \text{ and } B) \text{ implies } C)$, $((\text{not } E) \text{ implies } B)$, and $D$. Prove $(C \text{ or } E)$.

## 1.B Sets and Functions

Set theory is a large subject in its own right. We assume without discussion the existence of a sensible theory of sets and leave a full and rigorous development to books devoted to the subject. Our goal here to summarize the "intuitive" parts of set theory that we need for real analysis.

### SETS.

A **set** is a collection of elements; for example, $A = \{0,1,2,3\}$ is a set. This set has four elements, 0, 1, 2, and 3. The order in which they are listed is not relevant. A set can have other sets as elements. For example, $B = \{0,\{1,2\},3\}$ has three elements, one of which is the set $\{1,2\}$. Note that 1 is *not* an element of $B$, and that $A$ and $B$ are different.

We use $a \in A$ to denote that $a$ is an element of the set $A$ and $a \notin A$ to denote "not ($a \in A$)." The empty set $\varnothing$ is the set with no elements. We use the words *collection* and *family* as synonyms for sets. It is often clearer to talk about "a collection of sets" or "a family of sets" instead of "a set of sets." We say that two sets are equal if they have the same elements.

Given two sets $A$ and $B$, we say $A$ is a **subset** of $B$ if every element of $A$ is also an element of $B$. Formally, $A$ is a subset of $B$ if "$a \in A$ implies $a \in B$," or equivalently using quantifiers, $a \in B$ for all $a \in A$. If $A$ is a subset of $B$, then we write $A \subset B$. This allows the possibility that $A = B$. It also allows the possibility that $A$ has no elements, that is, $A = \varnothing$. We say $A$ is a **proper subset** of $B$ is $A \subset B$ and $A \neq B$. Notice that "$A \subset B$ and $B \subset A$" if and only if "$A = B$." Thus, if we want to prove that two sets, $A$ and $B$, are equal, it is equivalent to prove the two statements $A \subset B$ and $B \subset A$.

You should recognize that there is a distinction between membership in a set and a subset of a set. For the sets $A$ and $B$ defined at the beginning of this section, observe that $\{1,2\} \subset A$ and $\{1,2\} \in B$. The set $\{1,2\}$ is not a subset of $B$ nor an element of $A$. However, $\big\{\{1,2\}\big\} \subset B$.

There are a number of ways to combine sets to obtain new sets. The two most important are **union** and **intersection**. The union of two sets $A$ and $B$ is the set of all elements that are in $A$ or in $B$, and it is denoted $A \cup B$. Formally, $x \in A \cup B$ if and only if $x \in A$ or $x \in B$. The intersection of two sets $A$ and $B$ is the set of all elements that are both in $A$ and in $B$, and it is denoted $A \cap B$. Formally, $x \in A \cap B$ if and only if $x \in A$ and $x \in B$. Using our example, we have

$$A \cup B = \{0,1,2,\{1,2\},3\} \quad \text{and} \quad A \cap B = \{0,3\}.$$

Similarly, we may have an infinite family of sets $A_\gamma$ indexed by another set $\Gamma$. What this means is that for every element $\gamma$ of the set $\Gamma$, we have a set $A_\gamma$ indexed by that element. For example, for $n$ a positive integer, let $A_n$ be the set of positive numbers that divide $n$, so that $A_{12} = \{1,2,3,4,6,12\}$ and $A_{13} = \{1,13\}$. Then this collection $A_n$ is an infinite family of sets indexed by the positive integers, $\mathbb{N}$.

For infinite families of sets, intersection and union are defined formally in the same way. The union is

$$\bigcup_{\gamma \in \Gamma} A_\gamma = \{x : \text{there is a } \gamma \in \Gamma \text{ such that } x \in A_\gamma\}$$

and the intersection is

$$\bigcap_{\gamma \in \Gamma} A_\gamma = \{x : x \in A_\gamma \text{ for every } \gamma \in \Gamma\}.$$

In a particular situation, we are often working with a given set and subsets of it, such as the set of integers $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ and its subsets. We call this set our **universal set**. Once we have a universal set, say $U$, and a subset, say $A \subset U$, we can define the **complement** of $A$ to be the collection of all elements of $U$ that are not in $A$. The complement is denoted $A'$. Notice that the universal set can change from problem to problem, and that this will change the complement.

Given a universal set $U$, we can specify a subset of $U$ as all elements of $U$ with a certain property. For example, we may define the set of all the integers that are divisible by two. We write this formally as $\{x \in \mathbb{Z} \text{ so that } 2 \text{ divides } x\}$. It is traditional to use a vertical bar $|$ or a colon : for "so that," so that we can write the set of even integers as $2\mathbb{Z} = \{x \in \mathbb{Z} \mid 2 \text{ divides } x\}$. Similarly, we can write the complement of $A$ in a universal set $U$ as

$$A' = \{x \in U : x \notin A\}.$$

Given two sets $A$ and $B$, we define the **relative complement** of $B$ in $A$, denoted $A \setminus B$, to be

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

Notice that $B$ need not be a subset of $A$. Thus, we can talk about the relative complement of $2\mathbb{Z}$ in $\{0,1,2,3\}$, namely

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

In our example, $A \setminus B = \{1,2\}$. Curiously, $B \setminus A = \{\{1,2\}\}$, the set consisting of the single element $\{1,2\}$.

Finally, we need the idea of the **Cartesian product** of two sets, denoted $A \times B$. This is the set of ordered pairs $\{(a,b) : a \in A \text{ and } b \in B\}$. For example,

$$\{0,1,2\} \times \{2,4\} = \{(0,2),(1,2),(2,2),(0,4),(1,4),(2,4)\}.$$

More generally, if $A_1, \ldots, A_n$ is a finite collection of sets, the Cartesian product is written $A_1 \times \cdots \times A_n$ or $\prod_{i=1}^{n} A_i$, and consists of all ***n*-tuples** $a = (a_1, a_2, \ldots, a_n)$ such that $a_i \in A_i$ for $1 \le i \le n$. If $A_i = A$ is the same set for each $i$, then we write $A^n$ for the product of $n$ copies of $A$. For example, $\mathbb{R}^3$ consists of all triples $(x,y,z)$ with arbitrary real coefficients $x,y,z$. There is also a notion of the product of an infinite family of sets. We will not have any need of it, but we warn the reader that such infinite products raise subtle questions about the nature of sets.

## FUNCTIONS.

In practice, a function $f$ from $A$ to $B$ is a rule that assigns an element $f(a) \in B$ to each element $a \in A$. Such a rule may be very complicated with many different cases. In set theory, a very general definition of function is given that does not require the use of undefined terms such as *rule*. This definition specifies a function in terms of its graph, which is a subset of $A \times B$ with a special property. We provide the definition here. However, we will usually define functions by rules in the standard fashion.

**1.B.1. DEFINITION.** Given two nonempty sets $A$ and $B$, a **function** $f$ from $A$ to $B$ is a subset of $A \times B$, denoted $G(f)$, so that

(1) for each $a \in A$, there is some $b \in B$ so that $(a,b) \in G(f)$,
(2) for each $a \in A$, there is only one $b \in B$ so that $(a,b) \in G(f)$.

That is, for each $a \in A$, there is *exactly one* element $b \in B$ with $(a,b) \in G(f)$. We then write $f(a) = b$. A concise way to specify the function $f$ and the sets $A$ and $B$ all at once is to write $f : A \to B$. We call $G(f)$ the **graph of the function** $f$.

The property of a subset of $A \times B$ that makes it the graph of a function is that $\{b \in B : (a,b) \in G(f)\}$ has precisely one element for each $a \in A$. This is the "vertical line test" for functions.

We can think of $f$ as the rule that sends $a \in A$ to the unique point $b \in B$ such that $(a,b) \in G(f)$; and we write $f(a) = b$. Notice that $f(a)$ is an element of $B$ while $f$ is the name of the function as a whole. Sometimes we will use such convenient expressions as "the function $x^2$." This really means "the function that sends $x$ to $x^2$ for all $x$ such that $x^2$ makes sense."

We call $A$ the **domain** of the function $f : A \to B$ and $B$ is the **codomain**. Far more important than the codomain is the **range** of $f$, which is

$$\mathrm{ran}(f) := \{b \in B : b = f(a) \text{ for some } a \in A\}.$$

If $f$ is a function from $A$ into $B$ and $C \subset A$, the **image** of $C$ under $f$ is

$$f(C) := \{b \in B : \text{ there is some } c \in C \text{ so that } f(c) = b\}.$$

The range of $f$ is $f(A)$.

Notice that the notation $f(r)$ has two possible meanings, depending on whether $r$ is an element of $A$ or a subset of $A$. The standard practice of using lowercase letters for elements and uppercase letters for sets makes this notation clear in practice.

The same caveat is applied to the notation $f^{-1}$. If $f$ maps $A$ into $B$, the **inverse image** of $C \subset B$ under $f$ is

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

Note that $f^{-1}$ is not used here as a function from $B$ to $A$. Indeed, the domain of $f^{-1}$ is the set of all subsets of $B$, and the codomain consists of all subsets of $A$. Even if $C = \{b\}$ is a single point, $f^{-1}(\{b\})$ may be the empty set or it may be very large. For example, if $f(x) = \sin x$, then $f^{-1}(\{0\}) = \{n\pi : n \in \mathbb{Z}\}$ and $f^{-1}(\{y : |y| > 1\}) = \varnothing$.

**1.B.2. DEFINITION.** A function $f$ of $A$ into $B$ maps $A$ **onto** $B$ or $f$ is **surjective** if $\operatorname{ran}(f) = B$. In other words, for each $b \in B$, there is *at least one* $a \in A$ such that $f(a) = b$. Similarly, if $D \subset B$, say that $f$ maps $A$ **onto** $D$ if $D \subset \operatorname{ran}(f)$.

A function $f$ of $A$ into $B$ is **one-to-one** or **injective** if $f(a_1) = f(a_2)$ implies that $a_1 = a_2$ for $a_1, a_2 \in A$. In other words, for each $b$ in the range of $f$, there is *at most one* $a \in A$ such that $f(a) = b$.

A function from $A$ to $B$ that is both one-to-one and onto is called a **bijection**.

Suppose that $f : A \to B$, $\operatorname{ran}(f) \subset B_0 \subset B$ and $g : B_0 \to C$; then the **composition** of $g$ and $f$ is the function $g \circ f(a) = g(f(a))$ from $A$ into $C$.

A function is one-to-one if it passes a "horizontal line test." In this context, we can interpret $f^{-1}$ as a function from $B$ to $A$. This notion has a number of important consequences. In particular, when the ordered pairs in $G(f)$ are interchanged, the new set is the graph of a function known as the **inverse function** of $f$.

**1.B.3. LEMMA.** *If $f : A \to B$ is a one-to-one function, then there is a unique one-to-one function $g : f(A) \to A$ so that*

$$g(f(a)) = a \text{ for all } a \in A \quad \text{and} \quad f(g(b)) = b \text{ for all } b \in f(A).$$

*We call $g$ the inverse function of $f$ and denote it by $f^{-1}$.*

**PROOF.** Let $H \subset f(A) \times A$ be defined by

$$H = \{(b,a) \in f(A) \times A : (a,b) \in G(f)\}.$$

By the definition of $f(A)$, for each $b \in f(A)$, there is an $a \in A$ with $(a,b) \in G(f)$. Thus, $(b,a) \in H$, showing $H$ satisfies property (1) of a function.

Suppose $(b,a_1)$ and $(b,a_2)$ are in $H$. Then $(a_1,b)$ and $(a_2,b)$ are in $G(f)$; that is, $f(a_1) = b$ and $f(a_2) = b$. Since $f$ is one-to-one, $a_1 = a_2$. This confirms that $H$ has property (2); and so $H$ is the graph of a function $g : f(A) \to A$.

Suppose that $g(b_1) = g(b_2)$. Then there is some $a \in A$ so that $(b_1,a)$ and $(b_2,a)$ are in $G(g) = H$. Thus, $(a,b_1)$ and $(a,b_2)$ are in $G(f)$. But $f$ is a function, so by property (2) for $G(f)$, $b_1 = b_2$. Thus, $g$ is one-to-one.

Finally, observe that if $a \in A$, and $b = f(a)$, then $(b,a) \in G(g)$, so $g(b) = a$ and thus $g(f(a)) = g(b) = a$. Similarly, $f(g(b)) = b$ for all $b \in f(A)$. ∎

We say that two functions are equal if they have the same domains and the same codomains and if they agree at every point. So $f : A \to B$ and $g : A \to B$ are equal if $f(a) = g(a)$ for all $a \in A$.

We can express the relation between a one-to-one function and its inverse in terms of the identity maps. The **identity map** on a set $A$ is $\mathrm{id}_A(a) = a$ for $a \in A$. When only one set $A$ is involved, we use id instead of $\mathrm{id}_A$.

**1.B.4. COROLLARY.** *If $f : A \to B$ is a bijection, then $f^{-1}$ is a bijection and it is the unique function $g : B \to A$ so that $g \circ f = \mathrm{id}_A$ and $f \circ g = \mathrm{id}_B$. That is, $f^{-1}(f(a)) = a$ for all $a \in A$ and $f(f^{-1}(b)) = b$ for all $b \in B$.*

## Exercises for Section 1.B

**A.** Which of the following statements is true? Prove or give a counterexample.

(a) $(A \cap B) \subset (B \cup C)$
(b) $(A \cup B') \cap B = A \cap B$
(c) $(A \cap B') \cup B = A \cup B$
(d) $A \setminus B = B \setminus A$
(e) $(A \cup B) \setminus (A \cap B) = (A \setminus B) \cup (B \setminus A)$
(f) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
(g) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
(h) If $(A \cap C) \subset (B \cap C)$, then $(A \cup C) \subset (B \cup C)$.

**B.** How many different sets are there that may be described using two sets $A$ and $B$ and as many intersections, unions, complements and parentheses as desired?
HINT: First show that there are four minimal nonempty sets of this type.

**C.** What is the Cartesian product of the empty set with another set?

**D.** The **power set** $P(X)$ of a set $X$ is the set consisting of all subsets of $X$, including $\varnothing$.

(a) Find a bijection between $P(X)$ and the set of all functions $f : X \to \{0, 1\}$.
(b) How many different subsets of $\{1, 2, 3, \ldots, n\}$ are there?
HINT: Count the functions in part (a).

**E.** Let $f$ be a function from $A$ into $X$, and let $Y, Z \subset X$. Prove the following:

(a) $f^{-1}(Y \cap Z) = f^{-1}(Y) \cap f^{-1}(Z)$
(b) $f^{-1}(Y \cup Z) = f^{-1}(Y) \cup f^{-1}(Z)$
(c) $f^{-1}(X) = A$
(d) $f^{-1}(Y') = f^{-1}(Y)'$

**F.** Let $f$ be a function from $A$ into $X$, and let $B, C \subset A$. Prove the following statements. One of these statements may be sharpened to an equality. Prove it, and show by example that the others may be proper inclusions.

(a) $f(B \cap C) \subset f(B) \cap f(C)$
(b) $f(B \cup C) \subset f(B) \cup f(C)$
(c) $f(B) \subset X$
(d) If $f$ is one-to-one, then $f(B') \subset f(B)'$.

**G.** (a) What should a *two-to-one* function be?
(b) Give an example of a two-to-one function from $\mathbb{Z}$ onto $\mathbb{Z}$.

**H.** Suppose that $f, g, h$ are functions from $\mathbb{R}$ into $\mathbb{R}$. Prove or give a counterexample to each of the following statements. HINT: Only one is true.

(a) $f \circ g = g \circ f$
(b) $f \circ (g + h) = f \circ g + f \circ h$
(c) $(f + g) \circ h = f \circ h + g \circ h$

**I.** Suppose that $f : A \to B$ and $g : B \to A$ satisfy $g \circ f = \mathrm{id}_A$. Show that $f$ is one-to-one and $g$ is onto.

## 1.C The Role of Proofs

Mathematics is all about proofs. Mathematicians are not as much interested in *what* is true as in *why* it is true. For example, you were taught in high school that the roots of the quadratic equation $ax^2 + bx + c = 0$ are $\dfrac{-b \pm \sqrt{b^2 - 4ac}}{2a}$ provided that $a \neq 0$. A serious class would not have been given this as a fact to be memorized. It would have been justified by the technique of *completing the square*. This raises the formula from the realm of magic to the realm of understanding.

There are several important reasons for teaching this argument. The first goes beyond intellectual honesty and addresses the real point, which is that you shouldn't accept mathematics (or science) on faith. The essence of scientific thought is understanding why things work out the way they do.

Second, the formula itself does not help you do anything beyond what it is designed to accomplish. It is no better than a quadratic solver button that could be built into your calculator. The numbers $a, b, c$ go into a black box and two numbers come out or they don't—you might get an error message if $b^2 - 4ac < 0$. At this stage, you have no way of knowing if the calculator gave you a reasonable answer, or why it might give an error. If you know where the formula comes from, you can analyze all of these issues clearly.

Third, knowledge of the proof makes further progress a possibility. The creation of a new proof about something that you don't yet know is much more difficult than understanding the arguments someone else has written down. Moreover, understanding these arguments makes it easier to push further. It is for this reason that we can make progress. As Isaac Newton once said, "If I have seen further than others, it is by standing on the shoulders of giants." The first step toward proving things for yourself is to understand how others have done it before.

Fourth, if you understand that the *idea* behind the quadratic formula is completing the square, then you can always recover the quadratic formula whenever you forget it. This nugget of the proof is a useful method of data compression that saves you the trouble of memorizing a bunch of arcane formulae.

It is our hope that most students reading this book already have had some introduction to proofs in their earlier courses. If this is not the case, the examples in this section will help. This may be sufficient to tackle the basic material in this book. But be warned that some parts of this book require significant sophistication on the part of the reader.

### DIRECT PROOFS.

We illustrate several proof techniques that occur frequently. The first is **direct proof**. In this technique, one takes a statement, usually one asserting the existence of some mathematical object, and proceeds to verify it. Such an argument may amount to a computation of the answer. On the other hand, it might just show the existence of the object without actually computing it. The crucial distinction for existence proofs is between those that are **constructive proofs**—that is, those that give you a method or algorithm for finding the object—and those that are **nonconstructive proofs**—that is, they don't tell you how to find it. Needless to say, constructive proofs do something more than nonconstructive ones, but they sometimes take more work.

Every real number $x$ has a decimal expansion $x = a_0.a_1a_2a_3\dots$, where $a_i$ are integers and $0 \le a_i \le 9$ for all $i \ge 1$. This will be discussed thoroughly in Chapter 2. This expansion is **eventually periodic** if there are integers $N$ and $d > 0$ so that $a_{n+d} = a_n$ for $n \ge N$.

Occasionally a direct proof is just a straightforward calculation or verification.

**1.C.1. THEOREM.** *If the decimal expansion of a real number x is eventually periodic, then x is rational.*

**PROOF.** Suppose that $N$ and $d > 0$ are given so that $a_{n+d} = a_n$ for $n \ge N$. Compute $10^N x$ and $10^{N+d}x$ and observe that

$$\begin{aligned} 10^{N+d}x &= b.a_{N+1+d}a_{N+2+d}a_{N+3+d}a_{N+4+d}\cdots \\ &= b.a_{N+1}a_{N+2}a_{N+3}a_{N+4}\cdots \\ 10^N x &= c.a_{N+1}a_{N+2}a_{N+3}a_{N+4}\dots, \end{aligned}$$

where $b$ and $c$ are integers that you can easily compute. Subtracting the second equation from the first yields

$$(10^{N+d} - 10^N)x = b - c.$$

Therefore, $x = \dfrac{b-c}{10^{N+d} - 10^N}$ is a rational number. ■

The converse of this statement is also true. We will prove it by an existential argument that does not actually exhibit the exact answer, although the argument does provide a method for finding the exact answer. The next proof is definitely more sophisticated than a computational proof. It still, like the last proof, has the advantage of being constructive.

We need a simple but very useful fact.

**1.C.2. PIGEONHOLE PRINCIPLE.**
*If $n+1$ items are divided into n categories, then at least two of the items are in the same category.*

This is evident after a little thought, and we do not attempt to provide a formal proof. Note that it has variants that may also be useful. If $nd+1$ objects are divided into $n$ categories, then at least one category contains $d+1$ items. Also, if infinitely many items are divided into finitely many categories, then at least one category has infinitely many items.

**1.C.3. THEOREM.** *If $x$ is rational, then the decimal expansion of $x$ is eventually periodic.*

**PROOF.** Since $x$ is rational, we may write it as $x = \frac{p}{q}$, where $p, q$ are integers and $q > 0$. When an integer is divided by $q$, we obtain another integer with a remainder in the set $\{0, 1, \dots, q-1\}$. Consider the remainders $r_k$ when $10^k$ is divided by $q$ for $0 \le k \le q$. There are $q+1$ numbers $r_k$, but only $q$ possible remainders. By the Pigeonhole Principle, there are two integers $0 \le k < k+d \le q$ so that $r_k = r_{k+d}$. Therefore, $q$ divides $10^{k+d} - 10^k$ exactly, say $qm = 10^{k+d} - 10^k$.

Now compute

$$\frac{p}{q} = \frac{pm}{qm} = \frac{pm}{10^{k+d} - 10^k} = 10^{-k} \frac{pm}{10^d - 1}.$$

Divide $10^d - 1$ into $pm$ to obtain quotient $a$ with remainder $b$, $0 \le b < 10^d - 1$. So

$$x = \frac{p}{q} = 10^{-k}\left(a + \frac{b}{10^d - 1}\right),$$

where $0 \le b < 10^d - 1$. Write $b = b_1 b_2 \dots b_d$ as a decimal number with exactly $d$ digits even if the first few are zero. For example, if $d = 4$ and $b = 13$, we will write $b = 0013$. Then consider the periodic (or repeating) decimal

$$r = 0.b_1 b_2 \dots b_d b_1 b_2 \dots b_d b_1 b_2 \dots b_d \dots.$$

Using the proof of Theorem 1.C.1, we find that $(10^d - 1)r = b$ and thus $r = \frac{b}{10^d - 1}$. Observe that $10^k x = a + r = a.b_1 b_2 \dots b_d \dots$ has a repeating decimal expansion. The decimal expansion of $x = 10^{-k}(a + r)$ begins repeating every $d$ terms after the first $k$. Therefore, this expansion is eventually periodic. ■

## PROOF BY CONTRADICTION.

The second common proof technique is generally called **proof by contradiction**. Suppose that we wish to verify statement $A$. Now either $A$ is true or it is false. We assume that $A$ is false and make a number of logical deductions until we establish as true something that is clearly false. No false statement can be deduced from a logical sequence of deductions based on a valid hypothesis. So our hypothesis that $A$ is false must be incorrect, whence $A$ is true.

Here is a well-known example of this type.

**1.C.4. THEOREM.** *$\sqrt{3}$ is an irrational number.*

**PROOF.** Suppose to the contrary that $\sqrt{3} = a/b$, where $a, b$ are positive integers with no common factor. (This proviso of no common factor is crucial to setting the stage correctly. Watch for where it gets used.) Manipulating the equation, we obtain

$$a^2 = 3b^2.$$

When the number $a$ is divided by 3, it leaves a remainder $r \in \{0,1,2\}$. Let us write $a = 3k + r$. Then

$$a^2 = (3k+r)^2 = 3(3k^2+2kr)+r^2 = \begin{cases} 9k^2 & \text{if } r=0 \\ 3(3k^2+2k)+1 & \text{if } r=1 \\ 3(3k^2+4k+1)+1 & \text{if } r=2 \end{cases}$$

Observe that $a^2$ is a multiple of 3 only when $a$ is a multiple of 3. Therefore, we can write $a = 3c$ for some integer $c$. So $9c^2 = 3b^2$. Dividing by 3 yields $b^2 = 3c^2$.

Repeating exactly the same reasoning, we deduce that $b = 3d$ for some integer $d$. It follows that $a$ and $b$ do have a common factor 3, contrary to our assumption. The reason for this contradiction was the incorrect assumption that $\sqrt{3}$ was rational. Therefore, $\sqrt{3}$ is irrational. ∎

The astute reader might question why a fraction may be expressed in lowest terms. This is an easy fact that does not depend on deeper facts such as unique factorization into primes. It is merely the observation that if $a$ and $b$ have a common factor, then after it is factored out, one obtains a new fraction $a_1/b_1$ with a smaller denominator. This procedure must terminate by the time the denominator is reduced to 1, if not sooner. A very crude estimate of how many times the denominator can be factored is $b$ itself.

The same reasoning is commonly applied to verify "$A$ implies $B$." It is enough to show that "$A$ and not $B$" is always false. For then if $A$ is true, it follows that not $B$ is false, whence $B$ is true. This is usually phrased as follows: $A$ is given as true. Assume that $B$ is false. If we can make a sequence of logical deductions leading to a statement that is evidently false, then given that $A$ is true, our assumption that $B$ was false is itself incorrect. Thus $B$ is true.

### PROOF BY INDUCTION.

The Principle of Induction is the mathematical version of the domino effect.

**1.C.5. PRINCIPLE OF INDUCTION.** Let $P(n)$, $n \geq 1$, be a sequence of statements. Suppose that we can verify the following two statements:

(1) $P(1)$ is true.
(2) If $n > 1$ and $P(k)$ is true for $1 \leq k < n$, then $P(n)$ is true.

Then $P(n)$ is true for each $n \geq 1$.

We note that there is nothing special about starting at $n = 1$. For example, we can also start at $n = 0$ if the statements are numbered beginning at 0. You may have seen step (2) replaced by

(2′)If $n > 1$ and $P(n-1)$ is true, then $P(n)$ is true.

This requires a stronger dependence on the previous statements and thus is a somewhat weaker principle. However, it is frequently sufficient.

Most students reading this book will have seen how to verify statements like $\sum_{k=1}^{n} k^3 = \big(\sum_{k=1}^{n} k\big)^2$ by induction. As a quick warmup, we outline the proof that the sum of the first $n$ odd numbers is $n^2$, that is, $\sum_{k=1}^{n}(2k-1) = n^2$. If $n = 1$, then both sides are 1 and hence equal. Suppose the statement if true for $n-1$, so that $\sum_{k=1}^{n-1}(2k-1) = (n-1)^2$. Then

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

By induction, the statement is true for all integers $n \geq 1$.

Next, we provide an example that requires a bit more work and relies on the stronger version of induction. In fact, this example require two steps to get going, not just one.

**1.C.6. THEOREM.** *The **Fibonacci sequence** is given recursively by*

$$F(0) = F(1) = 1 \quad \textit{and} \quad F(n) = F(n-1) + F(n-2) \quad \textit{for all} \quad n \geq 2.$$

*Let* $\tau = \dfrac{1+\sqrt{5}}{2}$. *Then* $F(n) = \dfrac{\tau^{n+1} - (-\tau)^{-n-1}}{\sqrt{5}}$ *for all* $n \geq 0$.

**PROOF.** The statements are $P(n)$: $F(n) = \dfrac{\tau^{n+1} - (-\tau)^{-n-1}}{\sqrt{5}}$. Before we begin, observe that

$$\tau^2 = \Big(\frac{1+\sqrt{5}}{2}\Big)^2 = \frac{6+2\sqrt{5}}{4} = \frac{3+\sqrt{5}}{2} = \tau + 1.$$

Therefore, $\tau$ is a root of $x^2 - x - 1 = 0$. Now dividing by $\tau$ and rearranging yields

$$-\frac{1}{\tau} = 1 - \tau = \frac{1-\sqrt{5}}{2}.$$

Consider $n = 0$. It is generally better to begin with the complicated side of the equation and simplify it.

$$\frac{\tau^1 - (-\tau)^{-1}}{\sqrt{5}} = \frac{1}{\sqrt{5}}\Big(\frac{1+\sqrt{5}}{2} - \frac{1-\sqrt{5}}{2}\Big) = \frac{2\sqrt{5}}{2\sqrt{5}} = 1$$

This verifies the first step $P(0)$.

Right away we have a snag compared with a standard induction. Each $F(n)$ for $n \geq 2$ is determined by the two previous terms. But $F(1)$ does not fit into this pattern. It must also be verified separately.

$$\frac{\tau^2 - (-\tau)^{-2}}{\sqrt{5}} = \frac{1}{\sqrt{5}} \frac{(6+2\sqrt{5}) - (6-2\sqrt{5})}{4} = \frac{4\sqrt{5}}{4\sqrt{5}} = 1$$

This verifies statement $P(1)$.

Now consider the case $P(n)$ for $n \geq 2$, assuming that the statements $P(k)$ are known to be true for $0 \leq k < n$. In particular, they are valid for $k = n-1$ and $k = n-2$. Therefore,

$$\begin{aligned}
F(n) &= F(n-1) + F(n-2) \\
&= \frac{\tau^n - (-\tau)^{-n}}{\sqrt{5}} + \frac{\tau^{n-1} - (-\tau)^{1-n}}{\sqrt{5}} \\
&= \frac{\tau^{n-1}(\tau+1) - (-\tau)^{-n}(1-\tau)}{\sqrt{5}} \\
&= \frac{\tau^{n-1}(\tau^2) - (-\tau)^{-n}(-\tau^{-1})}{\sqrt{5}} = \frac{\tau^{n+1} - (-\tau)^{-n-1}}{\sqrt{5}}.
\end{aligned}$$

Thus $P(n)$ follows from knowing $P(n-1)$ and $P(n-2)$. The Principle of Induction now establishes that $P(n)$ is valid for each $n \geq 0$. ■

We will several times need a slightly stronger form of induction known as **recursion**. Simply put, the Principle of Recursion states that after an induction argument has been established, one has *all* of the statements $P(n)$. This undoubtedly seems to be what induction says. The difference is a subtle point of logic. Induction guarantees that each statement $P(n)$ is true, one at a time. To take all infinitely many of them at once requires a bit more. In order to deal with this rigorously, one needs to discuss the axioms of set theory, which takes us outside of the scope of this book. However it is intuitively believable, and we will take this as valid.

## Exercises for Section 1.C

**A.** Let $a \neq 0$. Prove that the quadratic equation $ax^2 + bx + c = 0$ has real solutions if and only if the discriminant $b^2 - 4ac$ is nonnegative. HINT: Complete the square.

**B.** Prove that the following numbers are irrational.
(a) $\sqrt[3]{2}$ (b) $\log_{10} 3$ (c) $\sqrt{3} + \sqrt[3]{7}$ (d) $\sqrt{6} - \sqrt{2} - \sqrt{3}$

**C.** Prove by induction that $\sum_{k=1}^{n} k^3 = \left(\sum_{k=1}^{n} k\right)^2 = (n(n+1)/2)^2$.

**D.** The **binomial coefficient** $\binom{n}{k}$ is $n!/(k!(n-k)!)$. Prove by induction that $\sum_{k=0}^{n} \binom{n}{k} = 2^n$. HINT: First prove that $\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1}$.

**E.** Let $A$ and $B$ be $n \times n$ matrices. Prove that $AB$ is invertible if and only if both $A$ and $B$ are invertible. HINT: Use direct algebraic calculations.

**F.** Prove by induction that every integer $n \geq 2$ factors as the product of prime numbers.
HINT: You need the statements $P(k)$ for all $2 \leq k < n$ here.

**G.** (a) Prove directly that if $a, b \geq 0$, then $\frac{a+b}{2} \geq \sqrt{ab}$.
(b) If $a_1, \dots, a_{2^n} \geq 0$, show by induction that $\dfrac{a_1 + \cdots + a_{2^n}}{2^n} \geq \sqrt[2^n]{a_1 a_2 \dots a_{2^n}}$.

(c) If $a_1, \ldots, a_m$ are positive numbers, choose $2^n \geq m$ and set $a_i = \dfrac{a_1 + \cdots + a_m}{m}$ for $m < i \leq 2^n$. Apply part (b) to deduce the **arithmetic mean–geometric mean inequality**, $\dfrac{a_1 + \cdots + a_m}{m} \geq \sqrt[m]{a_1 a_2 \ldots a_m}$.

**H.** Fix an integer $N \geq 2$. Consider the remainders $q(n)$ obtained by dividing the Fibonacci number $F(n)$ by $N$, so that $0 \leq q(n) < N$. Prove that this sequence is periodic with period $d \leq N^2$ as follows:

(a) Show that there are integers $0 \leq i < j \leq N^2$ such that $q(i) = q(j)$ and $q(i+1) = q(j+1)$. HINT: Pigeonhole.
(b) Show that if $q(i+d) = q(i)$ and $q(i+1+d) = q(i+1)$, then $q(n+d) = q(n)$ for all $n \geq i$. HINT: Use the recurrence relation for $F(n)$ and induction.
(c) Show that if $q(i+d) = q(i)$ and $q(i+1+d) = q(i+1)$, then $q(n+d) = q(n)$ for all $n \geq 0$. HINT: Work backward using the recurrence relation.

**I.** **The Binomial Theorem.** Prove by induction that $(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$ for all real numbers $x$ and $y$. HINT: Exercise D is the special case $x = y = 1$. Imitate its proof.

**J.** Consider the following "proof" by induction. We will argue that all students receive the same mark in calculus. *Let $P(n)$ be the statement that every set of $n$ students receives the same mark. This is evidently valid for $n = 1$. Now look at larger $n$. Suppose that $P(n-1)$ is true. Given a group of $n$ people, apply the induction hypothesis to all but the last person in the group. The students in this smaller group all have the same mark. Now repeat this argument with all but the first person. Combining these two facts, we find that all $n$ students have the same mark. By induction, all students have the same mark.* This is patently absurd, and you are undoubtedly ready to refute this by saying that Paul has a much lower mark than Mary. But you must find the mistake in the induction argument, not just in the conclusion.
HINT: The mistake is not $P(1)$, and $P(73)$ does imply $P(74)$.

## 2.A The Arithmetic of the Real Numbers

We outline the construction of the real numbers using infinite decimal expansions, including the necessary operations. These operations extend the familiar ones on the rational numbers and our definitions of the new operations on the reals will use the existing operations on the rationals. We focus on the order and addition of real numbers, and the additive inverse, i.e., the negative of a real number.

As discussed in the previous section, we will define a real number by using an infinite decimal expansion such as

$$\begin{aligned}
\frac{1}{3} &= 0.33333333333333333333333333333333333333333333333333\ldots \\
\frac{1}{4} &= 0.25000000000000000000000000000000000000000000000000\ldots \\
\sqrt{3} &= 1.73205080756887729352744634150587236694280525381038\ldots \\
\pi &= 3.141592653589793238462643383279502884197169399375105\ldots \\
e &= 2.718281828459045235360287471352662497757247093699959\ldots
\end{aligned}$$

and, in general,

$$x = a_0.a_1a_2a_3a_4a_5a_6a_7a_8a_9a_{10}a_{11}a_{12}a_{10}a_{11}a_{12}a_{13}a_{14}a_{15}a_{16}a_{17}a_{18}\cdots.$$

That is, an **infinite decimal expansion** is a function from $\{0\} \cup \mathbb{N}$ into $\mathbb{Z}$ so that for all $n \geq 1$, $n$ is mapped into $\{0,1,\ldots,9\}$. Such an expansion has, a priori, no connection to the rational numbers or to any kind of infinite series.

To relate infinite decimal expansions to our geometric idea of the real line, start with a line and mark two points on it; and call the left-hand one 0 and the right-hand one 1. Then we can easily construct points for every integer $\mathbb{Z}$, equally spaced along the line. Now divide each interval from an integer $n$ to $n+1$ into 10 equal pieces, marking the cuts as $n.1$, $n.2$, …, $n.9$. Proceed in this way, cutting each interval of length $10^{-k}$ into 10 equal intervals of length $10^{-k-1}$ and mark the endpoints by the corresponding number with $k+1$ decimals. In this way, all finite decimals are placed on the line.

Be warned that, by this construction, the point usually thought of as $-5/4$ will be marked $-2.75$, for example, because we think of this as $-2$ + .75. It turns out that some parts of the construction are simpler if we do this. After we have finished the construction, we will revert to the standard notation for negative decimals.

It seems clear that, given a number $a_0.a_1a_2a_3\ldots$, there will be a point on this line, call it $x$, with the property that for each positive integer $k$, $x$ lies in the interval between the two rational numbers $y = a_0.a_1\ldots a_k$ and $y + 10^{-k}$. For example,

$$3.141592653589 \leq \pi \leq 3.141592653590. \tag{2.A.1}$$

The decimal expansion of $x$ up to the $k$th decimal approximates $x$ to an accuracy of at least $10^{-k}$. We will use point of view to define the ordering on real numbers To make this precise, we'll have to define the ordering of real numbers.

Is there any point in distinguishing between the real number $x$ and its infinite decimal expansion? It might seem that every real number should have a unique infinite decimal expansion, but this is not quite true. What about the real numbers determined by the expansions $1.000000000\dots$ and $0.999999999\dots$? Call them 1 and $z$, respectively. Clearly these are *different* infinite decimal expansions. However, for each positive integer $k$, we expect

$$1 - 10^{-k} = 0.\underbrace{99999999999999}_{k} \leq z \leq 1.$$

Thus the difference between $z$ and 1 is arbitrarily small. It would create quite an unintuitive line if we decided to make $z$ and 1 different real numbers. To fit in with our intuition, we must agree that $z = 1$. That means that some real numbers (precisely all those numbers with a finite decimal expansion) have two different expansions, one ending in an infinite string of zeros, and the other ending with an infinite string of nines. For example, $0.12500\dots = 0.12499999\dots$.

This is an equivalence relation (see Exercise 1.3.B). Formally, each real number is an equivalence class of infinite decimal expansions given by the identification defined above. Each equivalence class will have one or two elements, depending on whether or not the number is divisible by some power of 10. The set of all real numbers is denoted by $\mathbb{R}$.

To recognize the rationals as a subset of the reals, we need a function that sends a fraction $a/b$ to an infinite decimal expansion. Making the identification described above, this function is injective, sending different fractions to different infinite decimal expansions (see Exercise 2.A.D). The rational numbers are distinguished among all real numbers by the fact that their decimal expansions are eventually periodic. See Theorems 1.C.1 and 1.C.3 in Chapter 1 if this is unfamiliar. In this section, we call this function $F$; afterwards, we will just think of $\mathbb{Q}$ as a subset of $\mathbb{R}$, i.e., identify $a/b$ with the associated real number.

What we need to do next is to extend the ordering and the addition and multiplication operations on $\mathbb{Q}$ to all of $\mathbb{R}$. The following theorem gives all of the properties that we expect these new operations to have. However, there are many details to check.

**2.A.2. Theorem.** *The relation $<$ on $\mathbb{R}$ satisfies:*

(1) *For any $a, b, c \in \mathbb{R}$, if $a < b$ and $b < c$, then $a < c$.*
(2) *For any $a, b \in \mathbb{R}$, exactly one of the following is true:*

  (1) *$a = b$,*
  (2) *$a < b$,*
  (3) *$b < a$.*

(4) *If $r, s \in \mathbb{Q}$, then $r < s$ if and only if $F(r) < F(s)$.*

*The operation + satisfies:*

(1) *For all* $a,b \in \mathbb{R}$, $a+b=b+a$.
(2) *For all* $a,b,c \in \mathbb{R}$, $(a+b)+c=a+(b+c)$.
(3) *There is an element* $0_{\mathbb{R}} \in \mathbb{R}$ *so that for all* $a \in \mathbb{R}$, $a+0_{\mathbb{R}}=a$.
(4) *For each* $a \in \mathbb{R}$, *there is an element, called* $-a$, *so that* $a+(-a)=0_{\mathbb{R}}$.
(5) *For all* $r,s \in \mathbb{Q}$, $F(r+s)=F(r)+F(s)$.

*The operation* $\times$ *satisfies:*

(6) *For all* $a,b \in \mathbb{R}$, $a \times b = b \times a$.
(7) *For all* $a,b,c \in \mathbb{R}$, $(a \times b) \times c = a \times (b \times c)$.
(8) *There is an element* $1_{\mathbb{R}} \in \mathbb{R}$ *so that for all* $a \in \mathbb{R}$, $a \times 1_{\mathbb{R}} = a$.
(9) *For each* $a \in \mathbb{R}\backslash\{0_{\mathbb{R}}\}$, *there is an element, called* $a^{-1}$, *in* $\mathbb{R}$ *so that* $a \times a^{-1} = 1_{\mathbb{R}}$.
(10) *For all* $r,s \in \mathbb{Q}$, $F(r \cdot s)=F(r) \times F(s)$.

*The two operations + and* $\times$ *together satisfy:*

(11) *For all* $a,b,c \in \mathbb{R}$, $a \times (b+c) = a \times b + a \times c$.

*The arithmetic operations relate to the order:*

(12) *For all* $a,b,c \in \mathbb{R}$, *if* $a > b$ *then* $a+c > b+c$.
(13) *For all* $a,b,c \in \mathbb{R}$, *if* $a > b$ *and* $c > 0$ *then* $ac > bc$.

*The Archimedean property is satisfied:*

(14) *For all* $a \in \mathbb{R}$, *if* $a \geq 0$ *and for all* $n \in \mathbb{N}$, $F(1/n) > a$, *then* $a = 0$.

There are many properties that follow easily from these. For example, $0_{\mathbb{R}}$ is unique and $F(0)=0_{\mathbb{R}}$, so we can identify $0_{\mathbb{R}}$ with 0. Similarly, $1_{\mathbb{R}}$ is unique and can be identified with $1 \in \mathbb{Q}$.

Instead of proving all the parts of this theorem, we explain how the ordering and addition are defined and sketch some of the arguments used to prove the required properties.

First, we have a built-in order on the line given by the placement of the points. This extends the natural order on the finite decimals. Notice that between any two distinct finite decimal numbers, there are (infinitely many) other finite decimal numbers. Now if $x$ and $y$ are distinct real numbers given by infinite decimal expansions, these expansions will differ at some finite point. This enables us to find finite decimals in between them. Because we know how an infinite decimal expansion should compare to its finite decimal approximants [using equations such as (2.A.1)], we can determine which of $x$ or $y$ is larger. For example, if

$$x = 2.7342118284590452354000064338325028841971693993\ldots \quad (2.A.3)$$

$$y = 2.7342118284590452353999928747135224977572470936\ldots \quad (2.A.4)$$

then $y < x$ because

$$y < 2.7342118284590452353999993 < 2.7342118284590452354000000 < x.$$

In fact, because their decimal expansions come from different real numbers, knowing that, in the first digit where they differ, that digit of $y$ is less than the corresponding digit of $x$ forces $y < x$.

We should also verify the order properties in Theorem 2.A.2.

Second, we should extend the arithmetic properties of the rational numbers to all real numbers—namely addition, multiplication, and their inverse operations—and verify all of the field axioms. This is done by making all the operations consistent with the order. For example, if $x = x_0, x_1 x_2 \dots$ and $y = y_0.y_1 y_2 \dots$ are real numbers and $k$ is a positive integer, then we have the finite decimal approximants

$$a_k = x_0.x_1 \dots x_k \le x \le a_k + 10^{-k} \quad \text{and} \quad b_k = y_0.y_1 \dots y_k \le y \le b_k + 10^{-k},$$

and so we want to have

$$a_k + b_k \le x + y \le a_k + b_k + 2 \cdot 10^{-k}. \tag{2.A.5}$$

Since the lefthand and righthand sum use only rational numbers, we know what they are, and this determines the sum $x + y$ to an accuracy of $2 \cdot 10^{-k}$, for each $k$.

However, computing the exact sum of two infinite decimals is more subtle. The first digit of $x + y$ may not be determined exactly after any fixed finite number of steps, even though the sum can be determined to any desired accuracy. To see why this is the case, consider

$$\begin{aligned} x &= 0.\overbrace{999999\dots999999}^{10^{15}\text{ nines}}\overbrace{0123456789\dots0123456789}^{10^4\text{ repetitions}}31415\dots \\ y &= 0.\underbrace{999999\dots999999}_{10^{15}\text{ nines}}\underbrace{9876543210\dots9876543210}_{10^4\text{ repetitions}}a9066\dots. \end{aligned}$$

When we add $x + y$ using the first $k$ decimal digits for any $k \le 10^{15}$, we obtain

$$1.\overbrace{999999\dots999999}^{k-1\text{ nines}}8 \le x + y \le 2.\overbrace{000000\dots000000}^{k\text{ zeros}}.$$

Taking $k = 10^{15}$ does not determine if the first digit of $x + y$ is 1 or 2, even though we know the sum to an accuracy of $2 \cdot 10^{-10^{15}} = 2/10^{1,000,000,000,000,000}$. When we proceed with the computation using one more digit, we obtain

$$1.\overbrace{999999\dots999999}^{10^{15}-1\text{ nines}}89 \le x + y \le 1.\overbrace{999999\dots999999}^{10^{15}-1\text{ nines}}91.$$

All of a sudden, not only is the first digit determined, but so are the next $10^{15} - 1$ digits.

A new period of uncertainty now occurs, again because of the problem that a long string of nines can *roll over* to a string of zeros like the odometer in a car. After using another $10^5$ digits, we obtain a different result depending on whether $a \le 4$, $a = 5$ or 6, or $a \ge 7$. When $a = 4$, we get

$$1.\overbrace{9999\ldots9999}^{10^{15}-1 \text{ nines}}8\overbrace{9999\ldots9999}^{10^4 \text{ nines}}7 \le x+y \le 1.\overbrace{9999\ldots9999}^{10^{15}-1 \text{ nines}}8\overbrace{9999\ldots9999}^{10^4 \text{ nines}}9.$$

So the digits are now determined for another $10^4+1$ places. When $a = 7$, we obtain

$$1.\overbrace{9999\ldots9999}^{10^{15}-1 \text{ nines}}9\overbrace{000\ldots0000}^{10^4 \text{ zeros}}0 \le x+y \le 1.\overbrace{9999\ldots9999}^{10^{15}-1 \text{ nines}}9\overbrace{000\ldots0000}^{10^4 \text{ zeros}}2.$$

Again, the next $10^4+1$ digits are now determined. However, when $a = 5$ or $a = 6$, these digits of the sum are still ambiguous. For $a = 5$, we have

$$1.\overbrace{9999\ldots9999}^{10^{15}-1 \text{ nines}}8\overbrace{9999\ldots9999}^{10^4 \text{ nines}}8 \le x+y \le 1.\overbrace{9999\ldots9999}^{10^{15}-1 \text{ nines}}9\overbrace{000\ldots0000}^{10^4 \text{ zeros}}0,$$

so the $10^{15}$-th decimal digit is still undetermined.

The important thing to recognize is that these difficulties are not a serious impediment to defining the sum of two real numbers using infinite decimals. Suppose that, no matter how large $k$ is, looking at the first $k$ digits of $x$ and $y$ does not tell us if the first digit of $x+y$ is a 1 or a 2. In terms of Equation (2.A.5), this means that, for each $k$, the interval from $a_k+b_k$ to $a_k+b_k+2\cdot10^{-k}$ contains 2. Since the length of the intervals goes to zero, it seems intuitively clear that the only real number in all of these intervals is 2.

In general, by considering *all* of the digits of $x$ and $y$, we can write down a definition of $x+y$ as an infinite decimal. We may not be able to specify an algorithm to compute the sum, but then we cannot represent all of even one infinite decimal expansion in a computer either.

In real life, knowing the sum to, say, within $2\cdot10^{-15}$ is much the same as knowing it to 15 decimal places (in fact, marginally better). So we are content, on both theoretical and practical grounds, that we have an acceptable working model of addition.

Because of our non-standard definition of the infinite decimal expansions of negative numbers, constructing the negative of an infinite decimal is not just a matter of flipping the sign. If $x = x_0.x_1x_2\ldots$ represents a real number, we can define

$$-x = (-x_0-1).(9-x_1)(9-x_2)\ldots.$$

Then one can see that

$$x+(-x) = -1.999\ldots.$$

The right hand side is one of the two infinite decimal expansions for 0, so $-x$ is an additive inverse for $x$.

The issues are similar for the other arithmetic operations: multiplication and multiplicative inverses. It is crucial that these operations are consistent with order, as this means that they are also continuous (respect limits). Carrying out all the details of this program is tedious but not especially difficult.

The key points of this section are that we can define real numbers as infinite decimal expansions (with some identifications) and that we can define the order and all the field operations in terms of infinite decimals. Moreover, the result fits our intuitive picture of the real line, so we have the order and arithmetic properties that we expect.

## Exercises for Section 2.A

**A.** If $x \neq y$, explain an algorithm to decide if $x < y$ or $y < x$. Does your method break down if $x = 0.9999\ldots$ and $y = 1.0000\ldots$?

**B.** If $a < b$ and $x < y$, is $ax < by$? What additional order hypotheses make the conclusion correct?

**C.** Verify Property 5 of Theorem 2.A.2, i.e., show that if $p, q \in \mathbb{Q}$, then $F(p) + F(q) = F(p+q)$.

**D.** Define precisely the infinite decimal expansion associated to a fraction $a/b \in \mathbb{Q}$. Show that this function is one-to-one as a map from $\mathbb{Q}$ into the real numbers.

## 3.A The Number *e*

Recall from calculus the formula

$$e = 1 + 1 + \frac{1}{2} + \frac{1}{6} + \frac{1}{24} + \cdots = \sum_{k=0}^{\infty} \frac{1}{k!} \approx 2.7182818285\ldots$$

If you don't recall this, you can review the section on Taylor series in your calculus book or study Example 10.1.4 later in this book. With $a_k = 1/k!$, we obtain

$$\lim_{k\to\infty} \frac{a_{k+1}}{a_k} = \lim_{k\to\infty} \frac{1}{k+1} = 0.$$

Thus this series converges by the Ratio Test. The limit is called $e$.

There are other ways to compute $e$. We give one such well-known formula and verify that it has the same limit.

**3.A.1. PROPOSITION.** *Consider the sequences*

$$b_n = \Big(1 + \frac{1}{n}\Big)^n \quad \textit{and} \quad c_n = \Big(1 + \frac{1}{n}\Big)^{n+1}.$$

*These sequences are monotone increasing and decreasing, respectively, and*

$$\lim_{n\to\infty} \Big(1 + \frac{1}{n}\Big)^n = \lim_{n\to\infty} \Big(1 + \frac{1}{n}\Big)^{n+1} = e.$$

**PROOF.** We need the inequality:

$$(1+x)^n > (1+nx) \quad \text{for} \quad x > -1 \quad \text{and} \quad n \geq 1.$$

To see this, let $f(t) = (1+t)^n - (1+nt)$. Notice that $f(0) = 0$ and

$$f'(t) = n\big((1+t)^{n-1} - 1\big) \quad \text{for} \quad t > -1.$$

Thus $f$ is decreasing on $(-1,0]$ and increasing on $[0,\infty)$, and so takes its minimum value at 0.

As $c_n = \big(1 + \frac{1}{n}\big)b_n$, we have $b_n < c_n$ for $n \geq 1$. Now compute

$$\frac{b_{n+1}}{c_n} = \left(\frac{1 + \frac{1}{n+1}}{1 + \frac{1}{n}}\right)^{n+1} = \left(\frac{n^2 + 2n}{n^2 + 2n + 1}\right)^{n+1}$$
$$= \left(1 - \frac{1}{(n+1)^2}\right)^{n+1} > 1 - \frac{n+1}{(n+1)^2} = \frac{n}{n+1}.$$

Therefore,

$$b_{n+1} > \frac{n}{n+1} c_n = b_n.$$

Similarly, inverting the preceding identity,

$$\frac{c_n}{b_{n+1}} = \left(1+\frac{1}{n^2+2n}\right)^{n+1} > 1+\frac{n+1}{n^2+2n} > \frac{n+2}{n+1}$$

whence $c_{n+1} = \dfrac{n+2}{n+1} b_{n+1} < c_n$.

It follows that $(b_n)_{n=1}^{\infty}$ is a monotone increasing sequence bounded above by $c_1$. Thus it converges to some limit $L$ by Theorem 2.6.1. Similarly, $(c_n)_{n=1}^{\infty}$ is a monotone decreasing sequence bounded below by $b_1$, and

$$\lim_{n\to\infty} c_n = \lim_{n\to\infty} \left(1+\tfrac{1}{n}\right) b_n = L.$$

Next we use the Binomial Theorem to estimate the terms $b_n$:

$$\begin{aligned} b_n = \left(1+\frac{1}{n}\right)^n &= \sum_{k=0}^{n} \binom{n}{k} \frac{1}{n^k} \\ &= \sum_{k=0}^{n} \frac{1}{k!} \frac{n}{n} \frac{(n-1)}{n} \cdots \frac{(n-k+1)}{n} \\ &= \sum_{k=0}^{n} \frac{1}{k!} \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right) \cdots \left(1-\frac{k-1}{n}\right). \end{aligned}$$

It follows that $b_n < s_n = \sum_{k=0}^{n} 1/k!$, and therefore

$$L = \lim_{n\to\infty} b_n \le \lim_{n\to\infty} s_n = e.$$

On the other hand, this also provides the lower bound. If $p$ is a fixed integer, then

$$\begin{aligned} b_n &> \sum_{k=0}^{p} \frac{1}{k!} \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right) \cdots \left(1-\frac{k-1}{n}\right) \\ &> \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right) \cdots \left(1-\frac{p-1}{n}\right) s_p. \end{aligned}$$

Thus

$$L = \lim_{n\to\infty} b_n \ge \lim_{n\to\infty} \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right) \cdots \left(1-\frac{p-1}{n}\right) s_p = s_p$$

for all $p \ge 1$. So

$$L \ge \lim_{p\to\infty} s_p = e.$$

We obtain

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

■

Infinite series have unexpected uses. We use the series for $e$ to prove that it is not a rational number.

**3.A.2. THEOREM.** *e is irrational.*

**PROOF.** Suppose that $e$ is rational and can be written as $e = p/q$ in lowest terms, where $p$ and $q$ are positive integers. We look for a contradiction. Compute:

$$(q-1)!p = q!\frac{p}{q} = q!e = q!\sum_{k=0}^{\infty}\frac{1}{k!} = \sum_{k=0}^{q}\frac{q!}{k!} + \sum_{k=q+1}^{\infty}\frac{q!}{k!}.$$

Rearranging, we get

$$\sum_{k=q+1}^{\infty}\frac{q!}{k!} = (q-1)!p - \sum_{k=1}^{q}\frac{q!}{k!}.$$

Since both terms on the right-hand are integers, the left-hand side must also be an integer. However, using the properties of limits, we have

$$\begin{aligned}0 < \sum_{k=q+1}^{\infty}\frac{q!}{k!} &= \frac{1}{q+1} + \frac{1}{(q+1)(q+2)} + \frac{1}{(q+1)(q+2)(q+3)} + \cdots \\ &< \frac{1}{q+1} + \frac{1}{(q+1)^2} + \frac{1}{(q+1)^3} + \cdots \\ &= \sum_{n\geq 1}\Big(\frac{1}{q+1}\Big)^n = \frac{1/(q+1)}{1-1/(q+1)} = \frac{1}{q} \leq 1.\end{aligned}$$

Thus $0 < \sum\limits_{k=q+1}^{\infty}\frac{q!}{k!} < 1$, contradicting the conclusion that this is an integer. ■

## Exercises for Section 3.A

**A.** (a) Show that $\sum_{k=0}^{\infty} x^k/k!$ converges for each $x \in \mathbb{R}$. The limit is called $e^x$.
(b) Show that $\lim\limits_{n\to\infty}(1+x/n)^n = e^x$. HINT: Compare with the series in part (a).

**B.** Compute the sum of the series $\sum_{n=1}^{\infty}\frac{1}{n!}(1^2+2^2+\cdots+n^2)$.
HINT: Express $\sum_{k=1}^{n} k^2$ in the form $An(n-1)(n-2)+Bn(n-1)+Cn+D$.

**C.** Evaluate $\sum_{n=0}^{\infty}\frac{1}{(2n+1)!}$. HINT: This is the difference of two known series.

**D.** Evaluate $\lim\limits_{n\to\infty}\big(1+\frac{1}{n}\big)^{n^2}\big(1+\frac{1}{n+1}\big)^{-(n+1)^2}$.

**E.** Decide whether $\sum_{n=1}^{\infty} n!/n^n$ converges.

**F.** Combine the two previous exercises to compute $\lim\limits_{n\to\infty}\sqrt[n]{n!}/n$.

**G.** Decide if $\sum_{n=1}^{\infty}(\log n)^{-\log n}$ converges. HINT: Show that $(\log n)^{\log n} = n^{\log\log n}$.

**H.** Show that $\sqrt{2}+e$ is irrational.

**I.** (a) If $a$ is rational, find $\lim\limits_{n\to\infty}\sin(n!\pi a)$.
(b) Show that $n!\sum_{k=n+1}^{\infty} 1/k! < 1/n$.
(c) If $a = e/2$, compute $\lim\limits_{n\to\infty}\sin((2n)!\pi a)$.

## 3.B Summation by Parts

This section contains another convergence test. The proof utilizes a rearrangement technique called **summation by parts**, which is analogous to integration by parts.

### 3.B.1. Summation by Parts Lemma.

*Suppose $(x_n)$ and $(y_n)$ are sequences of real numbers and define $X_n = \sum_{k=1}^{n} x_k$ and $Y_n = \sum_{k=1}^{n} y_k$. Then*

$$\sum_{n=1}^{m} x_n Y_n + \sum_{n=1}^{m} X_n y_{n+1} = X_m Y_{m+1}.$$

**Proof.** The argument is essentially an exercise in reindexing summations. Let $X_0 = 0$ and notice that the left-hand side (LHS) equals

$$\begin{aligned} \text{LHS} &= \sum_{n=1}^{m} (X_n - X_{n-1}) Y_n + \sum_{n=1}^{m} X_n (Y_{n+1} - Y_n) \\ &= \sum_{n=1}^{m} X_n Y_n - \sum_{n=1}^{m} X_{n-1} Y_n + \sum_{n=1}^{m} X_n Y_{n+1} - \sum_{n=1}^{m} X_n Y_n. \\ &= -X_0 Y_1 + X_m Y_{m+1} = X_m Y_{m+1}. \end{aligned}$$

■

Thus, provided that $\lim_{m\to\infty} X_m Y_{m+1}$ exists, the two series $\sum x_n Y_n$ and $\sum X_n y_n$ either both converge or both diverge.

### 3.B.2. Dirichlet's Test.

*Suppose that $(a_n)_{n=1}^{\infty}$ is a sequence of real numbers with bounded partial sums:*

$$\Big|\sum_{k=1}^{n} a_k\Big| \le M < \infty \quad \textit{for all} \quad n \ge 1.$$

*If $(b_n)_{n=1}^{\infty}$ is a sequence of positive numbers decreasing monotonically to 0, then the series $\sum_{n=1}^{\infty} a_n b_n$ converges. Moreover, $\big|\sum_{n=1}^{\infty} a_n b_n\big| \le 2Mb_1$.*

**Proof.** We use the Summation by Parts Lemma to rewrite $a_n b_n$. Let $x_n = a_n$ for all $n$; and set $y_1 = b_1$ and $y_n = b_n - b_{n-1}$ for $n > 1$. Define $X_n$ and $Y_n$ as in the lemma. Note that $y_n < 0$ for $n > 1$, and that there is a telescoping sum

$$Y_n = b_1 + (b_2 - b_1) + \cdots + (b_n - b_{n-1}) = b_n.$$

Hence $a_n b_n = x_n Y_n$.

Notice that $|X_n| = \big|\sum_{k=1}^{n} a_k\big| \le M$ for all $n$. Since $|X_nY_{n+1}| \le M|b_{n+1}|$, the Squeeze Theorem shows that $\lim_{n\to\infty} X_nY_{n+1} = 0$. Furthermore,

$$\sum_{k=1}^{n} |X_k y_{k+1}| \le \sum_{k=1}^{n} M|y_{k+1}| = M\big(b_1 + \sum_{k=1}^{n} b_k - b_{k+1}\big) = M(2b_1 - b_{n+1}) \le 2Mb_1.$$

Thus $\sum_{k=1}^{\infty} X_k y_{k+1}$ converges absolutely. Using the Summation by Parts Lemma, convergence follows from

$$\sum_{n=1}^{\infty} a_n b_n = \lim_{m\to\infty} \sum_{n=1}^{m} x_n Y_n = \lim_{m\to\infty} X_m Y_m - \sum_{n=1}^{m} X_n y_{n+1} = -\sum_{k=1}^{\infty} X_k y_{k+1}.$$

Moreover, $\big|\sum_{n=1}^{\infty} a_n b_n\big| \le 2Mb_1$. ■

**3.B.3. EXAMPLE.** Consider the series $\sum_{n=1}^{\infty} \frac{\sin n\theta}{n}$ for $-\pi \le \theta \le \pi$. At the points $\theta = 0$, $\pi$, and $2\pi$, the series is 0. For $\theta$ in $(0, 2\pi)\backslash\{\pi\}$, we will show that this series converges conditionally.

Let $a_n = \sin n\theta$ and $b_n = \frac{1}{n}$. To evaluate the partial sums of the $a_n$s, use the trigonometric identities: $\cos(k \pm \frac{1}{2})\theta = \cos k\theta \cos\frac{\theta}{2} \mp \sin k\theta \sin\frac{\theta}{2}$. Thus

$$2\sin k\theta \sin\tfrac{\theta}{2} = \cos(k-\tfrac{1}{2})\theta - \cos(k+\tfrac{1}{2})\theta.$$

Therefore, we obtain a telescoping sum

$$\begin{aligned}\sum_{k=1}^{n} a_k = \sum_{k=1}^{n} \sin k\theta &= \frac{1}{2\sin\frac{\theta}{2}} \sum_{k=1}^{n} \cos(k-\tfrac{1}{2})\theta - \cos(k+\tfrac{1}{2})\theta \\ &= \frac{1}{2}\csc\tfrac{\theta}{2}\big(\cos\tfrac{\theta}{2} - \cos(n+\tfrac{1}{2})\theta\big).\end{aligned}$$

Consequently,

$$\Big|\sum_{k=1}^{n} a_k\Big| \le \csc\tfrac{1}{2}\theta < \infty \quad \text{for all} \quad n \ge 1.$$

So Dirichlet's Test applies and the series converges.

We will show that convergence is conditional by comparing the terms with the harmonic series. To see this, we may suppose that $0 < \theta \le \pi/2$ because $|\sin(\pm k\theta)| = |\sin(\pm k(\pi - \theta))|$. Out of every two consecutive terms $\sin k\theta$, at most one term has absolute value less than $|\sin\theta|/2$. Indeed, if there is an integer multiple of $\pi$ so that $|m\pi - 2k\theta| < \theta/2$, then

$$\frac{\theta}{2} < |m\pi - (2k-1)\theta| = |(m - 2k\theta) + \theta| < \frac{3\theta}{2} < \pi - \frac{\theta}{2}.$$

We conclude that either $|\sin(2k\theta)| \geq |\sin(\theta/2)|$ or $|\sin((2k-1)\theta)| \geq |\sin(\theta/2)|$. Therefore

$$\frac{|\sin(2k-1)\theta|}{2k-1} + \frac{|\sin 2k\theta|}{2k} \geq \frac{|\sin\theta/2|}{2k}.$$

We conclude that

$$\sum_{k=1}^{2n} \frac{|\sin k\theta|}{k} \geq \frac{|\sin\theta|/2}{2} \sum_{k=1}^{n} \frac{1}{k}.$$

By comparison with a multiple of the harmonic series, this series diverges.

## Exercises for Section 3.B

**A.** Show that the Alternating Series Test is a special case of Dirichlet's Test.

**B.** Use summation by parts to prove **Abel's Test**: Suppose that $\sum_{n=1}^{\infty} a_n$ converges and $(b_n)$ is a monotonic convergent sequence. Show that $\sum_{n=1}^{\infty} a_n b_n$ converges.

**C.** Determine the values of $\theta$ for which the series $\sum_{n=1}^{\infty} \frac{e^{in\theta}}{\log(n+1)}$ converges.

**D.** Let $a_n = (-1)^k/n$ for $(k-1)^2 < n \leq k^2$ and $k \geq 1$. Decide if the series $\sum_{n=1}^{\infty} a_n$ converges.

## 5.A Bounded Variation

We introduce a family of functions which are closely related to monotone functions, and share many of their good properties. These functions play a central role in Riemann–Stieltjes integration, which we deal with in Section 6.D.

**5.A.1. DEFINITION.** Given a function $f : [a,b] \to \mathbb{R}$ and a partition of $[a,b]$, say $P = \{a = x_0 < x_1 < \cdots < x_n = b\}$, then the **variation** of $f$ over $P$ is

$$V(f,P) = \sum_{i=1}^{n} |f(x_i) - f(x_{i-1})|.$$

The **(total) variation** of $f$ on $[a,b]$ is

$$V_a^b f = \sup\{V(f,P) : P \text{ a partition of } [a,b]\}.$$

We say $f$ is of **bounded variation** on $[a,b]$ if $V_a^b f$ is finite.

**5.A.2. REMARKS.** Fix a partition $P$ and a point $c \notin P$, say $c \in (x_{j-1}, x_j)$. Then

$$|f(x_j) - f(x_{j-1})| \le |f(x_j) - f(c)| + |f(c) - f(x_{j-1})|.$$

Therefore $V(f,P) \le V(f, P \cup \{c\})$. It follows by induction that if $R$ is a refinement of $P$, then $V(f,P) \le V(f,R)$.

Another easy observation using $P = \{a,b\}$ is that $|f(b) - f(a)| \le V_a^b f$.

**5.A.3. EXAMPLES.**

(1) If $f : [a,b] \to \mathbb{R}$ is monotone increasing, and $P = \{a = x_0 < x_1 < \cdots < x_n = b\}$, then we get a telescoping sum

$$V(f,P) = \sum_{i=1}^{n} f(x_i) - f(x_{i-1}) = f(b) - f(a).$$

Hence $V_a^b f = f(b) - f(a) < \infty$. So monotone functions have bounded variation.

(2) If $f$ is Lipschitz with constant $C$ on $[a,b]$, and $P = \{a = x_0 < x_1 < \cdots < x_n = b\}$, then

$$V(f,P) = \sum_{i=1}^{n} |f(x_i) - f(x_{i-1})| \le \sum_{i=1}^{n} C(x_i - x_{i-1}) = C(b-a).$$

Therefore $f$ has bounded variation, and $V_a^b f \le C(b-a)$.

(3) The functions of bounded variation form a vector space $BV[a,b]$. Indeed, if $f$ and $g$ have bounded variation on $[a,b]$, and $c, d \in \mathbb{R}$, then

$$\begin{aligned} V(cf+dg,P) &= \sum_{i=1}^{n} |(cf+dg)(x_i)-(cf+dg)(x_{i-1})| \\ &\leq |c|\sum_{i=1}^{n}|f(x_i)-f(x_{i-1})| + |d|\sum_{i=1}^{n}|g(x_i)-g(x_{i-1})| \\ &= |c|V(f,P)+|d|V(g,P). \end{aligned}$$

Hence $V_a^b(cf+dg) \leq |c|V_a^b f + |d|V_a^b g < \infty$.

(4) Define $g:[0,1/\pi]\to\mathbb{R}$ by $g(0)=0$ and $g(x)=\sin(1/x)$ for $x>0$. We claim that $V_0^{1/\pi}g=+\infty$. Fix $n>1$ and let $P_n$ be the partition $\{0,\frac{2}{\pi(2n)},\frac{2}{\pi(2n-1)},\cdots,\frac{2}{3\pi},\frac{2}{2\pi}\}$. As $g$ takes values $0,-1,0,1,0,\ldots,-1,0$ on the elements of $P_n$, for each interval $[x_{j-1},x_j]$, $|g(x_j)-g(x_{j-1})|=1$. Thus, $V(f,P_n)=2n$ and $V_0^{1/\pi}f=+\infty$.

(5) Define $h:[0,1/\pi]\to\mathbb{R}$ by $h(0)=0$ and $h(x)=x\sin(1/x)$ for $x>0$. Even this $h$ is not of bounded variation. Using the partition $P$ above, some work will show that $V(f,P)$ is a multiple of $\sum_{k=1}^{n}1/k$, which diverges.

(6) Finally, we define $k:[0,1/\pi]\to\mathbb{R}$ by $k(0)=0$ and $k(x)=x^2\sin(1/x)$ for $x>0$. Observe that $k$ has a bounded derivative, $|k'(x)|=|2x\sin(1/x)-\cos(1/x)|\leq 3$. Thus $k$ is Lipschitz and therefore has bounded variation.

**5.A.4. LEMMA.** *If $a<b<c$ and $f:[a,c]\to\mathbb{R}$ is given, then*

$$V_a^c f = V_a^b f + V_b^c f.$$

**PROOF.** First, notice that if $P$ is a partition of $[a,b]$ and $Q$ is a partition of $[b,c]$, then,

$$V(f,P\cup Q)=V(f,P)+V(f,Q).$$

Let $P$ be a partition of $[a,c]$. Let $P'=P\cup\{b\}$, $P_1=P'\cap[a,b]$, and $P_2=P'\cap[b,c]$. Then $P_1$ is a partition of $[a,b]$, $P_2$ is a partition of $[b,c]$, and $P'=P_1\cup P_2$. Thus,

$$V(f,P)\leq V(f,P')=V(f,P_1)+V(f,P_2)\leq V_a^b f+V_b^c f.$$

Taking the supremum over all partitions $P$, we have $V_a^c f\leq V_a^b f+V_b^c f$.

For the reverse inequality, let $\varepsilon>0$. There is a partition $P_1$ of $[a,b]$ so that $V(f,P_1)\geq V_a^b f-\varepsilon/2$. and a partition $P_2$ of $[b,c]$ so that $V(f,P_2)\geq V_b^c f-\varepsilon/2$. Thus, if $P=P_1\cup P_2$, then the remark above shows that

$$V_a^c f\geq V(f,P)=V(f,P_1)+V(f,P_2)\geq V_a^b f+V_b^c f-\varepsilon$$

As $\varepsilon$ is arbitrary, this shows $V_a^c f\geq V_a^b f+V_b^c f$. ■

The key result about functions of bounded variation is the following characterization.

**5.A.5. THEOREM.** *If $f : [a,b] \to \mathbb{R}$ is of bounded variation on $[a,b]$, then there are increasing functions $g,h : [a,b] \to \mathbb{R}$ so that $f = g - h$. Moreover, $g$ and $h$ can be chosen so that, for each $x \in [a,b]$, $V_a^x f = V_a^x g + V_a^x h$.*

**PROOF.** Define $v(x) = V_a^x f$. Notice that by the previous lemma, if $a \le x < y \le b$, then $v(y) - v(x) = V_x^y f \ge 0$. Define $g = (v + f)/2$ and $h = (v - f)/2$. It is evident that $f(x) = g(x) - h(x)$. If $x < y$, then by Remark 5.A.2,

$$g(y) - g(x) = \frac{1}{2}(v(y) + f(y) - v(x)) \ge \frac{1}{2}(V_x^y f - |f(y) - f(x)|) \ge 0.$$

So $g$ is increasing. Similarly, $h$ is increasing. Finally, for $x \in [a,b]$,

$$V_a^x g + V_a^x h = g(x) + h(x) - g(a) - h(a) = v(x) - v(a) = V_a^x f.$$

■

The functions $g$ and $h$ are, in a suitable sense, minimal; see Exercise 5.A.B.

We obtain some immediate consequences by transferring results about monotone functions to functions of bounded variation. If we combine Proposition 5.7.2, Corollary 5.7.3 and Theorem 5.7.5, we obtain:

**5.A.6. COROLLARY.** *If $f$ is a function of bounded variation on $[a,b]$, then one-sided limits of $f$ exist at each point $c \in (a,b)$. So the only discontinuities of $f$ are jump discontinuities, and there are only countably many.*

## Exercises for Section 5.A

**A.** Define $h : [0, 1/\pi] \to \mathbb{R}$ by $h(0) = 0$ and $h(x) = x \sin(1/x)$ for $x > 0$. Prove that $h$ is not of bounded variation on $[0, 1/\pi]$.

**B.** Suppose that $f : [a,b] \to \mathbb{R}$ is of bounded variation on $[a,b]$ and $f = k - l$ where $k,l : [a,b] \to \mathbb{R}$ are increasing and $k(a) = -l(a) = f(a)/2$. If $g$ and $h$ are the functions from Thereom 5.A.5, then show that $k - g$ and $l - h$ are increasing.

**C.** Show that if $f : [a,b] \to \mathbb{R}$ is continuous and of bounded variation, then $f = g - h$ where $g,h$ are continuous and monotone increasing on $[a,b]$. HINT: Show that for $f$ continuous, $v(x) = V_a^x f$ is continuous.

**D.** If $f : [a,b] \to \mathbb{R}$ is differentiable and there is a constant $C > 0$ so that $|f'(x)| \le C$ for all $x \in (a,b)$, then $f$ is of bounded variation. HINT: use the Mean Value Theorem.

**E.** Suppose $g$ is of bounded variation on $[a,b]$ and $\varphi : [c,d] \to [a,b]$ is increasing, continuous, and onto. Show that $G = g \circ \varphi$ has bounded variation, and relate $V_a^b g$ and $V_c^d G$ by a formula.

**F.** (a) If $f : [a,b] \to \mathbb{R}$ has bounded variation, show that $|f|$ also has bounded variation
(b) Show that if $f$ and $g$ have bounded variation on $[a,b]$, then $h(x) = \max\{f(x), g(x)\}$ and $k(x) = \min\{f(x), g(x)\}$ have bounded variation. HINT: use (a) and $f \pm g \in BV[a,b]$.

**G.** If $f$ is Riemann integrable on $[a,b]$, define $F(x) = \int_a^x f(t)\,dt$. Prove $F$ has bounded variation.

**H.** Fix $a,b > 0$. Define $h : [0, 1/\pi] \to \mathbb{R}$ by $h(0) = 0$ and $h(x) = x^a \sin(x^{-b})$ for $x > 0$. Prove that $h$ has bounded variation on $[0, 1/\pi]$ if and only if $a \ge b + 1$. HINT: compute the derivative.

**I.** Suppose that $f_n$ are functions on $[a,b]$ with $V_a^b f_n \le M$ for all $n \ge 1$. Show that if the limit $f(x) = \lim_{n\to\infty} f_n(x)$ exists for all $x \in [a,b]$, then $f$ also has bounded variation.

## 6.A Wallis's Product and Stirling's Formula

In this section, we will obtain Stirling's formula, an elegant asymptotic formula for $n!$, using only basic calculus. However, to get a sharp result, the estimates must be done quite carefully. By a **sharp inequality**, we mean an inequality that cannot be improved. For example, $|\sin(x)/x| \le 1$ for $x > 0$ is sharp but $\tan^{-1}(x) < 2$ is not. These estimates lead us to a general method of numerical integration, which we develop in the next section.

We estimate $n!$ by approximating the integral of $\log x$ using the **trapezoidal rule**, a modification of Riemann sums. First verify that $\int \log x\,dx = x\log x - x$ by differentiating. Notice that the second derivative of $\log x$ is $-x^{-2}$, which is negative for all $x > 0$. So the graph of $\log x$ is curving downward (i.e., this function is concave). Rather than using the rectangular approximants used in Riemann sums, we can do significantly better by using a trapezoid. In other words, we approximate the curve $\log x$ for $x$ between $k-1$ and $k$ by the straight line segment connecting $(k-1, \log(k-1))$ to $(k, \log k)$. By the convexity, this line lies strictly below $\log x$ and thus has a smaller area.

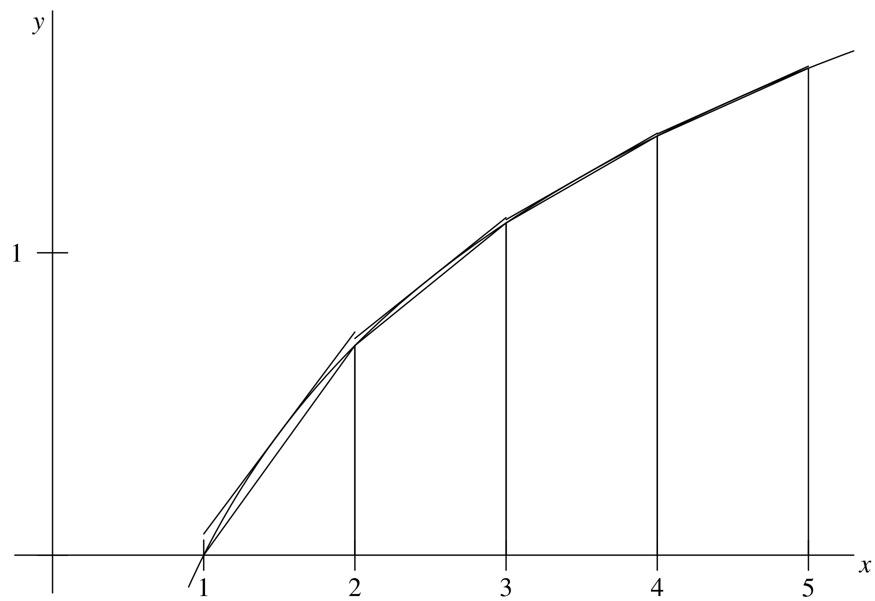


**FIG. 6.1** The graph of $y = \log x$ with the tangent approximation above and the trapezoidal approximation below.

The area of a trapezoid is the base times the average height. Therefore

$$\frac{\log(k-1) + \log k}{2} < \int_{k-1}^{k} \log x\,dx.$$

Sum this from 2 to $n$ to obtain

$$\sum_{k=2}^{n-1} \log k + \tfrac{1}{2}\log n < \int_1^n \log x\,dx = n\log n - (n-1).$$

This may be rearranged to compute the error as

$$E_n := (n+\tfrac{1}{2})\log n - (n-1) - \log n!.$$

To bound this error, let us estimate $\int_{k-1}^{k} \log x\,dx$ from above by another trapezoid, as shown in Figure 6.1. The tangent line to $\log x$ at $x = k - \frac{1}{2}$ lies strictly above the curve because $\log x$ is concave. This yields a trapezoid with an average height of $\log(k-\frac{1}{2})$. Therefore,

$$\begin{aligned}\varepsilon_k &:= \int_{k-1}^{k} \log x\,dx - \frac{\log(k-1)+\log k}{2} \\ &< \log(k-\tfrac{1}{2}) - \frac{1}{2}\log(k-1)k \\ &= \frac{1}{2}\log\frac{(k-\frac{1}{2})^2}{k^2-k} = \frac{1}{2}\log\big(1+\tfrac{1}{4(k^2-k)}\big).\end{aligned}$$

Now for $h > 0$,

$$\log(1+h) = \int_1^{1+h} 1/x\,dx < \int_1^{1+h} 1\,dx = h.$$

Using this to simplify our formula for $\varepsilon_k$, we obtain

$$\varepsilon_k < \frac{1}{8(k^2-k)} = \frac{1}{8}\Big(\frac{1}{k-1} - \frac{1}{k}\Big).$$

Observe that this produces a telescoping sum

$$E_n = \sum_{k=2}^{n} \varepsilon_k < \frac{1}{8}\sum_{k=2}^{n}\Big(\frac{1}{k-1} - \frac{1}{k}\Big) = \frac{1}{8}\big(1-\frac{1}{n}\big).$$

Consequently, $E_n$ is a monotone increasing sequence that is bounded above by $1/8$. Therefore, $\lim_{n\to\infty} E_n = E$ exists.

To put this in the desired form, note that

$$\log n! + n - (n+\tfrac{1}{2})\log n = 1 - E_n.$$

Exponentiating, we obtain

$$\lim_{n\to\infty} \frac{n!}{n^n\sqrt{n}e^{-n}} = e^{1-E}. \tag{6.A.1}$$

Rearranging and using only the estimate $0 \le E \le 1/8$, we have the useful estimate known as **Stirling's inequality**:

$$e^{7/8}\left(\frac{n}{e}\right)^n\sqrt{n} < n! < e\left(\frac{n}{e}\right)^n\sqrt{n}.$$

To get Stirling's formula, we must evaluate $e^{1-E}$ exactly. The first step is an exercise in integration that leads to a useful formula for $\pi$.

**6.A.2. EXAMPLE.** To begin, we wish to compute $I_n = \int_0^\pi \sin^n x\,dx$. We use integration by parts and induction. If $n \ge 2$,

$$\begin{aligned} I_n &= \int_0^\pi \sin^n x\,dx = \int_0^\pi \sin^{n-1} x \sin x\,dx \\ &= -\sin^{n-1} x \cos x\Big|_0^\pi + \int_0^\pi (n-1)\sin^{n-2} x\cos^2 x\,dx \\ &= (n-1)\int_0^\pi \sin^{n-2} x(1-\sin^2 x)\,dx = (n-1)(I_{n-2} - I_n). \end{aligned}$$

Solving for $I_n$, we obtain a **recursion formula**

$$I_n = \frac{n-1}{n} I_{n-2}.$$

Rather than repeatedly integrating by parts, we use this formula. For example,

$$I_6 = \frac{5}{6} I_4 = \frac{5}{6}\frac{3}{4} I_2 = \frac{5}{6}\frac{3}{4}\frac{1}{2} I_0 = \frac{5\pi}{16}.$$

Since the formula drops the index by 2 each time, we end up at a multiple of $I_0 = \pi$ if $n$ is even and a multiple of $I_1 = 2$ when $n$ is odd. Indeed,

$$\begin{aligned} I_{2n} &= \frac{1}{2}\cdot\frac{3}{4}\cdot\frac{5}{6}\cdots\frac{2n-3}{2n-2}\cdot\frac{2n-1}{2n}\pi \\ I_{2n+1} &= \frac{2}{3}\cdot\frac{4}{5}\cdot\frac{6}{7}\cdots\frac{2n-2}{2n-1}\cdot\frac{2n}{2n+1}2. \end{aligned}$$

Since $0 \le \sin x \le 1$ for $x \in [0,\pi]$, we have $\sin^{2n+2} x \le \sin^{2n+1} x \le \sin^{2n} x$. Therefore, $I_{2n+2} \le I_{2n+1} \le I_{2n}$. Divide through by $I_{2n}$ and use the preceding formula to obtain

$$\frac{2n+1}{2n+2} \le \frac{2^2\cdot 4^2\cdot 6^2\cdots(2n-2)^2\cdot(2n)^2}{1\cdot 3^2\cdot 5^2\cdots(2n-1)^2\cdot(2n+1)}\,\frac{2}{\pi} \le 1.$$

Rearranging this slightly and taking the limit, we obtain **Wallis's product**

$$\frac{\pi}{2} = \lim_{n\to\infty}\frac{2^2\cdot 4^2\cdot 6^2\cdots(2n-2)^2\cdot(2n)^2}{1\cdot 3^2\cdot 5^2\cdots(2n-1)^2\cdot(2n+1)}.$$

To this end, set $a_n = \dfrac{n!}{n^n\sqrt{n}e^{-n}}$. Then

$$\begin{aligned}
\frac{a_n^2}{a_{2n}} &= \frac{(n!)^2}{n^{2n+1}e^{-2n}}\frac{(2n)^{2n}\sqrt{2n}e^{-2n}}{(2n)!} = \sqrt{\frac{2}{n}}\,\frac{(2^n n!)^2}{(2n)!}\\
&= \sqrt{\frac{2}{n}}\,\frac{2^2\cdot 4^2\cdot 6^2\cdots(2n-4)^2\cdot(2n-2)^2\cdot(2n)^2}{1\cdot 2\cdot 3\cdot 4\cdots(2n-2)\cdot(2n-1)\cdot(2n)}\\
&= \sqrt{\frac{2}{n}}\,\frac{2\cdot 4\cdot 6\cdots(2n-4)\cdot(2n-2)\cdot(2n)}{1\cdot 3\cdot 5\cdots(2n-3)\cdot(2n-1)}\\
&= \sqrt{\frac{2(2n+1)}{n}}\sqrt{\frac{2^2\cdot 4^2\cdot 6^2\cdots(2n-4)^2\cdot(2n-2)^2\cdot(2n)^2}{3^2\cdot 5^2\cdot 7^2\cdots(2n-3)^2\cdot(2n-1)^2\cdot(2n+1)}}.
\end{aligned}$$

Combining (6.A.1) with our knowledge of Wallis's product,

$$e^{1-E} = \lim_{n\to\infty}\frac{a_n^2}{a_{2n}} = 2\sqrt{\frac{\pi}{2}} = \sqrt{2\pi}.$$

So we obtain the following:

### 6.A.3. Stirling's Formula.

$$\lim_{n\to\infty}\frac{n!}{n^n e^{-n}\sqrt{2\pi n}} = 1.$$

## Exercises for Section 6.A

**A.** Evaluate $\lim\limits_{n\to\infty}\dfrac{\sqrt[n]{n!}}{n}$.

**B.** Decide whether $\sum\limits_{n\geq 0}\dfrac{1}{2^{2n+k}\sqrt{n}}\dbinom{2n+k}{n}$ converges.

**C.** The **gamma function** is defined by $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\,dt$ for all $x > 0$.

(a) Prove that this improper integral has a finite value for all $x > 0$.
(b) Prove that $\Gamma(x+1) = x\Gamma(x)$. HINT: Integrate by parts.
(c) Prove by induction that $\Gamma(n+1) = n!$ for $n \geq 1$.
(d) Calculate $\Gamma(\frac{1}{2})$. HINT: Substitute $t = u^2$, write the square as a double integral, and convert to polar coordinates.

## 6.B The Trapezoidal Rule

The trapezoidal approximation method for computing integrals is a refinement of the Riemann sums that generally is more accurate. Consider a continuous function $f(x)$ on $[a,b]$ and a uniform partition $x_k = a + \frac{k(b-a)}{n}$ for $0 \le k \le n$. The idea of the trapezoidal approximation is to estimate the area $\int_{x_{k-1}}^{x_k} f(x)\,dx$ by the trapezoid with vertices $(x_{k-1}, f(x_{k-1}))$ and $(x_k, f(x_k))$. This has base $(b-a)/n$ and average height $(f(x_{k-1})+f(x_k))/2$. The approximate area is therefore

$$\begin{aligned} A_n &= \sum_{k=1}^{n} \Big(\frac{b-a}{n}\Big)\Big(\frac{f(x_{k-1})+f(x_k)}{2}\Big) \\ &= \frac{b-a}{n}\Big(\tfrac{1}{2}f(x_0)+f(x_1)+f(x_2)+\cdots+f(x_{n-1})+\tfrac{1}{2}f(x_n)\Big). \end{aligned}$$

The crucial questions about this, or any approximation, are, first, How close is it to the true answer, and second, How much easier is it to work with? Ideally, an approximation is both very close to the true answer and much easier to work with. In practice, there is usually a trade-off between these two properties. We return to this issue in detail when we consider approximating functions by polynomials (Chapter 10) and by Fourier series (Chapter 14).

For the trapezoidal rule, so long as the function is $C^2$, we can use the Mean Value Theorem twice to obtain a good estimate for the error.

### 6.B.1. Trapezoidal Rule.

*Suppose $f$ is a $C^2$ function on $[a,b]$ and let $\|f''\|_\infty = \sup\{|f''(x)| : x \in [a,b]\}$. Then the trapezoidal approximants $A_n$ satisfy*

$$\Big|\int_a^b f(x)\,dx - A_n\Big| \le \frac{(b-a)^3\|f''\|_\infty}{12n^2}.$$

**Proof.** Let $F(x) = \displaystyle\int_a^x f(t)\,dt$. For the interval $[x_{k-1}, x_k]$, set

$$\varepsilon_k = F(x_k) - F(x_{k-1}) - \frac{b-a}{2n}\big(f(x_k) - f(x_{k-1})\big).$$

Define $c = \frac{1}{2}(x_{k-1}+x_k)$ and consider the function

$$G(t) = F(c+t) - F(c-t) - t\big(f(c+t)+f(c-t)\big) - Bt^3.$$

Notice that $G(0) = 0$. We are of course interested in $t_0 = (b-a)/2n$. So we choose the constant $B$ so that $G(t_0) = 0$, namely $B = \varepsilon_k / t_0^3$.

By Rolle's Theorem, there is a point $t_1 \in (0, t_0)$ so that $G'(t_1) = 0$. By the Fundamental Theorem of Calculus and the chain rule,

$$\big(F(c+t) - F(c-t)\big)' = f(c+t) + f(c-t).$$

Therefore,

$$G'(t) = t\big(f'(c+t) - f'(c-t)\big) - 3Bt^2.$$

Substituting $t_1$ and solving for $B$ yields

$$B = \frac{f'(c+t_1) - f'(c-t_1)}{3t_1}.$$

An application of the Mean Value Theorem produces a point $t_2 \in (0, t_1)$ so that

$$\frac{f'(c+t_1) - f'(c-t_1)}{2t_1} = f''(t_2)$$

and hence $B = 2f''(t_2)/3$. Consequently,

$$|\varepsilon_k| = |Bt_0^3| = \Big|\frac{(b-a)^3 f''(t_2)}{12n^3}\Big| \le \frac{(b-a)^3 \|f''\|_\infty}{12n^3}.$$

Summing from 1 to $n$ yields

$$\Big|\int_a^b f(x)\,dx - A_n\Big| \le \sum_{k=1}^{n} |\varepsilon_k| \le \frac{(b-a)^3 \|f''\|_\infty}{12n^2}. \qquad \blacksquare$$

## Exercises for Section 6.B

**A.** Estimate the choice of $n$ needed to guarantee an approximation of $\int_0^1 e^{x^2}\,dx$ to 4 decimals accuracy using the trapezoidal rule.

**B.** For every Riemann integrable function, $f$, show that the trapezoidal rule yields a sequence which converges to $\int_a^b f(x)\,dx$.

**C.** **Simpson's Rule**. This method for estimating integrals is based on approximating the function by a parabola passing through three points on the graph of $f$. Given a uniform partition $P = \{a = x_0 < \ldots < x_{2n} = b\}$, let $y_k = f(x_k)$ for all $k$.

(a) Find the parabola passing through $(x_{2k-2}, y_{2k-2})$, $(x_{2k-1}, y_{2k-1})$ and $(x_{2k}, y_{2k})$.
(b) Find the integral of this parabola from $x_{2k-2}$ to $x_{2k}$.
(c) Sum these areas to obtain Simpson's rule for approximating $\int_a^b f(x)\,dx$:

$$A_n = \frac{b-a}{3n}\big(y_0 + 4y_1 + 2y_2 + 4y_3 + \cdots + 2y_{2n-2} + 4y_{2n-1} + y_{2n}\big).$$

## 6.C Measure Zero and Lebesgue's Theorem

As we have already said, it is possible to build a more powerful theory of integration, based on the ideas of measure theory. The first step toward this theory, and the crucial notion for this section, is the idea of a set of measure zero. This is a reasonable condition for a set to be "small," although the condition can have surprising and unintuitive properties. Using this idea, we obtain a characterization of precisely which functions are Riemann integrable.

If $U = (c,d)$ is an open interval, we write $|U| = d - c$ for its length.

**6.C.1. DEFINITION.** A subset $A$ of $\mathbb{R}$ has **measure zero** if for every $\varepsilon > 0$, there is a countable family of intervals $\{U_n = (c_n, d_n) : n \geq 1\}$ such that $A \subset \bigcup_{n\geq 1} U_n$ and $\sum_{n\geq 1} |U_n| = \sum_{n\geq 1} d_n - c_n < \varepsilon$.

A subset $A$ of $(a,b)$ has **content zero** if for every $\varepsilon > 0$, there is a *finite* family of intervals $\{U_n = (c_n, d_n) : 1 \leq n \leq N\}$ such that $A \subset \bigcup_{n=1}^{N} U_n$ and $\sum\limits_{n=1}^{N} |U_n| < \varepsilon$.

### 6.C.2. EXAMPLES.

(1) The set $\mathbb{Q}$ of all rational numbers has measure zero. To see this, write $\mathbb{Q}$ as a list $r_1, r_2, r_3, \ldots$. Given $\varepsilon > 0$, let $U_n = (r_n - 2^{-n-1}\varepsilon, r_n + 2^{-n-1}\varepsilon)$. Evidently, $\bigcup_{n\geq 1} U_n$ contains $\mathbb{Q}$ and $\sum_{n\geq 1} |U_n| = \sum_{n\geq 1} 2^{-n}\varepsilon = \varepsilon$. So $\mathbb{Q}$ has measure zero.

However, $\mathbb{Q}$ does not have content zero because any *finite* collection of intervals covering $\mathbb{Q}$ can miss only finitely many points in $\mathbb{R}$. Consequently, at least one would be infinite! In fact, the set of rational points in $[0,1]$ will not have content zero either for the same reason combined with the next example.

(2) Suppose that an interval $[a,b]$ is covered by open intervals $\{U_n\}$. Now $[a,b]$ is compact by the Heine–Borel Theorem (4.4.6). Moreover by the Borel–Lebesgue Theorem (9.2.3), the open cover $\{U_n\}$ has a finite subcover, say $U_1, \ldots, U_N$. Now this *finite* set of intervals pieces together to cover an interval of length $b - a$. From this we can easily show that $\sum_{n=1}^{N} d_n - c_n \geq b - a$. Consequently, $[a,b]$ does not have measure zero. (It has measure $b - a$.)

(3) In Example 4.4.8, we constructed the Cantor set. We showed there that it has measure zero and indeed has content zero. It also shows that an uncountable set can have measure zero.

### 6.C.3. PROPOSITION.

(1) *If $A$ has measure zero, and $B \subset A$, then $B$ has measure zero.*
(2) *If $A_n$ are sets of measure zero for $n \geq 1$, then $\bigcup_{n\geq 1} A_n$ has measure zero.*
(3) *Every countable set has measure zero.*
(4) *If $A$ is compact and has measure zero, then it has content zero.*

**PROOF.** (1) is trivial. For (2), suppose that $\varepsilon > 0$. We modify the argument in Example 6.C.2 (1). Choose a collection of open intervals $\{U_{nm} : m \geq 1\}$ covering $A_n$ such that $\sum_{m\geq 1} |U_{nm}| < 2^{-n}\varepsilon$. Then $\{U_{nm} : m,n \geq 1\}$ covers $\bigcup_{n\geq 1} A_n$ and the combined lengths of these intervals is $\sum_{n\geq 1}\sum_{m\geq 1} |U_{nm}| < \sum_{n\geq 1} 2^{-n}\varepsilon = \varepsilon$. (3) now follows from (2) and the observation that a single point has measure zero.

(4) Suppose that $A$ is compact and $\{U_n\}$ is an open cover of intervals with $\sum_{n\geq 1} |U_n| < \varepsilon$. By the Borel–Lebesgue Theorem (9.2.3), this cover has a finite subcover. This finite cover has total length less than $\varepsilon$ also. Therefore, $A$ has content zero. ■

**6.C.4. DEFINITION.** A property is valid **almost everywhere** (a.e.) if the set of points where it fails has measure zero.

For example, $f = g$ a.e. means that $\{x : f(x) \neq g(x)\}$ has measure zero. And $f(x) = \lim_{n\to\infty} f_n(x)$ a.e. means that this limit exists and equals $f(x)$ except on a set of measure zero.

It is important to distinguish between two similar but distinct statements. If $f$ is continuous almost everywhere, then the set of points of discontinuity has measure zero. If $f$ equals a continuous function almost everywhere, then there would be a continuous function $g$ so that $\{x : f(x) \neq g(x)\}$ is measure zero. This is a much stronger property. For example, the characteristic function of $[0,1]$ has only two points of discontinuity in $\mathbb{R}$ and thus is continuous almost everywhere. But any continuous function will differ from this on a whole interval near 0 and another near 1. Since these intervals are not measure zero (see Example 6.C.2), the continuous function is not equal to the characteristic function almost everywhere.

The next result is very appealing because it provides a simple description of exactly which functions are integrable. But in practice, it is often enough to know that piecewise continuous functions are integrable. It will be more useful to build a more powerful integral, as we do in Section 9.D.

We need a notion of the "size" of a discontinuity. There is a global version of this concept, known as the modulus of continuity, which we will develop later (see Definition 10.4.2).

**6.C.5. DEFINITION.** The **oscillation** of a function $f$ over an interval $I$ is defined as $\operatorname{osc}(f,I) = \sup\{|f(x) - f(y)| : x,y \in I\}$. Then set

$$\operatorname{osc}(f,x) = \inf_{r>0} \operatorname{osc}\big(f,(x-r,x+r)\big).$$

It is easy (Exercise 6.C.A) to prove that $f$ is continuous at $x$ if and only if $\operatorname{osc}(f,x) = 0$.

**6.C.6. LEBESGUE'S THEOREM.**
*A bounded function on $[a,b]$ is Riemann integrable if and only if it is continuous almost everywhere.*

**PROOF.** First suppose that $f$ is Riemann integrable. Then for each $k \geq 1$, there is a finite partition $P_k$ of $[a,b]$ so that $U(f,P_k), -L(f,P_k) < 4^{-k}$. Let $u_k$ and $l_k$ be the step functions that are constant on the intervals of $P_k$ and bound $f$ from above and below (as in Exercise 6.3.G) so that $\int_a^b u_k(x) - l_k(x)\,dx < 4^{-k}$.

The set $B_k = \{x : u_k(x) - l_k(x) \geq 2^{-k}\}$ is the union of certain intervals of the partition $P_k$, say $J_{k,i}$ for $i \in S_k$. Compute

$$4^{-k} > \int_a^b u_k(x) - l_k(x)\,dx \geq \sum_{i \in S_k} 2^{-k} |J_{k,i}|.$$

Then $\sum\limits_{i \in S_k} |J_{k,i}| < 2^{-k}$. Let $A_1 = \bigcap_{k \geq 1} \operatorname{int} B_k$ and let $A_2 = \bigcup_{k \geq 1} P_k$. Observe that $\{\operatorname{int} J_{k,i} : i \in S_k\}$ covers $A_1$ for each $k$. As these intervals have length summing to less than $2^{-k}$, it follows that $A_1$ has measure zero. Since $A_2$ is countable, it also has measure zero. Thus the set $A = A_1 \cup A_2$ has measure zero.

For any $x \notin A$ and any $\varepsilon > 0$, choose $k$ so that $2^{-k} < \varepsilon$ and $x \notin B_k$. Then

$$u_k(x) - l_k(x) < 2^{-k} < \varepsilon.$$

As $x$ is not a point in $P_k$, it is an interior point of some interval $J$ of this partition. Choose $r > 0$ so that $(x-r, x+r) \subset J$. For any $y$ with $|x-y| < r$, we have

$$l_k(x) \leq f(y),\ f(x) \leq u_k(x).$$

Thus $|f(x) - f(y)| < \varepsilon$ and so $f$ is continuous at $x$.

Conversely, suppose that $f$ is continuous almost everywhere on $[a,b]$ and is bounded by $M$. Let

$$A_k = \{x \in [a,b] : \operatorname{osc}(f,x) \geq 2^{-k}\}.$$

Then each $A_k$ has measure zero, and the set of points of discontinuity is $A = \bigcup_{k \geq 1} A_k$. Observe that $A_k$ is closed. Indeed suppose that $x$ is the limit of a sequence $(x_n)$ with all $x_n$ in $A_k$. By definition, there are points $y_n$ and $z_n$ with $|x_n - y_n| < 1/n$ and $|x_n - z_n| < 1/n$ such that $|f(y_n) - f(z_n)| \geq 2^{-k} - 2^{-n}$. Therefore,

$$\lim_{n \to \infty} y_n = \lim_{n \to \infty} z_n = x.$$

Consequently, $\operatorname{osc}(f,x) \geq 2^{-k}$.

By Proposition 6.C.3 (4), each $A_k$ has content zero. Cover $A_k$ with a finite number of open intervals $J_{k,i}$ such that $\sum_i |J_{k,i}| < 2^{-k}$. The complement $X$ consists of a finite number of closed intervals on which $\operatorname{osc}(f,x) < 2^{-k}$ for all $x \in X$. Thus there is an open interval $J_x$ containing $x$ so that $\operatorname{osc}(f, J_x) < 2^{-k}$. Notice that $X$ is a closed and bounded subset of $[a,b]$ and so is compact. The collection $\{J_x : x \in X\}$ is an open cover of $X$. By the Borel–Lebesgue Theorem (9.2.3), there is a finite subcover $J_{x_1}, \ldots, J_{x_p}$. Let $P_k$ be the finite partition consisting of all the endpoints of all of these intervals together with the endpoints of each $J_{k,i}$.

Let us estimate the upper and lower sums for this partition. As usual, let $M_j(f,P_k)$ and $m_j(f,P_k)$ be the supremum and infimum of $f$ over the $j$th interval, namely $I_{k,j} = [x_{k,j-1}, x_{k,j}]$. These intervals split into two groups, those contained in $X$ and those contained in $U_k = \bigcup_i J_{k,i}$. For the first group, the oscillation is less than $2^{-k}$ and thus $M_j - m_j \le 2^{-k}$. For the second group, the total length of the intervals is at most $2^{-k}$. Combining these estimates, we obtain

$$\begin{aligned} U(f,P_k) - L(f,P_k) &= \sum_{I_{k,i}\subset X} (M_i - m_i)\Delta_i + \sum_{I_{k,i}\subset U_k} (M_i - m_i)\Delta_i \\ &\le \sum_{I_{k,i}\subset X} 2^{-k}\Delta_i + \sum_{I_{k,i}\subset U_k} 2M\Delta_i \\ &\le 2^{-k}(b-a) + 2M2^{-k} = 2^{-k}(b-a+2M). \end{aligned}$$

Therefore, $\lim_{k\to\infty} U(f,P_k) - L(f,P_k) = 0$, and so $f$ is integrable by Riemann's condition. ■

You may want to review the discontinuous functions in Section 5.2 to see which are continuous almost everywhere and which are equal to a continuous function almost everywhere.

## Exercises for Section 6.C

**A.** Prove that $f$ is continuous at $x$ if and only if $\operatorname{osc}(f,x) = 0$.

**B.** If measure zero sets were defined using closed intervals instead of open intervals, show that one obtains the same sets.

**C.** If $A \subset \mathbb{R}$ has measure zero, what is $\operatorname{int}(A)$? Is $\overline{A}$ also measure zero?

**D.** If $A$ has measure zero and $B$ is countable, show that $A+B = \{a+b : a \in A, b \in B\}$ has measure zero.

**E.** Consider the set $C'$ obtained using the construction of the Cantor set in Example 4.4.8 but removing $2^{n-1}$ intervals of length $4^{-n}$ (instead of length $3^{-n}$) at the $n$th stage.

(a) Show that $C'$ is closed and has no interior.
(b) Show that $C'$ is not measure zero.
HINT: Any cover of $C'$ together with the intervals removed covers $[0,1]$.

**F.** Let $D$ be the set of numbers $x \in [0,1]$ with a decimal expansion containing no odd digits. Prove that $D$ has measure zero. HINT: Cover $D$ with some intervals of length $10^{-n}$.

**G.** Suppose that $f$ and $g$ are both Riemann integrable on $[a,b]$. Use Lebesgue's Theorem to prove that $fg$ is Riemann integrable. HINT: Compare with Exercise 6.3.N.

**H.** Show that $A \subset \mathbb{R}$ has content zero if and only if $\overline{A}$ is compact and has measure zero.
HINT: Show that an unbounded set cannot have content zero. Note that a finite open cover of $A$ contains most of $\overline{A}$.

**I.** Define a relation on functions on $[a,b]$ by $f \sim g$ if the set $\{x : f(x) \neq g(x)\}$ has measure zero. Prove that this is an equivalence relation.

## 6.D Riemann-Stieltjes Integration

Riemann-Stieltjes integration is a generalization of Riemann integration. The essential change is to weight the intervals using a function $g$, called the integrator, so that the interval $[c,d]$ has weight $g(d)-g(c)$. Taking $g(x)=x$ will yield the usual Riemann integral. This weighting allows integrals that combine aspects of both discrete summation and the usual Riemann integral. One application of Riemann-Stieltjes integration is in probability theory, mixing discrete and continuous probability. The cost of this increased generality is that it requires some careful work with partitions.

**6.D.1. DEFINITION.** Consider bounded functions $f,g:[a,b]\to\mathbb{R}$. Given a partition of $[a,b]$, $P=\{a=x_0<x_1<\cdots<x_n=b\}$, and an evaluation sequence for $P$, $X=(x_1',x_2',\dots,x_n')$ with $x_j'\in[x_{j-1},x_j]$, we define the **Riemann-Stieltjes sum**, or briefly, the **R-S sum**, for $f$ with respect to $g$ using $P$ and $X$, as

$$I_g(f,P,X)=\sum_{j=1}^{n} f(x_j')[g(x_j)-g(x_{j-1})].$$

Say $f$ is **Riemann-Stieltjes integrable** with respect to $g$, denoted $f\in\mathscr{R}(g)$, if there is a number $L$ so that for all $\varepsilon>0$, there is a partition $P_\varepsilon$ so that, for all partitions $P$ with $P\supseteq P_\varepsilon$ and all evaluation sequences $X$ for $P$, we have

$$\big|I_g(f,P,X)-L\big|<\varepsilon.$$

In this case, the **Riemann-Stieltjes integral** of $f$ with respect to $g$ is $L=\int_a^b f\,dg$.

Before developing the general properties of this integral, we compare this definition to the corresponding condition for the Riemann integral, Condition (4) in Theorem 6.3.6. Here, we pick a partition $P_\varepsilon$ and only consider partitions containing $P_\varepsilon$, whereas, for the Riemann integral, we pick a number $\delta$ and have to consider all partitions with mesh size less than $\delta$. By putting a particular number $x$ into $P_\varepsilon$, we can restrict attention to partitions containing $x$, something we cannot accomplish by specifying a mesh size. To see the difference, look at Example 6.D.2 (2).

**6.D.2. EXAMPLES.**

(1) If we take $g(x)=x$, this is just the Riemann integral.

(2) Pick $c\in(a,b)$ and $s,t\in\mathbb{R}$. Let $g(x)=s$ on $[a,c)$ and $g(x)=s+t$ on $[c,b]$; i.e. $g(x)=s\chi_{[a,c)}+(s+t)\chi_{[c,b]}$. Suppose that $f$ is a bounded function on $[a,b]$. For $\delta>0$, consider a partition $P=\{a=x_0<\cdots<x_{k-1}=c-\delta<x_k=c<\cdots<x_n=b\}$, and pick an evaluation sequence $X$ for $P$. Then $g(x_j)-g(x_{j-1})=0$ except for $g(x_k)-g(x_{k-1})=t$, so

$$I_g(f,P,X)=\sum_{j=1}^{n} f(x_j')[g(x_j)-g(x_{j-1})]=tf(x_k').$$

Since $x'_k$ is an arbitrary point in $[c-\delta,c]$, we see that this will have a definite limit precisely when $f$ is continuous from the left at $c$, and in this case, $\int_a^b f\,dg = tf(c)$.

Observe that it is important that $c$ be in the partition. Take $f = \chi_{(c,b]}$. This function is left continuous at $c$, but is not continuous from the right. Suppose that $P$ is a partition with very fine mesh, but with $x_{k-1} < c < x_k$. Then either $x'_k$ is in $[x_{k-1},c]$ or in $(c,x_k]$. As above, $I_g(f,P,X) = tf(x'_k)$, which is either 0 or $t$, depending on the location of $x'_k$. No matter how small the mesh of $P$ is, these two possible values for the partial sums prevent the existence of a limiting value $L$ over this family of partitions. However, since we are allowed to specify that $c$ belongs to $P$, say by setting $P_0 = \{a,c,b\}$, we find that $I_g(f,P,X) = 0$ for all $P \supset P_0$. Hence $\int_a^b f\,dg = 0$.

(3) Define $g(x)$ on $[0,1]$ by $g(0) = 0$ and $g(x) = x + 2^{-n}$ on $(2^{-n-1}, 2^{-n}]$ for $n \geq 0$. Let $f(x)$ be continuous on $[0,1]$. We observe that $g(x) = x + \sum_{i=1}^{\infty} 2^{-i}\chi_{[2^{-i},1]}$. From properties of the integral which we develop, and the previous examples, we get

$$\int_a^b f\,dg = \int_a^b f(x)\,dx + \sum_{i=1}^{\infty} \frac{1}{2^i}\int_a^b f\,d\chi_{[1/2^i,1]} = \int_a^b f(x)\,dx + \sum_{i=1}^{\infty} \frac{1}{2^i} f(\tfrac{1}{2^i}).$$

We collect together some basic facts about Riemann-Stieltjes integrals.

**6.D.3. THEOREM.** *If $f_1$ and $f_2$ are in $\mathscr{R}(g)$ on $[a,b]$, and $c_1, c_2 \in \mathbb{R}$, then $c_1 f_1 + c_2 f_2 \in \mathscr{R}(g)$ on $[a,b]$ and*

$$\int_a^b c_1 f_1 + c_2 f_2\,dg = c_1 \int_a^b f_1\,dg + c_2 \int_a^b f_2\,dg.$$

*If $f \in \mathscr{R}(g_1)$ and $f \in \mathscr{R}(g_2)$ on $[a,b]$, and $d_1, d_2 \in \mathbb{R}$, then $f \in \mathscr{R}(d_1 g_1 + d_2 g_2)$ on $[a,b]$ and*

$$\int_a^b f\,d(d_1 g_1 + d_2 g_2) = d_1 \int_a^b f\,dg_1 + d_2 \int_a^b f\,dg_2.$$

*Finally, if $a < b < c$ and $f$ is R-S integrable with respect to $g$ on both $[a,b]$ and $[b,c]$ then it is R-S integrable with respect to $g$ on $[a,c]$ and*

$$\int_a^c f\,dg = \int_a^b f\,dg + \int_b^c f\,dg.$$

**PROOF.** We prove the first result and leave the second and third as exercises. Consider any partition $P = \{a = x_0 < x_1 < \cdots < x_n = b\}$ and evaluation sequence $X = (x'_1, \ldots, x'_n)$ for $P$. We have

$$\begin{aligned} I_g(c_1 f_1 + c_2 f_2, P, X) &= \sum_{j=1}^{n} (c_1 f_1(x'_j) + c_2 f_2(x'_j))(g(x_j) - g(x_{j-1})) \\ &= c_1 \sum_{j=1}^{n} f_1(x'_j)(g(x_j) - g(x_{j-1})) + c_2 \sum_{j=1}^{n} f_2(x'_j)(g(x_j) - g(x_{j-1})) \\ &= c_1 I_g(f_1, P, X) + c_2 I_g(f_2, P, X) \end{aligned}$$

Let $\varepsilon > 0$. Pick $\delta > 0$ so that $|c_1|\delta + |c_2|\delta < \varepsilon$. Since $f_i \in R(g)$ for $i = 1,2$, there are partitions $P_i$ so that, for all partitions $P$ with $P \supseteq P_i$ and all evaluation sequences $X$ for $P$,

$$\left| I_g(f_i, P, X) - \int_a^b f_i \, dg \right| < \delta.$$

Thus, for any partition $P$ with $P \supseteq P_1 \cup P_2$, and for any evaluation sequence $X$,

$$\begin{aligned} &\left| I_g(c_1 f_1 + c_2 f_2, P, X) - \left( c_1 \int_a^b f_1 \, dg + c_2 \int_a^b f_2 \, dg \right) \right| \\ &\quad \le |c_1| \left| I_g(f_1, P, X) - \int_a^b f_1 \, dg \right| + |c_2| \left| I_g(f_2, P, X) - \int_a^b f_2 \, dg \right| < (|c_1| + |c_2|)\delta < \varepsilon. \end{aligned}$$

The result follows. ■

We avoided upper sums and lower sums so far, because for general integrators where $g(x_j) - g(x_{j-1})$ can change sign, one has to choose carefully to get the largest and smallest sums. When the integrator is monotone increasing, then upper and lower sums have properties similar to upper and lower sums for Riemann integrals.

**6.D.4. DEFINITION.** Let $f$ and $g$ be bounded functions on $[a,b]$. Let $P = \{x_0 < x_1 < \cdots < x_n\}$ be a partition of $[a,b]$ and recall that

$$M_j(f,P) = \sup_{x_{j-1} \le x \le x_j} f(x) \qquad \text{and} \qquad m_j(f,P) = \inf_{x_{j-1} \le x \le x_j} f(x).$$

Define the **upper sum with respect to** $g$ and the **lower sum with respect to** $g$ to be

$$\begin{aligned} U_g(f,P) &= \sum_{i=1}^n M_i(f,P)[g(x_i) - g(x_{i-1})] \\ L_g(f,P) &= \sum_{i=1}^n m_i(f,P)[g(x_i) - g(x_{i-1})] \end{aligned}$$

When $g$ is increasing, it is easy to verify that for any evaluation sequence $X$,

$$L_g(f,P) \le I_g(f,P,X) \le U_g(f,P). \tag{6.D.5}$$

This is not true if $g$ is not increasing. Also, it is possible for the above inequalities to be strict. For example, if $f(x) = x$ on $[0,1)$ and $f(1) = 0$ and $g(x) = x$, then $U_g(f,P) > I_g(f,P,X)$ for all choices of $P$ and $X$.

As with the Riemann integral, we have

**6.D.6. REFINEMENT LEMMA.**
*Let $f, g$ be bounded functions on $[a,b]$, and assume that $g$ is increasing. Let $P$ and $R$ be partitions of $[a,b]$ with $P \subset R$. Then*

$$L_g(f,P) \le L_g(f,R) \le U_g(f,R) \le U_g(f,P).$$

The proof is exactly the same as the Refinement Lemma (6.3.2) for the Riemann integral. The key points are that if $x_{j-1}, x_j \in P$ and $[x_{j-1}, x_j] \cap R = [t_k, \dots, t_l]$, then

$$g(x_j) - g(x_{j-1}) = \sum_{i=k+1}^{l} g(t_i) - g(t_{i-1})$$

and, for any $i$ between $k+1$ and $l$,

$$m_j(f,P) \le m_i(f,R) \le M_i(f,R) \le M_j(f,P).$$

**6.D.7. COROLLARY.** *Let $f, g$ be bounded functions on $[a,b]$ and assume that $g$ is increasing. If $P$ and $Q$ are any two partitions of $[a,b]$,*

$$L_g(f,P) \le U_g(f,Q).$$

This follows because

$$L_g(f,P) \le L_g(f, P \cup Q) \le U_g(f, P \cup Q) \le U_g(f,Q).$$

Thus the collection of lower sums is bounded above by any upper sum, and vice versa. So in keeping with the notation for Riemann integration, we define

**6.D.8. DEFINITION.** Let $f, g$ be bounded functions on $[a,b]$ with $g$ increasing. Set $L_g(f) = \sup_P L_g(f,P)$ and $U_g(f) = \inf_P U_g(f,P)$.

We have the following analogue of Theorem 6.3.6.

**6.D.9. RIEMANN-STIELTJES CONDITION.**
*Let $f, g$ be bounded functions on $[a,b]$ and assume that $g$ is monotone increasing. The following are equivalent:*

(1) *$f$ is Riemann-Stieltjes integrable with respect to $g$,*
(2) *$U_g(f) = L_g(f)$, and*
(3) *for each $\varepsilon > 0$, there is a partition $P$ so that $U_g(f,P) - L_g(f,P) < \varepsilon$.*

*Moreover, if (2) holds, then the common value is $\int f\,dg$.*

**PROOF.** The argument that (3) implies (2) is immediate.

Suppose that (2) holds and let $L = L_g(f) = U_g(f)$. Let $\varepsilon > 0$. We can find two partitions $P_1$ and $P_2$ so that $U_g(f,P_1) < L + \varepsilon$ and $L_g(f,P_2) > L - \varepsilon$. Let $P_\varepsilon = P_1 \cup P_2$ be their common refinement. By the Refinement Lemma,

$$L - \varepsilon < L_g(f,P_2) \le L_g(f,P_\varepsilon) \le U_g(f,P_\varepsilon) \le U_g(f,P_1) < L + \varepsilon.$$

Suppose that $Q$ is any partition with $Q \supseteq P_\varepsilon$ and $X$ is any evaluation sequence for $Q$. Using (6.D.5) and the Refinement Lemma again,

$$L - \varepsilon < L_g(f,P_\varepsilon) \le L_g(f,Q) \le I_g(f,Q,X) \le U_g(f,Q) \le U_g(f,P_\varepsilon) < L + \varepsilon.$$

Thus, (1) holds and $L = \int_a^b f\,dg$.

Now suppose that (1) holds with $L = \int_a^b f\,dg$. Let $\varepsilon > 0$. Find a partition $P = P_{\varepsilon/4}$ satisfying the definition of R-S integrability for $\varepsilon/4$; i.e., $|I(f,Q,X) - L| < \varepsilon/4$ for any refinement $Q \supset P$ and any evaluation sequence $X$ for $Q$. We may assume that $D := g(b) - g(a) > 0$, as otherwise $U_g(f,P) = L_g(f,P) = 0$.

Let $P = \{x_0 < x_1 \dots < x_n\}$. For each $j$, choose points $s_j, t_j$ in $[x_{j-1}, x_j]$ so that

$$f(s_j) < m_j(f,P) + \frac{\varepsilon}{4nD} \quad \text{and} \quad f(t_j) > M_j(f,P) - \frac{\varepsilon}{4nD}.$$

Letting $S = (s_1, \dots, s_n)$ and $T = (t_0, \dots, t_n)$, we have

$$\begin{aligned} L - \frac{\varepsilon}{4} < I_g(f,P,S) &= \sum_{j=1}^n f(s_j)[g(x_j) - g(x_{j-1})] \\ &< \sum_{j=1}^n \Big(m_j(f,P) + \frac{\varepsilon}{4nD}\Big)[g(x_j) - g(x_{j-1})] \\ &\le L(f,P) + \sum_{j=1}^n \frac{\varepsilon}{4nD} D = L(f,P) + \frac{\varepsilon}{4}. \end{aligned}$$

Hence $L(f,P) > L - \varepsilon/2$. Similarly, $U(f,P) < L + \varepsilon/2$. So $U(f,P) - L(f,P) < \varepsilon$. This establishes (3). ■

Next we show that there are many integrable functions.

**6.D.10. Theorem.** *If $f$ is continuous on $[a,b]$, and $g$ is monotone increasing, then $f$ is Riemann-Stieltjes integrable with respect to $g$ on $[a,b]$; i.e., $f \in \mathscr{R}(g)$.*

**Proof.** We assume $D := g(b) - g(a) > 0$, as otherwise $g$ is constant and every bounded function is in $\mathscr{R}(g)$. Fix $\varepsilon > 0$. As $f$ is continuous and $[a,b]$ compact, $f$ is uniformly continuous by Theorem 5.5.9. Thus, there is $\delta > 0$ so that $x, y \in [a,b]$ and $|x - y| < \delta$ implies $|f(x) - f(y)| \le \varepsilon/D$.

Let $P$ be a partition of $[a,b]$ with $\operatorname{mesh}(P) < \delta$. Then $M_j(f,P) - m_j(f,P) \le \varepsilon/D$. If $P = \{a = x_0 < x_1 < \dots < x_n = b\}$, then

$$\begin{aligned} U_g(f,P) - L_g(f,P) &= \sum_{j=1}^n \big(M_j(f,P) - m_j(f,P)\big)[g(x_j) - g(x_{j-1})] \\ &\le \frac{\varepsilon}{D} \sum_{j=1}^n g(x_j) - g(x_{j-1}) = \varepsilon. \end{aligned}$$

As $g$ is increasing, we get a telescoping sum. By the R-S condition, $f \in \mathscr{R}(g)$. ■

It is now possible to expand these results to a wider class of integrators. Specifically, if $g$ has bounded variation, then we can apply the results from Section 5.A. While not the most general integrators, they have great practical importance.

**6.D.11. COROLLARY.** *If $f : [a,b] \to \mathbb{R}$ is continuous and $g : [a,b] \to \mathbb{R}$ is of bounded variation, then $f \in \mathscr{R}(g)$ on $[a,b]$.*

**PROOF.** Using Theorem 5.A.5, write $g = g_1 - g_2$, where $g_i$ are increasing functions on $[a,b]$. By the previous theorem, $f \in \mathscr{R}(g_i)$ for $i = 1,2$. Thus by Theorem 6.D.3, $f \in \mathscr{R}(g_1 - g_2) = \mathscr{R}(g)$. ■

An important change from the Riemann integral lies in bounds for integrals. A first guess might be that if $f$ is bounded by $M$ on an interval $[a,b]$, then $\int_a^b f\,dg$ would be bounded by $M(g(b) - g(a))$. While this is true if $g$ is increasing on $[a,b]$, it is not true in general. For example, suppose $a < c < d < b$, and let $g = \chi_{[c,d)}$. By Example 6.D.2 (2), every continuous function is in $\mathscr{R}(g)$, and $\int_a^b f\,dg = f(c) - f(d)$. In particular, the integral need not be bounded above by $M\big(g(b) - g(a)\big) = 0$.

The problem is that $g(x_j) - g(x_{j-1})$ changes sign over the interval $[a,b]$, and we need to replace $g(b) - g(a)$ with the total variation of $g$.

**6.D.12. THEOREM.** *Let $g$ have bounded variation on $[a,b]$. If $f \in \mathscr{R}(g)$ is bounded by $M$ on $[a,b]$, then*

$$\Big| \int_a^b f\,dg \Big| \le M \cdot V_a^b g.$$

**PROOF.** For any partition $P = \{a = x_0 < x_1 < \cdots < x_n = b\}$ and any evaluation sequence $X$ for $P$, we have

$$|I_g(f,P,X)| \le \sum_{i=1}^{n} M|g(x_i) - g(x_{i-1})| \le M \cdot V_a^b g.$$

By choosing a sequence of partitions $P_n$ and evaluation sequences $X_n$ so that $I_g(f,P_n,X_n)$ converges to $\int_a^b f\,dg$, we obtain the result. ■

The following two theorems are the analogues of the change of variables or substitution formula, Equation 6.4.7, and the integration by parts formula, Equation 6.4.5. They are valid for general integrators.

**6.D.13. THEOREM.** *Suppose $f,g$ are bounded on $[a,b]$ and $\varphi : [c,d] \to [a,b]$ is a strictly increasing, continuous function onto $[a,b]$. Let $F = f \circ \varphi$ and $G = g \circ \varphi$. If $f \in \mathscr{R}(g)$ on $[a,b]$, then $F \in \mathscr{R}(G)$ on $[c,d]$ and*

$$\int_c^d F\,dG = \int_a^b f\,dg.$$

**PROOF.** Consider any partition $P = \{c = x_0 < x_1 < \cdots < x_n = d\}$ of $[c,d]$ and evaluation sequence $X = (x_1', \ldots, x_n')$ for $P$. Since $\varphi$ is strictly increasing and onto, it is one-to-one, $\varphi(c) = a$, and $\varphi(d) = b$. Thus, $\varphi(P)$ is a partition of $[a,b]$ and $\varphi(X) = (\varphi(x_1'), \ldots, \varphi(x_n'))$ is an evaluation sequence for $\varphi(P)$. Moreover,

$$I_G(F,P,X) = \sum_{j=1}^{n} f(\varphi(x_j'))[g(\varphi(x_j)) - g(\varphi(x_{j-1}))] = I_g(f, \varphi(P), \varphi(X)).$$

Let $\varepsilon > 0$. As $f \in \mathscr{R}(g)$, there is a partition $Q_\varepsilon$ so that for all partitions $Q$ of $[a,b]$ with $Q \supseteq Q_\varepsilon$ and all evaluations sequences $Y$ for $Q$, we have

$$\left| I_g(f,Q,Y) - \int_a^b f\,dg \right| < \varepsilon.$$

Let $P_\varepsilon = \{\varphi^{-1}(x) : x \in Q_\varepsilon\}$. If $P$ is a partition of $[c,d]$ with $P \supseteq P_\varepsilon$, then $\varphi(P) \supseteq Q_\varepsilon$. Using the above two displayed equations, we have, for all partitions $P$ with $P \supseteq P_\varepsilon$ and all evaluation sequences $X$ for $P$,

$$\left| I_G(F,P,X) - \int_a^b f\,dg \right| < \varepsilon.$$

By the definition, $F \in \mathscr{R}(G)$ on $[c,d]$ and $\int_a^b F\,dG = \int_a^b f\,dg$. ■

**6.D.14. THEOREM.** *Let $f, g$ be bounded functions on $[a,b]$. If $f \in \mathscr{R}(g)$ on $[a,b]$, then $g \in \mathscr{R}(f)$ on $[a,b]$ and*

$$\int_a^b f\,dg + \int_a^b g\,df = g(b)f(b) - g(a)f(a).$$

**PROOF.** Consider any partition $P = \{a = x_0 < x_1 < \cdots < x_n = b\}$ of $[a,b]$ and evaluation sequence $X = (x_1', \ldots, x_n')$ for $P$. We need to relate $I_f(g,P,X)$ with $I_g(f,Q,Y)$ for a suitable choice of partition $Q$ and evaluation sequence $Y$. By definition,

$$I_f(g,P,X) = \sum_{j=1}^{n} g(x_j')(f(x_j) - f(x_{j-1})).$$

We also have a telescoping sum

$$A := f(b)g(b) - f(a)g(a) = \sum_{j=1}^{n} f(x_j)g(x_j) - \sum_{j=1}^{n} f(x_{j-1})g(x_{j-1}).$$

Subtracting the first equation from the second, we have

$$A - I_f(g,P,X) = \sum_{j=1}^{n} f(x_j)(g(x_j) - g(x_j')) + \sum_{j=1}^{n} f(x_{j-1})(g(x_j') - g(x_{j-1})).$$

This pair of summations is just $I_g(f,P',Y)$, where

$$P' = P \cup X = \{a = x_0 \le x'_1 \le x_1 \le x'_2 \le x_2 \le \cdots \le x'_n \le x_n = b\}$$

and $Y$ is the evaluation sequence $(x_0, x_1, x_1, x_2, x_2, \ldots, x_{n-1}, x_{n-1}, x_n)$. If it happens that some $x'_i$ equals one of the endpoints $x_{i-1}$ or $x_i$, then some of the terms disappear, but the result is the same. Consequently, for any $P$ and $X$, there is a partition $P' \supset P$ and an evaluation sequence $Y$ so that

$$I_f(g,P,X) + I_g(f,P',Y) = f(b)g(b) - f(a)g(a) = A. \tag{6.D.15}$$

Let $\varepsilon > 0$. As $f \in \mathscr{R}(g)$, there is a partition $P_\varepsilon$ so that for all partitions $P$ of $[a,b]$ with $P \supseteq P_\varepsilon$ and all evaluation sequences $Y$ for $P$, we have

$$\Big| I_g(f,P,Y) - \int_a^b f\,dg \Big| < \varepsilon.$$

Let $L = A - \int_a^b f\,dg$. Suppose that $P \supseteq P_\varepsilon$ and $X$ is an evaluation sequence. Let $P' = P \cup X$ and let $Y$ be chosen as in (6.D.15). Then

$$\big| I_f(g,P,X) - L \big| = \Big| (A - I_g(f,P',Y)) - \big(A - \int_a^b f\,dg\big) \Big| < \varepsilon.$$

Therefore $g \in \mathscr{R}(f)$ on $[a,b]$ and that

$$\int_a^b g\,df = \big(f(b)g(b) - f(a)g(a)\big) - \int_a^b f\,dg.$$

■

Finally, we show that, for a smooth function $g$, the Riemann-Stieltjes integral with respect to $g$ is really just a Riemann integral.

**6.D.16. THEOREM.** *Suppose that $f \in \mathscr{R}(g)$ on $[a,b]$, where $g$ is $C^1$. Then $fg'$ is Riemann integrable on $[a,b]$ and*

$$\int_a^b f\,dg = \int_a^b f(x)g'(x)\,dx.$$

**PROOF.** Let $\varepsilon > 0$ and $L = \int_a^b f\,dg$. Let $P_\varepsilon$ be a partition so that $\big| I_g(f,P,X) - L \big| < \varepsilon$ for every partition $P \supset P_\varepsilon$ and any evaluation sequence. Consider such a partition $P = \{a = x_0 < x_1 < \cdots < x_n = b\}$ of $[a,b]$. By the Mean Value Theorem (6.2.2), for each $j$, there is a point $x'_j$ so that

$$g'(x'_j) = \frac{g(x_j) - g(x_{j-1})}{x_j - x_{j-1}}.$$

Take $X = \{x'_j : 1 \le j \le n\}$. Then

$$
\begin{aligned}
I_g(f,P,X) &= \sum_{j=1}^{n} f(x'_j)[g(x_j)-g(x_{j-1})] \\
&= \sum_{j=1}^{n} f(x'_j)g'(x'_j)[x_j - x_{j-1}] = I_x(fg',P,X).
\end{aligned}
$$

It follows that $fg'$ is R-S integrable with respect to $x$. By Exercise 6.D.F below, $fg'$ is Riemann integrable, and

$$
\int_a^b f(x)g'(x)\,dx = \int_a^b fg'\,dx = \int_a^b f\,dg. \qquad \blacksquare
$$

## Exercises for Section 6.D

**A.** Prove parts (2) and (3) of Theorem 6.D.3.

**B.** Let $f(x) = \operatorname{sign}(x)$, i.e., $f(0) = 0$ and $f(x) = x/|x|$ for $x \neq 0$. Show that if $g$ is the Heaviside step function, then $\int_{-1}^{1} f\,dg$ does not exist.

**C.** Suppose that $f,g$ are bounded on $[a,b]$ and $g$ is increasing on $[a,b]$. For any partition $P$ and any $\varepsilon > 0$, find evaluation sequences $X$ and $Y$ so that $I_g(f,P,X) > U_g(f,P) - \varepsilon$ and $I_g(f,P,Y) < L_g(f,P) + \varepsilon$.

**D.** Suppose $f$ is bounded on $[a,b]$, and continuous on except at $c$. Let $g$ be a function of bounded variation on $[a,b]$ which is continuous at $c$. Show that $f \in \mathscr{R}(g)$.

**E.** Suppose that $\sum_{i\geq 1} r_i$ is a convergent series of positive numbers and $x_i$ is a sequence of real numbers in $(a,b]$. Define $g(x) = \sum_{i=1}^{\infty} r_i H(x - x_i)$, where $H$ is the Heaviside step function. If $f$ is continuous on $[a,b]$, show that $f \in \mathscr{R}(g)$ and $\int_a^b f\,dg = \sum_{i=1}^{\infty} r_i f(x_i)$.

**F.** Suppose that $f$ is R-S integrable on $[a,b]$ with respect to $g(x) = x$ and has integral $L$. Prove that $f$ is Riemann integrable by finding a $\delta > 0$ for a given $\varepsilon > 0$ so that $|I(f,P,X) - L| < \varepsilon$ whenever $\operatorname{mesh}(P) < \delta$. HINT: Use the MVT and the fact that $g'$ is bounded.

**G.** Suppose that $g$ has bounded variation on $[a,b]$, and let $h(x) = V_a^x g$ for $x \in [a,b]$. If $f \in \mathscr{R}(g)$ on $[a,b]$, show that $f \in \mathscr{R}(h)$.

**H.** (a) Show that if $f \in \mathscr{R}(g)$, then $f^2 \in \mathscr{R}(g)$. HINT: Bound $I_g(f^2,P,X) - I_g(f^2,P,Y)$ in terms of $I_g(f,P,X) - I_g(f,P,Y)$.
(b) Show that if $f_1, f_2 \in \mathscr{R}(g)$, then $fg \in \mathscr{R}(g)$. HINT: Write $fg$ using squares.

**I.** Suppose that $g$ is a function on $[a,b]$ so that $C[a,b] \subset \mathscr{R}(g)$. Show that

$$
\sup\Big\{ \int_a^b f\,dg : f \in C[a,b], \|f\|_\infty \leq 1 \Big\} = V_a^b(g).
$$

**J.** Define $g : [0,1/\pi] \to \mathbb{R}$ by $g(0) = 0$ and $g(x) = x^{1/2}\sin(1/x)$ for $x > 0$. Find a continuous function $f : [0,1/\pi] \to \mathbb{R}$ so that $f \notin \mathscr{R}(g)$. Thus, in Corollary 6.D.11, we cannot change '$g$ of bounded variation' to '$g$ continuous'.

**K.** Define $h$ on $[0,1/\pi]$ by $h(0) = 0$ and $h(x) = x\sin(1/x)$ for $x > 0$. Find $f \in C[0,1/\pi]$ so that $f \notin \mathscr{R}(h)$. HINT: $h$ is differentiable on $[\varepsilon, 1/\pi]$. Try $f(x) = \cos(1/x)/\log x$.

## 7.A The $L^p$ norms

The $L^p$ norms on $C[a,b]$ for $1 \le p < \infty$ were defined in Example 7.1.5 by

$$\|f\|_p = \left( \int_a^b |f(x)|^p \, dx \right)^{1/p}.$$

The point of this section is to prove that these really are norms by proving the triangle inequality, which is the only part of the definition not established in Section 7.1.

There are other variants of the $L^p$ norms that are handled in exactly the same way. Here are two important examples.

### 7.A.1. Examples.

(1) For $1 \le p < \infty$, $\ell^p$ consists of the set of all infinite sequences $\mathbf{a} = (a_n)$ such that

$$\|\mathbf{a}\|_p = \Big( \sum_{n=1}^{\infty} |a_n|^p \Big)^{1/p} < \infty$$

and we call this norm the $\ell^p$ norm. Clearly, it is positive definite and homogeneous.

We can deduce the triangle inequality here from the main example by identifying the sequence $\mathbf{a}$ with the function $f$ on $[0,\infty)$ given by

$$f(x) = a_n \quad \text{for} \quad n-1 \le x < n, \quad n \ge 1.$$

This function has the property that $\|f\|_p = \|\mathbf{a}\|_p$. So we can deduce our inequality from the main case. You probably noticed that this function is not continuous. However, if we prove the result for piecewise continuous functions, we shall be able to apply the result here. (Alternatively, we could doctor this function so that we still obtain the correct integral and the function is continuous. But this is not as natural.)

(2) Let $w(x)$ be a strictly positive piecewise continuous function on $[a,b]$. Define a norm on $C[a,b]$ or on $PC[a,b]$, the space of piecewise continuous functions, by

$$\|f\|_{L^p(w)} = \left( \int_a^b |f(x)|^p w(x) \, dx \right)^{1/p}.$$

We call this the $L^p(w)$ norm. The standard case takes $w(x) = 1$. So if we prove the result for $L^p(w)$, then it will follow for $L^p$ and for $\ell^p$.

First we need an easy lemma from calculus.

**7.A.2. Lemma.** *Let $A, B > 0$. Then*

$$A^t B^{1-t} \le tA + (1-t)B \quad \textit{for all} \quad 0 < t < 1.$$

*Moreover, equality holds for some (or all) $t$ only if $A = B$.*

**PROOF.** Since $A > 0$, write $A = e^a$, where $a = \log A$. Similarly, $B = e^b$ for $b = \log B$. Substituting these formulas for $A$ and $B$, we have to prove that

$$e^{ta}e^{(1-t)b} = e^{ta+(1-t)b} \leq te^a + (1-t)e^b.$$

But the preceding inequality follows from the convexity of the exponential function on $\mathbb{R}$ (see Figure 7.2). As the exponential function *is* convex (apply Exercise 6.2.L), we are done. ■

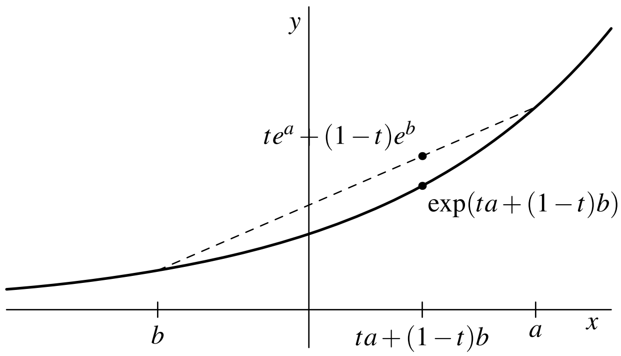


**FIG. 7.2** The convexity of $e^x$.

We can now prove one of the most important inequalities of analysis, the Hölder inequality.

### 7.A.3. HÖLDER'S INEQUALITY.

*Let $w$ be a positive function on an interval $[a,b]$. Let $f \in L^p(w)$ and $g \in L^q(w)$ where $1 < p < \infty$ and $\dfrac{1}{p} + \dfrac{1}{q} = 1$. Then*

$$\left| \int_a^b f(x)g(x)w(x)\,dx \right| \leq \|f\|_{L^p(w)} \|g\|_{L^q(w)}.$$

**PROOF.** We may assume that $f$ and $g$ are nonzero because both sides are easily seen to be 0 otherwise. Let

$$A = \frac{|f(x)|^p}{\|f\|^p_{L^p(w)}} \quad \text{and} \quad B = \frac{|g(x)|^q}{\|g\|^q_{L^q(w)}}.$$

Then take $t = 1/p$, so that $1 - t = 1/q$, and apply the lemma. We obtain

$$\frac{|f(x)|}{\|f\|_{L^p(w)}} \frac{|g(x)|}{\|g\|_{L^q(w)}} = A^t B^{1-t} \leq tA + (1-t)B = \frac{|f(x)|^p}{p\|f\|^p_{L^p(w)}} + \frac{|g(x)|^q}{q\|g\|^q_{L^q(w)}}.$$

We now multiply by $w(x)$ and integrate from $a$ to $b$.

$$\left| \int_a^b \frac{f(x)g(x)}{\|f\|_{L^p(w)}\|g\|_{L^q(w)}} w(x)\,dx \right| \le \int_a^b \frac{|f(x)|}{\|f\|_{L^p(w)}} \frac{|g(x)|}{\|g\|_{L^q(w)}} w(x)\,dx$$
$$\le \frac{1}{p}\int_a^b \frac{|f(x)|^p}{\|f\|^p_{L^p(w)}} w(x)\,dx + \frac{1}{q}\int_a^b \frac{|g(x)|^q}{\|g\|^q_{L^q(w)}} w(x)\,dx = \frac{1}{p}+\frac{1}{q} = 1$$

Multiplying by $\|f\|^p_{L^p(w)}\|g\|^q_{L^q(w)}$ gives the inequality. ■

## 7.A.4. Examples.

(1) $$\int_a^b f(x)g(x)\,dx \le \left(\int_a^b f(x)^p\,dx\right)^{1/p}\left(\int_a^b g(x)^q\,dx\right)^{1/q}$$

for all continuous functions $f$ and $g$ on $[a,b]$. This is just the standard weight $w(x) = 1$.

(2) If we take $p = 2$, then $q = 2$ and we obtain the Cauchy–Schwarz inequality, Theorem 7.4.4, for $L^2(w)$.

(3) As in Example 7.A.1(1), take $w(x) = 1$ on $[0,\infty)$ and consider $f(x) = a_n$ and $g(x) = b_n$ on $[n-1,n]$ for $n \ge 1$, where $(a_n) \in \ell^p$ and $(b_n) \in \ell^q$. Then

$$\sum_{n=1}^{\infty} a_n b_n = \int_0^\infty f(x)g(x)\,dx \le \left(\int_a^b f(x)^p\,dx\right)^{1/p}\left(\int_a^b g(x)^q\,dx\right)^{1/q}$$
$$= \left(\sum_{n=1}^{\infty}|a_n|^p\right)^{1/p}\left(\sum_{n=1}^{\infty}|b_n|^q\right)^{1/q}.$$

Again, this reduces to the Cauchy–Schwarz inequality if $p = q = 2$.

The triangle inequality now follows by another trick.

## 7.A.5. Minkowski's Inequality.

*The triangle inequality holds for $L^p(w)$, that is,*

$$\left(\int_a^b |f(x)+g(x)|^p w(x)dx\right)^{1/p} \le \left(\int_a^b |f(x)|^p w(x)dx\right)^{1/p} + \left(\int_a^b |g(x)|^p w(x)dx\right)^{1/p}.$$

**Proof.** Notice that $\frac{1}{q} = 1 - \frac{1}{p} = \frac{p-1}{p}$. So $q = \frac{p}{p-1}$. Thus

$$\|\,|f|^{p-1}\|_{L^q(w)} = \left(\int_a^b |f(x)|^{(p-1)q} w(x)\,dx\right)^{1/q}$$
$$= \left(\int_a^b |f(x)|^p w(x)\,dx\right)^{1/q} = \|f\|^{p/q}_{L^p(w)} = \|f\|^{p-1}_{L^p(w)}.$$

We use this equality in the last line of the next calculation, which estimates the norm of $f+g$ in $L^p(w)$,

$$\begin{aligned}
\|f+g\|^p_{L^p(w)} &= \int_a^b |f(x)+g(x)|^{p-1}|f(x)+g(x)|w(x)\,dx \\
&\le \int_a^b |f(x)+g(x)|^{p-1}|f(x)|w(x)\,dx + \int_a^b |f(x)+g(x)|^{p-1}|g(x)|w(x)\,dx \\
&\le \|\,|f+g|^{p-1}\|_{L^q(w)}\big(\|f\|_{L^p(w)} + \|g\|_{L^p(w)}\big) \\
&= \|f+g\|^{p-1}_{L^p(w)}\big(\|f\|_{L^p(w)} + \|g\|_{L^p(w)}\big).
\end{aligned}$$

Now divide by $\|f+g\|^{p-1}_{L^p(w)}$ to obtain $\|f+g\|_{L^p(w)} \le \|f\|_{L^p(w)} + \|g\|_{L^p(w)}$. ∎

## Exercises for Section 7.A

**A.** If $a_n \ge 0$ and $b_n \ge 0$ and $0 < t < 1$, show that $\sum\limits_{n=1}^{\infty} a_n^t b_n^{1-t} \le \big(\sum\limits_{n=1}^{\infty} a_n\big)^t \big(\sum\limits_{n=1}^{\infty} b_n\big)^{1-t}$.
HINT: Rework the Hölder inequality.

**B.** Show that the Hölder inequality is sharp in the sense that for every $f \in L^p(w)$, there is a nonzero $g \in L^q(w)$ so that $\big|\int_a^b f(x)g(x)w(x)\,dx\big| = \|f\|_{L^p(w)}\|g\|_{L^q(w)}$.
HINT: For nonzero $f$, set $g(x) = |f(x)|^p/f(x)$ if $f(x) \ne 0$ and $g(x) = 0$ if $f(x) = 0$.

**C.** Let $f(t)$ be a continuous, strictly increasing function on $[0,\infty)$ with $f(0) = 0$. Recall that $f$ has an inverse function $g$ with the same properties by Theorem 5.7.6. Define $F(x) = \int_0^x f(t)\,dt$ and $G(x) = \int_0^x g(t)\,dt$.

(a) Prove **Young's Inequality**: $xy \le F(x) + G(y)$ for all $x, y \ge 0$, with equality if and only if $f(x) = y$. HINT: Sketch $f$ and find two regions with areas $F(x)$ and $G(y)$.
(b) Take $f(x) = x^{p-1}$ for $1 < p < \infty$. What inequality do you get?

**D.** Let $f_n = n\chi_{[\frac{1}{n},\frac{2}{n}]}$ be defined on $[0,1]$ for $n \ge 2$.

(a) Show that $f_n$ converges pointwise to 0.
(b) Show that $f_n$ does not converge in $L^p[0,1]$ for any $1 \le p < \infty$.

**E.** (a) If $f \in C[0,1]$ and $1 \le r \le s < \infty$, show that $\|f\|_1 \le \|f\|_r \le \|f\|_s \le \|f\|_\infty$. HINT: Let $p = s/r$ and think of $\int_0^1 |f|^r \cdot 1\,dx$ as the product of an $L^p$ function and an $L^q$ function.
(b) Hence show that if $f_n \in C[0,1]$ converge uniformly to $f$, then they also converge in the $L^p$ norm for all $1 \le p < \infty$.

**F.** Let $f_n = \frac{1}{n}\chi_{[0,n^n]}$ be defined on $[0,\infty)$ for $n \ge 1$.

(a) Show that $f_n$ converges uniformly to 0.
(b) Show that $f_n$ does not converge in $L^p[0,1]$ for any $1 \le p < \infty$.

**G.** (a) If $\mathbf{a} = (a_k) \in \ell_1$ and $1 \le r \le s < \infty$, show that $\|\mathbf{a}\|_\infty \le \|\mathbf{a}\|_s \le \|\mathbf{a}\|_r \le \|\mathbf{a}\|_1$.
HINT: Show that for $p \ge 1$, $\sum_{k=1}^n |b_k|^p \le \big(\sum_{k=1}^n |b_k|\big)^p$.
(b) Hence show that $\ell_1 \subset \ell_r \subset \ell_s \subset \ell_\infty$.

**H.** Find continuous functions $f$ and $g$ on $[0,\infty)$ so that $f$ is in $L^1[0,\infty)$ but not in $L^2[0,\infty)$, and $g$ is in $L^2[0,\infty)$ but not in $L^1[0,\infty)$.

## 8.A Term by term differentiation

The purpose of this section is to provide another proof of Theorem 8.5.3 showing that term by term differentiation of power series is legitimate. Our original proof worked well because we integrate the series, and integration of series is better behaved than differentiation. We retrieve the result on differentiation by integrating the putative power series for the derivative. The proof we give here actually differentiates the power series by using a very clever trick and the Weierstrass M-test (8.4.7).

**8.A.1. TERM-BY-TERM DIFFERENTIATION OF POWER SERIES.** *If* $f(x) = \sum\limits_{n=0}^{\infty} a_n x^n$ *has radius of convergence* $R > 0$*, then* $\sum\limits_{n=1}^{\infty} n a_n x^{n-1}$ *has radius of convergence R, f is differentiable on* $(-R,R)$*, and*

$$f'(x) = \sum_{n=1}^{\infty} n a_n x^{n-1} \quad \textit{for} \quad x \in (-R,R).$$

**PROOF.** By Hadamard's Theorem 8.5.1, we know that

$$\frac{1}{R} = \limsup_{n\to\infty} |a_n|^{1/n}.$$

So the radius of convergence of the series $\sum_{n=1}^{\infty} n a_n x^{n-1} = \sum_{k=0}^{\infty} (k+1) a_{k+1} x^k$ is also $R$ because

$$\limsup_{k\to\infty} ((k+1)|a_{k+1}|)^{1/k} = \limsup_{k\to\infty} (k+1)^{1/k} \big(|a_{k+1}|^{1/k+1}\big)^{(k+1)/k} = \frac{1}{R}.$$

This works just as well for the series $\sum\limits_{n=1}^{\infty} n|a_n| x^{n-1}$. So in particular, if $0 < r < R$, we have

$$\sum_{n=1}^{\infty} n|a_n| r^{n-1} < \infty \quad \text{for all} \quad 0 \le r < R.$$

Here is the clever trick. For $0 < r < R$, define $F(x,y)$ on $[-r,r] \times [-r,r]$ by

$$F(x,y) = \sum_{n=1}^{\infty} a_n (x^{n-1} + x^{n-2}y + \cdots + xy^{n-2} + y^{n-1}) = \sum_{n=1}^{\infty} u_n(x,y)$$

where $u_n(x,y) = a_n(x^{n-1} + x^{n-2}y + \cdots + xy^{n-2} + y^{n-1})$. Observe that

$$\|u_n\|_\infty = \sup_{|x|\le r, |y|\le r} |u_n(x,y)| = |a_n| r^{n-1}.$$

Thus

$$\sum_{n=1}^{\infty} \|u_n\|_\infty = \sum_{n=1}^{\infty} n|a_n| r^{n-1} < \infty.$$

Hence this series converges uniformly on $[-r,r]\times[-r,r]$ by the Weierstrass M-test (8.4.7). It follows that $F(x,y)$ is continuous on $[-r,r]\times[-r,r]$.

Next observe that

$$F(x,x)=\sum_{n=1}^{\infty} na_n x^{n-1}.$$

While if $x\neq y$ and $n\geq 1$, we have

$$\frac{y^n-x^n}{y-x}=x^{n-1}+x^{n-2}y+\cdots+xy^{n-2}+y^{n-1}.$$

The power series converges absolutely, and therefore we can rearrange terms and obtain the same sum. Hence for $x\neq y$,

$$\begin{aligned}
F(x,y)&=\sum_{n=1}^{\infty} a_n(x^{n-1}+x^{n-2}y+\cdots+xy^{n-2}+y^{n-1})\\
&=\sum_{n=1}^{\infty} a_n\frac{y^n-x^n}{y-x}=\frac{1}{y-x}\sum_{n=0}^{\infty} a_n(y^n-x^n)\\
&=\frac{1}{y-x}\sum_{n=0}^{\infty} a_n y^n-\frac{1}{y-x}\sum_{n=0}^{\infty} a_n x^n\\
&=\frac{f(y)-f(x)}{y-x}.
\end{aligned}$$

We have shown that

$$F(x,y)=\begin{cases}\sum\limits_{n=1}^{\infty} na_n x^{n-1} & \text{for}\quad x=y\\[2mm] \dfrac{f(y)-f(x)}{y-x} & \text{for}\quad x\neq y.\end{cases}$$

Since $F$ is continuous, we obtain

$$\begin{aligned}
f'(x)&=\lim_{y\to x}\frac{f(y)-f(x)}{y-x}=\lim_{y\to x}F(x,y)\\
&=F(x,x)=\sum_{n=1}^{\infty} na_n x^{n-1}.
\end{aligned}$$

Therefore $f$ is differentiable, and the derivative is given by the term by term derivative of the power series. ■

## 8.B Abel's Theorem

When a power series converges at an endpoint of the interval of convergence, we can conclude something a lot stronger about the convergence on the whole interval.

### 8.B.1. Abel's Theorem.

*Suppose that* $\sum_{k=0}^{\infty} a_k$ *converges. Then the power series* $\sum_{k=0}^{\infty} a_k x^k$ *converges uniformly on* $[0,1]$ *to a continuous function* $f(x)$.

**Proof.** Since $\sum_{k=0}^{\infty} a_k$ converges, $\lim_{k\to\infty} a_k = 0$. Hence by Hadamard's Theorem 8.5.1, the power series $\sum_{k=0}^{\infty} a_k x^k$ has radius of convergence at least 1. Let $f(x)$ be the sum of this series for $-1 < x \le 1$. We know that convergence is uniform on $[-r,r]$ for $r < 1$, but Hadamard's Theorem does not tell us about convergence near $x = 1$.

Since $\sum_{k=0}^{\infty} a_k$ converges, the partial sums form a Cauchy sequence. Thus given $\varepsilon > 0$, there is an integer $N$ so that

$$\Big|\sum_{k=n+1}^{m} a_k\Big| < \varepsilon/2 \quad \text{for all} \quad N \le n < m.$$

We make use of the Dirichlet Test 3.B.2 for summation of a series using summation by parts. Fix $n \ge N$, and set $b_k = x^{n+k}$ for any $x \in [0,1)$. Since this sequence decreases monotonically to 0, Dirichlet's Test shows that $\sum_{k=1}^{\infty} a_{n+k} b_k = \sum_{k=1}^{\infty} a_{n+k} x^{n+k}$ converges, say to a function $f_n(x)$. In addition, Dirichlet's Test provides an estimate for the size of the sum, namely that $|f_n(x)| \le 2(\varepsilon/2) b_1 \le \varepsilon$. This estimate is independent of $x$, so we obtain

$$\sup_{0\le x<1} \Big|f(x) - \sum_{k=0}^{n} a_k x^k\Big| = \sup_{0\le x<1} |f_n(x)| \le \varepsilon.$$

We also have

$$\Big|f(1) - \sum_{k=0}^{n} a_k\Big| = \lim_{m\to\infty} \Big|\sum_{k=n+1}^{m} a_k\Big| \le \varepsilon/2.$$

This establishes that the power series converges uniformly to $f(x)$ on the whole interval $[0,1]$. In particular, $f(x)$ is continuous by Theorem 8.2.1. ■

**8.B.2. Corollary.** *If* $f(x) = \sum_{k=0}^{\infty} a_k x^k$ *is a power series with finite radius of convergence* $r$ *and* $\sum_{k=0}^{\infty} a_k r^k$ *converges, then* $\sum_{k=0}^{\infty} a_k r^k = \lim_{x\to r^-} f(x)$.

**Proof.** This reduces to Abel's Theorem for the series $\sum_{k=0}^{\infty}(a_k r^k)$. ■

## 9.A Connectedness

The Intermediate Value Theorem (5.6.1) shows that continuous real-valued functions map intervals onto intervals (see Corollary 5.6.2). This should mean that intervals have some special property not shared by other subsets of the line. This fact may appear to be dependent on the order structure. However, this property can be described in topological terms that allows a generalization to higher dimensions.

**9.A.1. DEFINITION.** A subset $A$ of a metric space $X$ is **not connected** if there are *disjoint* open sets $U$ and $V$ such that $A \subset U \cup V$ and $A \cap U \neq \varnothing \neq A \cap V$. Otherwise, the set $A$ is said to be **connected**.

A set $A$ is **totally disconnected** if for every pair of distinct points $x, y \in A$, there are two *disjoint* open sets $U$ and $V$ so that $x \in U$, $y \in V$ and $A \subset U \cup V$.

**9.A.2. EXAMPLES.**

(1) $A = [-1,0] \cup [3,4]$ is not connected in $\mathbb{R}$ because $U = (-2,1)$ and $V = (2,5)$ are disjoint open sets which each intersect $A$ and jointly contain $A$.

(2) The set $A = \{x \in \mathbb{R} : x \neq 0\}$ is not connected because $U = (-\infty, 0)$ and $V = (0, +\infty)$ provide the necessary separation. The sets $U$ and $V$ need not be a positive distance apart. Even the small gap created by omitting the origin disconnects the set.

(3) $\mathbb{Q}$ is totally disconnected. For if $x < y \in \mathbb{Q}$, choose an irrational number $z$ with $x < z < y$. Then $U = (-\infty, z) \ni x$ and $V = (z, \infty) \ni y$ are disjoint open sets such that $U \cup V$ contains $\mathbb{Q}$.

(4) It is more difficult to construct a closed set that is totally disconnected. However, the Cantor set $C$ is an example. Suppose that $x < y \in C$. Choose $N$ so that $y - x > 3^{-N}$. By Example 4.4.8, $C$ is obtained by repeatedly removing middle thirds from each interval. At the $N$th stage, $C$ is contained in the set $S_N$ that consists of $2^N$ intervals of length $3^{-N}$. Thus $x$ and $y$ belong to distinct intervals in $S_N$. Therefore, there is a point $z$ in the complement of $S_N$ such that $x < z < y$. Then $U = (-\infty, z) \ni x$ and $V = (z, \infty) \ni y$ are disjoint open sets such that $U \cup V$ contains $S_N$ and therefore contains $C$.

To obtain an example of a connected set, we must appeal to the Intermediate Value Theorem.

**9.A.3. PROPOSITION.** *The interval $[a,b]$ is connected.*

**PROOF.** Suppose to the contrary that there are disjoint open sets $U$ and $V$ such that $[a,b] \cap U$ and $[a,b] \cap V$ are both nonempty, and $[a,b] \subset U \cup V$. Define a function $f$ on $[a,b]$ by

$$f(x) = \begin{cases} 1 & \text{if} \quad x \in [a,b] \cap U \\ -1 & \text{if} \quad x \in [a,b] \cap V. \end{cases}$$

We claim that this function is continuous. Indeed, let $x$ be any point in $[a,b]$ and let $\varepsilon > 0$. If $x \in U$, then since $U$ is open, there is a positive number $r > 0$ such that $B_r(x)$ is contained in $U$. Thus $|x-y| < r$ implies that

$$|f(y) - f(x)| = |1-1| = 0 < \varepsilon.$$

So $f$ is continuous at $x$. Similarly, if $x \in V$, there is some ball $B_r(x)$ contained in $V$; and thus $|x-y| < r$ implies that

$$|f(y) - f(x)| = |-1-(-1)| = 0 < \varepsilon.$$

This accounts for every point in $[a,b]$, and therefore $f$ is a continuous function.

Thus we have constructed a continuous function on $[a,b]$ with range equal to $\{-1,1\}$. This contradicts the Intermediate Value Theorem. So our assumption that $[a,b]$ is not connected must be wrong. Therefore, the interval is connected. ■

To obtain a useful criterion for connectedness, we need to know that it is a property that is preserved by continuous functions. This is readily established using the topological characterization of continuity based on open sets, Theorem 9.1.7 (3).

**9.A.4. THEOREM.** *The continuous image of a connected set is connected.*

**PROOF.** This theorem is equivalent to the contrapositive statement: *If the image of a continuous function is not connected, then the domain is not connected.* To prove this, suppose that $f$ is a continuous function from a metric space $(X,\rho)$ into $(Y,\sigma)$ such that the range $f(X)$ is not connected. Then there are disjoint open sets $U$ and $V$ in $Y$ such that $f(X) \cap U$ and $f(S) \cap V$ are both nonempty, and $f(X) \subset U \cup V$.

Let $S = f^{-1}(U)$ and $T = f^{-1}(V)$. By Theorem 9.1.7, the inverse image of an open set is open; and thus both $S$ and $T$ are open. Since $U$ and $V$ are disjoint and cover $f(X)$, each $x \in X$ has either $f(x) \in U$ or $f(x) \in V$ but not both. Therefore, $S$ and $T$ are disjoint and cover $X$. Finally, since $U \cap f(X)$ is nonempty, it follows that $S \cap X$ is also nonempty; and likewise, $T \cap X$ is not empty. This shows that $X$ is not connected, as required. ■

**9.A.5. EXAMPLE.** In Section 5.6, we defined a path in $\mathbb{R}^n$. Similarly, we define a **path** in any metric space as a continuous image of a closed interval. A path is connected because a closed interval is connected. In particular, if $f$ is a continuous function of $[a,b]$ into $\mathbb{R}^n$, the graph $G(f) = \{(x, f(x)) : a \le x \le b\}$ is connected.

**9.A.6. DEFINITION.** A subset $S$ of a metric space $X$ is said to be **path connected** if for each pair of points $x$ and $y$ in $S$, there is a **path** in $S$ connecting $x$ to $y$, meaning that there is a continuous function $f$ from $[0,1]$ into $S$ such that $f(0) = x$ and $f(1) = y$.

**9.A.7. COROLLARY.** *Every path connected set is connected.*

**PROOF.** Let $S$ be a path connected set, and suppose to the contrary that it is not connected. Then there are disjoint open sets $U$ and $V$ such that

$$S \subset U \cup V \quad \text{and} \quad S \cap U \neq \varnothing \neq S \cap V.$$

Pick points $u \in S \cap U$ and $v \in S \cap V$. Since $S$ is path connected, there is a path $P$ connecting $u$ to $v$ in $S$. It follows that $P \subset U \cup V$. And $P \cap U$ contains $u$ and $P \cap V$ contains $v$. Thus both are nonempty. Consequently,the path $P$ is not connected. This contradicts the previous example. Therefore, $S$ must be connected. ■

## 9.A.8. EXAMPLES.

(1) $\mathbb{R}^n$ is path connected and therefore connected.

(2) Recall that a subset of $\mathbb{R}^n$ is *convex* if the straight line between any two points in the set belongs to the set. Therefore, any convex subset of $\mathbb{R}^n$ is path connected.

(3) The annulus $\{x \in \mathbb{R}^2 : 1 \leq \|x\| \leq 2\}$ is path connected and hence connected.

(4) There are connected sets that are not path connected. An example is the compact set

$$S = Y \cup G = \{(0,t) : |t| \leq 1\} \cup \{(x, \sin \tfrac{1}{x}) : 0 < x \leq 1\}.$$

To see that $S$ is connected, suppose that $U$ and $V$ are two disjoint open sets such that $S$ is contained in $U \cup V$. The point $(0,0)$ belongs to one of them, which may be called $U$. Therefore, $U$ intersects the line $Y$. Since $Y$ is path connected, it must be completely contained in $U$ (for if $V$ intersects $Y$, this would show that $Y$ was disconnected).

Since $U$ is open, it contains a ball of positive radius $r > 0$ around the origin. But $(0,0)$ is a limit point of $G$, and therefore $U$ contains points of $G$. Since $G$ is a graph, it is path connected. So by the same argument, $U$ contains $G$. Therefore, $V \cap S$ is empty. This shows that $S$ is connected.

If $S$ were path connected, there would be a continuous function $f$ of $[0,1]$ into $S$ with $f(0) = (0,0)$ and $f(1) = (1, \sin 1)$. Let $a = \sup\{x : f(x) \in Y\}$. Since $f$ is continuous, there is a $\delta > 0$ such that $\|f(x) - f(a)\| \leq 0.5$ if $|x - a| \leq \delta$. Set $b = a + \delta$. Let $f(a) = (0,y)$ and $f(b) = (u, \sin \frac{1}{u}) =: \mathbf{u}$. A similar argument shows that there is a point $a < c < b$ such that $f(c) = (t, \sin \frac{1}{t}) =: \mathbf{t}$, where $t \leq \dfrac{u}{1+2\pi u}$. Therefore, $f([c,b])$ is a connected subset of $S$ containing both $\mathbf{t}$ and $\mathbf{u}$. The graph $G' = \{(x, \sin \frac{1}{x}) : t \leq x \leq u\}$ is connected, but removing any point will disconnect it. So $f([c,b])$ contains $G'$. However, $\sin \frac{1}{x}$ takes both values $\pm 1$ on $[t,u]$, and so there is a point $\mathbf{v} = (v, \sin \frac{1}{v})$ on $G'$ with $\|\mathbf{v} - f(a)\| > |\sin \frac{1}{v} - y| \geq 1$. This contradicts the estimate $\|f(x) - f(a)\| \leq 0.5$ for all $x \in [c,b]$. Therefore, $S$ is not path connected.

## Exercises for Section 9.A

**A.** Show that the only connected subsets of $\mathbb{R}$ are intervals (which may be finite or infinite and may or may not include the endpoints).

**B.** Show that if $S$ is a connected subset of $X$, then so is $\overline{S}$.

**C.** Let $A \subset \mathbb{R}^m$ and $B \subset \mathbb{R}^n$ be connected sets. Prove that $A \times B$ is connected in $\mathbb{R}^{m+n}$.
HINT: Recall that $A \times B = \{(a,b) : a \in A,\, b \in B\}$.

**D.** Show that a connected open set $U \subset \mathbb{R}^n$ is path connected. HINT: Fix $u_0 \in U$. The set $V$ of points in $U$ which are path connected to $u_0$ is open, as is the set $W$ of points in $U$ which are not path connected to $u_0$.

**E.** Find an example of a decreasing sequence $A_1 \supset A_2 \supset A_3 \cdots$ of closed connected sets in $\mathbb{R}^2$ such that $\bigcap_{k \ge 1} A_k$ is not connected.

**F.** Let $A_1 \supset A_2 \supset A_3 \cdots$ be a decreasing sequence of connected compact subsets of $\mathbb{R}^n$. Show that $\bigcap_{k \ge 1} A_k$ is connected. HINT: Show that if $U \cup V$ is an open set containing the intersection, then it contains some $A_n$ for $n$ sufficiently large.

**G.** Show that there is no continuous bijection of $[-1,1]$ onto the circle.
HINT: Such a function would map $[-1,0) \cup (0,1]$ onto a connected set.

**H.** Let $(Y,\rho)$ be a subset of a metric space $(X,\rho)$. Show that $Y$ is connected (as a subset of itself) if and only if it is connected in $X$.
HINT: If $U$ and $V$ are nonempty disjoint open subsets of $Y$ with $Y = U \cup V$, let $f(y) = \max\{\operatorname{dist}(y,U), \operatorname{dist}(y,V)\}$. Then $U_1 = \bigcup\{B_{f(y)/2}(y) : y \in U\}$ is open in $X$.

## 9.B Metric Completion

Completeness is such a powerful property that we attempt to work in a complete context whenever possible. The purpose of this section is to show that every metric space may be naturally imbedded in a unique complete metric space known as its completion. We do this by an efficient, if sneaky, argument based on the completeness of $C_b(X)$, the space of bounded real functions on $X$.

In principle, this method could be applied to the rational numbers to obtain the real numbers in a base independent way. You should be concerned that our argument uses the real numbers in its proof. Because of this circularity, we cannot construct the real numbers this way. Nevertheless, we can show that the real line is unique and hence the various different constructions we have mentioned all give the same object. This puts the real numbers on a firmer theoretical footing.

In the next section, the completeness theorem will be applied to $C[a,b]$ with the $L^p$ norms to obtain the $L^p$ spaces. The completeness of these normed spaces allows us to use the full power of analysis in studying them.

**9.B.1. DEFINITION.** Let $(X,\rho)$ be a metric space. A **completion** of $(X,\rho)$ is a complete metric space $(Y,d)$ together with a map $T$ of $X$ into $Y$ such that $d(Tx_1,Tx_2) = \rho(x_1,x_2)$ for all $x_1,x_2 \in X$ (an **isometry**) and $TX$ is dense in $Y$.

When the metric space and its completion are concretely represented, often the map $T$ is an inclusion map, as in the next example. By an **inclusion map**, we mean that $T : X \to Y$, where $X \subset Y$ and $T$ is given by $T(x) = x$. Except for changing the codomain, $T$ is the identity map on $X$.

**9.B.2. EXAMPLE.** Let $A$ be an arbitrary subset of $\mathbb{R}^n$ with the usual Euclidean distance. Then the closure $\overline{A}$ of $A$ is a complete metric space (see Exercise 4.3.H). Evidently, the distance functions coincide on these two sets, so the inclusion map is trivially an isometry. As $A$ is dense in $\overline{A}$, this is a completion of $A$.

**9.B.3. THEOREM.** *Every metric space $(X,\rho)$ has a completion.*

**PROOF.** Pick a point $x_0 \in X$. For each $x \in X$, define $f_x(y) = \rho(x,y) - \rho(x_0,y)$. If $\rho(y_1,y_2) < \varepsilon/2$, then $|\rho(x,y_1) - \rho(x,y_2)| \le \rho(y_1,y_2) < \varepsilon/2$. So

$$|f_x(y_1) - f_x(y_2)| \le |\rho(x,y_1) - \rho(x,y_2)| + |\rho(x_0,y_1) - \rho(x_0,y_2)| < \varepsilon.$$

Therefore, $f_x$ is (uniformly) continuous.

Observe that $f_{x_1}(y) - f_{x_2}(y) = \rho(x_1,y) - \rho(x_2,y)$. By the triangle inequality, $\rho(x_1,y) \le \rho(x_1,x_2) + \rho(x_2,y)$, and thus $f_{x_1}(y) - f_{x_2}(y) \le \rho(x_1,x_2)$. Interchanging $x_1$ and $x_2$ yields $\|f_{x_1} - f_{x_2}\|_\infty \le \rho(x_1,x_2)$. However, substituting $y = x_2$ shows that this is an equality. In particular as $f_{x_0} = 0$, we see that $f_x$ is bounded. Thus $Tx = f_x$ is an isometric imbedding of $X$ into $C_b(X)$.

Let $F = TX = \{f_x : x \in X\}$, and consider $\overline{F}$, the closure of $F$ in $C_b(X)$. This is a closed subset of a complete normed space, and therefore it is complete. Evidently, $\overline{F}$ is a metric completion of $X$. ■

It is important that this completion is unique in a natural sense. This will justify our use of the terminology *the completion* and will allow us to consider $X$ as a subset of $C$ without explicit use of the isometry $T$.

To prove uniqueness, we need a significant result about continuous functions on the metric completion. To simplify notation, we drop the map $T$ and consider $X$ as a subset of its completion $C$, with the same notation for the two metrics. A continuous function $g$ on $C$ is a **continuous extension** of a function $f$ on $X$ if $g|_X = f$.

### 9.B.4. Extension Theorem.

*Let $(X,\rho)$ be a metric space with metric completion $(C,\rho)$. Let $f$ be a uniformly continuous function from $(X,\rho)$ into a complete metric space $(Y,\sigma)$. Then $f$ has a unique uniformly continuous extension $g$ mapping $C$ into $Y$.*

**Proof.** Let $c$ be a point in $C$. Since $X$ is dense in $C$, we may choose a sequence $(x_n)$ in $X$ converging to $c$. We claim that $(f(x_n))$ is a Cauchy sequence in $Y$. Indeed, let $\varepsilon > 0$ be given. By the uniform continuity of $f$, there is a $\delta > 0$ so that $\rho(x,x') < \delta$ implies $\sigma(f(x),f(x')) < \varepsilon$. Since $(x_n)$ is Cauchy, there is an integer $N$ so that $\sigma(x_m,x_n) < \delta$ for all $m,n \geq N$. Hence $\sigma(f(x_m),f(x_n)) < \varepsilon$ for all $m,n \geq N$. Therefore, $(f(x_n))$ is Cauchy and we may define $g(c)$ to be $\lim_{n\to\infty} f(x_n)$. Moreover, this is the *only* possible choice for a continuous extension of $f$.

We must verify that $g$ is well defined, meaning that any other sequence converging to $c$ will yield the same value for $g(c)$. Consider a second sequence $(x'_n)$ in $X$ with $\lim_{n\to\infty} x'_n = c$. Then $(x_1,x'_1,x_2,x'_2,\dots)$ is another sequence converging to $c$. By the previous paragraph, $(f(x_1),f(x'_1),f(x_2),f(x'_2),\dots)$ is a Cauchy sequence in $Y$. Therefore $\lim_{n\to\infty} f(x'_n) = \lim_{n\to\infty} f(x_n)$; and $g$ is well defined.

To verify that $g$ is continuous, let $\varepsilon > 0$. Let $\delta > 0$ be chosen as before. Suppose that $c,d$ are points in $C$ with $\rho(c,d) < \delta$ and that $(x_n)$ and $(x'_n)$ are sequences in $X$ converging to $c$ and $d$, respectively. Then $\lim_{n\to\infty} \rho(x_n,x'_n) = \rho(c,d) < \delta$. Thus there is an $N$ so that $\rho(x_n,x'_n) < \delta$ for all $n \geq N$. It follows that

$$\sigma(g(c),g(d)) = \lim_{n\to\infty} \sigma(f(x_n),f(x'_n)) \leq \varepsilon.$$

Consequently, $g$ is uniformly continuous. ■

**9.B.5. Corollary.** *The metric completion $(C,\rho)$ of $(X,\rho)$ is unique in the sense that if $S$ is an isometry of $X$ into another completion $(Y,\sigma)$, then there is a unique isometry $S'$ of $C$ onto $Y$ extending $S$.*

**Proof.** As in Theorem 9.B.4, we consider $X$ as a subset of $C$ with the same metric. The map $S$ is an isometry and thus is uniformly continuous. Let $S'$ be the unique continuous extension of $S$ from $C$ into $Y$. If $c,d \in C$ with $c = \lim_{n\to\infty} x_n$ and $d = \lim_{n\to\infty} y_n$, then

$$\sigma(S'c, S'd) = \lim_{n\to\infty} \sigma(Sx_n, Sy_n) = \lim_{n\to\infty} \rho(x_n, y_n) = \rho(c,d).$$

So $S'$ is also an isometry.

Conversely, there is an isometry $T'$ of $Y$ into $C$ extending the map $S^{-1}$ from $SX$ onto $X$. Notice that $T'S'$ is therefore a continuous extension of the identity map on $X$ and hence equals the identity map on $C$. Likewise, $S'T'$ is the identity on $Y$. In particular, for each $y$ in $Y$, $S'(T'y) = y$ and so $S$ maps $C$ onto $Y$. ■

## Exercises for Section 9.B

**A.** Show that the metric completion of a normed vector space is a complete normed vector space.
HINT: Use the Extension Theorem to extend the vector space operations.

**B.** Prove that the map taking each bounded uniformly continuous function on a metric space $X$ to its continuous extension on the completion $C$ is an isometry from $BUC(X)$ *onto* $BUC(C)$, where $BUC(X)$ is the normed space of bounded uniformly continuous functions on $X$.

**C.** Show by example that there is a bounded continuous function on $(0,1)$ that does not extend to a continuous function on the completion.

**D.** Let $(X,\rho)$ and $(Y,\sigma)$ be metric spaces with completions $(C,\rho)$ and $(D,\sigma)$.

(a) Show that $d\big((x_1,y_1),(x_2,y_2)\big) = \rho(x_1,x_2) + \sigma(y_1,y_2)$ is a metric on $X \times Y$.
(b) Show that $C \times D$ is the completion of $X \times Y$.

**E.** Let $(X,\rho)$ be the metric on all finite sequences of letters with the dictionary metric of Example 9.1.J. Describe the metric completion.
HINT: Consider the natural extension of this metric to all infinite words.

**F.** Show that the completion of $(X,\rho)$ is compact if and only if $X$ is totally bounded.

**G.** **The $p$-adic numbers**. Fix a prime number $p$. For each $n \in \mathbb{Z}$, let $\operatorname{ord}_p(n)$ denote the largest integer $d$ such that $p^d$ divides $n$. Extend this function to all rational numbers by setting $\operatorname{ord}_p(a/b) = \operatorname{ord}_p(a) - \operatorname{ord}_p(b)$.

(a) Show that $\operatorname{ord}_p(r)$ is independent of the representation of $r$ as a fraction.
(b) Set $|r|_p = p^{-\operatorname{ord}_p(r)}$ and $\rho_p(r,s) = |r-s|_p$. Prove that $\rho_p$ is a metric on $\mathbb{Q}$.
HINT: In fact, $\rho_p(r,t) \le \max\{\rho_p(r,s), \rho_p(s,t)\}$.
(c) Let $\mathbb{Q}_p$ be the metric completion of $(\mathbb{Q},\rho_p)$. Prove that addition extends to a continuous operation on $\mathbb{Q}_p$.
(d) Show that $|xy|_p = |x|_p|y|_p$. Hence establish continuous extension of multiplication to $\mathbb{Q}_p$.
(e) Verify that every nonzero element of $\mathbb{Q}_p$ has an inverse.
(f) Why doesn't this complete field containing $\mathbb{Q}$ contradict the uniqueness of the reals?

**H.** We further develop properties of the $p$-adic numbers. This requires a bit of number theory.

(a) If $r \in \mathbb{Q}$ and $|r|_p = p^{-k}$, find $a_k \in \{0,1,\ldots,p-1\}$ so that $|r - a_k p^k|_p \le p^{-k-1}$.
HINT: Write $r = p^k \frac{a}{b}$ where $\gcd(p,b) = 1$. By the Euclidean algorithm, we can write $1 = mb + np$. Let $a_k \equiv am \pmod{p}$.
(b) Hence show that every element $r \in \mathbb{Q}$ with $|r|_p \le 1$ is the limit in the $\rho_p$ metric of a sequence of integers of the form $a_0 + a_1 p + \cdots + a_n p^n$, where $a_k \in \{0,1,\ldots,p-1\}$.
(c) Extend (b) to every element $x \in \mathbb{Q}_p$ with $|x|_p \le 1$. Hence deduce that $\mathbb{Z}_p$, the closure of $\mathbb{Z}$ in $\mathbb{Q}_p$, consists of all these points.
(d) Prove that $\mathbb{Z}_p$ is compact.

## 9.C Uniqueness of the Real Number System.

We apply the results of the previous section to the real number system. The real numbers have several important properties, which were discussed at length in Chapter 2. To establish uniqueness in some sense, we need to be explicit about the required structure. First, $\mathbb{R}$ should be a field that contains the rational numbers $\mathbb{Q}$. In addition to the algebraic operations, it is important to recognize that $\mathbb{R}$ has an order extending the order on $\mathbb{Q}$. Second, $\mathbb{R}$ should be complete in some sense. The sense that fits well with analysis is completeness as a metric space and containing $\mathbb{Q}$ as a dense subset. A subtlety that immediately presents itself is, Where does this metric take its values? Fortunately, the metric on $\mathbb{Q}$ takes values in $\mathbb{Q}$, but any completions of $\mathbb{Q}$ will have to take metric values in $\mathbb{R}$! So this would still appear to depend on the original construction of $\mathbb{R}$.

Therefore, it seems prudent to introduce a different and more algebraic notion of completeness, using order. A field is an **ordered field** if it contains a subset $P$ of positive elements with the properties that

(i) every element belongs to exactly one of $P$, $\{0\}$ or $-P$; and
(ii)whenever $x, y \in P$, then $x+y$ and $xy$ belong to $P$.

Then we define $x < y$ if $y - x \in P$. An ordered field is **order complete** if every nonempty set that is bounded above has a least upper bound. This notion doesn't mention metric and thus is free from the difficulty mentioned previously about where the metric takes its values. The density of the rationals is also addressed by order. The rationals are **order dense** if whenever $x < y$, there is a rational $r$ so that $x < r < y$.

### 9.C.1. UNIQUENESS OF $\mathbb{R}$.

(1) *The rational numbers $\mathbb{Q}$ have a unique metric completion with a unique field structure extending the field operations on $\mathbb{Q}$ making addition and multiplication continuous. The order structure is also unique.*

(2) *The rational numbers are contained as an order dense subfield of a unique order complete field.*

**PROOF.** (1) The metric space of real numbers $\mathbb{R}$ is a metric completion of $\mathbb{Q}$, by the completeness of $\mathbb{R}$ (Theorem 2.8.5). By Corollary 9.B.5, it follows that if $\mathbb{S}$ is any metric space completion of $\mathbb{Q}$, then there is an isometry $T$ of $\mathbb{R}$ onto $\mathbb{S}$ that extends the identity map on $\mathbb{Q}$. The Extension Theorem guarantees that there is only one way to extend addition and multiplication on $\mathbb{Q}$ to a continuous function on $\mathbb{S}$. One such way is to transfer the operations on $\mathbb{R}$ by $Tx + Ty = T(x+y)$ and $Tx \cdot Ty = T(xy)$. Uniqueness says that this is the only way, and thus the field operations on $\mathbb{S}$ are equivalent to those on $\mathbb{R}$. The positive elements of $\mathbb{R}$ are exactly those nonzero elements that have a square root. Since this is determined by the algebra structure, we see that order is also preserved.

(2) The real numbers are order complete by the Least Upper Bound Principle (2.3.3), and the rationals are order dense by construction (see Exercise 2.2.D). Suppose that $\mathbb{S}$ is another order complete field containing $\mathbb{Q}$ as an order dense subfield. Define a map $T$ from $\mathbb{R}$ to $\mathbb{S}$ by defining $T(x)$ to the least upper bound in $\mathbb{S}$ of $\{r \in \mathbb{Q} : r < x\}$. Notice that $Tq = q$ for $q \in \mathbb{Q}$. Indeed, $q$ is an upper bound for $\{r \in \mathbb{Q} : r < q\}$. If $Tq < q$, the order density of $\mathbb{Q}$ would imply that there is an $r \in \mathbb{Q}$ such that $Tq < r < q$, contradicting the fact that $Tq$ is the least upper bound for this set.

Next we show that $T$ is a bijection. Clearly, if $x \leq y$, then $Tx \leq Ty$. If $x < y$ in $\mathbb{R}$, then there is a rational $r$ with $x < r < y$. Consequently, $Tx < r < Ty$ and so $T$ is one-to-one and preserves the order. Now suppose that $s \in \mathbb{S}$ and let $S = \{r \in \mathbb{Q} : r < s\}$. This set has a least upper bound $x$ in $\mathbb{R}$. Then $Tx$ is the least upper bound of $S$ in $\mathbb{S}$. Again $s$ is an upper bound by definition and if $Tx < s$, the order density would yield a rational $r$ with $Tx < r < s$. Hence $Tx = s$ and the map $T$ is onto.

Let us verify that $T$ preserves addition and multiplication. The fact that $P + P \subset P$ means that addition preserves order: if $x_1 < x_2$ amd $y_1 < y_2$, then $x_1 + y_1 < x_2 + y_2$. Thus in both $\mathbb{R}$ and $\mathbb{S}$, we have that

$$x + y = \sup\{r + s : r, s \in \mathbb{Q},\ r < x,\ s < y\}.$$

Since $T$ preserves order and therefore sups, it follows that $Tx + Ty = T(x + y)$. Similarly, if $x, y$ are *positive*,

$$xy = \sup\{rs : r, s \in \mathbb{Q} \cap P,\ r < x,\ s < y\}.$$

So $(Tx)(Ty) = T(xy)$ for $x, y \in P$. The rest follows since

$$(-x)y = x(-y) = -(xy) \qquad \text{and} \qquad (-x)(-y) = xy$$

allow the extension of multiplication to the whole field. This establishes that the map $T$ preserves all of the field operations and the order. ■

So Part (2) establishes, without reference to any properties based on our particular construction of $\mathbb{R}$, that there is only one field that has the properties we want. This we call the real numbers. We could, for example, define the real numbers using expansions in base 2 or base 3. We have implicitly assumed up to now that these constructions yield the same object. We now know this to be the case.

There are other constructions of the real numbers that do not depend on a base. This has a certain esthetic appeal but does not in itself address the uniqueness question. One approach is to consider Cauchy sequences of rational numbers as representing points (see Exercise 2.8.L). This is essentially the metric completion, but with a subtle difference that a metric is only defined after the space $\mathbb{R}$ is constructed (as the set of equivalence classes), and takes values in the positive part of $\mathbb{R}$ itself. So there is no *a priori* notion of general metric space using a real valued metric. We have to decide when two Cauchy sequences should represent the same point. This is done in much the same way as we proved the Extension Theorem (9.B.4). If you take two Cauchy sequences $(r_k)$ and $(s_k)$ in $\mathbb{Q}$, and combine them as $(r_1, s_1, r_2, s_2, \dots)$, then this new sequence is Cauchy if and only if the two sequences have the same limit. So we say that the sequences are equivalent when this combined sequence is Cauchy, thereby avoiding the need to talk about convergence at all. The ideas in the proof of the Extension Theorem make it possible to extend addition and multiplication to this completion of $\mathbb{Q}$, making it into an ordered field.

In 1858, Dedekind described a formal construction of the real numbers that did not require the use of any base nor any notion of limit at all. He noticed that for each real number $x$, there was an associated set $S_x = \{r \in \mathbb{Q} : r < x\}$ of rational numbers. This determines a different set of rational numbers for each real $x$. Of course, we defined these sets using the real numbers. But we can turn it around. Dedekind considered all sets $S$ of rational numbers that have the properties

(1) $S$ is a nonempty subset of $\mathbb{Q}$ that is bounded above,
(2) $S$ does not contain its upper bound, and
(3) if $s \in S$ and $r < s$ for $r \in \mathbb{Q}$, then $r \in S$.

These sets are known as **Dedekind cuts**. He then associated a point $\mathbf{x}$ to each of these sets. For convenience, call the set $\mathbf{x}$ as well. In particular, each rational number $r$ is associated to the set $\mathbf{r} = \{q \in \mathbb{Q} : q < r\}$.

We can then go on to define orders, arithmetic operations, and limits. Order is defined by set inclusion: $\mathbf{x} \leq \mathbf{y}$ if $\mathbf{x} \subseteq \mathbf{y}$. It is not hard to deduce from Dedekind's axioms that either $\mathbf{x} < \mathbf{y}$, $\mathbf{x} = \mathbf{y}$ or $\mathbf{x} > \mathbf{y}$. One major payoff of this construction is that the least upper bound property is easy to establish. If $A$ is a non-empty subset of $\mathbb{R}$ which is bounded about by $M$, then the supremum is $\sup A = \bigcup_{\mathbf{x} \in A} \mathbf{x}$. It is clear that this union is a set satisfying Dedekind's axioms, and it is the least upper bound.

Addition also has an easy definition:

$$\mathbf{x} + \mathbf{y} = \{q + r : q \in \mathbf{x},\ r \in \mathbf{y}\}.$$

Additive inverses are a bit subtle:

$$-\mathbf{x} = \{r \in \mathbb{Q} : \text{there is an } s \notin \mathbf{x} \text{ such that } r + s < 0\}.$$

Multiplication is trickier, but for $\mathbf{x} > \mathbf{0}$ and $\mathbf{y} > \mathbf{0}$, we can define

$$\mathbf{xy} = \{r \in \mathbb{Q} : r \leq 0\} \cup \{rs : r > 0,\ s > 0, \text{ and } r \in \mathbf{x},\ s \in \mathbf{y}\}.$$

With a bit of work, one can verify all of the ordered field axioms.

This somewhat artificial construction finally freed the definition of $\mathbb{R}$ from reliance on intuitive notions.

Both of these constructions yield a complete structure; Dedekind's is order complete and Cauchy's is metrically complete. But there remains in both cases the tedious, but basically easy, task of defining addition and multiplication and verifying all of the ordered field properties. Indeed, we did not verify all of these details for our decimal construction either.

Another practical approach is to take *any* construction of the real numbers and show that every number has a decimal expansion. This is an alternate route to the uniqueness theorem. The reader interested in the details of these foundational issues should consult Landau's *Foundations of Analysis*.

## 9.D The $L^p$ Spaces and Abstract Integration

In Section 7.A, the $L^p$ norms for $p \geq 1$ were introduced and shown to be norms via the Hölder and Minkowski inequalities. However, these norms were put on $C[a,b]$, and it is easy to show that $C[a,b]$ is not complete in any of these norms. The completion process of Section 9.B may be used to remedy this situation. One caveat to the reader is that there is a better, although more involved, method of defining the $L^p$ spaces using measure theory. Measure theory provides a powerful way to develop not only integration on $\mathbb{R}$, but also integration on many other spaces, and provides a setting for probability theory. Nevertheless, we are able to achieve some useful things by our abstract process, including recognizing $L^p$ as a space of functions (or, more precisely, as equivalence classes of functions).

**9.D.1. DEFINITION.** The space $L^p[a,b]$ is the completion of $C[a,b]$ in the $L^p$ norm $\|f\|_p = \Big( \int_a^b |f(x)|^p \, dx \Big)^{1/p}$.

At this point, we know that $L^p[a,b]$ is a complete normed space by Exercise 9.B.A. In particular, $\|f\|_p$ is the distance in $L^p(a,b)$ to the zero function. However, the elements of this space are no longer represented by functions. We will rectify that in this section. The first step is to extend our integration theory to $L^p$ spaces.

Keeping in mind that we expect to represent elements of $L^p(a,b)$ as functions, we will write elements as $f, g$, and so on. In particular, we shall say that $f \geq 0$ if it is the limit in the $L^p$ norm of a sequence of positive continuous functions $f_n$. So we say $f \leq g$ if $g - f \geq 0$.

**9.D.2. THEOREM.** *The Riemann integral has a unique continuous extension to $L^1(a,b)$ denoted by $\int f$. It has the properties*

(1) $\int sf + tg = s \int f + t \int g$ *for all $f, g \in L^1(a,b)$ and $s,t \in \mathbb{R}$.*
(2) $\big| \int f \big| \leq \int |f| = \|f\|_1$ *for $f \in L^1(a,b)$.*
(3) *If $f \geq 0$, then $\int f \geq 0$.*

**PROOF.** The Riemann integral on $C[a,b]$ satisfies the properties (1)–(3). In particular, it is uniformly continuous in the $L^1$ norm by (2) since for $f, g \in C[a,b]$,

$$\left| \int_a^b f(x)\, dx - \int_a^b g(x)\, dx \right| \leq \int_a^b |f(x) - g(x)|\, dx = \|f - g\|_1.$$

Therefore, by the Extension Theorem, there is a unique continuous extension to $L^1(a,b)$, which we call $\int f$. Properties (1)–(3) follow by taking limits. ■

The integral defined in this manner is called the **Daniell integral**. The alternative is the **Lebesgue integral**, which is constructed using measure theory. Although the constructions are different, the two integrals have the same theory. They are powerful for two reasons. The first reason is that the $L^p$ spaces of integrable functions are complete. For the Lebesgue integral, this is a theorem, proved using properties of the integral; whereas in our case, the complete space comes first and the properties of the integral are a theorem. The second reason is that there are much better limit theorems than for the Riemann integral. Here is first of these. The best limit theorem, the Dominated Convergence Theorem, we leave as an exercise.

This result, like Theorem 2.6.1, is traditionally called the Monotone Convergence Theorem. It is usually easy to distinguish between the two, as one result applies to sequences of real numbers and the other to sequences of functions.

### 9.D.3. Monotone Convergence Theorem.

*Suppose $(f_n) \in L^1(a,b)$ is an increasing sequence such that $\sup_{n\geq 1} \int f_n < \infty$. Then $f_n$ converges in the $L^1(a,b)$ norm to an element $f$ and $\int f = \lim\limits_{n\to\infty} \int f_n$.*

**Proof.** By Property (2) of Theorem 9.D.2, $\int f_n$ is an increasing sequence of real numbers that is bounded above. Hence by the Monotone Convergence Theorem for real numbers, $L = \lim\limits_{n\to\infty} \int f_n$ exists. So if $\varepsilon > 0$, there is an integer $N$ so that

$$L - \varepsilon < \int f_N \leq \int f_n \leq L \quad \text{for all} \quad n \geq N.$$

Thus when $N \leq m \leq n$,

$$\|f_n - f_m\| = \int |f_n - f_m| = \int f_n - f_m = \int f_n - \int f_m < \varepsilon.$$

Therefore, $(f_n)$ is Cauchy in $L^1(a,b)$. It follows that $f = \lim\limits_{n\to\infty} f_n$ exists. Since the integral is continuous, $\int f = \lim\limits_{n\to\infty} \int f_n$. ■

The preceding analysis would appear to deal only with $L^1(a,b)$. For $L^p(a,b)$, we obtain much the same result by extending the Hölder inequality. Minkowski's inequality, which is just the triangle inequality, is immediately satisfied because the metric $d(f,g) = \|f-g\|_p$ satisfies the triangle inequality.

**9.D.4. Theorem.** *Let $1 < p,q < \infty$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$. If $f \in L^p(a,b)$ and $g \in L^q(a,b)$, then $fg$ 'belongs' to $L^1(a,b)$ and $\left|\int fg\right| \leq \|f\|_p \|g\|_q$.*

**Proof.** We make $L^p(a,b) \times L^q(a,b)$ into a complete metric space with the metric $d\big((f_1,g_1),(f_2,g_2)\big) = \|f_1 - f_2\|_p + \|g_1 - g_2\|_q$. By Exercise 9.B.D, this is the completion of $C[a,b] \times C[a,b]$ with the same metric. By the Hölder inequality (7.A.3), the map $\Phi(f,g) = fg$ maps $C[a,b] \times C[a,b]$ into $L^1(a,b)$ and

$$\|fg\|_1 = \int_a^b |f(x)g(x)|\,dx \leq \|f\|_p \|g\|_q.$$

This map is not uniformly continuous on the whole product, but it is uniformly continuous on balls. So fix $r \geq 0$ and consider

$$X_r = \big\{(f,g) \in C[a,b] \times C[a,b] : \|f\|_p \leq r \text{ and } \|g\|_q \leq r\big\}.$$

Then for $(f_1,g_1)$ and $(f_2,g_2)$ in $X_r$,

$$\begin{aligned}\|\Phi(f_1,g_1) - \Phi(f_2,g_2)\|_1 &\leq \|f_1(g_1-g_2)\|_1 + \|(f_1-f_2)g_2\|_1 \\ &\leq \|f_1\|_p \|g_1-g_2\|_q + \|f_1-f_2\|_p \|g_2\|_q \\ &\leq r\big(\|f_1-f_2\|_p + \|g_1-g_2\|_q\big) = r\,d\big((f_1,g_1),(f_2,g_2)\big).\end{aligned}$$

So $\Phi$ is uniformly continuous on $X_r$. By the Extension Theorem (9.B.4), $\Phi$ extends to a uniformly continuous function on the completion of $X_r$, which is its closure in $L^p(a,b) \times L^q(a,b)$, namely

$$\overline{X_r} = \big\{(f,g) \in L^p(a,b) \times L^q(a,b) : \|f\|_p \leq r \text{ and } \|g\|_q \leq r\big\}.$$

In particular, if $f \in L^p(a,b)$ and $g \in L^q(a,b)$, then $\Phi(f,g) = fg$ belongs to $L^1(a,b)$. Moreover, $\|\Phi(f,g)\|_1 \leq \|f\|_p \|g\|_q$ by continuity since this holds on $C[a,b] \times C[a,b]$. Moreover, since $\int$ is uniformly continuous on $L^1(a,b)$, we obtain Hölder's inequality:

$$\big|\textstyle\int fg\big| = \big|\textstyle\int \Phi(f,g)\big| \leq \|\Phi(f,g)\|_1 \leq \|f\|_p \|g\|_q. \qquad \blacksquare$$

We turn to the problem of representing each element of $L^p$ as a function. We already know we must allow discontinuous functions, as $C[a,b]$ is not complete in the $L^p$ norm. But this allows the possibility of having a function that is zero except at a countable set of points. Such a function will have integral zero, and then the $L^p$ norm will not be positive definite. So we must identify this function with the zero function.

In general, we must identify functions that are equal on "negligble" sets. That is, each element of $L^p$ will be an equivalence class of functions that are equal almost everywhere, in the sense of Section 6.C. We return to this issue after proving the next theorem.

Measure theory has something useful to add to this picture, since there is a notion of measurable function that provides a natural set to choose the $L^p$ functions from. However, this approach cannot get around the essential difficulty that an element of $L^p(a,b)$ is not actually a function.

**9.D.5. Theorem.** *If $f \in L^p(a,b)$, we may choose a sequence $(f_n)$ in $C[a,b]$ converging to $f$ in $L^p(a,b)$ so that $\lim\limits_{n\to\infty} f_n(x)$ converges almost everywhere* (a.e.) *to a function $f(x)$. Moreover, if $(g_n)$ is another sequence converging to $f$ in $L^1(a,b)$ and pointwise to $g(x)$* a.e.*, then $g(x) = f(x)$* a.e.

**Proof.** Choose any sequence $(f_n)$ in $C[a,b]$ such that $\|f - f_n\|_p < 4^{-n}$ for $n \geq 1$. Let

$$U_n = \{x \in [a,b] : |f_n(x) - f_{n+1}(x)| > 2^{-n}\}.$$

This is an open set since $|f_n - f_{n+1}|$ is continuous. Hence there are disjoint intervals $(c_i, d_i)$ so that $U_n = \bigcup_{i \geq 1}(c_i, d_i)$. Now

$$\|f_n - f_{n+1}\|_p \leq \|f_n - f\|_p + \|f - f_{n+1}\|_p < 2 \cdot 4^{-n}.$$

On the other hand,

$$\|f_n - f_{n+1}\|_p^p \geq \sum_{i \geq 1} \int_{c_i}^{d_i} |f_n(x) - f_{n+1}(x)|^p \, dx \geq 2^{-np} \sum_{i \geq 1} d_i - c_i.$$

Therefore,

$$\sum_{i \geq 1} d_i - c_i < (2 \cdot 4^{-n})^p 2^{np} = 2^{(1-n)p}.$$

Let $E = \bigcap_{k \geq 1} \bigcup_{n \geq k} U_n$. This set may be covered by the open intervals making up $\bigcup_{n \geq k} U_n$, which have total length at most $\sum_{n \geq k} 2^{(1-n)p} = 2^{-kp}\left(\frac{2^p}{1-2^{-p}}\right)$. This converges to 0 as $k$ increases and thus may be made less than any given $\varepsilon > 0$. Therefore, $E$ has measure zero.

If $x \in [a,b] \setminus E$, then there is an integer $N = N(x)$ so that $x$ is not in $\bigcup_{n \geq N} U_n$. Hence $|f_n(x) - f_{n+1}(x)| \leq 2^{-n}$ for all $n \geq N$. Thus if $N \leq n \leq m$,

$$|f_n(x) - f_m(x)| \leq \sum_{k=n}^{m-1} |f_k(x) - f_{k+1}(x)| \leq \sum_{k=n}^{m-1} 2^{-k} < 2^{1-n}.$$

Therefore $(f_n(x))$ is a Cauchy sequence, and $\lim\limits_{n \to \infty} f_n(x)$ converges, say to $f(x)$. That is, this sequence converges almost everywhere.

Now suppose that $(g_n)$ is another sequence in $C[a,b]$ converging to $f$ in the $L^p$ norm and converging almost everywhere to a function $g$. Drop to a subsequence $(g_{n_k})$ so that $\|g_{n_k} - f\|_p < 4^{-k}$ for $k \geq 1$. Let

$$V_k = \{x \in [a,b] : |g_{n_k}(x) - f_k(x)| > 2^{-k}\}.$$

This is an open set because $|g_{n_k} - f_k|$ is continuous. Now

$$\|g_{n_k} - f_k\|_p \leq \|g_{n_k} - f\|_p + \|f - f_n\|_p < 2 \cdot 4^{-k}.$$

Arguing as previously, the intervals making up $V_k$ have length at most $2^{(1-n)p}$. Consequently, as before, the set $F = \bigcap_{k \geq 1} \bigcup_{n \geq k} V_n$ has measure zero.

If $x \in [a,b] \setminus F$, then there is an $N = N(x)$ so that $x$ is not in $\bigcup_{n \geq N} V_n$. Hence $\lim\limits_{k \to \infty} |g_{n_k}(x) - f_k(x)| = 0$. Now $f_k(x)$ converges to $f(x)$ for $x \in [a,b] \setminus E$. Therefore, $g(x) = f(x)$ for every $x \in [a,b] \setminus (E \cup F)$. Thus $g = f$ a.e. ■

One of the subtleties of $L^p$ spaces is that the elements of $L^p[a,b]$ are not really functions. The preceding theorem shows that they may be represented by functions but that representation is not unique. Two functions that differ on a set of measure zero both represent the same element of $L^p[a,b]$. The correct way to handle this is to identify elements of $L^p[a,b]$ with an equivalence class of functions that agree almost everywhere. (See Section 1.3.)

Another point that bears repeating is that in this approach, the integral of these limit functions is determined abstractly. The Lebesgue integral develops a method of integrating these functions in terms of their values. This is an important viewpoint that we do not address here. We refer the interested reader to the books [11], [12], or [13].

## Exercises for Section 9.D

**A.** Show that every piecewise continuous function on $[a,b]$ is simultaneously the limit in $L^1(a,b)$ and the pointwise limit everywhere of a sequence of continuous functions.

**B.** (a) Show that the characteristic function $\chi_U$ of any open set $U \subset [a,b]$ is the increasing limit of a sequence of continuous functions.
(b) Hence show that $\chi_U$ belongs to $L^1(a,b)$. If $U$ is the union of countably many disjoint open intervals $U_n = (c_n, d_n)$, show that $\int \chi_U = \|\chi_U\|_1 = \sum_n |U_n|$.

**C.** Suppose that $f \in L^1[a,b]$ and $f \geq 0$ and $\int f = 0$. Prove that $f = 0$ a.e.
HINT: Apply the Monotone Convergence Theorem (MCT) with $f_n = nf$.

**D.** (a) Suppose that $A \subset [a,b]$ has measure zero. Show that there is a decreasing sequence of open sets $U_n$ containing $A$ such that $\lim\limits_{n\to\infty} \int \chi_{U_n} = 0$.
(b) If $f$ is a positive element of $L^1(a,b)$ that is represented by a function $f(x)$ such that $f(x) = 0$ a.e., prove that $\|f\|_1 = 0$. HINT: Let $A_n = \{x : 0 < f(x) < n\}$. Consider $f\chi_{A_n}$.
(c) Suppose $f \in L^1(a,b)$ is represented by a function $f(x)$ such that $f(x) = 0$ a.e. Prove that $\int |f| = 0$ (i.e. $f = 0$ in $L^1(a,b)$).

**E.** Suppose that $f_n \in C[a,b]$ is a decreasing sequence of continuous functions converging pointwise to 0. Prove that $\lim\limits_{n\to\infty} \int_a^b f_n(x)\,dx = 0$. HINT: Exercise 8.1.I

**F.** (a) Show that if $f,g,h \in L^1[a,b]$, then $f \vee g := \frac{1}{2}(f + g + |f - g|)$ satisfies $f \leq f \vee g$ and $g \leq f \vee g$. If $f \leq h$ and $g \leq h$, show that $f \vee g \leq h$. That is, $f \vee g = \max\{f,g\}$.
(b) Show that $f \wedge g = \min\{f,g\}$ may be defined as $f \wedge g = -(-f \vee -g)$.

**G.** **Dominated Convergence Theorem**. Suppose that $f_n, g \in L^1[a,b]$ and $|f_n| \leq g$ for all $n \geq 1$. Also suppose that $\lim\limits_{n\to\infty} f_n(x) = f(x)$ a.e.

(a) Define $u_{mn} = f_m \vee f_{m+1} \vee \cdots \vee f_n$ and $l_{mn} = f_m \wedge f_{m+1} \wedge \cdots \wedge f_n$. Show that $u_m = \lim\limits_{n\to\infty} u_{mn}$ belongs to $L^1(a,b)$, as does $l_m = \lim\limits_{n\to\infty} l_{mn}$. HINT: MCT
(b) Show that $u_m$ is a decreasing sequence of functions converging to $f$ almost everywhere, and $l_m$ is an increasing sequence converging to $f$ almost everywhere.
(c) Show that $\lim\limits_{m\to\infty} \int u_m = \lim\limits_{m\to\infty} \int l_m = \int f$. HINT: MCT
(d) Prove that $\lim\limits_{n\to\infty} \int f_n = \int f$. HINT: $l_n \leq f_n \leq u_n$

**H.** **Lebesgue Measure**. Let $\Sigma$ consist of all subsets $A$ of $[a,b]$ such that $\chi_A \in L^1(a,b)$ (i.e., there is an $f \in L^1(a,b)$ with $f(x) = \chi_A$ a.e.). Define $m : \Sigma \to [0,+\infty)$ by $m(A) = \int \chi_A$.

(a) Show that if $A, B \in \Sigma$, then $A \cap B$, $A \cup B$ and $A \setminus B$ belong to $\Sigma$.
(b) Show that $m(A \cap B) + m(A \cup B) = m(A) + m(B)$ for all $A, B \in \Sigma$.
(c) Show that $m(A) = 0$ if and only if $A$ has measure zero.

(d) **Countable Additivity.** If $A_n$ are disjoint sets in $\Sigma$, show that $\bigcup_{n\geq 1} A_n$ belongs to $\Sigma$, and $m\big(\bigcup_{n\geq 1} A_n\big) = \sum_{n\geq 1} m(A_n)$.

(e) Show that $\Sigma$ contains all open and all closed subsets of $[a,b]$; and verify that $m((c,d)) = m([c,d]) = d - c$.

# Index

A **boldface** page number indicates a definition or primary entry.