ANALYSIS OF FUNCTIONS OF A SINGLE VARIABLE
A DETAILED DEVELOPMENT

LAWRENCE W. BAGGETT

University of Colorado

OCTOBER 29, 2006

For Christy
My Light

PREFACE

I have written this book primarily for serious and talented mathematics scholars , seniors or first-year graduate students, who by the time they finish their schooling should have had the opportunity to study in some detail the great discoveries of our subject. What did we know and how and when did we know it? I hope this book is useful toward that goal, especially when it comes to the great achievements of that part of mathematics known as analysis. I have tried to write a complete and thorough account of the elementary theories of functions of a single real variable and functions of a single complex variable. Separating these two subjects does not at all jive with their development historically, and to me it seems unnecessary and potentially confusing to do so. On the other hand, functions of several variables seems to me to be a very different kettle of fish, so I have decided to limit this book by concentrating on one variable at a time.

Everyone is taught (told) in school that the area of a circle is given by the formula $A = \pi r^2$. We are also told that the product of two negatives is a positive, that you cant trisect an angle, and that the square root of 2 is irrational. Students of natural sciences learn that $e^{i\pi} = -1$ and that $\sin^2 + \cos^2 = 1$. More sophisticated students are taught the Fundamental Theorem of calculus and the Fundamental Theorem of Algebra. Some are also told that it is impossible to solve a general fifth degree polynomial equation by radicals. On the other hand, very few people indeed have the opportunity to find out precisely why these things are really true, and at the same time to realize just how intellectually deep and profound these "facts" are. Indeed, we mathematicians believe that these facts are among the most marvelous accomplishments of the human mind. Engineers and scientists can and do commit such mathematical facts to memory, and quite often combine them to useful purposes. However, it is left to us mathematicians to share the basic knowledge of why and how, and happily to us this is more a privilege than a chore. A large part of what makes the verification of such simple sounding and elementary truths so difficult is that we of necessity must spend quite a lot of energy determining what the relevant words themselves really mean. That is, to be quite careful about studying mathematics, we need to ask very basic questions: What is a circle? What are numbers? What is the definition of the area of a set in the Euclidean plane? What is the precise definition of numbers like π, i, and e? We surely cannot prove that $e^{i\pi} = -1$ without a clear definition of these particular numbers. The mathematical analysis story is a long one, beginning with the early civilizations, and in some sense only coming to a satisfactory completion in the late nineteenth century. It is a story of ideas, well worth learning.

There are many many fantastic mathematical truths (facts), and it seems to me that some of them are so beautiful and fundamental to human intellectual development, that a student who wants to be called a mathematician, ought to know how to explain them, or at the very least should have known how to explain them at some point. Each professor might make up a slightly different list of such truths. Here is mine:

(1) The square root of 2 is a real number but is not a rational number.
(2) The formula for the area of a circle of radius r is $A = \pi r^2$.
(3) The formula for the circumference of a circle of radius r is $C = 2\pi r$.
(4) $e^{i\pi} = -1$.
(5) The Fundamental Theorem of Calculus, $\int_a^b f(t)\,dt = F(b) - F(a)$.

(6) The Fundamental Theorem of Algebra, every nonconstant polynomial has at least one root in the complex numbers.
(7) It is impossible to trisect an arbitrary angle using only a compass and straight edge.

Other mathematical marvels, such as the fact that there are more real numbers than there are rationals, the set of all sets is not a set, an arbitrary fifth degree polynomial equation can not be solved in terms of radicals, a simple closed curve divides the plain into exactly two components, there are an infinite number of primes, etc., are clearly wonderful results, but the seven in the list above are really of a more primary nature to me, an analyst, for they stem from the work of ancient mathematicians and except for number 7, which continues to this day to evoke so-called disproofs, have been accepted as true by most people even in the absence of precise "arguments" for hundreds if not thousands of years. Perhaps one should ruminate on why it took so long for us to formulate precise definitions of things like numbers and areas?

Only with the advent of calculus in the seventeenth century, together with the contributions of people like Euler, Cauchy, and Weierstrass during the next two hundred years, were the first six items above really proved, and only with the contributions of Galois in the early nineteenth century was the last one truly understood.

This text, while including a traditional treatment of introductory analysis, specifically addresses, as kinds of milestones, the first six of these truths and gives careful derivations of them. The seventh, which looks like an assertion from geometry, turns out to be an algebraic result that is not appropriate for this course in analysis, but in my opinion it should definitely be presented in an undergraduate algebra course. As for the first six, I insist here on developing precise mathematical definitions of all the relevant notions, and moving step by step through their derivations. Specifically, what are the definitions of $\sqrt{2},\ A,\ \pi,\ r,\ r^2,\ C,\ 2,\ e,\ i,$, and -1? My feeling is that mathematicians should understand exactly where these concepts come from in precise mathematical terms, why it took so long to discover these definitions, and why the various relations among them hold.

The numbers $-1, 2,$ and i can be disposed of fairly quickly by a discussion of what exactly is meant by the real and complex number systems. Of course, this is in fact no trivial matter, having had to wait until the end of the nineteenth century for a clear explanation, and in fact I leave the actual proof of the existence of the real numbers to an appendix. However, a complete mathematics education ought to include a study of this proof, and if one finds the time in this analysis course, it really should be included here. Having a definition of the real numbers to work with, i.e., having introduced the notion of least upper bound, one can relatively easily prove that there is a real number whose square is 2, and that this number can not be a rational number, thereby disposing of the first of our goals. All this is done in Chapter I. Maintaining the attitude that we should not distinguish between functions of a real variable and functions of a complex variable, at least at the beginning of the development, Chapter I concludes with a careful introduction of the basic properties of the field of complex numbers.

unlike the elementary numbers $-1, 2,$ and i, the definitions of the real numbers e and π are quite a different story. In fact, one cannot make sense of either e or π until a substantial amount of analysis has been developed, for they both are necessarily defined somehow in terms of a limit process. I have chosen to define e here as the limit of the rather intriguing sequence $\{(1+\frac{1}{n})^n\}$, in some ways the first nontrivial example of a convergent sequence, and this is presented in Chapter II. Its relation to logarithms and exponentials, whatever they are, has to be postponed to Chapter IV. Chapter II also contains a section on the elementary topological properties (compactness, limit points, etc.) of the real and complex numbers as well as a thorough development of infinite series.

To define π as the ratio of the circumference of a circle to its diameter is attractive, indeed was quite acceptable to Euclid, but is dangerously imprecise unless we have at the outset a clear definition of what is meant by the length of a curve, e.g., the circumference of a circle. That notion is by no means trivial, and in fact it only can be carefully treated in a development of analysis well after other concepts. Rather, I have chosen to define π here as the smallest positive zero of the sine function. Of course, I have to define the sine function first, and this is itself quite deep. I do it using power series functions, choosing to avoid the common definition of the trigonometric functions in terms of " wrapping" the real line around a circle, for that notion again requires a precise definition of arc length before it would make sense. I get to arc length eventually, but not until Chapter VI.

In Chapter III I introduce power series functions as generalizations of polynomials, specifically the three power series functions that turn out to be the exponential, sine, and cosine functions. From these definitions it follows directly that $\exp iz = \cos z + i \sin z$ for every complex number z. Here is a place where allowing the variable to be complex is critical, and it has cost us nothing. However, even after establishing that there is in fact a smallest positive zero of the sine function (which we decide to call π, since we know how we want things to work out), one cannot at this point deduce that $\cos\pi = -1$, so that the equality $e^{i\pi} = -1$ also has to wait for its derivation until Chapter IV. In fact, more serious, we have no knowledge at all at this point of the function e^z for a complex exponent z. What does it mean to raise a real number, or even an integer, to a complex exponent? The very definition of such a function has to wait.

Chapter III also contains all the standard theorems about continuous functions, culminating with a lengthy section on uniform convergence, and finally Abel's fantastic theorem on the continuity of a power series function on the boundary of its disk of convergence.

The fourth chapter begins with all the usual theorems from calculus, Mean Value Theorem, Chain Rule, First Derivative Test, and so on. Power series functions are shown to be differentiable, from which the law of exponents emerges for the power series function exp. Immediately then, all of the trigonometric and exponential identities are also derived. We observe that $e^r = \exp(r)$ for every rational number r, and we at last can define consistently e^z to be the value of the power series function $\exp(z)$ for any complex number z. From that, we establish the equation $e^{i\pi} = -1$. Careful proofs of Taylor's Remainder Theorem and L'Hopital's Rule are given, as well as an initial approach to the general Binomial Theorem for non-integer exponents.

It is in Chapter IV that the first glimpse of a difference between functions of a real variable and functions of a complex variable emerges. For example, one of the results in this chapter is that every differentiable, real-valued function of a complex variable must be a constant function, something that is certainly not true for functions of a real variable. At the end of this chapter, I briefly slip into the realm of real-valued functions of two real variables. I introduce the definition of differentiability of such a function of two real variables, and then derive the initial relationships among the partial derivatives of such a function and the derivative of that function thought of as a function of a complex variable. This is obviously done in preparation for Chapter VII where holomorphic functions are central.

Perhaps most well-understood by math majors is that computing the area under a curve requires Newton's calculus, i.e., integration theory. What is often overlooked by students is that the very definition of the concept of area is intimately tied up with this integration theory. My treatment here of integration differs from most others in that the class of functions defined as integrable are those that are uniform limits of step functions. This is a smaller collection of functions than those that are Riemann-integrable, but they suffice for my purposes, and this approach serves to emphasize the importance of uniform convergence. In particular, I include careful proofs of the Fundamental Theorem of Calculus, the integration by substitution theorem, the integral form of Taylor's Remainder Theorem, and the complete proof of the general Binomial Theorem.

Not wishing to delve into the set-theoretic complications of measure theory, I have chosen only to define the area for certain "geometric" subsets of the plane. These are those subsets bounded above and below by graphs of continuous functions. Of course these suffice for most purposes, and in particular circles are examples of such geometric sets, so that the formula $A = \pi r^2$ can be established for the area of a circle of radius r. Chapter V concludes with a development of integration over geometric subsets of the plane. Once again, anticipating later needs, we have again strayed into some investigations of functions of two real variables.

Having developed the notions of arc length in the early part of Chapter VI, including the derivation of the formula for the circumference of a circle, I introduce the idea of a contour integral, i.e., integrating a function around a curve in the complex plane. The Fundamental Theorem of Calculus has generalizations to higher dimensions, and it becomes Green's Theorem in 2 dimensions. I give a careful proof in Chapter VI, just over geometric sets, of this rather complicated theorem.

Perhaps the main application of Green's Theorem is the Cauchy Integral Theorem, a result about complex-valued functions of a complex variable, that is often called the Fundamental Theorem of Analysis. I prove this theorem in Chapter VII. From this Cauchy theorem one can deduce the usual marvelous theorems of a first course in complex variables, e.g., the Identity Theorem, Liouville's Theorem, the Maximum Modulus Principle, the Open Mapping Theorem, the Residue Theorem, and last but not least our mathematical truth number 6, the Fundamental Theorem of Algebra. That so much mathematical analysis is used to prove the fundamental theorem of algebra does make me smile. I will leave it to my algebraist colleagues to point out how some of the fundamental results in analysis require substantial algebraic arguments.

The overriding philosophical point of this book is that many analytic assertions in mathematics are intellectually very deep; they require years of study for most people to understand; they demonstrate how intricate mathematical thought is and how far it has come over the years. Graduates in mathematics should be proud of the degree they have earned, and they should be proud of the depth of their understanding and the extremes to which they have pushed their own intellect. I love teaching these students, that is to say, I love sharing this marvelous material with them.

Larry Baggett
29 October 2006

CONTENTS

CHAPTER I
THE REAL AND COMPLEX NUMBERS
DEFINITION OF THE NUMBERS 1, i, AND $\sqrt{2}$

In order to make precise sense out of the concepts we study in mathematical analysis, we must first come to terms with what the "real numbers" are. Everything in mathematical analysis is based on these numbers, and their very definition and existence is quite deep. We will, in fact, not attempt to demonstrate (prove) the existence of the real numbers in the body of this text, but will content ourselves with a careful delineation of their properties, referring the interested reader to an appendix for the existence and uniqueness proofs.

Although people may always have had an intuitive idea of what these real numbers were, it was not until the nineteenth century that mathematically precise definitions were given. The history of how mathematicians came to realize the necessity for such precision in their definitions is fascinating from a philosophical point of view as much as from a mathematical one. However, we will not pursue the philosophical aspects of the subject in this book, but will be content to concentrate our attention just on the mathematical facts. These precise definitions are quite complicated, but the powerful possibilities within mathematical analysis rely heavily on this precision, so we must pursue them. Toward our primary goals, we will in this chapter give definitions of the symbols (numbers) $-1,\ i$, and $\sqrt{2}$.

The main points of this chapter are the following:

(1) The notions of **least upper bound** (*supremum*) and **greatest lower bound** (*infimum*) of a set of numbers,
(2) The definition of the **real numbers** $\mathbb{R}$,
(3) the formula for the sum of a **geometric progression** (Theorem 1.9),
(4) the **Binomial Theorem** (Theorem 1.10), and
(5) the **triangle inequality** for complex numbers (Theorem 1.15).

THE NATURAL NUMBERS AND THE INTEGERS

We will take for granted that we understand the existence of what we call the *natural numbers*, i.e., the set $\mathbb{N}$ whose elements are the numbers $1, 2, 3, 4, \ldots$. Indeed, the two salient properties of this set are that (a) there is a frist element (the natural number 1), and (b) for each element n of this set there is a "very next" one, i.e., an immediate successor. We assume that the algebraic notions of sum and product of natural numbers is well-defined and familiar. These operations satisfy three basic relations:

BASIC ALGEBRAIC RELATIONS.

(1) (Commutativity) $n + m = m + n$ and $n \times m = m \times n$ for all $n, m \in \mathbb{N}$.
(2) (Associativity) $n + (m + k) = (n + m) + k$ and $n \times (m \times k) = (n \times m) \times k$ for all $n, m, k \in \mathbb{N}$.
(3) (Distributivity) $n \times (m + k) = n \times m + n \times k$ for all $n, m, k \in \mathbb{N}$.

We also take as given the notion of one natural number being larger than another one. $2 > 1$, $5 > 3$, $n + 1 > n$, etc. We will accept as true the **axiom of mathematical induction,** that is, the following statement:

AXIOM OF MATHEMATICAL INDUCTION. *Let S be a subset of the set $\mathbb{N}$ of natural numbers. Suppose that*

(1) $1 \in S$.

(2) *If a natural number k is in S, then the natural number $k+1$ also is in S.*

Then $S = \mathbb{N}$. That is, every natural number n belongs to S.

REMARK. The axiom of mathematical induction is for our purposes frequently employed as a method of proof. That is, if we wish to show that a certain proposition holds for all natural numbers, then we let S denote the set of numbers for which the proposition is true, and then, using the axiom of mathematical induction, we verify that S is all of $\mathbb{N}$ by showing that S satisfies both of the above conditions.
Mathematical induction can also be used as a method of definition. That is, using it, we can define an infinite number of objects $\{O_n\}$ that are indexed by the natural numbers. Think of S as the set of natural numbers for which the object O_n is defined. We check first to see that the object O_1 is defined. We check next that, if the object O_k is defined for a natural number k, then there is a prescribed procedure for defining the object O_{k+1}. So, by the axiom of mathematical induction, the object is defined for all natural numbers. This method of defining an infinite set of objects is often referred to as sl recursive definition, or *definition by recursion.*
As an example of recursive definition, let us carefully define *exponentiation.*

DEFINITION. Let a be a natural number. We define inductively natural numbers a^n as follows: $a^1 = a$, and, whenever a^k is defined, then a^{k+1} is defined to be $a \times a^k$.

. The set S of all natural numbers for which a^n is defined is therefore all of $\mathbb{N}$. For, a^1 is defined, and if a^k is defined there is a prescription for defining a^{k+1}. This "careful" definition of a^n may seem unnecessarily detailed. Why not simply define a^n as $a \times a \times a \times a \ldots \times a$ n times? The answer is that the $\ldots$, though suggestive enough, is just not mathematically precise. After all, how would you explain what $\ldots$ means? The answer to that is that you invent a recursive definition to make the intuitive meaning of the $\ldots$ mathematically precise. We will of course use the symbol $\ldots$ to simplify and shorten our notation, but keep in mind that, if pressed, we should be able to provide a careful definition.

Exercise 1.1. (a) Derive the three laws of exponents for the natural numbers:
$a^{n+m} = a^n \times a^m$.
HINT: Fix a and m and use the axiom of mathematical induction.
$a^{n\times m} = (a^m)^n$.
HINT: Fix a and m and use the axiom of mathematical induction.
$(a \times b)^n = a^n \times b^n$.
HINT: Fix a and b and use the axiom of mathematical induction.
(b) Define inductively numbers $\{S_i\}$ as follows: $S_1 = 1$, and if S_k is defined, then S_{k+1} is defined to be $S_k + k + 1$. Prove, by induction, that $S_n = n(n+1)/2$. Note that we could have defined S_n using the $\ldots$ notation by $S_n = 1 + 2 + 3 + \ldots + n$.
(c) Prove that

$$1 + 4 + 9 + 16 + \ldots + n^2 = \frac{n(n+1)(2n+1)}{6}.$$

(d) Make a recursive definition of $n! = 1 \times 2 \times 3 \times \ldots \times n$. $n!$ is called n factorial.

There is a slightly more general statement of the axiom of mathematical induction, which is sometimes of use.

GENERAL AXIOM OF MATHEMATICAL INDUCTION. *Let S be a subset of the set $\mathbb{N}$ of natural numbers, and suppose that S satisfies the following conditions*

(1) *There exists a natural number k_0 such that $k_0 \in S$.*

(2) *If S contains a natural number k, then S contains the natural number $k+1$.*

Then S contains every natural number n that is larger than or equal to k_0.

From the fundamental set $\mathbb{N}$ of natural numbers, we construct the set $\mathbb{Z}$ of all integers. First, we simply create an additional number called 0 that satisfies the equations $0+n=n$ for all $n \in \mathbb{N}$ and $0 \times n = 0$ for all $n \in \mathbb{N}$. The word "create" is, for some mathematicians, a little unsettling. In fact, the idea of zero did not appear in mathematics until around the year 900. It is easy to see how the so-called natural numbers came by their name. Fingers, toes, trees, fish, etc., can all be counted, and the very concept of counting is what the natural numbers are about. On the other hand, one never needed to count zero fingers or fish, so that the notion of zero as a number easily could have only come into mathematics at a later time, a time when arithmetic was becoming more sophisticated. In any case, from our twenty-first century viewpoint, 0 seems very understandable, and we won't belabor the fundamental question of its existence any further here.

Next, we introduce the so-called *negative numbers*. This is again quite reasonable from our point of view. For every natural number n, we let $-n$ be a number which, when added to n, give 0. Again, the question of whether or not such negative numbers exist will not concern us here. We simply create them.

In short, we will take as given the existence of a set $\mathbb{Z}$, called the *integers*, which comprises the set $\mathbb{N}$ of natural numbers, the additional number 0, and the set $-\mathbb{N}$ of all negative numbers. We assume that addition and multiplication of integers satisfy the three basic algebraic relations of commutativity, associativity, and distributivity stated above. We also assume that the following additional relations hold:

$$(-n) \times (-k) = n \times k, \text{ and } (-n) \times k = n \times (-k) = -(n \times k)$$

for all natural numbers n and k.

THE RATIONAL NUMBERS

Next, we discuss the set $\mathbb{Q}$ of rational numbers, which we ordinarily think of as quotients k/n of integers. Of course, we do not allow the "second" element n of the quotient k/n to be 0. Also, we must remember that there isn't a 1-1 correspondence between the set $\mathbb{Q}$ of all rational numbers and the set of all such quotients k/n. Indeed, the two distinct quotients 2/3 and 6/9 represent the same rational number. To be precise, the set $\mathbb{Q}$ is a collection of equivalence classes of ordered pairs (k,n) of integers, for which the second component of the pair is not 0. The equivalence relation among these ordered pairs is this:

$$(k,n) \equiv (k',n') \text{ if } k \times n' = n \times k'.$$

We will not dwell on this possibly subtle definition, but will rather accept the usual understanding of the rational numbers and their arithmetic properties. In particular, we will represent them as quotients rather than as ordered pairs, and, if r is a rational number, we will write $r = k/n$, instead of writing r as the equivalence class containing the ordered pair (k, n). As usual, we refer to the first integer in the quotient k/n as the *numerator* and the second (nonzero) integer in the quotient k/n as the *denominator* of the quotient. The familiar definitions of sum and product for rational numbers are these:

$$\frac{k}{n} + \frac{k'}{n'} = \frac{kn' + nk'}{nn'}$$

and

$$\frac{k}{n} \times \frac{k'}{n'} = \frac{kk'}{nn'}.$$

Addition and multiplication of rational numbers satisfy the three basic algebraic relations of commutativity, associativity and distributivity stated earlier.

We note that the integers $\mathbb{Z}$ can be identified in an obvious way as a subset of the rational numbers $\mathbb{Q}$. Indeed, we identify the integer k with the quotient $k/1$. In this way, we note that $\mathbb{Q}$ contains the two numbers $0 \equiv 0/1$ and $1 \equiv 1/1$. Notice that any other quotient k/n that is equivalent to $0/1$ must satisfy $k = 0$, and any other quotient k/n that is equivalent to $1/1$ must satisfy $k = n$. Remember, $k/n \equiv k'/n'$ if and only if $kn' = k'n$.

The set $\mathbb{Q}$ has an additional property not shared by the set of integers $\mathbb{Z}$. It is this: For each nonzero element $r \in \mathbb{Q}$, there exists an element $r' \in \mathbb{Q}$ for which $r \times r' = 1$. Indeed, if $r = k/n \neq 0$, then $k \neq 0$, and we may define $r' = n/k$. Consequently, the set $\mathbb{Q}$ of all rational numbers is what is known in mathematics as a field.

DEFINITION. A *field* is a nonempty set F on which there are defined two binary operations, addition (+) and multiplication (×), such that the following six axioms hold:

(1) Both addition and multiplication are commutative and associative.

(2) Multiplication is distributive over addition; i.e.,

$$x \times (y + z) = x \times y + x \times z$$

for all $x, y, z \in F$.

(3) There exists an element in F, which we will denote by 0, that is an identity for addition; i.e., $x + 0 = x$ for all $x \in F$.

(4) There exists a **nonzero** element in F, which we will denote by 1, that is an identity for multiplication; i.e., $x \times 1 = x$ for all $x \in F$.

(5) If $x \in F$, then there exists a unique element $y \in F$ such that $x + y = 0$. This element y is called the *additive inverse* of x and is denoted by $-x$.

(6) If $x \in F$ and $x \neq 0$, then there exists a unique element $y \in F$ such that $x \times y = 1$. This element y is called the *multiplicative inverse* of x and is denoted by x^{-1}.

REMARK. There are many examples of fields. (See the exercise below.) They all share certain arithmetic properties, which can be derived from the axioms above. If x is an element of a field F, then according to one of the axioms above, we have that $1 \times x = x$. (Note that this "1" is the multiplicative identity of the field F and not the natural number 1.) However, it is tempting to write $x + x = 2 \times x$ in the field F. The "2" here is not à priori an element of F, so that the equation $x + x = 2 \times x$ is not really justified. This is an example of a situation where a careful recursive definition can be useful.

DEFINITION. If x is an element of a field F, define inductively elements $n\cdot x \equiv nx$ of F by $1\cdot x = x$, and, if $k\cdot x$ is defined, set $(k+1)\cdot x = x + k\cdot x$. The set S of all natural numbers n for which $n\cdot x$ is defined is therefore, by the axiom of mathematical induction, all of $\mathbb{N}$.

Usually we will write nx instead of $n\cdot x$. Of course, nx is just the element of F obtained by adding x to itself n times: $nx = x + x + x + \ldots + x$.

Exercise 1.2. (a) Justify for yourself that the set $\mathbb{Q}$ of all rational numbers is a field. That is, carefully verify that all six of the axioms hold.
(b) Let F_7 denote the seven elements $\{0,1,2,3,4,5,6\}$. Define addition and multiplication on F_7 as ordinary addition and multiplication mod 7. Prove that F_7 is a field. (You may assume that axioms (1) and (2) hold. Check only conditions (3)–(6).) Show in addition that $7x = 0$ for every $x \in F_7$.
(c) Let F_9 denote the set consisting of the nine elements $\{0,1,2,3,4,5,6,7,8\}$. Define addition and multiplication on F_9 to be ordinary addition and multiplication mod 9. Show that F_9 is not a field. Which of the axioms fail to hold?
(d) Show that the set $\mathbb{N}$ of natural numbers is not a field. Which of the field axioms fail to hold? Show that the set $\mathbb{Z}$ of all integers is not a field. Which of the field axioms fail to hold?

Exercise 1.3. Let F be any field. Verify that the following arithmetic properties hold in F.
(a) $0 \times x = 0$ for all $x \in F$.
HINT: Use the distributive law and the fact that $0 = 0 + 0$.
(b) If x and y are nonzero elements of F, then $x \times y$ is nonzero. And, the multiplicative inverse of $x \times y$ satisfies $(x \times y)^{-1} = x^{-1} \times y^{-1}$.
(c) $(-1) \times x = (-x)$ for all $x \in F$.
(d) $(-x) \times (-y) = x \times y$ for all $x, y \in F$.
(e) $x \times x - y \times y = (x - y) \times (x + y)$.
(f) $(x + y) \times (x + y) = x \times x + 2 \cdot x \times y + y \times y$.

DEFINITION. Let F be a field, and let x be a nonzero element of F.
For each natural number n, we define inductively an element x^n in F as follows: $x^1 = x$, and, if x^k is defined, set $x^{k+1} = x \times x^k$. Of course, x^n is just the product of n x's.
Define x^0 to be 1.
For each natural number n, define x^{-n} to be the multiplicative inverse $(x^n)^{-1}$ of the element x^n.
Finally, we define 0^m to be 0 for every positive integer m, and we leave 0^{-n} and 0^0 undefined.

We have therefore defined x^m for every nonzero x and every integer $m \in \mathbb{Z}$.

Exercise 1.4. Let F be a field. Derive the following laws of exponents:
(a) $x^{n+m} = x^n \times x^m$ for all nonzero elements $x \in F$ and all integers n and m.
HINT: Fix $x \in F$ and $m \in \mathbb{Z}$ and use induction to derive this law for all natural numbers n. Then use the fact that in any field $(x \times y)^{-1} = x^{-1} \times y^{-1}$.
(b) $x^{n\times m} = (x^m)^n$ for all nonzero $x \in F$ and all $n, m \in \mathbb{Z}$.
(c) $(x \times y)^n = x^n \times y^n$ for all nonzero $x, y \in F$ and all $n \in \mathbb{Z}$.

From now on, we will indicate multiplication in a field by juxtaposition; i.e., $x \times y$ will be denoted simply as xy. Also, we will use the standard fractional notation to indicate multiplicative inverses. For instance,

$$xy^{-1} = x\frac{1}{y} = \frac{x}{y}.$$

THE REAL NUMBERS

What are the real numbers? From a geometric point of view (and a historical one as well) real numbers are quantities, i.e., lengths of segments, areas of surfaces, volumes of solids, etc. For example, once we have settled on a unit of length, i.e., a segment whose length we call 1, we can, using a compass and straightedge, construct segments of any rational length k/n. In some obvious sense then, the rational numbers are real numbers. Apparently it was an intellectual shock to the Pythagoreans to discover that there are some other real numbers, the so-called irrational ones. Indeed, the square root of 2 is a real number, since we can construct a segment the square of whose length is 2 by making a right triangle each of whose legs has length 1. (By the Pythagorean Theorem of plane geometry, the square of the hypotenuse of this triangle must equal 2.) And, Pythagoras proved that there is no rational number whose square is 2, thereby establishing that there are real numbers tha are not rational. See part (c) of Exercise 1.9.

Similarly, the area of a circle of radius 1 should be a real number; i.e., π should be a real number. It wasn't until the late 1800's that Hermite showed that π is not a rational number. One difficulty is that to define π as the area of a circle of radius 1 we must first define what is meant by the " area" of a circle, and this turns out to be no easy task. In fact, this naive, geometric approach to the definition of the real numbers turns out to be unsatisfactory in the sense that we are not able to prove or derive from these first principles certain intuitively obvious arithmetic results. For instance, how can we multiply or divide an area by a volume? How can we construct a segment of length the cube root of 2? And, what about negative numbers?

Let us begin by presenting two properties we expect any set that we call the real numbers ought to possess.

Algebraic Properties

We should be able to add, multiply, divide, etc., real numbers. In short, we require the set of real numbers to be a field.

Positivity Properties

The second aspect of any set we think of as the real numbers is that it has some notion of direction, some notion of positivity. It is this aspect that will allow us to "compare" numbers, e.g., one number is larger than another. The mathematically precise way to discuss this notion is the following.

DEFINITION. A field F is called an *ordered field* if there exists a subset $P \subseteq F$ that satisfies the following two properties:

(1) If $x, y \in P$, then $x + y$ and xy are in P.

(2) If $x \in F$, then one and only one of the following three statements is true. (i) $x \in P$, (ii) $-x \in P$, and (iii) $x = 0$. (This property is known as the *law of tricotomy*.)

The elements of the set P are called *positive* elements of F, and the elements x for which $-x$ belong to P are called *negative* elements of F.

As a consequence of these properties of P, we may introduce in F a notion of order.

DEFINITION. If F is an ordered field, and x and y are elements of F, we say that $x < y$ if $y - x \in P$. We say that $x \leq y$ if either $x < y$ or $x = y$.
We say that $x > y$ if $y < x$, and $x \geq y$ if $y \leq x$.

An ordered field satisfies the familiar laws of inequalities. They are consequences of the two properties of the set P.

Exercise 1.5. Using the positivity properties above for an ordered field F, together with the axioms for a field, derive the familiar laws of inequalities:
(a) (Transitivity) If $x < y$ and $y < z$, then $x < z$.
(b) (Adding like inequalities) If $x < y$ and $z < w$, then $x + z < y + w$.
(c) If $x < y$ and $a > 0$, then $ax < ay$.
(d) If $x < y$ and $a < 0$, then $ay < ax$.
(e) If $0 < a < b$ and $0 < c < d$, then $ac < bd$.
(f) Verify parts (a) through (e) with $<$ replaced by $\leq$.
(g) If x and y are elements of F, show that one and only one of the following three relations can hold: (i) $x < y$, (ii) $x > y$, (iii) $x = y$.
(h) Suppose x and y are elements of F, and assume that $x \leq y$ and $y \leq x$. Prove that $x = y$.

Exercise 1.6. (a) If F is an ordered field, show that $1 \in P$; i.e., that $0 < 1$.
HINT: By the law of tricotomy, only one of the three possibilities holds for 1. Rule out the last two.
(b) Show that F_7 of Exercise 1.2 is not an ordered field; i.e., there is no subset $P \subseteq F_7$ such that the two positivity properties can hold.
HINT: Use part (a) and positivity property (1).
(c) Prove that $\mathbb{Q}$ is an ordered field, where the set P is taken to be the usual set of positive rational numbers. That is, P consists of those rational numbers a/b for which both a and b are natural numbers.
(d) Suppose F is an ordered field and that x is a nonzero element of F. Show that for all natural numbers n $nx \neq 0$.
(e) Show that, in an ordered field, every nonzero square is positive; i.e., if $x \neq 0$, then $x^2 \in P$.

We remarked earlier that there are many different examples of fields, and many of these are also ordered fields. Some fields, though technically different from each other, are really indistinguishable from the algebraic point of view, and we make this mathematically precise with the following definition.

DEFINITION. Let F_1 and F_2 be two ordered fields, and write P_1 and P_2 for the set of positive elements in F_1 and F_2 respectively. A 1-1 correspondence J between F_1 and F_2 is called an *isomorphism* if

(1) $J(x + y) = J(x) + J(y)$ for all $x, y \in F_1$.
(2) $J(xy) = J(x)J(y)$ for all $x, y \in F_1$.
(3) $x \in P_1$ if and only if $J(x) \in P_2$.

REMARK. In general, if A_1 and A_2 are two algebraic systems, then a 1-1 correspondence between A_1 and A_2 is called an *isomorphism* if it converts the algebraic structure on A_1 into the corresponding algebraic structure on A_2.

Exercise 1.7. (a) Let F be an ordered field. Define a function $J : \mathbb{N} \to F$ by $J(n) = n \cdot 1$. Prove that J is an isomorphism of $\mathbb{N}$ onto a subset $\widetilde{\mathbb{N}}$ of F. That is, show that this correspondence is one-to-one and converts addition and multiplication in $\mathbb{N}$ into addition and multiplication in F. Give an example to show that this result is not true if F is merely a field and not an ordered field.
(b) Let F be an ordered field. Define a function $J : \mathbb{Q} \to F$ by $J(k/n) = k \cdot 1 \times (n \cdot 1)^{-1}$. Prove that J is an isomorphism of the ordered field $\mathbb{Q}$ onto a subset $\widetilde{\mathbb{Q}}$ of the ordered field F. Conclude that every ordered field F contains a subset that is isomorphic to the ordered field $\mathbb{Q}$.

REMARK. Part (b) of the preceding exercise shows that the ordered field $\mathbb{Q}$ is the smallest possible ordered field, in the sense that every other ordered field contains an isomorphic copy of $\mathbb{Q}$. However, as mentioned earlier, the ordered field $\mathbb{Q}$ cannot suffice as the set of real numbers. There is no rational number whose square is 2, and we want the square root of 2 to be a real number. See Exercise 1.9 below.
What extra property is there about an ordered field F that will allow us to prove that numbers like $\sqrt{2}$, π, and so on are elements of F? It turns out that the extra property we need is related to a quite subtle point concerning upper and lower bounds of sets. It gives us some initial indication that the known-to-be subtle concept of a **limit** may be fundamental to our very notion of what the real numbers are.

DEFINITION. If S is a subset of an ordered field F, then an element $x \in F$ is called an *upper bound* for S if $x \geq y$ for every $y \in S$. An element z is called a *lower bound* for S if $z \leq y$ for every $y \in S$.
A subset S of an ordered field F is called *bounded above* if it has an upper bound; it is called *bounded below* if it has a lower bound; and it is called *bounded* if it has both an upper bound and a lower bound.
An element M is called the *least upper bound* or *supremum* of a set S if it is an upper bound for S and if $M \leq x$ for every other upper bound x of S. That is, M is less than or equal to any other upper bound of S.
Similarly, an element m is called the *greatest lower bound* or *infimum* of S if it is a lower bound for S and if $z \leq m$ for every other lower bound z of S. That is, m is greater than or equal to any other lower bound of S.

Clearly, the supremum and infimum of a set S are unique. For instance, if M and M' are both least upper bounds of a set S, then they are both upper bounds of S. We would then have $M \leq M'$ and $M' \leq M$. Therefore, by part (h) of Exercise 1.5, $M = M'$.
It is important to keep in mind that an upper bound of a set S need not be an element of S, and in particular, the least upper bound of S **may or may not** actually belong to S.
If M is the supremum of a set S, we denote M by $\sup S$. If m is the infimum of a set S, we denote it by $\inf S$.

Exercise 1.8. (a) Suppose S is a nonempty subset of an ordered field F and that x is an element of F. What does it mean to say that "x is not an upper bound for S?''
(b) Let F be an ordered field, and let S be the empty set, thought of as a subset of F. Prove that every element $x \in F$ is an upper bound for S and that every element $y \in F$ is a lower bound for S.

HINT: If not, then what?
(c) If $S = \emptyset$, show that S has no least upper bound and no greatest lower bound.

REMARK. The preceding exercise shows that peculiar things about upper and lower bounds happen when S is the empty set. One point is that just because a set has an upper bound does not mean it has to have a least upper bound. That is, no matter which upper bound we choose, there is always another one that is strictly smaller. This is a very subtle point, and it is in fact quite difficult to give a simple concrete example of this phenomenon. See the remark following Theorem 1.6. However, part (d) of the next exercise contains the seed of an example.

Exercise 1.9. A natural number a is called *even* if there exists a natural number c such that $a = 2c$, and a is called *odd* if there exists a natural number c such that $a = 2c + 1$.
(a) Prove by induction that every natural number is either odd or even.
(b) Prove that a natural number a is even if and only if $a^2 = a \times a$ is even.
(c) Prove that there is no element x of $\mathbb{Q}$ whose square is 2. That is, the square root of 2 is not a rational number.
HINT: Argue by contradiction. Suppose there is a rational number k/n for which $k^2/n^2 = 2$, and assume, as we may, that the natural numbers k and n have no common factor. Observe that k must be even, and then observe that n also must be even.
(d) Let S be the set of all positive rational numbers x for which $x^2 = x \times x < 2$. Prove that S has an upper bound and a lower bound. Can you determine whether or not S has a least upper bound?

The existence of least upper bounds and greatest lower bounds of bounded sets turns out to be the critical idea in defining the real numbers. It is precisely the existence of such suprema and infimas that enables us to define as real numbers quantities such as $\sqrt{2}$, π, e, and so on.

DEFINITION. An ordered field F is called *complete* if every nonempty subset S of F that has an upper bound has a least upper bound.

REMARK. Although $\mathbb{Q}$ is an ordered field, we will see that it is not a complete ordered field. In fact, the answer to part (d)(of Exercise 1.9 is no. The set described there, though bounded above, has no least upper bound. In fact, it was one of nineteenth century mathematicians' major achievements to prove the following theorem.

THEOREM 1.1. *There exists a complete ordered field.*

We leave the proof of this theorem to the appendix.
Perhaps the most reassuring result along these lines is the following companion theorem, whose proof we also leave to the appendix.

THEOREM 1.2. *If F_1 and F_2 are two complete ordered fields, then they are isomorphic.*

Taken together, the content of the two preceding theorems is that, up to isomorphism, there exists one and only one complete ordered field. For no other reason that that, this special field should be an important object in mathematics. Our definition of the real numbers is then the following:

DEFINITION. By the set $\mathbb{R}$ of *real numbers* we mean the (unique) complete ordered field.

PROPERTIES OF THE REAL NUMBERS

THEOREM 1.3. *The set $\mathbb{R}$ contains a subset that is isomorphic to the ordered field $\mathbb{Q}$ of rational numbers, and hence subsets that are isomorphic to $\mathbb{N}$ and $\mathbb{Z}$.*

REMARK. The proof of Theorem 1.3 is immediate from part (b) of Exercise 1.7. In view of this theorem, we will simply think of the natural numbers, the integers, and the rational numbers as subsets of the real numbers.

Having made a definition of the set of real numbers, it is incumbent upon us now to verify that this set $\mathbb{R}$ satisfies our intuitive notions about the reals. Indeed, we will show that $\sqrt{2}$ is an element of $\mathbb{R}$ and hence is a real number (as plane geometry indicates it should be), and we will show in later chapters that there are elements of $\mathbb{R}$ that agree with our intuition about e and π. Before we can proceed to these tasks, we must establish some special properties of the field $\mathbb{R}$. The first, the next theorem, is simply an analog for lower bounds of the least upper bound condition that comes from the completeness property.

THEOREM 1.4. *If S is a nonempty subset of $\mathbb{R}$ that is bounded below, then there exists a greatest lower bound for S.*

PROOF. Define T to be the set of all real numbers x for which $-x \in S$. That is, T is the set $-S$. We claim first that T is bounded above. Thus, let m be a lower bound for the set S, and let us show that the number $-m$ is an upper bound for T. If $x \in T$, then $-x \in S$. So, $m \leq -x$, implying that $-m \geq x$. Since this is true for all $x \in T$, the number $-m$ is an upper bound for T.

Now, by the completeness assumption, T has a least upper bound M_0. We claim that the number $-M_0$ is the greatest lower bound for S. To prove this, we must check two things. First, we must show that $-M_0$ is a lower bound for S. Thus, let y be an element of S. Then $-y \in T$, and therefore $-y \leq M_0$. Hence, $-M_0 \leq y$, showing that $-M_0$ is a lower bound for S.

Finally, we must show that $-M_0$ is the greatest lower bound for S. Thus, let m be a lower bound for S. We saw above that this implies that $-m$ is an upper bound for T. Hence, because M_0 is the least upper bound for T, we have that $-m \geq M_0$, implying that $m \leq -M_0$, and this proves that $-M_0$ is the infimum of the set S.

The following is the most basic and frequently used property of least upper bounds. It is our first glimpse of " limits." Though the argument is remarkably short and sweet, it will provide the mechanism for many of our later proofs, so master this one.

THEOREM 1.5. *Let S be a nonempty subset of $\mathbb{R}$ that is bounded above, and Let M_0 denote the least upper bound of S; i.e., $M_0 = \sup S$. Then, for any positive real number ϵ there exists an element t of S such that $t > M_0 - \epsilon$.*

PROOF. Let $\epsilon > 0$ be given. Since $M_0 - \epsilon < M_0$, it must be that $M_0 - \epsilon$ is not an upper bound for S. (M_0 is necessarily less than or equal to any other upper bound of S.) Therefore, there exists an element $t \in S$ for which $t > M_0 - \epsilon$. This is exactly what the theorem asserts.

Exercise 1.10. (a) Let S be a nonempty subset of $\mathbb{R}$ which is bounded below, and let m_0 denote the infimum of S. Prove that, for every positive δ, there exists an element s of S such that $s < m_0 + \delta$. Mimic the proof to Theorem 1.5.
(b) Let S be any bounded subset of $\mathbb{R}$, and write $-S$ for the set of negatives of the elements of S. Prove that $\sup(-S) = -\inf S$.
(c) Use part (b) to give an alternate proof of part (a) by using Theorem 1.5 and a minus sign.

Exercise 1.11. (a) Let S be the set of all real numbers x for which $x < 1$. Give an example of an upper bound for S. What is the least upper bound of S? Is $\sup S$ an element of S?
(b) Let S be the set of all $x \in \mathbb{R}$ for which $x^2 \leq 4$. Give an example of an upper bound for S. What is the least upper bound of S? Does $\sup S$ belong to S?

We show now that $\mathbb{R}$ contains elements other than the rational numbers in $\mathbb{Q}$. Of course this holds for any complete ordered field. The next theorem makes this quite explicit.

THEOREM 1.6. *If x is a positive real number, then there exists a positive real number y such that $y^2 = x$. That is, every positive real number x has a positive square root in $\mathbb{R}$. Moreover, there is only one positive square root of x.*

PROOF. Let S be the set of positive real numbers t for which $t^2 \leq x$. Then S is nonempty Indeed, If $x > 1$, then 1 is in S because $1^2 = 1 \times 1 < 1 \times x = x$. And, if $x \leq 1$, then x itself is in S, because $x^2 = x \times x \leq 1 \times x = x$.
Also, S is bounded above. In fact, the number $1 + x/2$ is an upper bound of S. Indeed, arguing by contradiction, suppose there were a t in S such that $t > 1+x/2$. Then

$$x \geq t^2 > (1 + x/2)^2 = 1 + x + x^2/4 > x,$$

which is a contradiction. Therefore, $1 + x/2$ is an upper bound of S, and so S is bounded above.
Now let $y = \sup S$. We wish to show that $y^2 = x$. We show first that $y^2 \leq x$, and then we will show that $y^2 \geq x$. It will then follow from the tricotomy law that $y^2 = x$. We prove both these inequalities by contradiction.
So, assume first that $y^2 > x$, and write α for the positive number $y^2 - x$. Let ϵ be the positive number $\alpha/(2y)$, and, using Theorem 1.5, choose a $t \in S$ such that $t > y - \epsilon$. Then $y + t \leq (2y)$, and $y - t < \epsilon = \alpha/2y$. So,

$$\begin{aligned}\alpha &= y^2 - x \\ &= y^2 - t^2 + t^2 - x \\ &\leq y^2 - t^2 \\ &= (y+t)(y-t) \\ &\leq 2y(y-t) \\ &< 2y\epsilon \\ &< 2y \times \frac{\alpha}{2y} \\ &= \alpha,\end{aligned}$$

which is a contradiction. Therefore y^2 is not greater than x.

Now we show that y^2 is not less than x. Again, arguing by contradiction, suppose it is, and let ϵ be the positive number $x - y^2$. Choose a positive number δ that is less than y and also less than $\epsilon/(3y)$. Let $s = y + \delta$. Then s is not in S, whence $s^2 > x$, so that we must have

$$\begin{aligned}\epsilon &= x - y^2 \\ &= x - s^2 + s^2 - y^2 \\ &\leq s^2 - y^2 \\ &= (s+y)(s-y) \\ &= (2y+\delta)\delta \\ &< 3y\delta \\ &< \epsilon,\end{aligned}$$

which again is a contradiction.

This completes the proof that $y^2 = x$; i.e., that x has a positive square root.

Finally, if y' were another positive number for which ${y'}^2 = x$, we show that $y = y'$ by ruling out the other two cases: $y < y'$ and $y > y'$. For instance, if $y < y'$, then we would have that $y^2 < {y'}^2$, giving that

$$x = y^2 < {y'}^2 = x,$$

implying that $x < x$, and this is a contradiction.

DEFINITION. If x is a positive real number, then the symbol $\sqrt{x}$ will denote the unique positive number y for which $y^2 = x$. Of course, $\sqrt{0}$ denotes the number 0.

REMARK. Part (c) of Exercise 1.9 shows that the field $\mathbb{Q}$ contains no number whose square is 2, and Theorem 1.6 shows that the field $\mathbb{R}$ does contain a number whose square is 2. We have therefore "proved" that the real numbers is a larger set than the rational numbers. It may come as a surprise to learn that we only now have been able to prove that. Look back through the chapter to be sure. It follows also that $\mathbb{Q}$ itself is not a complete ordered field. If it were, it would be isomorphic to $\mathbb{R}$, by Theorem 1.2, so that it would have to contain a square root of 2, which it does not.

DEFINITION. A real number x that is not a rational number, i.e., is not an element of the subset $\mathbb{Q}$ of $\mathbb{R}$, is called an *irrational number*.

Exercise 1.12. (a) Prove that every positive real number has exactly 2 square roots, one positive ($\sqrt{x}$) and the other negative ($-\sqrt{x}$).
(b) Prove that if x is a negative real number, then there is no real number y such that $y^2 = x$.
(c) Prove that the product of a nonzero rational number and an arbitrary irrational number must be irrational. Show by example that the sum and product of irrational numbers can be rational.

INTERVALS AND APPROXIMATION

We introduce next into the set of real numbers some geometric concepts, namely, a notion of distance between numbers. Of course this had to happen, for geometry is the very basis of mathematics.

DEFINITION. The absolute value of a real number x is denoted by $|x|$ and is defined as follows:
(i) $|0| = 0$.
(ii) If $x > 0$ then $|x| = x$.
(iii) If $x < 0$ $(-x > 0)$ then $|x| = -x$.
We define the *distance* $d(x, y)$ between two real numbers x and y by $d(x, y) = |x-y|$.

Obviously, such definitions of absolute value and distance can be made in any ordered field.

Exercise 1.13. Let x and y be real numbers.
(a) Show that $|x| \geq 0$, and that $x \leq |x|$.
(b) Prove the Triangle Inequality for absolute values.

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

HINT: Check the three cases $x + y > 0$, $x + y < 0$, and $x + y = 0$.
(c) Prove the so-called ' ' backward" triangle inequality.

$$|x - y| \geq ||x| - |y||.$$

HINT: Write $|x| = |(x - y) + y|$, and use part (b).
(d) Prove that $|xy| = |x||y|$.
(e) Prove that $|x| = \sqrt{x^2}$ for all real numbers x.
(f) Prove the Triangle Inequality for the distance function. That is, show that

$$d(x, y) \leq d(x, z) + d(z, y)$$

for all $x, y, z \in \mathbb{R}$.

Exercise 1.14. (a) Prove that $x = y$ if $|x - y| < \epsilon$ for every positive number ϵ.
HINT: Argue by contradiction. Suppose $x \neq y$, and take $\epsilon = |x - y|/2$.
(b) Prove that $x = y$ if and only if $x - y \leq \epsilon$ and $y - x \leq \epsilon$ for every positive ϵ.

DEFINITION. Let a and b be real numbers for which $a < b$. By the *open interval (a,b)* we mean the set of all real numbers x for which $a < x < b$, and by the *closed interval [a,b]* we mean the set of all real numbers x for which $a \leq x \leq b$.
By (a, ∞) we mean the set of all real numbers x for which $a < x$, and by $[a, \infty)$ we mean the set of all real numbers x for which $a \leq x$.
Analogously, we define $(-\infty, b)$ and $(-\infty, b]$ to be respectively the set of all real numbers x for which $x < b$ and the set of all real numbers x for which $x \leq b$.

Exercise 1.15. (a) Show that the intersection of two open intervals either is the empty set or it is again an open interval.
(b) Show that $(a, b) = (-\infty, b) \cap (a, \infty)$.
(c) Let y be a fixed real number, and let ϵ be a positive number. Show that the inequality $|x - y| < \epsilon$ is equivalent to the pair of inequalities

$$y - \epsilon < x \,\text{and}\, x < y + \epsilon;$$

i.e., show that x satisfies the first inequality if and only if it satisfies the two latter ones. Deduce that $|x - y| < \epsilon$ if and only if x is in the open interval $(y - \epsilon, y + \epsilon)$.

Here is one of those assertions that seems like an obvious fact. However, it requires a proof which we only now can give, for it depends on the completeness axiom, and in fact is false in some ordered fields.

THEOREM 1.7. *Let $\mathbb{N}$ denote the set of natural Numbers, thought of as a subset of $\mathbb{R}$. Then $\mathbb{N}$ is not bounded above.*

PROOF. Suppose false. Let M be an upper bound for the nonempty set $\mathbb{N}$, and let M_0 be the least upper bound for $\mathbb{N}$. Taking ϵ to be the positive number $1/2$, and applying Theorem 1.5, we have that there exists an element k of $\mathbb{N}$ such that $M_0 - 1/2 < k$. But then $M_0 - 1/2 + 1 < k + 1$, or, $M_0 + 1/2 < k + 1$. So $M_0 < k + 1$. But $M_0 \geq k + 1$ because M_0 is an upper bound for $\mathbb{N}$. We have thus arrived at a contradiction, and the theorem is proved.

REMARK. As mentioned above, there do exist ordered fields F in which the subset $\mathbb{N}$ **is** bounded above. Such fields give rise to what is called "nonstandard analysis," and they were first introduced by Abraham Robinson in 1966. The fact that $\mathbb{R}$ is a complete ordered field is apparently crucial to be able to conclude the intuitively clear fact that the natural numbers have no upper bound.

The next exercise presents another intuitively obvious fact, and this one is in some real sense the basis for many of our upcoming arguments about limits. It relies on the preceding theorem, is in fact just a corollary, so it has to be considered as a rather deep property of the real numbers; it is not something that works in every ordered field.

Exercise 1.16. Prove that if ϵ is a positive real number, then there exists a natural number N such that $1/N < \epsilon$.

The next theorem and exercise show that the set $\mathbb{Q}$ of rational numbers is "everywhere dense" in the field $\mathbb{R}$. That is, every real number can be approximated arbitrarily closely by rational numbers. Again, we point out that this result holds in any complete ordered field, and it is the completeness that is critical.

THEOREM 1.8. *Let $a < b$ be two real numbers. Then there exists a rational number $r = p/q$ in the open interval (a, b). In fact, there exist infinitely many rational numbers in the interval (a, b).*

PROOF. If $a < 0$ and $b > 0$, then taking $r = 0$ satisfies the first statement of the theorem. Assume first that $a \geq 0$ and $b > a$. Let n be a natural number for which $1/n$ is less than the positive number $b - a$. (Here, we are using the completeness of the field, because we are referring to Theorem 1.7, where completeness was vital.) If $a = 0$, then $b = b - a$. Setting $r = 1/n$, we would have that $a < r < b$. So, again, the first part of the theorem would be proved in that case.

Suppose then that $a > 0$, and choose the natural number q to be such that $1/q$ is less than the minimum of the two positive numbers a and $b - a$. Now, because the number aq is not an upper bound for the set $\mathbb{N}$, we may let p be the smallest natural number that is larger than aq. Set $r = p/q$.

We have first that $aq < p$, implying that $a < p/q = r$. Also, because p is the smallest natural number larger than aq, we must have that $p - 1 \leq aq$. Therefore, $(p-1)/q < a$, or $(p/q)-(1/q) < a$, implying that $r = p/q \leq a+1/q < a+(b-a) = b$. Hence, $a < r$ and $r < b$, and the first statement of the theorem is proved when both a and b are nonnegative.

If both a and b are nonpositive, then both $-b$ and $-a$ are nonnegative, and, using the first part of the proof, we can find a rational number r such that $-b < r < -a$. So, $a < -r < b$, and the first part of the theorem is proved in this case as well.

Clearly, we may replace b by r and repeat the argument to obtain another rational r_1 such that $a < r_1 < r < b$. Then, replacing b by r_1 and repeating the argument, we get a third rational r_2 such that $a < r_2 < r_1 < r < b$. Continuing this procedure would lead to an infinite number of rationals, all between a and b. This proves the second statement of the theorem.

Exercise 1.17. (a) Let $\epsilon > 0$ be given, and let k be a nonnegative integer. Prove that there exists a rational number p/q such that

$$k\epsilon < p/q < (k+1)\epsilon.$$

(b) Let x be a positive real number and let ϵ be a positive real number. Prove that there exists a rational number p/q such that $x - \epsilon < p/q < x$. State and prove an analogous result for negative numbers x.

Exercise 1.18. (a) If a and b are real numbers with $a < b$, show that there is an irrational number x (not a rational number) between a and b, i.e., with $a < x < b$. HINT: Apply Theorem 1.8 to the numbers $a\sqrt{2}$ and $b\sqrt{2}$.
(b) Conclude that within every open interval (a, b) there is a rational number and an irrational number. Are there necessarily infinitely many rationals and irrationals in (a, b)?

The preceding exercise shows the "denseness" of the rationals and the irrationals in the reals. It is essentially clear from this that every real number is arbitrarily close to a rational number and an irrational one.

THE GEOMETRIC PROGRESSION AND THE BINOMIAL THEOREM

There are two special algebraic identities that hold in $\mathbb{R}$ (in fact in any field F whatsoever) that we emphasize. They are both proved by mathematical induction. The first is the formula for the sum of a geometric progression.

THEOREM 1.9. (Geometric Progression) Let x be a real number, and let n be a natural number. Then,

(1) If $x \neq 1$, then

$$\sum_{j=0}^{n} x^j = \frac{1 - x^{n+1}}{1 - x}.$$

(2) If $x = 1$, then

$$\sum_{j=0}^{n} x^j = n + 1.$$

PROOF. The second claim is clear, since there are $n + 1$ summands and each is equal to 1.
We prove the first claim by induction. Thus, if $n = 1$, then the assertion is true, since

$$\sum_{j=0}^{1} x^j = x^0 + x^1 = 1 + x = (1 + x)\frac{1 - x}{1 - x} = \frac{1 - x^2}{1 - x}.$$

Now, supposing that the assertion is true for the natural number k, i.e., that

$$\sum_{j=0}^{k} x^j = \frac{1 - x^{k+1}}{1 - x},$$

let us show that the assertion holds for the natural number $k+1$. Thus

$$\begin{aligned}\sum_{j=0}^{k+1} x^j &= \sum_{j=0}^{k} x^j + x^{k+1}\\ &= \frac{1-x^{k+1}}{1-x} + x^{k+1}\\ &= \frac{1-x^{k+1}+x^{k+1}-x^{k+2}}{1-x}\\ &= \frac{1-x^{k+1+1}}{1-x},\end{aligned}$$

which completes the proof.

The second algebraic formula we wish to emphasize is the Binomial Theorem. Before stating it, we must introduce some useful notation.

DEFINITION. Let n be a natural number. As earlier in this chapter, we define $n!$ as follows:

$$n! = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1.$$

For later notational convenience, we also define $0!$ to be 1.
If k is any integer for which $0 \leq k \leq n$, we define the *binomial coefficient* $\binom{n}{k}$ by

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n \times (n-1) \times (n-2) \times \ldots \times (n-k+1)}{k!}.$$

Exercise 1.19. (a) Prove that $\binom{n}{0} = 1$, $\binom{n}{1} = n$ and $\binom{n}{n} = 1$.
(b) Prove that

$$\binom{n}{k} \leq \frac{2n^k}{2^k}$$

for all natural numbers n and all integers $0 \leq k \leq n$.
(c) Prove that

$$\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1}$$

for all natural numbers n and all integers $1 \leq k \leq n$.

THEOREM 1.10. (Binomial Theorem) If $x, y \in \mathbb{R}$ and n is a natural number, then

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

PROOF. We shall prove this theorem by induction. If $n = 1$, then the assertion is true, for $(x+y)^1 = x+y$ and

$$\sum_{k=0}^{1} \binom{1}{k} x^k y^{1-k} = \binom{1}{0} x^0 y^1 + \binom{1}{1} x^1 y^0 = x+y.$$

Now, assume that the assertion holds for the natural number j; i.e.,

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

and let us prove that the assertion holds for the natural number $j+1$. We will make use of part (c) of Exercise 1.19. We have that

$$\begin{aligned}
(x+y)^{j+1} &= (x+y)(x+y)^j \\
&= (x+y)\sum_{k=0}^{j} \binom{j}{k} x^k y^{j-k} \\
&= x\sum_{k=0}^{j} \binom{j}{k} x^k y^{j-k} + y\sum_{k=0}^{j} \binom{j}{k} x^k y^{j-k} \\
&= \sum_{k=0}^{j} \binom{j}{k} x^{k+1} y^{j-k} + \sum_{k=0}^{j} \binom{j}{k} x^k y^{j+1-k} \\
&= \sum_{k=0}^{j-1} \binom{j}{k} x^{k+1} y^{j-k} + \binom{j}{j} x^{j+1} y^0 \\
&\quad + \sum_{k=1}^{j} \binom{j}{k} x^k y^{j+1-k} + \binom{j}{0} x^0 y^{j+1} \\
&= x^{j+1} + \sum_{k=1}^{j} \binom{j}{k-1} x^k y^{j+1-k} \\
&\quad + \sum_{k=1}^{j} \binom{j}{k} x^k y^{j+1-k} + y^{j+1} \\
&= x^{j+1} + \sum_{k=1}^{j} (\binom{j}{k-1} + \binom{j}{k}) x^k y^{j+1-k} + y^{j+1} \\
&= x^{j+1} + \sum_{k=1}^{j} \binom{j+1}{k} x^k y^{j+1-k} + y^{j+1} \\
&= \binom{j+1}{j+1} x^{j+1} y^0 + \sum_{k=1}^{j} \binom{j+1}{k} x^k y^{j+1-k} + \binom{j+1}{0} x^0 y^{j+1} \\
&= \sum_{k=0}^{j+1} \binom{j+1}{k} x^k y^{j+1-k},
\end{aligned}$$

which shows that the assertion of the theorem holds for the natural number $j+1$. This completes the proof.

The next exercise is valid in any ordered field, but, since we are mainly interested in the order field $\mathbb{R}$, we state everything in terms of that field.

Exercise 1.20. (a) If x and y are positive real numbers, and if n and k are natural numbers with $k \leq n$, show that $(x+y)^n \geq \binom{n}{k} x^k y^{n-k}$.

(b) For any positive real number x and natural number n, show that $(1+x)^n \geq 1+nx$.
(c) For any real number $x > -1$ and natural number n, prove that $(1+x)^n \geq 1+nx$.
HINT: Do not try to use the binomial theorem as in part (b); it won't work because the terms are not all positive; prove this directly by induction.

There is one more important algebraic identity, which again can be proved by induction. It is actually just a corollary of the geometric progression formula.

THEOREM 1.11. *If $x, y \in \mathbb{R}$ and n is a natural number, then*

$$x^n - y^n = (x-y)(\sum_{j=0}^{n-1} x^j y^{n-1-j}.$$

PROOF. If $n = 1$ the theorem is clear. Suppose it holds for a natural number k, and let us prove the identity for the natural number $k+1$. We have

$$\begin{aligned} x^{k+1} - y^{k+1} &= x^{k+1} - x^k y + x^k y - y^{k+1} \\ &= (x-y)x^k + y(x^k - y^k) \\ &= (x-y)x^k + y(x-y)(\sum_{j=0}^{k-1} x^j y^{k-1-j}) \\ &= (x-y)x^k + (x-y)(\sum_{j=0}^{k-1} x^j y^{k-j} \\ &= (x-y)(x^k y^{k-k} + \sum_{j=0}^{k-1} x^j y^{k-j}) \\ &= (x-y)(\sum_{j=0}^{k} x^j y^{k-j}), \end{aligned}$$

which shows that the assertion holds for the natural number $k+1$. So, by induction, the theorem is proved.

Exercise 1.21. Let x and y be real numbers.
(a) Let n be an odd natural number; i.e., $n = 2k+1$ for some natural number k. Show that

$$x^n + y^n = (x+y)(\sum_{j=0}^{n-1} (-1)^j x^j y^{n-1-j}.$$

HINT: Write $x^n + y^n = x^n - (-y)^n$.
(b) Show that x^2+y^2 can not be factored into a product of the form $(ax+by)(cx+dy)$ for any choices of real numbers $a, b, c,$ and d.

Using the Binomial Theorem together with the preceding theorem, we may now investigate the existence of nth roots of real numbers. This next theorem is definitely not valid in any ordered field, for it again depends on the completeness property.

THEOREM 1.12. *Let n be a natural number and let x be a positive real number. Then there exists a unique positive real number y such that $y^n = x$; i.e., x has a unique positive nth root.*

PROOF. Note first that if $0 \leq t < s$, then $t^n < s^n$. (To see this, argue by induction, and use part (e) of Exercise 1.5.) Using this, we mimic the proof of Theorem 1.6. Thus, let S be the set of all positive real numbers t for which $t^n \leq x$. Then S is nonempty and bounded above. Indeed, if $x \geq 1$, then $1 \in S$, while if $x < 1$, then x itself is in S. Therefore, S is nonempty. Also, using part (b) of Exercise 1.20, we see that $1 + (x/n)$ is an upper bound for S. For, if $t > 1 + x/n$, then

$$t^n > (1 + (x/n))^n \geq 1 + n(x/n) > x.$$

Now let $y = \sup S$, and let us show that $y^n = x$. We rule out the other two possibilities. First, if $y^n > x$, let ϵ be the positive number $y^n - x$, and define ϵ' to be the positive number $\epsilon/(ny^{n-1})$. Then, using Theorem 1.5, choose $t \in S$ so that $y - \epsilon' < t \leq y$. (Theorem 1.5 is where the completeness of the ordered field $\mathbb{R}$ is crucial.) We have

$$\begin{aligned}\epsilon &= y^n - x \\ &= y^n - t^n + t^n - x \\ &\leq y^n - t^n \\ &= (y - t)(\sum_{j=0}^{n-1} y^j t^{n-1-j}) \\ &\leq (y - t)(\sum_{j=0}^{n-1} y^j y^{n-1-j}) \\ &= (y - t)(\sum_{j=0}^{n-1} y^{n-1} \\ &< \epsilon' n y^{n-1} \\ &= \epsilon,\end{aligned}$$

and this is a contradiction. Therefore, y^n is not greater than x.

Now, if $y^n < x$, let ϵ be the positive number $x - y^n$, and choose a $\delta > 0$ such that

$\delta < 1$ and $\delta < \epsilon/(y+1)^n$. Then, using the Binomial Theorem, we have that

$$\begin{aligned}(y+\delta)^n &= \sum_{k=0}^{n} \binom{n}{k} y^k \delta^{n-k} \\ &= y^n + \sum_{k=0}^{n-1} \binom{n}{k} y^k \delta^{n-k} \\ &= y^n + \delta \sum_{k=0}^{n-1} \binom{n}{k} y^k \delta^{n-1-k} \\ &< y^n + \delta \sum_{k=0}^{n} \binom{n}{k} y^k 1^{n-k} \\ &= y^n + \delta(y+1)^n \\ &= x - \epsilon + \delta(y+1)^n \\ &< x - \epsilon + \epsilon \\ &= x,\end{aligned}$$

implying that $y + \delta \in S$. But this is a contradiction, since $y = \sup S$. Therefore, y^n is not less than x, and so $y^n = x$.

We have shown the existence of a positive nth root of x. To see the uniqueness, suppose y and y' are two positive nth roots of x. Then

$$\begin{aligned}0 &= y^n - y'^n \\ &= (y - y')(\sum_{j=0}^{n-1} y^j y'^{n-j-1},\end{aligned}$$

which implies that either $y - y' = 0$ or $\sum_{j=0}^{n-1} y^j y'^{n-j-1} = 0$. Since this latter sum consists of positive terms, it cannot be 0, whence $y = y'$. This shows that there is but one positive nth root of x, and the theorem is proved.

Exercise 1.22. (a) Show that if $n = 2k$ is an even natural number, then every positive real number has exactly two distinct nth roots.

(b) If $n = 2k+1$ is an odd natural number, show that every real number has exactly one nth root.

(c) If n is a natural number greater than 1, prove that there is no rational number whose nth power equals 2, i.e., the nth root of 2 is not a rational number.

THE COMPLEX NUMBERS

It is useful to build from the real numbers another number system called the *complex numbers*. Although the real numbers $\mathbb{R}$ have many of the properties we expect, i.e., every positive number has a positive square root, every number has a cube root, and so on, there are somewhat less prominent properties that $\mathbb{R}$ fails to possess. For instance, negative numbers do not have square roots. This is actually a property that is missing in any ordered field, since every square is positive in an ordered field. See part (e) of Exercise 1.6. One way of describing this shortcoming on the

part of the real numbers is to note that the equation $1 + x^2 = 0$ has no solution in the real numbers. Any solution would have to be a number whose square is -1, and no real number has that property. As an initial extension of the set of real numbers, why not build a number system in which this equation has a solution?
We faced a similar kind of problem earlier on. In the set $\mathbb{N}$ there is no element j such that $j + n = n$ for all $n \in \mathbb{N}$. That is, there was no element like 0 in the natural numbers. The solution to the problem in that case was simply to "create" something called zero, and just adjoin it to our set $\mathbb{N}$. The same kind of solution exists for us now. Let us invent an additional number, this time denoted by i, which has the property that its square i^2 is -1. Because the square of any nonzero real number is positive, this new number i was traditionally referred to as an "imaginary" number. We simply adjoin this number to the set $\mathbb{R}$, and we will then have a number whose square is negative, i.e., -1. Of course, we will require that our new number system should still be a field; we don't want to give up our basic algebraic operations. There are several implications of this requirement: First of all, if y is any real number, then we must also adjoin to $\mathbb{R}$ the number $y \times i \equiv yi$, for our new number system should be closed under multiplication. Of course the square of iy will equal $i^2y^2 = -y^2$, and therefore this new number iy must also be imaginary, i.e., not a real number. Secondly, if x and y are any two real numbers, we must have in our new system a number called $x + yi$, because our new system should be closed under addition.

DEFINITION. Let i denote an object whose square $i^2 = -1$. Let $\mathbb{C}$ be the set of all objects that can be represented in the form $z = x + yi$, where both x and y are real numbers.
Define two operations $+$ and $\times$ on $\mathbb{C}$ as follows:

$$(x + yi) + (x' + y'i) = x + x' + (y + y')i,$$

and

$$(x + iy)(x' + iy') = xx' + xiy' + iyx' + iyiy' = xx' - yy' + (xy' + yx')i.$$

THEOREM 1.13.

(1) *The two operations $+$ and $\times$ defined above are commutative and associative, and multiplication is distributive over addition.*
(2) *Each operation has an identity: $(0 + 0i)$ is the identity for addition, and $(1 + 0i)$ is the identity for multiplication.*
(3) *The set $\mathbb{C}$ with these operations is a field.*

PROOF. We leave the proofs of Parts (1) and (2) to the following exercise. To see that $\mathbb{C}$ is a field, we need to verify one final condition, and that is to show that if $z = x + yi \neq 0 = 0 + 0i$, then there exists a $w = u + vi$ such that $z \times w = 1 = 1 + 0i$. Thus, suppose $z = x + yi \neq 0$. Then at least one of the two real numbers x and y must be nonzero, so that $x^2 + y^2 > 0$. Define a complex number w by

$$w = \frac{x}{x^2 + y^2} + \frac{-y}{x^2 + y^2} i.$$

We then have

$$\begin{aligned} z \times w &= (x+yi) \times (\frac{x}{x^2+y^2} + \frac{-y}{x^2+y^2} i \\ &= \frac{x^2}{x^2+y^2} - \frac{-y^2}{x^2+y^2} + (x\frac{-y}{x^2+y^2} + y\frac{x}{x^2+y^2})i \\ &= \frac{x^2+y^2}{x^2+y^2} + \frac{0}{x^2+y^2} i \\ &= 1 + 0i \\ &= 1, \end{aligned}$$

as desired.

Exercise 1.23. Prove parts (1) and (2) of Theorem 1.13.

One might think that these kinds of improvements of the real numbers will go on and on. For instance, we might next have to create and adjoin another object j so that the number i has a square root; i.e., so that the equation $i - z^2 = 0$ has a solution. Fortunately and surprisingly, this is not necessary, as we will see when we finally come to the Fundamental Theorem of Algebra in Chapter VII.
The subset of $\mathbb{C}$ consisting of the pairs $x + 0i$ is a perfect (isomorphic) copy of the real number system $\mathbb{R}$. We are justified then in saying that the complex number system extends the real number system, and we will say that a real number x is the same as the complex number $x + 0i$. That is, real numbers are special kinds of complex numbers. The complex numbers of the form $0 + yi$ are called *purely imaginary numbers.* Obviously, the only complex number that is both real and purely imaginary is the number $0 = 0 + 0i$. The set $\mathbb{C}$ can also be regarded as a 2-dimensional space, a plane, and it is also helpful to realize that the complex numbers form a 2-dimensional vector space over the field of real numbers.

DEFINITION. If $z = x + yi$, we say that the real number x is the *real part* of z and write $x = \Re(z)$. We say that the real number y is the *imaginary part* of z and write $y = \Im(z)$.
If $z = x+yi$ is a complex number, define the *complex conjugate* $\bar{z}$ of z by $\bar{z} = x-yi$.

The complex number i satisfies $i^2 = -1$, showing that the negative number -1 has a square root in $\mathbb{C}$, or equivalently that the equation $1 + z^2 = 0$ has a solution in $\mathbb{C}$. We have thus satisfied our initial goal of extending the real numbers. But what about other complex numbers? Do they have square roots, cube roots, nth roots? What about solutions to other kinds of equations than $1 + z^2$?

Exercise 1.24. (a) Prove that every complex number has a square root.
HINT: Let $z = a + bi$. Assume $w = x + yi$ satisfies $w^2 = z$, and just solve the two equations in two unknowns that arise.
(b) Prove that every quadratic equation $az^2 + bz + c = 0$, for a, b, and c complex numbers, has a solution in $\mathbb{C}$.
HINT: If $a = 0$, it is easy to find a solution. If $a \neq 0$, we need only find a solution to the equivalent equation

$$z^2 + \frac{b}{a}z + \frac{c}{a} = 0.$$

Justify the following algebraic manipulations, and then solve the equation.

$$\begin{aligned} z^2 + \frac{b}{a}z + \frac{c}{a} &= z^2 + \frac{b}{a}z + \frac{b^2}{4a^2} - \frac{b^2}{4a^2} + \frac{c}{a} \\ &= (z + \frac{b}{2a})^2 - \frac{b^2}{4a^2} + \frac{c}{a}. \end{aligned}$$

What about this new field $\mathbb{C}$? Does every complex number have a cube root, a fourth root, does every equation have a solution in $\mathbb{C}$? A natural instinct would be to suspect that $\mathbb{C}$ takes care of square roots, but that it probably does not necessarily have higher order roots. However, the content of the Fundamental Theorem of Algebra, to be proved in Chapter VII, is that every equation of the form $P(z) = 0$, where P is a nonconstant polynomial, has a solution in $\mathbb{C}$. This immediately implies that every complex number c has an nth root, for any solution of the equation $z^n - c = 0$ would be an nth root of c.
The fact that the Fundamental Theorem of Algebra is true is a good indication that the field $\mathbb{C}$ is a "good" field. But it's not perfect.

THEOREM 1.14. *In no way can the field $\mathbb{C}$ be made into an ordered field. That is, there exists no subset P of $\mathbb{C}$ that satisfies the two positivity axioms.*

PROOF. Suppose $\mathbb{C}$ were an ordered field, and write P for its set of positive elements. Then, since every square in an ordered field must be in P (part (e) of Exercise 1.6), we must have that $-1 = i^2$ must be in P. But, by part (a) of Exercise 1.6, we also must have that 1 is in P, and this leads to a contradiction of the law of tricotomy. We can't have both 1 and -1 in P. Therefore, $\mathbb{C}$ is not an ordered field.

Although we may not define when one complex number is smaller than another, we can define the absolute value of a complex number and the distance between two of them.

DEFINITION. If $z = x + yi$ is in $\mathbb{C}$, we define the *absolute value* of z by

$$|z| = \sqrt{x^2 + y^2}.$$

We define the *distance* $d(z, w)$ between two complex numbers z and w by
$d(z, w) = |z - w|$.
If $c \in \mathbb{C}$ and $r > 0$, we define the *open disk of radius r around c*, and denote it by $B_r(c)$, by

$$B_r(c) = \{z \in \mathbb{C} : |z - c| < r\}.$$

The *closed disk* of radius r around c is denoted by $\overline{B}_r(c)$ and is defined by

$$\overline{B}_r(c) = \{z \in \mathbb{C} : |z - c| \leq r\}.$$

We also define open and closed *punctured* disks $B'_r(c)$ and $\overline{B}'_r(c)$ around c by

$$B'_r(c) = \{z : 0 < |z - c| < r\}$$

and

$$\overline{B}'_r(c) = \{z : 0 < |z - c \leq r\}.$$

These punctured disks are just like the regular disks, except that they do not contain the central point c.
More generally, if S is any subset of $\mathbb{C}$, we define the *open neighborhood of radius r* around S, denoted by $N_r(S)$, to be the set of all z such that there exists a $w \in S$ for which $|z-w| < r$. That is, $N_r(S)$ is the set of all complex numbers that are within a distance of r of the set S. We define the *closed neighborhood* of radius r around S, and denote it by $\overline{N}_r(S)$, to be the set of all $z \in \mathbb{C}$ for which there exists a $w \in S$ such that $|z-w| \le r$.

Exercise 1.25. (a) Prove that the absolute value of a complex number z is a nonnegative real number. Show in addition that $|z|^2 = z\bar{z}$.
(b) Let x be a real number. Show that the absolute value of x is the same whether we think of x as a real number or as a complex number.
(c) Prove that $\max(|\Re(z)|, |\Im(z)|) \le |z| \le |\Re(z)| + |\Im(z)|$. Note that this just amounts to verifying that

$$\max(|x|,|y|) \le \sqrt{x^2+y^2} \le |x|+|y|$$

for any two real numbers x and y.
(d) For any complex numbers z and w, show that $\overline{z+w} = \bar{z} + \bar{w}$, and that $\bar{\bar{z}} = z$.
(e) Show that $z + \bar{z} = 2\Re(z)$ and $z - \bar{z} = 2i\Im(z)$.
(f) If $z = a + bi$ and $w = a' + b'i$, prove that $|zw| = |z||w|$.
HINT: Just compute $|(a+bi)(a'+b'i)|^2$.

The next theorem is in a true sense the most often used inequality of mathematical analysis. We have already proved the triangle inequality for the absolute value of real numbers, and the proof was not very difficult in that case. For complex numbers, it is not at all simple, and this should be taken as a good indication that it is a deep result.

THEOREM 1.15. (Triangle Inequality) If z and z' are two complex numbers, then

$$|z+z'| \le |z| + |z'|$$

and

$$|z-z'| \ge ||z| - |z'||.$$

PROOF. We use the results contained in Exercise 1.25.

$$\begin{aligned}
|z+z'|^2 &= (z+z')\overline{(z+z')} \\
&= (z+z')(\bar{z}+\bar{z'}) \\
&= z\bar{z} + z'\bar{z} + z\bar{z'} + z'\bar{z'} \\
&= |z|^2 + z'\bar{z} + \overline{z'\bar{z}} + |z'|^2 \\
&= |z|^2 + 2\Re(z'\bar{z}) + |z'|^2 \\
&\le |z|^2 + 2|\Re(z'\bar{z})| + |z'|^2 \\
&\le |z|^2 + 2|z'\bar{z}| + |z'|^2 \\
&= |z|^2 + 2|z'||z| + |z'|^2 \\
&= (|z| + |z'|)^2.
\end{aligned}$$

The Triangle Inequality follows now by taking square roots.

REMARK. The Triangle Inequality is often used in conjunction with what's called the "add and subtract trick." Frequently we want to estimate the size of a quantity like $|z-w|$, and we can often accomplish this estimation by adding and subtracting the same thing within the absolute value bars:

$$|z-w| = |z-v+v-w| \leq |z-v| + |v-w|.$$

The point is that we have replaced the estimation problem of the possibly unknown quantity $|z-w|$ by the estimation problems of two other quantities $|z-v|$ and $|v-w|$. It is often easier to estimate these latter two quantities, usually by an ingenious choice of v of course.

Exercise 1.26. (a) Prove the second assertion of the preceding theorem.
(b) Prove the Triangle Inequality for the distance function. That is, prove that

$$d(z,w) \leq d(z,v) + d(v,w)$$

for all $z, w, v \in \mathbb{C}$.
(c) Use mathematical induction to prove that

$$|\sum_{i=1}^{n} a_i| \leq \sum_{i=1}^{n} |a_i|.$$

It may not be necessary to point out that part (b) of the preceding exercise provides a justification for the name "triangle inequality." Indeed, part (b) of that exercise is just the assertion that the length of one side of a triangle in the plane is less than or equal to the sum of the lengths of the other two sides. Plot the three points $z, w,$ and v, and see that this interpretation is correct.

DEFINITION. A subset S of $\mathbb{C}$ is called *Bounded* if there exists a real number M such that $|z| \leq M$ for every z in S.

Exercise 1.27. Let S be a subset of $\mathbb{C}$. Let S_1 be the subset of $\mathbb{R}$ consisting of the real parts of the complex numbers in S, and let S_2 be the subset of $\mathbb{R}$ consisting of the imaginary parts of the elements of S. Prove that S is bounded if and only if S_1 and S_2 are both bounded.
HINT: Use Part (c) of Exercise 1.25.
(b) Let S be the unit circle in the plane, i.e., the set of all complex numbers $z = x + iy$ for which $|z| = 1$. Compute the sets S_1 and S_2 of part (a).

Exercise 1.28. (a) Verify that the formulas for the sum of a geometric progression and the binomial theorem (Theorems 1.9 and 1.10) are valid for complex numbers z and z'.
HINT: Check that, as claimed, the proofs of those theorems work in any field.
(b) Prove Theorem 1.11 for complex numbers z and z'.

CHAPTER II
THE LIMIT OF A SEQUENCE OF NUMBERS
DEFINITION OF THE NUMBER e.

This chapter contains the beginnings of the most important, and probably the most subtle, notion in mathematical analysis, i.e., the concept of a limit. Though Newton and Leibniz discovered the calculus with its tangent lines described as limits of secant lines, and though the Greeks were already estimating areas of regions by a kind of limiting process, the precise notion of limit that we use today was not formulated until the 19th century by Cauchy and Weierstrass.
The main results of this chapter are the following:

(1) The definition of the **limit of a sequence**,
(2) The definition of the real number e (Theorem 2.3),
(3) The **Squeeze Theorem** (Theorem 2.5),
(4) the **Bolzano Weierstrass Theorem** (Theorems 2.8 and 2.10),
(5) The **Cauchy Criterion** (Theorem 2.9),
(6) the definition of an **infinite series**,
(7) the **Comparison Test** (Theorem 2.17), and
(8) the **Alternating Series Test** (Theorem 2.18).

These are powerful basic results about limits that will serve us well in later chapters.

SEQUENCES AND LIMITS

DEFINITION. A *sequence* of real or complex numbers is defined to be a function from the set $\mathbb{N}$ of natural numbers into the set$\mathbb{R}$ or $\mathbb{C}$. Instead of referring to such a function as an assignment $n \rightarrow f(n)$, we ordinarily use the notation $\{a_n\}$, $\{a_n\}_1^\infty$, or $\{a_1, a_2, a_3, \dots\}$. Here, of course, a_n denotes the number $f(n)$.

REMARK. We expand this definition slightly on occasion to make some of our notation more indicative. That is, we sometimes **index** the terms of a sequence beginning with an integer other than 1. For example, we write $\{a_n\}_0^\infty$, $\{a_0, a_1, \dots\}$, or even $\{a_n\}_{-3}^\infty$.
We give next what is the most significant definition in the whole of mathematical analysis, i.e., what it means for a sequence to converge or to have a limit.

DEFINITION. Let $\{a_n\}$ be a sequence of real numbers and let L be a real number. The sequence $\{a_n\}$ is said to *converge* to L, or that L is the *limit* of $\{a_n\}$, if the following condition is satisfied. For every positive number ϵ, there exists a natural number N such that if $n \geq N$, then $|a_n - L| < \epsilon$.
In symbols, we say $L = \lim a_n$ or

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

We also may write $a_n \mapsto L$.
If a sequence $\{a_n\}$ of real or complex numbers converges to a number L, we say that the sequence $\{a_n\}$ is *convergent*.
We say that a sequence $\{a_n\}$ of real numbers *diverges* to $+\infty$ if for every positive number M, there exists a natural number N such that if $n \geq N$, then $a_n \geq M$. Note that we do **not** say that such a sequence is convergent.

Similarly, we say that a sequence $\{a_n\}$ of real numbers *diverges* to $-\infty$ if for every real number M, there exists a natural number N such that if $n \geq N$, then $a_n \leq M$.

The definition of convergence for a sequence $\{z_n\}$ of complex numbers is exactly the same as for a sequence of real numbers. Thus, let $\{z_n\}$ be a sequence of complex numbers and let L be a complex number. The sequence $\{z_n\}$ is said to *converge* to L, or that L is the *limit* of $\{z_n\}$, if the following condition is satisfied. For every positive number ϵ, there exists a natural number N such that if $n \geq N$, then $|z_n - L| < \epsilon$.

REMARKS. The natural number N of the preceding definition surely depends on the positive number ϵ. If ϵ' is a smaller positive number than ϵ, then the corresponding N' very likely will need to be larger than N. Sometimes we will indicate this dependence by writing $N(\epsilon)$ instead of simply N. It is always wise to remember that N depends on ϵ. On the other hand, the N or $N(\epsilon)$ in this definition is not unique. It should be clear that if a natural number N satisfies this definition, then any larger natural number M will also satisfy the definition. So, in fact, if there exists one natural number that works, then there exist infinitely many such natural numbers.

It is clear, too, from the definition that whether or not a sequence is convergent only depends on the "tail" of the sequence. Specifically, for any positive integer K, the numbers $a_1, a_2, \ldots, a_K$ can take on any value whatsoever without affecting the convergence of the entire sequence. We are only concerned with a_n's for $n \geq N$, and as soon as N is chosen to be greater than K, the first part of the sequence is irrelevant.

The definition of convergence is given as a fairly complicated sentence, and there are several other ways of saying the same thing. Here are two: For every $\epsilon > 0$, there exists a N such that, whenever $n \geq N$, $|a_n - L| < \epsilon$. And, given an $\epsilon > 0$, there exists a N such that $|a_n - L| < \epsilon$ for all n for which $n \geq N$. It's a good idea to think about these two sentences and convince yourself that they really do "mean" the same thing as the one defining convergence.

It is clear from this definition that we can't check whether a sequence converges or not unless we know the limit value L. The whole thrust of this definition has to do with estimating the quantity $|a_n - L|$. We will see later that there are ways to tell in advance that a sequence converges without knowing the value of the limit.

EXAMPLE 2.1. Let $a_n = 1/n$, and let us show that $\lim a_n = 0$. Given an $\epsilon > 0$, let us choose a N such that $1/N < \epsilon$. (How do we know we can find such a N?) Now, if $n \geq N$, then we have

$$|a_n - 0| = |\frac{1}{n}| = \frac{1}{n} \leq \frac{1}{N} < \epsilon,$$

which is exactly what we needed to show to conclude that $0 = \lim a_n$.

EXAMPLE 2.2. Let $a_n = (2n+1)/(1-3n)$, and let $L = -2/3$. Let us show that $L = \lim a_n$. Indeed, if $\epsilon > 0$ is given, we must find a N, such that if $n \geq N$ then $|a_n + (2/3)| < \epsilon$. Let us examine the quantity $|a_n + 2/3|$. Maybe we can make some estimates on it, in such a way that it becomes clear how to find the natural number N.

$$\begin{aligned}|a_n+(2/3)| &= |\frac{2n+1}{1-3n}+\frac{2}{3}| \\ &= |\frac{6n+3+2-6n}{3-9n}| \\ &= |\frac{5}{3-9n}| \\ &= \frac{5}{9n-3} \\ &= \frac{5}{6n+3n-3} \\ &\leq \frac{5}{6n} \\ &< \frac{1}{n},\end{aligned}$$

for all $n \geq 1$. Therefore, if N is an integer for which $N > 1/\epsilon$, then

$$|a_n + 2/3| < 1/n \leq 1/N < \epsilon,$$

whenever $n \geq N$, as desired. (How do we know that there exists a N which is larger than the number $1/\epsilon$?)

EXAMPLE 2.3. Let $a_n = 1/\sqrt{n}$, and let us show that $\lim a_n = 0$. Given an $\epsilon > 0$, we must find an integer N that satisfies the requirements of the definition. It's a little trickier this time to choose this N. Consider the positive number ϵ^2. We know, from Exercise 1.16, that there exists a natural number N such that $1/N < \epsilon^2$. Now, if $n \geq N$, then

$$|a_n - 0| = \frac{1}{\sqrt{n}} \leq \frac{1}{\sqrt{N}} = \sqrt{\frac{1}{N}} < \sqrt{\epsilon^2} = \epsilon,$$

which shows that $0 = \lim 1/\sqrt{n}$.

REMARK. A good way to attack a limit problem is to immediately examine the quantity $|a_n - L|$, which is what we did in Example 2.2 above. This is the quantity we eventually wish to show is less than ϵ when $n \geq N$, and determining which N to use is always the hard part. Ordinarily, some algebraic manipulations can be performed on the expression $|a_n - L|$ that can help us figure out exactly how to choose N. Just know that this process takes some getting used to, so practice!

Exercise 2.1. (a) Using the basic definition, prove that $\lim 3/(2n+7) = 0$.
(b) Using the basic definition, prove that $\lim 1/n^2 = 0$.
(c) Using the basic definition, prove that $\lim (n^2+1)/(n^2+100n) = 1$.
HINT: Use the idea from the remark above; i.e., examine the quantity $|a_n - L|$.
(d) Again, using the basic definition, prove that

$$\lim \frac{n+n^2i}{n-n^2i} = -1.$$

Remember the definition of the absolute value of a complex number.

(e) Using the basic definition, prove that

$$\lim \frac{n^3 + n^2 i}{1 - n^3 i} = i.$$

(f) Let $a_n = (-1)^n$. Prove that 1 is **not** the limit of the sequence $\{a_n\}$.
HINT: Suppose the sequence $\{a_n\}$ does converge to 1. Use $\epsilon = 1$, let N be the corresponding integer that exists in the definition, satisfying $|a_n - 1| < 1$ for all $n \geq N$, and then examine the quantity $|a_n - 1|$ for various n's to get a contradiction.

Exercise 2.2. (a) Let $\{a_n\}$ be a sequence of (real or complex) numbers, and let L be a number. Prove that $L = \lim a_n$ if and only if for every positive integer k there exists an integer N, such that if $n \geq N$ then $|a_n - L| < 1/k$.
(b) Let $\{c_n\}$ be a sequence of complex numbers, and suppose that $c_n \mapsto L$. If $c_n = a_n + b_n i$ and $L = a + bi$, show that $a = \lim a_n$ and $b = \lim b_n$. Conversely, if $a = \lim a_n$ and $b = \lim b_n$, show that $a + bi = \lim(a_n + b_n i)$. That is, a sequence $\{c_n = a_n + b_n i\}$ of complex numbers converges if and only if the sequence $\{a_n\}$ of the real parts converges and the sequence $\{b_n\}$ of the imaginary parts converges.
HINT: You need to show that, given some hypotheses, certain quantities are less than ϵ. Part (c) of Exercise 1.25 should be of help.

Exercise 2.3. (a) Prove that a constant sequence ($a_n \equiv c$) converges to c.
(b) Prove that the sequence $\{\frac{2n^2+1}{1-3n}\}$ diverges to $-\infty$.
(c) Prove that the sequence $\{(-1)^n\}$ does not converge to any number L.
HINT: Argue by contradiction. Suppose it does converge to a number L. Use $\epsilon = 1/2$, let N be the corresponding integer that exists in the definition, and then examine $|a_n - a_{n+1}|$ for $n \geq N$. Use the following useful add and subtract trick:

$$|a_n - a_{n+1}| = |a_n - L + L - a_{n+1}| \leq |a_n - L| + |L - a_{n+1}|.$$

EXISTENCE OF CERTAIN FUNDAMENTAL LIMITS

We have, in the preceding exercises, seen that certain specific sequences converge. It's time to develop some general theory, something that will apply to lots of sequences, and something that will help us actually evaluate limits of certain sequences.

DEFINITION. A sequence $\{a_n\}$ of real numbers is called *nondecreasing* if $a_n \leq a_{n+1}$ for all n, and it is called *nonincreasing* if $a_n \geq a_{n+1}$ for all n. It is called *strictly increasing* if $a_n < a_{n+1}$ for all n, and *strictly decreasing* if $a_n > a_{n+1}$ for all n.
A sequence $\{a_n\}$ of real numbers is called *eventually nondecreasing* if there exists a natural number N such that $a_n \leq a_{n+1}$ for all $n \geq N$, and it is called *eventually nonincreasing* if there exists a natural number N such that $a_n \geq a_{n+1}$ for all $n \geq N$. We make analogous definitions of "eventually strictly increasing" and "eventually strictly decreasing."

It is ordinarily very difficult to tell whether a given sequence converges or not; and even if we know in theory that a sequence converges, it is still frequently difficult to tell what the limit is. The next theorem is therefore very useful. It is also very fundamental, for it makes explicit use of the existence of a least upper bound.

THEOREM 2.1. *Let $\{a_n\}$ be a nondecreasing sequence of real numbers. Suppose that the set S of elements of the sequence $\{a_n\}$ is bounded above. Then the sequence $\{a_n\}$ is convergent, and the limit L is given by$L = \sup S = \sup a_n$.*
Analogously, if $\{a_n\}$ is a nonincreasing sequence that is bounded below, then $\{a_n\}$ converges to $\inf a_n$.

PROOF. We prove the first statement. The second is done analogously, and we leave it to an exercise. Write L for the supremum $\sup a_n$. Let ϵ be a positive number. By Theorem 1.5, there exists an integer N such that $a_N > L - \epsilon$, which implies that $L - a_N < \epsilon$. Since $\{a_n\}$ is nondecreasing, we then have that $a_n \geq a_N > L - \epsilon$ for all $n \geq N$. Since L is an upper bound for the entire sequence, we know that $L \geq a_n$ for every n, and so we have that

$$|L - a_n| = L - a_n \leq L - a_N < \epsilon$$

for all $n \geq N$. This completes the proof of the first assertion.

Exercise 2.4. (a) Prove the second assertion of the preceding theorem.
(b) Show that Theorem 2.1 holds for sequences that are eventually nondecreasing or eventually nonincreasing. (Re-read the remark following the definition of the limit of a sequence.)

The next exercise again demonstrates the “denseness” of the rational and irrational numbers in the set $\mathbb{R}$ of all real numbers.

Exercise 2.5. (a) Let x be a real number. Prove that there exists a sequence $\{r_n\}$ of rational numbers such that $x = \lim r_n$. In fact, show that the sequence $\{r_n\}$ can be chosen to be nondecreasing.
HINT: For example, for each n, use Theorem 1.8 to choose a rational number r_n between $x - 1/n$ and x.
(b) Let x be a real number. Prove that there exists a sequence $\{r'{}_n\}$ of irrational numbers such that $x = \lim r'_n$.
(c) Let $z = x + iy$ be a complex number. Prove that there exists a sequence $\{\alpha_n\} = \{\beta_n + i\gamma_n\}$ of complex numbers that converges to z, such that each β_n and each γ_n is a rational number.

Exercise 2.6. Suppose $\{a_n\}$ and $\{b_n\}$ are two convergent sequences, and suppose that $\lim a_n = a$ and $\lim b_n = b$. Prove that the sequence $\{a_n + b_n\}$ is convergent and that

$$\lim(a_n + b_n) = a + b.$$

HINT: Use an $\epsilon/2$ argument. That is, choose a natural number N_1 so that $|a_n - a| < \epsilon/2$ for all $n \geq N_1$, and choose a natural number N_2 so that $|b_n - b| < \epsilon/2$ for all $n \geq N_2$. Then let N be the larger of the two numbers N_1 and N_2.

The next theorem establishes the existence of four nontrivial and important limits. This time, the proofs are more tricky. Some clever idea will have to be used before we can tell how to choose the N.

THEOREM 2.2.

(1) *Let $z \in \mathbb{C}$ satisfy $|z| < 1$, and define $a_n = z^n$. then the sequence $\{a_n\}$ converges to* 0. *We write* $\lim z^n = 0$.

(2) *Let b be a fixed positive number greater than 1, and define $a_n = b^{1/n}$. See Theorem 1.11. Then* $\lim a_n = 1$. *Again, we write* $\lim b^{1/n} = 1$.

(3) *Let b be a positive number less than 1. Then* $\lim b^{1/n} = 1$.

(4) *If $a_n = n^{1/n}$, then* $\lim a_n = \lim n^{1/n} = 1$.

PROOF. We prove parts (1) and (2) and leave the rest of the proof to the exercise that follows. If $z = 0$, claim (1) is obvious. Assume then that $z \neq 0$, and let $\epsilon > 0$ be given. Let $w = 1/|z|$, and observe that $w > 1$. So, we may write $w = 1 + h$ for some positive h. (That step is the clever idea for this argument.) Then, using the Binomial Theorem, $w^n > nh$, and so $1/w^n < 1/(nh)$. See part (a) of Exercise 1.20. But then

$$|z^n - 0| = |z^n| = |z|^n = (1/w)^n = 1/w^n < 1/(nh).$$

So, if N is any natural number larger than $1/(\epsilon h)$, then

$$|z^n - 0| = |z^n| = |z|^n < \frac{1}{nh} \leq \frac{1}{Nh} < \epsilon$$

for all $n \geq N$. This completes the proof of the first assertion of the theorem.
To see part (2), write $a_n = b^{1/n} = 1 + x_n$, i.e., $x_n = b^{1/n} - 1$, and observe first that $x_n > 0$. Indeed, since $b > 1$, it must be that the nth root $b^{1/n}$ is also > 1. (Why?) Therefore, $x_n = b^{1/n} - 1 > 0$. (Again, writing $b^{1/n}$ as $1 + x_n$ is the clever idea.) Now, $b = b^{1/n^n} = (1 + x_n)^n$, which, again by the Binomial Theorem, implies that $b > 1 + nx_n$. So, $x_n < (b-1)/n$, and therefore

$$|b^{1/n} - 1| = b^{1/n} - 1 = x_n < \frac{b-1}{n} < \epsilon$$

whenever $n > \epsilon/(b-1)$, and this proves part (2).

Exercise 2.7. (a) Prove part (3) of the preceding theorem.
HINT: For $b \leq 1$, use the following algebraic calculation:

$$|b^{1/n} - 1| = b^{1/n}|1 - (1/b)^{1/n}| \leq |1 - (1/b)^{1/n}|,$$

and then use part (2) as applied to the positive number $1/b$.
(b) Prove part (4) of the preceding theorem. Explain why it does not follow directly from part (2).
HINT: Write $n^{1/n} = 1 + h_n$. Observe that $h_n > 0$. Then use the third term of the binomial theorem in the expansion $n = (1 + h_n)^n$.
(c) Construct an alternate proof to part (2) of the preceding theorem as follows: Show that the sequence $\{b^{1/n}\}$ is nonincreasing and bounded below by 1. Deduce, from Theorem 2.1, that the sequence converges to a number L. Now prove that L must be 1.

DEFINITION OF e

Part (4) of Theorem 2.2 raises an interesting point. Suppose we have a sequence $\{a_n\}$, like $\{n\}$, that is diverging to infinity, and suppose we have another sequence $\{b_n\}$, like $\{1/n\}$, that is converging to 0. What can be said about the sequence $\{a_n^{b_n}\}$? The base a_n is blowing up, while the exponent b_n is going to 0. In other words, there are two competing processes going on. If a_n is blowing up, then its powers ought to be blowing up as well. On the other hand, anything to the 0 power should be 1, so that, as the exponents of the elements of a sequence converge to 0, the sequence ought to converge to 1. This competition between the convergence of the base to infinity and the convergence of the exponent to 0 makes it subtle, if not impossibly difficult, to tell what the combination does. For the special case of part (4) of Theorem 2.2, the answer was 1, indicating that, in that case at least, the exponents going to 0 seem to be more important than the base going to infinity. One can think up all kinds of such examples: $\{(2^n)^{1/n}\}$, $\{(n!)^{1/n}\}$, $\{(n!)^{1/n^2}\}$, and so on. We will see later that all sorts of things can happen.

Of course there is the reverse situation. Suppose $\{a_n\}$ is a sequence of numbers that decreases to 1, and suppose $\{b_n\}$ is a sequence of numbers that diverges to infinity. What can we say about the sequence $\{{a_n}^{b_n}\}$? The base is tending to 1, so that one might expect that the whole sequence also would be converging to 1. On the other hand the exponents are blowing up, so that one might think that the whole sequence should blow up as well. Again, there are lots of examples, and they don't all work the same way. Here is perhaps the most famous such example.

THEOREM 2.3. (Definition of e.) For $n \geq 1$, define $a_n = (1 + 1/n)^n$. Then the sequence $\{a_n\}$ is nondecreasing and bounded above, whence it is convergent. (We will denote the limit of this special sequence by the letter e.)

PROOF. To see that $\{a_n\}$ is nondecreasing, it will suffice to prove that $a_{n+1}/a_n \geq 1$ for all n. In the computation below, we will use the fact (part (c)of Exercise 1.20) that if $x > -1$ then $(1 + x)^n \geq 1 + nx$. So,

$$\begin{aligned}
\frac{a_{n+1}}{a_n} &= \frac{(1+\frac{1}{n+1})^{n+1}}{(1+\frac{1}{n})^n} \\
&= \frac{(\frac{n+2}{n+1})^{n+1}}{(\frac{n+1}{n})^n} \\
&= \frac{n+1}{n}\frac{\frac{n+2}{n+1})^{n+1}}{(\frac{n+1}{n})^{n+1}} \\
&= \frac{n+1}{n}(\frac{n^2+2n}{n^2+2n+1})^{n+1} \\
&= \frac{n+1}{n}(1-\frac{1}{(n+1)^2})^{n+1} \\
&\geq \frac{n+1}{n}(1-(n+1)(\frac{1}{n+1})^2) \\
&= \frac{n+1}{n}(1-\frac{1}{n+1}) \\
&= \frac{n+1}{n}\frac{n}{n+1} \\
&= 1,
\end{aligned}$$

as desired.

We show next that $\{a_n\}$ is bounded above. This time, we use the binomial theorem, the geometric progression, and Exercise 1.19.

$$
\begin{aligned}
a_n &= (1+\frac{1}{n})^n \\
&= \sum_{k=0}^{n} \binom{n}{k} (\frac{1}{n})^k \\
&< \sum_{k=0}^{n} 2\frac{n^k}{2^k} (\frac{1}{n})^k \\
&= 2\sum_{k=0}^{n} (\frac{1}{2})^k \\
&= 2\frac{1-(\frac{1}{2})^{n+1}}{1-\frac{1}{2}} \\
&< 4,
\end{aligned}
$$

as desired.
That the sequence $\{a_n\}$ converges is now a consequence of Theorem 2.1.

REMARK. We have now defined the real number e. Its central role in mathematics is not at all evident yet; at this point we have no definition of exponential function, logarithm, or trigonometric functions. It does follow from the proof above that e is between 2 and 4, and with a little more careful estimates we can show that actually $e \leq 3$. For the moment, we will omit any further discussion of its precise value. Later, in Exercise 4.19, we will show that it is an irrational number.

PROPERTIES OF CONVERGENT SEQUENCES

Often, our goal is to show that a given sequence is convergent. However, as we study convergent sequences, we would like to establish various properties that they have in common. The first theorem of this section is just such a result.

THEOREM 2.4. *Suppose $\{a_n\}$ is a convergent sequence of real or complex numbers. Then the sequence $\{a_n\}$ forms a bounded set.*

PROOF. Write $L = \lim a_n$. Let ϵ be the positive number 1. Then, there exists a natural number N such that $|a_n - L| < 1$ for all $n \geq N$. By the backward triangle inequality, this implies that $||a_n| - |L|| < 1$ for all $n \geq N$, which implies that $|a_n| \leq |L| + 1$ for all $n \geq N$. This shows that at least the tail of the sequence is bounded by the constant $|L| + 1$.
Next, let K be a number larger than the finitely many numbers $|a_1|, \ldots, |a_{N-1}|$. Then, for any n, $|a_n|$ is either less than K or $|L| + 1$. Let M be the larger of the two numbers K and $|L| + 1$. Then $|a_n| < M$ for all n. Hence, the sequence $\{a_n\}$ is bounded.

Note that the preceding theorem is a partial converse to Theorem 2.1; i.e., a convergent sequence is necessarily bounded. Of course, not every convergent sequence must be either nondecreasing or nonincreasing, so that a full converse to theorem 2.1 is not true. For instance, take $z = -1/2$ in part (1) of Theorem 2.2. It converges to 0 all right, but it is neither nondecreasing nor nonincreasing.

Exercise 2.8. (a) Suppose $\{a_n\}$ is a sequence of real numbers that converges to a number a, and assume that $a_n \geq c$ for all n. Prove that $a \geq c$.
HINT: Suppose not, and let ϵ be the positive number $c - a$. Let N be a natural number corresponding to this choice of ϵ, and derive a contradiction.
(b) If $\{a_n\}$ is a sequence of real numbers for which $\lim a_n = a$, and if $a \neq 0$, then prove that $a_n \neq 0$ for all large enough n. Show in fact that there exists an N such that $|a_n| > |a|/2$ for all $n \geq N$.
HINT: Make use of the positive number $\epsilon = |a|/2$.

Exercise 2.9. (a) If $\{a_n\}$ is a sequence of positive real numbers for which $\lim a_n = a > 0$, prove that $\lim \sqrt{a_n} = \sqrt{a}$.
HINT: Multiply the expression $\sqrt{a_n} - \sqrt{a}$ above and below by $\sqrt{a_n} + \sqrt{a}$.
(b) If $\{a_n\}$ is a sequence of complex numbers, and $\lim a_n = a$, prove that $\lim |a_n| = |a|$.
HINT: Use the backward triangle inequality.

Exercise 2.10. Suppose $\{a_n\}$ is a sequence of real numbers and that $L = \lim a_n$. Let M_1 and M_2 be real numbers such that $M_1 \leq a_n \leq M_2$ for all n. Prove that $M_1 \leq L \leq M_2$.
HINT: Suppose, for instance, that $L > M_2$. Make use of the positive number $L - M_2$ to derive a contradiction.

We are often able to show that a sequence converges by comparing it to another sequence that we already know converges. The following exercise demonstrates some of these techniques.

Exercise 2.11. Let $\{a_n\}$ be a sequence of complex numbers.
(a) Suppose that, for each n, $|a_n| < 1/n$. Prove that $0 = \lim a_n$.
(b) Suppose $\{b_n\}$ is a sequence that converges to 0, and suppose that, for each n, $|a_n| < |b_n|$. Prove that $0 = \lim a_n$.

The next result is perhaps the most powerful technique we have for showing that a given sequence converges to a given number.

THEOREM 2.5. $\{$*Squeeze Theorem*) Suppose that $\{a_n\}$ is a sequence of real numbers and that $\{b_n\}$ and $\{c_n\}$ are two sequences of real numbers for which $b_n \leq a_n \leq c_n$ for all n. Suppose further that $\lim b_n = \lim c_n = L$. Then the sequence $\{a_n\}$ also converges to L.

PROOF. We examine the quantity $|a_n - L, |$ employ some add and subtract tricks, and make the following computations:

$$\begin{aligned} |a_n - L| &\leq |a_n - b_n + b_n - L| \\ &\leq |a_n - b_n| + |b_n - L| \\ &= a_n - b_n + |b_n - L| \\ &\leq c_n - b_n + |b_n - L| \\ &= |c_n - b_n| + |b_n - L| \\ &\leq |c_n - L| + |L - b_n| + |b_n - L|. \end{aligned}$$

So, we can make $|a_n - L| < \epsilon$ by making $|c_n - L| < \epsilon/3$ and $|b_n - L| < \epsilon/3$. So, let N_1 be a positive integer such that $|c_n - L| < \epsilon/3$ if $n \geq N_1$, and let N_2 be a positive integer so that $|b_n - L| < \epsilon/3$ if $n \geq N_2$. Then set $N = \max(N_1, N_2)$.

Clearly, if $n \geq N$, then both inequalities $|c_n - L| < \epsilon/3$ and $|b_n - L| < \epsilon/3$, and hence $|a_n - L| < \epsilon$. This finishes the proof.

The next result establishes what are frequently called the " limit theorems." Basically, these results show how convergence interacts with algebraic operations.

THEOREM 2.6. *Let $\{a_n\}$ and $\{b_n\}$ be two sequences of complex numbers with $a = \lim a_n$ and $b = \lim b_n$. Then*

(1) *The sequence $\{a_n + b_n\}$ converges, and*

$$\lim(a_n + b_n) = \lim a_n + \lim b_n = a + b.$$

(2) *The sequence $\{a_n b_n\}$ is convergent, and*

$$\lim(a_n b_n) = \lim a_n \lim b_n = ab.$$

(3) *If all the b_n's as well as b are nonzero, then the sequence $\{a_n/b_n\}$ is convergent, and*

$$\lim(\frac{a_n}{b_n} = \frac{\lim a_n}{\lim b_n} = \frac{a}{b}.$$

PROOF. Part (1) is exactly the same as Exercise 2.6. Let us prove part (2).
By Theorem 2.4, both sequences $\{a_n\}$ and $\{b_n\}$ are bounded. Therefore, let M be a number such that $|a_n| \leq M$ and $|b_n| \leq M$ for all n. Now, let $\epsilon > 0$ be given. There exists an N_1 such that $|a_n - a| < \epsilon/(2M)$ whenever $n \geq N_1$, and there exists an N_2 such that $|b_n - b| < \epsilon/(2M)$ whenever $n \geq N_2$. Let N be the maximum of N_1 and N_2. Here comes the add and subtract trick again.

$$\begin{aligned}
|a_n b_n - ab| &= |a_n b_n - a b_n + a b_n - ab| \\
&\leq |a_n b_n - a b_n| + |a b_n - ab| \\
&= |a_n - a||b_n| + |a||b - b_n| \\
&\leq |a_n - a|M + M|b_n - b| \\
&< \epsilon
\end{aligned}$$

if $n \geq N$, which shows that $\lim(a_n b_n) = ab$.
To prove part (3), let M be as in the previous paragraph, and let $\epsilon > 0$ be given. There exists an N_1 such that $|a_n - a| < (\epsilon|b|^2)/(4M)$ whenever $n \geq N_1$; there also exists an N_2 such that $|b_n - b| < (\epsilon|b|^2)/(4M)$ whenever $n \geq N_2$; and there exists an N_3 such that $|b_n| > |b|/2$ whenever $n \geq N_3$. (See Exercise 2.8.) Let N be the maximum of the three numbers N_1, N_2 and N_3. Then:

$$\begin{aligned}
|\frac{a_n}{b_n} - \frac{a}{b}| &= |\frac{a_n b - b_n a}{b_n b}| \\
&= |a_n b - b_n a|\frac{1}{|b_n b|} \\
&< |a_n b - b_n a|\frac{1}{|b|^2/2} \\
&\leq (|a_n - a||b| + |a||b_n - b|)\frac{2}{|b|^2} \\
&< (M|a_n - a| + M|b_n - b|)\frac{2}{|b|^2} \\
&< \epsilon
\end{aligned}$$

if $n \geq N$. This completes the proof.

REMARK. The proof of part (3) of the preceding theorem may look mysterious. Where, for instance, does this number $\epsilon|b|^2/4M$ come from? The answer is that one begins such a proof by examining the quantity $|a_n/b_n - a/b|$ to see if by some algebraic manipulation one can discover how to control its size by using the quantities $|a_n - a|$ and $|b_n - b|$. The assumption that $a = \lim a_n$ and $b = \lim b_n$ mean exactly that the quantities $|a_n - a|$ and $|b_n - b|$ can be controlled by requiring n to be large enough. The algebraic computation in the proof above shows that

$$|\frac{a_n}{b_n} - \frac{a}{b}| \leq (M|a_n - a| + M|b_n - b|)\frac{2}{|b|^2},$$

and one can then see exactly how small to make $|a_n - a|$ and $|b_n - b|$ so that $|a_n/b_n - a/b| < \epsilon$. Indeed, this is the way most limit proofs work.

Exercise 2.12. If possible, determine the limits of the following sequences by using Theorems 2.2, 2.3, 2.6, and the squeeze theorem 2.5.
(a) $\{n^{1/n^2}\}$.
(b) $\{(n^2)^{1/n}\}$.
(c) $\{(1+n)^{1/n}\}$.
(d) $\{(1+n^2)^{1/n^3}\}$.
(e) $\{(1+1/n)^{2/n}\}$.
(f) $\{(1+1/n)^{2n}\}$.
(g) $\{(1+1/n)^{n^2}\}$.
(h) $\{(1-1/n)^n\}$.
HINT: Note that

$$1 - 1/n = \frac{n-1}{n} = \frac{1}{\frac{n}{n-1}} = \frac{1}{\frac{n-1+1}{n-1}} = \frac{1}{1+\frac{1}{n-1}}.$$

(i) $\{(1-1/(2n))^{3n}\}$.
(j) $\{(n!)^{1/n}\}$.

SUBSEQUENCES AND CLUSTER POINTS

DEFINITION. Let $\{a_n\}$ be a sequence of real or complex numbers. A *subsequence* of $\{a_n\}$ is a sequence $\{b_k\}$ that is determined by the sequence $\{a_n\}$ together with a strictly increasing sequence $\{n_k\}$ of natural numbers. The sequence $\{b_k\}$ is defined by $b_k = a_{n_k}$. That is, the kth term of the sequence $\{b_k\}$ is the n_kth term of the original sequence $\{a_n\}$.

Exercise 2.13. Prove that a subsequence of a subsequence of $\{a_n\}$ is itself a subsequence of $\{a_n\}$. Thus, let $\{a_n\}$ be a sequence of numbers, and let $\{b_k\} = \{a_{n_k}\}$ be a subsequence of $\{a_n\}$. Suppose $\{c_j\} = \{b_{k_j}\}$ is a subsequence of the sequence $\{b_k\}$. Prove that $\{c_j\}$ is a subsequence of $\{a_n\}$. What is the strictly increasing sequence $\{m_j\}$ of natural numbers for which $c_j = a_{m_j}$?

Here is an interesting generalization of the notion of the limit of a sequence.

DEFINITION. Let $\{a_n\}$ be a sequence of real or complex numbers. A number x is called a *cluster point* of the sequence $\{a_n\}$ if there exists a subsequence $\{b_k\}$ of $\{a_n\}$ such that $x = \lim b_k$. The set of all cluster points of a sequence $\{a_n\}$ is called the *cluster set* of the sequence.

Exercise 2.14. (a) Give an example of a sequence whose cluster set contains two points. Give an example of a sequence whose cluster set contains exactly n points. Can you think of a sequence whose cluster set is infinite?
(b) Let $\{a_n\}$ be a sequence with cluster set S. What is the cluster set for the sequence $\{-a_n\}$? What is the cluster set for the sequence $\{a_n^2\}$?
(c) If $\{b_n\}$ is a sequence for which $b = \lim b_n$, and $\{a_n\}$ is another sequence, what is the cluster set of the sequence $\{a_n b_n\}$?
(d) Give an example of a sequence whose cluster set is empty.
(e) Show that if the sequence $\{a_n\}$ is bounded above, then the cluster set S is bounded above. Show also that if $\{a_n\}$ is bounded below, then S is bounded below.
(f) Give an example of a sequence whose cluster set S is bounded above but not bounded below.
(g) Give an example of a sequence that is not bounded, and which has exactly one cluster point.

THEOREM 2.7. *Suppose $\{a_n\}$ is a sequence of real or complex numbers.*

(1) (Uniqueness of limits) Suppose $\lim a_n = L$, and $\lim a_n = M$. Then $L = M$. That is, if the limit of a sequence exists, it is unique.
(2) If $L = \lim a_n$, and if $\{b_k\}$ is a subsequence of $\{a_n\}$, then the sequence $\{b_k\}$ is convergent, and $\lim b_k = L$. That is, if a sequence has a limit, then every subsequence is convergent and converges to that same limit.

PROOF. Suppose $\lim a_n = L \text{and} \lim a_n = M$. Let ϵ be a positive number, and choose N_1 so that $|a_n - L| < \epsilon/2$ if $n \geq N_1$, and choose N_2 so that $|a_n - M| < \epsilon/2$ if $n \geq N_2$. Choose an n larger than both $N_1 \text{and} N_2$. Then

$$|L - M| = |L - a_n + a_n - M| \leq |L - a_n| + |a_n - M| < \epsilon.$$

Therefore, since $|L - M| < \epsilon$ for every positive ϵ, it follows that $L - M = 0$ or $L = M$. This proves part (1).
Next, suppose $\lim a_n = L$ and let $\{b_k\}$ be a subsequence of $\{a_n\}$. We wish to show that $\lim b_k = L$. Let $\epsilon > 0$ be given, and choose an N such that $|a_n - L| < \epsilon$ if $n \geq N$. Choose a K so that $n_K \geq N$. (How?) Then, if $k \geq K$, we have $n_k \geq n_K \geq N$, whence $|b_k - L| = |a_{n_k} - L| < \epsilon$, which shows that $\lim b_k = L$. This proves part (2).

REMARK. The preceding theorem has the following interpretation. It says that if a sequence converges to a number L, then the cluster set of the sequence contains only one number, and that number is L. Indeed, if x is a cluster point of the sequence, then there must be some subsequence that converges to x. But, by part (2), everysubsequence converges to L. Then, by part (1), $x = L$. Part (g) of Exercise 2.14 shows that the converse of this theorem is not valid. that is, the cluster set may contain only one point, and yet the sequence is not convergent.
We give next what is probably the most useful fundamental result about sequences, the Bolzano-Weierstrass Theorem. It is this theorem that will enable us to derive many of the important properties of continuity, differentiability, and integrability.

THEOREM 2.8. (Bolzano-Weierstrass) Every bounded sequence $\{a_n\}$ of real or complex numbers has a cluster point. In other words, every bounded sequence has a convergent subsequence.

The Bolzano-Weierstrass Theorem is, perhaps not surprisingly, a very difficult theorem to prove. We begin with a technical, but very helpful, lemma.

LEMMA. *Let $\{a_n\}$ be a bounded sequence of real numbers; i.e., assume that there exists an M such that $|a_n| \leq M$ for all n. For each $n \geq 1$, let S_n be the set whose elements are $\{a_n, a_{n+1}, a_{n+2}, \dots\}$. That is, S_n is just the elements of the tail of the sequence from n on. Define $x_n = \sup S_n = \sup_{k\geq n} a_k$. Then*

(1) *The sequence $\{x_n\}$ is bounded (above and below).*
(2) *The sequence $\{x_n\}$ is non-increasing.*
(3) *The sequence $\{x_n\}$ converges to a number x.*
(4) *The limit x of the sequence $\{x_n\}$ is a cluster point of the sequence $\{a_n\}$. That is, there exists a subsequence $\{b_k\}$ of the sequence $\{a_n\}$ that converges to x.*
(5) *If y is any cluster point of the sequence $\{a_n\}$, then $y \leq x$, where x is the cluster point of part (4). That is, x is the maximum of all cluster points of the sequence $\{a_n\}$.*

PROOF OF THE LEMMA. Since x_n is the supremum of the set S_n, and since each element of that set is bounded between $-M$ and M, part (1) is immediate. Since $S_{n+1} \subseteq S_n$, it is clear that

$$x_{n+1} = \sup S_{n+1} \leq \sup S_n = x_n,$$

showing part (2).

The fact that the sequence $\{x_n\}$ converges to a number x is then a consequence of Theorem 2.1.

We have to show that the limit x of the sequence $\{x_n\}$ is a cluster point of $\{a_n\}$. Notice that $\{x_n\}$ may not itself be a subsequence of $\{a_n\}$, each x_n may or may not be one of the numbers a_k, so that there really is something to prove. In fact, this is the hard part of this lemma. To finish the proof of part (4), we must define an increasing sequence $\{n_k\}$ of natural numbers for which the corresponding subsequence $\{b_k\} = \{a_{n_k}\}$ of $\{a_n\}$ converges to x. We will choose these natural numbers $\{n_k\}$ so that $|x - a_{n_k}| < 1/k$. Once we have accomplished this, the fact that the corresponding subsequence $\{a_{n_k}\}$ converges to x will be clear. We choose the n_k's inductively. First, using the fact that $x = \lim x_n$, choose an n so that $|x_n - x| = x_n - x < 1/1$. Then, because $x_n = \sup S_n$, we may choose by Theorem 1.5 some $m \geq n$ such that $x_n \geq a_m > x_n - 1/1$. But then $|a_m - x| < 1/1$. (Why?) This m we call n_1. We have that $|a_{n_1} - x| < 1/1$.

Next, again using the fact that $x = \lim x_n$, choose another n so that $n > n_1$ and so that $|x_n - x| = x_n - x < 1/2$. Then, since this $x_n = \sup S_n$, we may choose another $m \geq n$ such that $x_n \geq a_m > x_n - 1/2$. This m we call n_2. Note that we have $|a_{n_2} - x| < 1/2$.

Arguing by induction, if we have found an increasing set $n_1 < n_2 < \dots < n_j$, for which $|a_{n_i} - x| < 1/i$ for $1 \leq i \leq j$, choose an n larger than n_j such that $|x_n - x| < 1/(j+1)$. Then, since $x_n = \sup S_n$, choose an $m \geq n$ so that $x_n \geq a_m > x_n - 1/(j+1)$. Then $|a_m - x| < 1/(j+1)$, and we let n_{j+1} be this m. It follows that $|a_{n_{j+1}} - x| < 1/(j+1)$.

So, by recursive definition, we have constructed a subsequence of $\{a_n\}$ that converges to x, and this completes the proof of part (4) of the lemma.
Finally, if y is any cluster point of $\{a_n\}$, and if $y = \lim a_{n_k}$, then $n_k \geq k$, and so $a_{n_k} \leq x_k$, implying that $x_k - a_{n_k} \geq 0$. Hence, taking limits on k, we see that $x - y \geq 0$, and this proves part (5).

Now, using the lemma, we can give the proof of the Bolzano-Weierstrass Theorem.

PROOF OF THEOREM 2.8. If $\{a_n\}$ is a sequence of real numbers, this theorem is an immediate consequence of part (4) of the preceding lemma.
If $a_n = b_n + c_n i$ is a sequence of complex numbers, and if $\{a_n\}$ is bounded, then $\{b_n\}$ and $\{c_n\}$ are both bounded sequences of real numbers. See Exercise 1.27. So, by the preceding paragraph, there exists a subsequence $\{b_{n_k}\}$ of $\{b_n\}$ that converges to a real number b. Now, the subsequence $\{c_{n_k}\}$ is itself a bounded sequence of real numbers, so there is a subsequence $\{c_{n_{k_j}}\}$ that converges to a real number c. By part (2) of Theorem 2.7, we also have that the subsequence $\{b_{n_{k_j}}\}$ converges to b. So the subsequence $\{a_{n_{k_j}}\} = \{b_{n_{k_j}} + c_{n_{k_j}} i\}$ of $\{a_n\}$ converges to the complex number $b + ci$; i.e., $\{a_n\}$ has a cluster point. This completes the proof.

There is an important result that is analogous to the Lemma above, and its proof is easily adapted from the proof of that lemma.

Exercise 2.15. Let $\{a_n\}$ be a bounded sequence of real numbers. Define a sequence $\{y_n\}$ by $y_n = inf_{k\geq n} a_k$. Prove that:
(a) $\{y_n\}$ is nondecreasing and bounded above.
(b) $y = \lim y_n$ is a cluster point of $\{a_n\}$.
(c) If z is any cluster point of $\{a_n\}$, then $y \leq z$. That is, y is the minimum of all the cluster points of the sequence $\{a_n\}$.
HINT: Let $\{\alpha_n\} = \{-a_n\}$, and apply the preceding lemma to $\{\alpha_n\}$. This exercise will then follow from that.

The Bolzano-Wierstrass Theorem shows that the cluster set of a bounded sequence $\{a_n\}$ is nonempty. It is also a bounded set itself.
The following definition is only for sequences of real numbers. However, like the Bolzano-Weierstrass Theorem, it is of very basic importance and will be used several times in the sequel.

DEFINITION. Let $\{a_n\}$ be a sequence of real numbers and let S denote its cluster set.
If S is nonempty and bounded above, we define $\limsup a_n$ to be the supremum $\sup S$ of S.
If S is nonempty and bounded below, we define $\liminf a_n$ to be the infimum $infS$ of S.
If the sequence $\{a_n\}$ of real numbers is not bounded above, we define $\limsup a_n$ to be ∞, and if $\{a_n\}$ is not bounded below, we define $\liminf a_n$ to be $-\infty$.
If $\{a_n\}$ diverges to ∞, then we define $\limsup a_n$ and $\liminf a_n$ both to be ∞. And, if $\{a_n\}$ diverges to $-\infty$, we define $\limsup a_n$ and $\liminf a_n$ both to be $-\infty$.
We call $\limsup a_n$ the *limit superior* of the sequence $\{a_n\}$, and $\liminf a_n$ the *limit inferior* of $\{a_n\}$.

Exercise 2.16. (a) Suppose $\{a_n\}$ is a bounded sequence of real numbers. Prove that the sequence $\{x_n\}$ of the lemma following Theorem 2.8 converges to $\limsup a_n$. Show also that the sequence $\{y_n\}$ of Exercise 2.15 converges to $\liminf a_n$.

(b) Let $\{a_n\}$ be a not necessarily bounded sequence of real numbers. Prove that

$$\limsup a_n = \inf_n \sup_{k\geq n} a_k = \lim_n \sup_{k\geq n} a_k.$$

and

$$\liminf a_n = \sup_n \inf_{k\geq n} a_k = \lim_n \inf k \geq n a_k.$$

HINT: Check all cases, and use the lemma following Theorem 2.8 and Exercise 2.15.
(c) Let $\{a_n\}$ be a sequence of real numbers. Prove that

$$\limsup a_n = -\liminf(-a_n).$$

(d) Give examples to show that all four of the following possibilities can happen.

(1) $\limsup a_n$ is finite, and $\liminf a_n = -\infty$.
(2) $\limsup a_n = \infty$ and $\liminf a_n$ is finite.
(3) $\limsup a_n = \infty$ and $\liminf a_n = -\infty$.
(4) both $\limsup a_n$ and $\liminf a_n$ are finite.

The notions of limsup and liminf are perhaps mysterious, and they are in fact difficult to grasp. The previous exercise describes them as the resultof a kind of two-level process, and there are occasions when this description is a great help. However, the limsup and liminf can also be characterized in other ways that are more reminiscent of the definition of a limit. These other ways are indicated in the next exercise.

Exercise 2.17. Let $\{a_n\}$ be a bounded sequence of real numbers with $\limsup a_n = L$ and $\liminf a_n = l$. Prove that L and l satisfy the following properties.
(a) For each $\epsilon > 0$, there exists an N such that $a_n < L + \epsilon$ for all $n \geq N$.
HINT: Use the fact that $\limsup a_n = L$ is the number x of the lemma following Theorem 2.8, and that x is the limit of a specific sequence $\{x_n\}$.
(b) For each $\epsilon > 0$, and any natural number k, there exists a natural number $j \geq k$ such that $a_j > L - \epsilon$. Same hint as for part (a).
(c) For each $\epsilon > 0$, there exists an N such that $a_n > l - \epsilon$ for all $n \geq N$.
(d) For each $\epsilon > 0$, and any natural number k, there exists a natural number $j > k$ such that $a_j < l + \epsilon$.
(e) Suppose L' is a number that satisfies parts (a) and (b). Prove that L' is the limsup of $\{a_n\}$.
HINT: Use part (a) to show that L' is greater than or equal to every cluster point of $\{a_n\}$. Then use part (b) to show that L' is less than or equal to some cluster point.
(f) If l' is any number that satisfies parts (c) and (d), show that l' is the liminf of the sequence $\{a_n\}$.

Exercise 2.18. (a) Let $\{a_n\}$ and $\{b_n\}$ be two bounded sequences of real numbers, and write $L = \limsup a_n$ and $M = \limsup b_n$. Prove that $\limsup(a_n + b_n) \leq \limsup a_n + \limsup b_n$.
HINT: Using part (a) of the preceding exercise, show that for every $\epsilon > 0$ there exists a N such that $a_n + b_n < L + M + \epsilon$ for all $n \geq N$, and conclude from this that every cluster point y of the sequence $\{a_n + b_n\}$ is less than or equal to $L + M$. This will finish the proof, since $\limsup(a_n + b_n)$ is a cluster point of that sequence.

(b) Again, let $\{a_n\}$ and $\{b_n\}$ be two bounded sequences of real numbers, and write $l = \liminf a_n$ and $m = \liminf b_n$. Prove that $\liminf(a_n+b_n) \geq \liminf a_n + \liminf b_n$. HINT: Use part (c) of the previous exercise.
(c) Find examples of sequences $\{a_n\}$ and $\{b_n\}$ for which $\limsup a_n = \limsup b_n = 1$, but $\limsup(a_n + b_n) = 0$.

We introduce next another property that a sequence can possess. It looks very like the definition of a convergent sequence, but it differs in a crucial way, and that is that this definition only concerns the elements of the sequence $\{a_n\}$ and not the limit L.

DEFINITION. A sequence $\{a_n\}$ of real or complex numbers is a *Cauchy* sequence if for every $\epsilon > 0$, there exists a natural number N such that if $n \geq N$ and $m \geq N$ then $|a_n - a_m| < \epsilon$.

REMARK. No doubt, this definition has something to do with limits. Any time there is a positive ϵ and an N, we must be near some kind of limit notion. The point of the definition of a Cauchy sequence is that there is no explicit mention of what the limit is. It isn't that the terms of the sequence are getting closer and closer to some number L, it's that the terms of the sequence are getting closer and closer to each other. This subtle difference is worth some thought.

Exercise 2.19. Prove that a Cauchy sequence is bounded. (Try to adjust the proof of Theorem 2.4 to work for this situation.)

The next theorem, like the Bolzano-Weierstrass Theorem, seems to be quite abstract, but it also turns out to be a very useful tool for proving theorems about continity, differentiability, etc. In the proof, the completeness of the set of real numbers will be crucial. This theorem is not true in ordered fields that are not complete.

THEOREM 2.9. (Cauchy Criterion) A sequence $\{a_n\}$ of real or complex numbers is convergent if and only if it is a Cauchy sequence.

PROOF. If $lima_n = a$ then given $\epsilon > 0$, choose N so that $|a_k - a| < \epsilon/2$ if $k \geq N$. From the triangle inequality, and by adding and subtracting a, we obtain that $|a_n - a_m| < \epsilon$ if $n \geq N$ and $m \geq N$. Hence, if $\{a_n\}$ is convergent, then $\{a_n\}$ is a Cauchy sequence.
Conversely, if $\{a_n\}$ is a cauchy sequence, then $\{a_n\}$ is bounded by the previous exercise. Now we use the fact that $\{a_n\}$ is a sequence of real or complex numbers. Let x be a cluster point of $\{a_n\}$. We know that one exists by the Bolzano-Weierstrass Theorem. Let us show that in fact this number x not only is a cluster point but that it is in fact the limit of the sequence $\{a_n\}$. Given $\epsilon > 0$, choose Nso that $|a_n - a_m| < \epsilon/2$ whenever both n and $m \geq N$. Let $\{a_{n_k}\}$ be a subsequence of $\{a_n\}$ that converges to x. Because $\{n_k\}$ is strictly increasing, we may choose a k so that $n_k > N$ and also so that $|a_{n_k} - x| < \epsilon/2$. Then, if $n \geq N$, then both n and this particular n_k are larger than or equal to N. Therefore, $|a_n - x| \leq |a_n - a_{n_k}| + |a_{n_k} - x| < \epsilon$. this completes the proof that $x = \lim a_n$.

A LITTLE TOPOLOGY

We now investigate some properties that subsets of $\mathbb{R}$ and $\mathbb{C}$ may possess. We will define "closed sets," "open sets," and "limit points" of sets. These notions are the rudimentary notions of what is called topology. As in earlier definitions, these topological ones will be enlightening when we come to continuity.

DEFINITION. Let S be a subset of $\mathbb{C}$. A complex number x is called a *limit point* of S if there exists a sequence $\{x_n\}$ of elements of S such that $x = \lim x_n$. A set $S \subseteq \mathbb{C}$ is called *closed* if every limit point of S belongs to S.

Every limit point of a set of real numbers is a real number. Closed intervals $[a, b]$ are examples of closed sets in $\mathbb{R}$, while open intervals and half-open intervals may not be closed sets. Similarly, closed disks $\overline{B_r(c)}$ of radius r around a point c in $\mathbb{C}$, and closed neighborhoods $\overline{N}_r(S)$ of radius r around a set $S \subseteq \mathbb{C}$, are closed sets, while the open disks or open neighborhoods are not closed sets. As a first example of a limit point of a set, we give the following exercise.

Exercise 2.20. Let S be a nonempty bounded set of real numbers, and let $M = \sup S$. Prove that there exists a sequence $\{a_n\}$ of elements of S such that $M = \lim a_n$. That is, prove that the supremum of a bounded set of real numbers is a limit point of that set. State and prove an analogous result for infs.
HINT: Use Theorem 1.5, and let ϵ run through the numbers $1/n$.

Exercise 2.21. (a) Suppose S is a set of real numbers, and that $z = a + bi \in \mathbb{C}$ with $b \neq 0$. Show that z is not a limit point of S. That is, every limit point of a set of real numbers is a real number.
HINT: Suppose false; write $a + bi = \lim x_n$, and make use of the positive number $|b|$.
(b) Let c be a complex number, and let $S = \overline{B}_r(c)$ be the set of all $z \in \mathbb{C}$ for which $|z - c| \leq r$. Show that S is a closed subset of $\mathbb{C}$.
HINT: Use part (b) of Exercise 2.9.
(c) Show that the open disk $B_r(0)$ is not a closed set in $\mathbb{C}$ by finding a limit point of $B_r(0)$ that is not in $B_r(0)$.
(d) State and prove results analogous to parts b and c for intervals in $\mathbb{R}$.
(e) Show that every element x of a set S is a limit point of S.
(f) Let S be a subset of $\mathbb{C}$, and let x be a complex number. Show that x is not a limit point of S if and only if there exists a positive number ϵ such that if $|y - x| < \epsilon$, then y is not in S. That is, $S \cap B_\epsilon(x) = \emptyset$.
HINT: To prove the " only if" part, argue by contradiction, and use the sequence $\{1/n\}$ as ϵ's.
(g) Let $\{a_n\}$ be a sequence of complex numbers, and let S be the set of all the a_n's. What is the difference between a cluster point of the sequence $\{a_n\}$ and a limit point of the set S?
(h) Prove that the cluster set of a sequence is a closed set.
HINT: Use parts (e) and (f).

Exercise 2.22. (a) Show that the set $\mathbb{Q}$ of all rational numbers is not a closed set. Show also that the set of all irrational numbers is not a closed set.
(b) Show that if S is a closed subset of $\mathbb{R}$ that contains $\mathbb{Q}$, then S must equal all of $\mathbb{R}$.

Here is another version of the Bolzano-Weierstrass Theorem, this time stated in terms of closed sets rather than bounded sequences.

THEOREM 2.10. *Let S be a bounded and closed subset of $\mathbb{C}$. Then every sequence $\{x_n\}$ of elements of S has a subsequence that converges to an element of S.*

PROOF. Let $\{x_n\}$ be a sequence in S. Since S is bounded, we know by Theorem 2.8 that there exists a subsequence $\{x_{n_k}\}$ of $\{x_n\}$ that converges to some number x. Since each x_{n_k} belongs to S, it follows that x is a limit point of S. Finally, because S is a closed subset of $\mathbb{C}$, it then follows that $x \in S$.

We have defined the concept of a closed set. Now let's give the definition of an open set.

DEFINITION. Let S be a subset of $\mathbb{C}$. A point $x \in S$ is called an *interior point* of S if there exists an $\epsilon > 0$ such that the open disk $B_\epsilon(x)$ of radius ϵ around x is entirely contained in S. The set of all interior points of S is denoted by S^0 and we call S^0 the *interior* of S.
A subset S of $\mathbb{C}$ is called an *open* subset of $\mathbb{C}$ if every point of S is an interior point of S; i.e., if $S = S^0$.
Analogously, let S be a subset of $\mathbb{R}$. A point $x \in S$ is called an *interior point* of S if there exists an $\epsilon > 0$ such that the open interval $(x - \epsilon, x + \epsilon)$ is entirely contained in S. Again, we denote the set of all interior points of S by S^0 and call S^0 the *interior* of S.
A subset S of $\mathbb{R}$ is called an *open* subset of $\mathbb{R}$ if every point of S is an interior point of S; i.e., if $S = S^0$.

Exercise 2.23. (a) Prove that an open interval (a, b) in $\mathbb{R}$ is an open subset of $\mathbb{R}$; i.e., show that every point of (a, b) is an interior point of (a, b).
(b) Prove that any disk $B_r(c)$ is an open subset of $\mathbb{C}$. Show also that the *punctured disk* $B'_r(c)$ is an open set, where $B'_r(c) = \{z : 0 < |z - c| < r\}$, i.e., evrything in the disk $B_r(c)$ except the central point c.
(c) Prove that the neighborhood $N_r(S)$ of radius r around a set S is an open subset of $\mathbb{C}$.
(d) Prove that no nonempty subset of $\mathbb{R}$ is an open subset of $\mathbb{C}$.
(e) Prove that the set $\mathbb{Q}$ of all rational numbers is not an open subset of $\mathbb{R}$. We have seen in part (a) of Exercise 2.22 that $\mathbb{Q}$ is not a closed set. Consequently it is an example of a set that is neither open nor closed. Show that the set of all irrational numbers is neither open nor closed.

We give next a useful application of the Bolzano-Weierstrass Theorem, or more precisely an application of Theorem 2.10. This also provides some insight into the structure of open sets.

THEOREM 2.11. *Let S be a closed and bounded subset of $\mathbb{C}$, and suppose S is a subset of an open set U. Then there exists an $r > 0$ such that the neighborhood $N_r(S)$ is contained in U. That is, every open set containing a closed and bounded set S actually contains a neighborhood of S.*

PROOF. If S is just a singleton $\{x\}$, then this theorem is asserting nothing more than the fact that x is in the interior of U, which it is if U is an open set. However, when S is an infinite set, then the result is more subtle. We argue by contradiction. Thus, suppose there is no such $r > 0$ for which $N_r(S) \subseteq U$. then for each positive integer n there must be a point x_n that is not in U, and a corresponding point

$y_n \in S$, such that $|x_n - y_n| < 1/n$. Otherwise, the number $r = 1/n$ would satisfy the claim of the theorem. Now, because the y_n's all belong to S, we know from Theorem 2.10 that a subsequence $\{y_{n_k}\}$ of the sequence $\{y_n\}$ must converge to a number $y \in S$. Next, we see that

$$|x_{n_k} - y| \leq |x_{n_k} - y_{n_k}| + |y_{n_k} - y|, < \frac{1}{n_k} + |y_{n_k} - y|,$$

and this quantity tends to 0. Hence, the subsequence $\{x_{n_k}\}$ of the sequence $\{x_n\}$ also converges to y.

Finally, because y belongs to S and hence to the open set U, we know that there must exist an $\epsilon > 0$ such that the entire disk $B_\epsilon(y) \subseteq U$. Then, since the subsequence $\{x_{n_k}\}$ converges to y, there must exist an_k such that $|x_{n_k} - y| < \epsilon$, implying that $x_{n_k} \in B_\epsilon(y)$, and hence belongs to U. But this is our contradiction, because all of the x_n's were not in U. So, the theorem is proved.

We give next a result that cliarifies to some extent the connection between open sets and closed sets. Always remember that there are sets that are neither open nor closed, and just because a set is not open **does not** mean that it is closed.

THEOREM 2.12. *A subset S of $\mathbb{C}$ ($\mathbb{R}$) is open if and only if its complement $\tilde{S} = \mathbb{C} \setminus S$ ($\mathbb{R} \setminus S$) is closed.*

PROOF. First, assume that S is open, and let us show that $\tilde{S}$ is closed. Suppose not. We will derive a contradiction. Suppose then that there is a sequence $\{x_n\}$ of elements of $\tilde{S}$ that converges to a number x that is not in $\tilde{S}$; i.e., x is an element of S. Since every element of S is an interior point of S, there must exist an $\epsilon > 0$ such that the entire disk $B_\epsilon(x)$ (or interval $(x - \epsilon, x + \epsilon)$) is a subset of S. Now, since $x = \lim x_n$, there must exist an$N$ such that $|x_n - x| < \epsilon$ for every $n \geq N$. In particular, $|x_N - x| < \epsilon$; i.e., x_N belongs to $B_\epsilon(x)$ (or $(x - \epsilon, x + \epsilon)$). This implies that $x_N \in S$. But $x_N \in \tilde{S}$, and this is a contradiction. Hence, if S is open, then $\tilde{S}$ is closed.

Conversely, assume that $\tilde{S}$ is closed, and let us show that S must be open. Again we argue by contradiction. Thus, assuming that S is not open, there must exist a point $x \in S$ that is not an interior point of S. Hence, for every $\epsilon > 0$ the disk $B_\epsilon(x)$ (or interval $(x - \epsilon, x + \epsilon)$) is not entirely contained in S. So, for each positive integer n, there must exist a point x_n such that $|x_n - x| < 1/n$ and $x_n \notin S$. It follows then that $x = \lim x_n$, and that each $x_n \in \tilde{S}$. Since $\tilde{S}$ is a closed set, we must have that $x \in \tilde{S}$. But $x \in S$, and we have arrived at the desired contradiction. Hence, if $\tilde{S}$ is closed, then S is open, and the theorem is proved.

The theorem below, the famous Heine-Borel Theorem, gives an equivalent and different description of closed and bounded sets. This description is in terms of open sets, whereas the original definitions were interms of limit points. Any time we can find two very different descriptions of the same phenomenon, we have found something useful.

DEFINITION. Let S be a subset of $\mathbb{C}$ (respectively $\mathbb{R}$). By an *open cover* of S we mean a sequence $\{U_n\}$ of open subsets of $\mathbb{C}$ (respectively $\mathbb{R}$) such that $S \subseteq \cup U_n$; i.e., for every $x \in S$there exists an n such that $x \in U_n$.

A subset S of $\mathbb{C}$ (respectively $\mathbb{R}$) is called *compact*, or is said to satisfy the *Heine-Borel property*, if every open cover of S has a finite subcover. That is, if $\{U_n\}$ is an open cover of S, then there exists an integer N such that $S \subseteq \cup_{n=1}^{N} U_n$. In other words, only a finite number of the open sets are necessary to cover S.

REMARK. The definition we have given here for a set being compact is a little less general from the one found in books on topology. We have restricted the notion of an open cover to be a sequence of open sets, while in the general setting an open cover is just a collection of open sets. The distinction between a sequence of open sets and a collection of open sets is genuine in general topology, but it can be disregarded in the case of the topological spaces $\mathbb{R}$ and $\mathbb{C}$.

THEOREM 2.13. (Heine-Borel Theorem) A subset S of $\mathbb{C}$ (respectively $\mathbb{R}$) is compact if and only if it is a closed and bounded set.

PROOF. We prove this theorem for subsets S of $\mathbb{C}$, and leave the proof for subsets of $\mathbb{R}$ to the exercises.
Suppose first that $S \subseteq \mathbb{C}$ is compact, i.e., satisfies the Heine-Borel property. For each positive integer n, define U_n to be the open set $B_n(0)$. Then $S \subseteq \cup U_n$, because $\mathbb{C} = \cup U_n$. Hence, by the Heine-Borel property, there must exist an N such that $S \subseteq \cup_{n=1}^{N} U_n$. But then $S \subseteq B_N(0)$, implying that S is bounded. Indeed, $|x| \leq N$ for all $x \in S$.
Next, still assuming that S is compact, we will show that S is closed by showing that $\tilde{S}$ is open. Thus, let x be an element of $\tilde{S}$. For each positive integer n, define U_n to be the complement of the closed set $\overline{B_{1/n}(x)}$. Then each U_n is an open set by Theorem 2.12, and we claim that $\{U_n\}$ is an open cover of S. Indeed, if $y \in S$, then $y \neq x$, and $|y - x| > 0$. Choose an n so that $1/n < |y - x|$. Then $y \notin \overline{B_{1/n}(x)}$, implying that $y \in U_n$. This proves our claim that $\{U_n\}$ is an open cover of S. Now, by the Heine-Borel property, there exists an N such that $S \subseteq \cup_{n=1}^{N} U_n$. But this implies that for every $z \in S$ we must have $|z - x| \geq 1/N$, and this implies that the disk $B_{1/N}(x)$ is entirely contained in $\tilde{S}$. Therefore, every element x of $\tilde{S}$ is an interior point of $\tilde{S}$. So, $\tilde{S}$ is open, whence S is closed. This finishes the proof that compact sets are necessarily closed and bounded.
Conversely, assume that S is both closed and bounded. We must show that S satisfies the Heine-Borel property. Suppose not. Then, there exists an open cover $\{U_n\}$ that has no finite subcover. So, for each positive integer n there must exist an element $x_n \in S$ for which $x_n \notin \cup_{k=1}^{n} U_k$. Otherwise, there would be a finite subcover. By Theorem 2.10, there exists a subsequence $\{x_{n_j}\}$ of $\{x_n\}$ that converges to an element x of S. Now, because $\{U_n\}$ is an open cover of S, there must exist an N such that $x \in U_N$. Because U_N is open, there exists an $\epsilon > 0$ so that the entire disk $B_\epsilon(x)$ is contained in U_N. Since $x = \lim x_{n_j}$, there exists a J so that $|x_{n_j} - x| < \epsilon$ if $j \geq J$. Therefore, if $j \geq J$, then $x_{n_j} \in U_N$. But the sequence $\{n_j\}$ is strictly increasing, so that there exists a $j' \geq J$ such that $n_{j'} > N$, and by the choice of the point $x_{n_{j'}}$, we know that $x_{n_{j'}} \notin \cup_{k=1}^{N} U_k$. We have arrived at a contradiction, and so the second half of the theorem is proved.

Exercise 2.24. (a) Prove that the union $A \cup B$ of two open sets is open and the intersection $A \cap B$ is also open.
(b) Prove that the union $A \cup B$ of two closed sets is closed and the intersection $A \cap B$ is also closed.
HINT: Use Theorem 2.12 and the set equations $\widetilde{A \cup B} = \widetilde{A} \cap \widetilde{B}$, and $\widetilde{A \cap B} = \widetilde{A} \cup \widetilde{B}$. These set equations are known as Demorgan's Laws.

(c) Prove that the union $A \cup B$ of two bounded sets is bounded and the intersection $A \cap B$ is also bounded.
(d) Prove that the union $A \cup B$ of two compact sets is compact and the intersection $A \cap B$ is also compact.
(e) Prove that the intersection of a compact set and a closed set is compact.
(f) Suppose S is a compact set in $\mathbb{C}$ and r is a positive real number. Prove that the closed neighborhood $\overline{N}_r(S)$ of radius r around S is compact.
HINT: To see that this set is closed, show that its coplement is open.

INFINITE SERIES

Probably the most interesting and important examples of sequences are those that arise as the partial sums of an infinite series. In fact, it will be infinite series that allow us to explain such things as trigonometric and exponential functions.

DEFINITION. Let $\{a_n\}_0^\infty$ be a sequence of real or complex numbers. By the *infinite series* $\sum a_n$ we mean the sequence $\{S_N\}$ defined by

$$S_N = \sum_{n=0}^{N} a_n.$$

The sequence $\{S_N\}$ is called the *sequence of partial sums* of the infinite series $\sum a_n$, and the infinite series is said to be *summable* to a number S, or to be *convergent,* if the sequence $\{S_N\}$ of partial sums converges to S. **The sum of an infinite series is the limit of its partial sums.**
An infinite series $\sum a_n$ is called *absolutely summable* or *absolutely convergent* if the infinite series $\sum |a_n|$ is convergent.
If $\sum a_n$ is not convergent, it is called *divergent.* If it is convergent but not absolutely convergent, it is called *conditionally convergent.*

A few simple formulas relating the a_n's and the S_N's are useful:

$$S_N = a_0 + a_1 + a_2 + \ldots + a_N,$$

$$S_{N+1} = S_N + a_{N+1},$$

and

$$S_M - S_K = \sum_{n=K+1}^{M} a_n = a_{K+1} + a_{K+2} + \ldots + A_M,$$

for $M > K$.

REMARK. Determining whether or not a given infinite series converges is one of the most important and subtle parts of analysis. Even the first few elementary theorems depend in deep ways on our previous development, particularly the Cauchy criterion.

THEOREM 2.14. *Let $\{a_n\}$ be a sequence of nonnegative real numbers. Then the infinite series $\sum a_n$ is summable if and only if the sequence $\{S_N\}$ of partial sums is bounded.*

PROOF. If $\sum a_n$ is summable, then $\{S_N\}$ is convergent, whence bounded according to Theorem 2.4. Conversely, we see from the hypothesis that each $a_n \geq 0$ that

$\{S_N\}$ is nondecreasing ($S_{N+1} = S_N + a_{N+1} \geq S_N$). So, if $\{S_N\}$ is bounded, then it automatically converges by Theorem 2.1, and hence the infinite series $\sum a_n$ is summable.

The next theorem is the first one most calculus students learn about infinite series. Unfortunately, it is often misinterpreted, so be careful! Both of the proofs to the next two theorems use Theorem 2.9, which again is a serious and fundamental result about the real numbers. Therefore, these two theorems must be deep results themselves.

THEOREM 2.15. *Let $\sum a_n$ be a convergent infinite series. Then the sequence $\{a_n\}$ is convergent, and $\lim a_n = 0$.*

PROOF. Because $\sum a_n$ is summable, the sequence $\{S_N\}$ is convergent and so is a Cauchy sequence. Therefore, given an $\epsilon > 0$, there exists an N_0 so that $|S_n - S_m| < \epsilon$ whenever both n and $m \geq N_0$. If $n > N_0$, let $m = n - 1$. We have then that $|a_n| = |S_n - S_m| < \epsilon$, which completes the proof.

REMARK. Note that this theorem is **not** an "if and only if" theorem. The harmonic series (part (b) of Exercise 2.26 below) is the standard counterexample. The theorem above is mainly used to show that an infinite series is **not** summable. If we can prove that the sequence $\{a_n\}$ does not converge to 0, then the infinite series $\sum a_n$ does not converge. The misinterpretation of this result referred to above is exactly in trying to apply the (false) converse of this theorem.

THEOREM 2.16. *If $\sum a_n$ is an absolutely convergent infinite series of complex numbers, then it is a convergent infinite series. (Absolute convergence implies convergence.)*

PROOF. If $\{S_N\}$ denotes the sequence of partial sums for $\sum a_n$, and if $\{T_N\}$ denotes the sequence of partial sums for $\sum |a_n|$, then

$$|S_M - S_N| = |\sum_{n=N+1}^{M} a_n| \leq \sum_{n=N+1}^{M} |a_n| = |T_M - T_N|$$

for all N and M. We are given that $\{T_N\}$ is convergent and hence it is a Cauchy sequence. So, by the inequality above, $\{S_N\}$ must also be a Cauchy sequence. (If $|T_N - T_M| < \epsilon$, then $|S_N - S_M| < \epsilon$ as well.) This implies that $\sum a_n$ is convergent.

Exercise 2.25. (The Infinite Geometric Series) Let z be a complex number, and define a sequence $\{a_n\}$ by $a_n = z^n$. Consider the infinite series $\sum a_n$. Show that $\sum_{n=0}^{\infty} a_n$ converges to a number S if and only if $|z| < 1$. Show in fact that $S = 1/(1-z)$, when $|z| < 1$.
HINT: Evaluate explicitly the partial sums S_N, and then take their limit. Show that $S_N = \frac{1-z^{N+1}}{1-z}$.

Exercise 2.26. (a) Show that $\sum_{n=1}^{\infty} \frac{1}{n(n+1)}$ converges to 1, by computing explicit formulas for the partial sums.
HINT: Use a partial fraction decomposition for the a_n's.
(b) (The Harmonic Series.) Show that $\sum_{n=1}^{\infty} 1/n$ diverges by verifying that $S_{2^k} > k/2$.

HINT: Group the terms in the sum as follows,

$$1+\frac{1}{2}+(\frac{1}{3}+\frac{1}{4})+(\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8})+(\frac{1}{9}+\frac{1}{10}+\ldots+\frac{1}{16})+\ldots,$$

and then estimate the sum of each group. Remember this example as an infinite series that diverges, despite the fact that is terms tend to 0.

The next theorem is the most important one we have concerning infinite series of numbers.

THEOREM 2.17. (Comparison Test) Suppose $\{a_n\}$ and $\{b_n\}$ are two sequences of nonnegative real numbers for which there exists a positive integer M and a constant C such that $b_n \leq Ca_n$ for all $n \geq M$. If the infinite series $\sum a_n$ converges, so must the infinite series $\sum b_n$.

PROOF. We will show that the sequence $\{T_N\}$ of partial sums of the infinite series $\sum b_n$ is a bounded sequence. Then, by Theorem 2.14, the infinite series $\sum b_n$ must be summable.
Write S_N for the Nth partial sum of the convergent infinite series $\sum a_n$. Because this series is summable, its sequence of partial sums is a bounded sequence. Let B be a number such that $S_N \leq B$ for all N. We have for all $N > M$ that

$$\begin{aligned} T_N &= \sum_{n=1}^{N} b_n \\ &= \sum_{n=1}^{M} b_n + \sum_{n=M+1}^{N} b_n \\ &\leq \sum_{n=1}^{M} b_n + \sum_{n=M+1}^{N} Ca_n \\ &= \sum_{n=1}^{M} b_n + C\sum_{n=M+1}^{N} a_n \\ &\leq \sum_{n=1}^{M} b_n + C\sum_{n=1}^{N} a_n \\ &\leq \sum_{n=1}^{M} b_n + CS_N \\ &\leq \sum_{n=1}^{M} b_n + CB, \end{aligned}$$

which completes the proof, since this final quantity is a fixed constant.

Exercise 2.27. (a) Let $\{a_n\}$ and $\{b_n\}$ be as in the preceding theorem. Show that if $\sum b_n$ diverges, then $\sum a_n$ also must diverge.
(b) Show by example that the hypothesis that the a_n's and b_n's of the Comparison Test are nonnegative can not be dropped.

Exercise 2.28. (The Ratio Test) Let $\{a_n\}$ be a sequence of positive numbers.

(a) If $\limsup a_{n+1}/a_n < 1$, show that $\sum a_n$ converges.
HINT: If $\limsup a_{n+1}/a_n = \alpha < 1$, let β be a number for which $\alpha < \beta < 1$. Using part (a) of Exercise 2.17, show that there exists an N such that for all $n > N$ we must have $a_{n+1}/a_n < \beta$, or equivalently $a_{n+1} < \beta a_n$, and therefore $a_{N+k} < \beta^k a_N$. Now use the comparison test with the geometric series $\sum \beta^k$.
(b) If $\liminf a_{n+1}/a_n > 1$, show that $\sum a_n$ diverges.
(c) As special cases of parts (a) and (b), show that $\{a_n\}$ converges if $\lim_n a_{n+1}/a_n < 1$, and diverges if $\lim_n a_{n+1}/a_n > 1$.
(d) Find two examples of infinite series' $\sum a_n$ of positive numbers, such that $\lim a_{n+1}/a_n = 1$ for both examples, and such that one infinite series converges and the other diverges.

Exercise 2.29. (a) Derive the Root Test: If $\{a_n\}$ is a sequence of positive numbers for which $\limsup a_n^{1/n} < 1$, then $\sum a_n$ converges. And, if $\liminf a_n^{1/n} > 1$, then $\sum a_n$ diverges.
(b) Let r be a positive integer. Show that $\sum 1/n^r$ converges if and only if $r \geq 2$.
HINT: Use Exercise 2.26 and the Comparison Test for $r = 2$.
(c) Show that the following infinite series are summable.

$$\sum 1/(n^2+1),\ \sum n/2^n,\ \sum a^n/n!,$$

for a any complex number.

Exercise 2.30. Let $\{a_n\}$ and $\{b_n\}$ be sequences of complex numbers, and let $\{S_N\}$ denote the sequence of partial sums of the infinite series $\sum a_n$. Derive the Abel Summation Formula:

$$\sum_{n=1}^{N} a_n b_n = S_N b_N + \sum_{n=1}^{N-1} S_n (b_n - b_{n+1}).$$

The Comparison Test is the most powerful theorem we have about infinite series of positive terms. Of course, most series do not consist entirely of positive terms, so that the Comparison Test is not enough. The next theorem is therefore of much importance.

THEOREM 2.18. (Alternating Series Test) Suppose $\{a_1, a_2, a_3, \dots\}$ is an alternating sequence of real numbers; i.e., their signs alternate. Assume further that the sequence $\{|a_n|\}$ is nonincreasing with $0 = \lim |a_n|$. Then the infinite series $\sum a_n$ converges.

PROOF. Assume, without loss of generality, that the odd terms a_{2n+1} of the sequence $\{a_n\}$ are positive and the even terms a_{2n} are negative. We collect some facts about the partial sums $S_N = a_1 + a_2 + \ldots + a_N$ of the infinite series $\sum a_n$.
1. Every even partial sum S_{2N} is less than the following odd partial sum $S_{2N+1} = S_{2N} + a_{2N+1}$, And every odd partial sum S_{2N+1} is greater than the following even partial sum $S_{2N+2} = S_{2N+1} + a_{2N+2}$.
2. Every even partial sum S_{2N} is less than or equal to the next even partial sum $S_{2N+2} = S_{2N} + a_{2N+1} + a_{2N+2}$, implying that the sequence of even partial sums $\{S_{2N}\}$ is nondecreasing.

3. Every odd partial sum S_{2N+1} is greater than or equal to the next odd partial sum $S_{2N+3} = S_{2N+1} + a_{2N+2} + a_{2N+3}$, implying that the sequence of odd partial sums $\{S_{2N+1}\}$ is nonincreasing.
4. Every odd partial sum S_{2N+1} is bounded below by S_2. For, $S_{2N+1} > S_{2N} \geq S_2$. And, every even partial sum S_{2N} is bounded above by S_1. For, $S_{2N} < S_{2N+1} \leq S_1$.
5. Therefore, the sequence $\{S_{2N}\}$ of even partial sums is nondecreasing and bounded above. That sequence must then have a limit, which we denote by S_e. Similarly, the sequence $\{S_{2N+1}\}$ of odd partial sums is nonincreasing and bounded below. This sequence of partial sums also must have a limit, which we denote by S_o.
Now

$$S_o - S_e = \lim S_{2N+1} - \lim S_{2N} = \lim(S_{2N+1} - S_{2N}) = \lim a_{2N+1} = 0,$$

showing that $S_e = S_o$, and we denote this common limit by S. Finally, given an $\epsilon > 0$, there exists an N_1 so that $|S_{2N} - S| < \epsilon$ if $2N \geq N_1$, and there exists an N_2 so that $|S_{2N+1} - S| < \epsilon$ if $2N + 1 \geq N_2$. Therefore, if $N \geq \max(N_1, N_2)$, then $|S_N - S| < \epsilon$, and this proves that the infinite series converges.

Exercise 2.31. (a) (The Alternating Harmonic Series) Show that $\sum_{n=1}^{\infty}(-1)^n/n$ converges, but that it is not absolutely convergent.
(b) Let $\{a_n\}$ be an alternating series, as in the preceding theorem. Show that the sum $S = \sum a_n$ is trapped between S_N and S_{N+1}, and that $|S - S_N| \leq |a_N|$.
(c) State and prove a theorem about "eventually alternating infinite series."
(d) Show that $\sum z^n/n$ converges if and only if $|z| \leq 1$, and $z \neq 1$.
HINT: Use the Abel Summation Formula to evaluate the partial sums.

Exercise 2.32. Let $s = p/q$ be a positive rational number.
(a) For each $x > 0$, show that there exists a unique $y > 0$ such that $y^s = x$; i.e., $y^p = x^q$.
(b) Prove that $\sum 1/n^s$ converges if $s > 1$ and diverges if $s \leq 1$.
HINT: Group the terms as in part (b) of Exercise 2.26.

THEOREM 2.19. (Test for Irrationality) Let x be a real number, and suppose that $\{p_N/q_N\}$ is a sequence of rational numbers for which $x = \lim p_N/q_N$ and $x \neq p_N/q_N$ for any N. If $\lim q_N|x - p_N/q_N| = 0$, then x is irrational.

PROOF. We prove the contrapositive statement; i.e., if $x = p/q$ is a rational number, then $\lim q_N|x - p_N/q_N| \neq 0$. We have

$$x - p_N/q_N = p/q - p_N/q_N = \frac{pq_N - qp_N}{qq_N}.$$

Now the numerator $pq_N - qp_N$ is not 0 for any N. For, if it were, then $x = p/q = p_N/q_N$, which we have assumed not to be the case. Therefore, since $pq_N - qp_N$ is an integer, we have that

$$|x - p_N/q_N| = |\frac{pq_N - qp_N}{qq_N}| \geq \frac{1}{|qq_N|}.$$

So,

$$q_N|x - p_N/q_N| \geq \frac{1}{|q|},$$

and this clearly does not converge to 0.

Exercise 2.33. (a) Let $x = \sum_{n=0}^{\infty}(-1)^n/2^n$. Prove that x is a rational number.
(b) Let $y = \sum_{n=0}^{\infty}(-1)^n/2^{n^2}$. Prove that y is an irrational number.
HINT: The partial sums of this series are rational numbers. Now use the preceding theorem and part (b) of Exercise 2.31.

CHAPTER III
FUNCTIONS AND CONTINUITY
DEFINITION OF THE NUMBER π.

The concept of a function is perhaps the most basic one in mathematical analysis. The objects of interest in our subject can often be represented as functions, and the " unknowns" in our equations are frequently functions. Therefore, we will spend some time developing and understanding various kinds of functions, including functions defined by polynomials, by power series, and as limits of other functions. In particular, we introduce in this chapter the elementary transcendental functions. We begin with the familiar set theoretical notion of a function, and then move quickly to their analytical properties, specifically that of continuity.

The main theorems of this chapter include:

1. The **Intermediate Value Theorem** (Theorem 3.6),
2. the theorem that asserts that a **continuous real-valued function on a compact set attains a maximum and minimum value** (Theorem 3.8),
3. A **continuous function on a compact set is uniformly continuous** (Theorem 3.9),
4. The **Identity Theorem for Power Series Functions** (Theorem 3.14),
5. The definition of the real number π,
6. The theorem that asserts that the **uniform limit of a sequence of continuous functions is continuous** (Theorem 3.17), and
7. the **Weierstrass M-Test** (Theorem 3.18).

FUNCTIONS

DEFINITION. Let S and T be sets. A *function* from S into T (notation $f : S \to T$) is a rule that assigns to each element x in S a unique element denoted by $f(x)$ in T.

It is useful to think of a function as a mechanism or black box. We use the elements of S as inputs to the function, and the outputs are elements of the set T.

If $f : S \to T$ is a function, then S is called the *domain* of f, and the set T is called the *codomain* of f. The *range* or *image* of f is the set of all elements y in the codomain T for which there exists an x in the domain S such that $y = f(x)$. We denote the range by $f(S)$. The codomain is the set of all potential outputs, while the range is the set of actual outputs.

Suppose f is a function from a set S into a set T. If $A \subseteq S$, we write $f(A)$ for the subset of T containing all the elements $t \in T$ for which there exists an $s \in A$ such that $t = f(s)$. We call $f(A)$ the *image* of A under f. Similarly, if $B \subseteq T$, we write $f^{-1}(B)$ for the subset of S containing all the elements $s \in S$ such that $f(s) \in B$, and we call the set $f^{-1}(B)$ the *inverse image* or *preimage* of B. The symbol $f^{-1}(B)$ is a little confusing, since it could be misinterpreted as the image of the set B under a function called f^{-1}. We will discuss inverse functions later on, but this notation is not meant to imply that the function f has an inverse.

If $f : S \to T$, then the *graph* of f is the subset G of the Cartesian product $S \times T$ consisting of all the pairs of the form $(x, f(x))$.

If $f : S \to \mathbb{R}$ is a function, then we call f a *real-valued* function, and if $f : S \to \mathbb{C}$, then we call f a *complex-valued* function. If $f : S \to \mathbb{C}$ is a complex-valued function, then for each $x \in S$ the complex number $f(x)$ can be written as $u(x) + iv(x)$, where

$u(x)$ and $v(x)$ are the real and imaginary parts of the complex number $f(x)$. The two real-valued functions $u : S \to \mathbb{R}$ and $v : S \to \mathbb{R}$ are called respectively the *real* and *imaginary* parts of the complex-valued function f.
If $f : S \to T$ and $S \subseteq \mathbb{R}$, then f is called a function of a *real variable*, and if $S \subseteq \mathbb{C}$, then f is called a function of a *complex variable*.
If the range of f equals the codomain, then f is called *onto*.
The function $f : S \to T$ is called one-to-one if $f(x_1) = f(x_2)$ implies that $x_1 = x_2$.

The domain of f is the set of x's for which $f(x)$ is defined. If we are given a function $f : S \to T$, we are free to regard f as having a smaller domain, i.e., a subset S' of S. Although this restricted function is in reality a different function, we usually continue to call it by the same name f. Enlarging the domain of a function, in some consistent manner, is often impossible, but is nevertheless frequently of great importance. The codomain of f is distinguished from the range of f, which is frequently a proper subset of the codomain. For example, since every real number is a complex number, any real-valued function $f : S \to \mathbb{R}$ is also a (special kind of) complex-valued function.
We consider in this book functions either of a real variable or of complex variable. that is, the domains of functions here will be subsets either of $\mathbb{R}$ or of $\mathbb{C}$. Frequently, we will indicate what kind of variable we are thinking of by denoting real variables with the letter x and complex variables with the letter z. Be careful about this, for this distinction is **not always** made.
Many functions, though not all by any means, are defined by a single equation:

$$y = 3x - 7,$$

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

$$x^2 + y^2 = 4,$$

(How does this last equation define a function?)

$$(1 - x^7 y^{11})^{2/3} = (x/(1 - y))^{8/17}.$$

(How does this equation determine a function?)
There are various types of functions, and they can be combined in a variety of ways to produce other functions. It is necessary therefore to spend a fair amount of time at the beginning of this chapter to present these definitions.

DEFINITION. If f and g are two complex-valued functions with the same domain S, i.e., $f : S \to \mathbb{C}$ and $g : S \to \mathbb{C}$, and if c is a complex number, we define $f + g,\ fg,\ f/g$ (if $g(x)$ is never 0), and cf by the familiar formulas:

$$(f + g)(x) = f(x) + g(x),$$

$$(fg)(x) = f(x)g(x),$$

$$(f/g)(x) = f(x)/g(x),$$

and

$$(cf)(x) = cf(x).$$

If f and g are real-valued functions, we define functions $\max(f,g)$ and $\min(f,g)$ by

$$[\max(f,g)](x)=\max(f(x),g(x))$$

(the maximum of the numbers $f(x)$ and $g(x)$), and

$$[\min(f,g)](x)=\min(f(x),g(x)),$$

(the minimum of the two numbers $f(x)$ and $g(x)$).
If f is either a real-valued or a complex-valued function on a domain S, then we say that f is *bounded* if there exists a positive number M such that $|f(x)|\leq M$ for all $x\in S$.

There are two special types of functions of a real or complex variable, the even functions and the odd functions. In fact, every function that is defined on all of $\mathbb{R}$ or $\mathbb{C}$ (or, more generally, any function whose domain S equals $-S$) can be written uniquely as a sum of an even part and an odd part. This decomposition of a general function into simpler parts is frequently helpful.

DEFINITION. A function f whose domain S equals $-S$, is called an *even* function if $f(-z)=f(z)$ for all z in its domain. It is called an *odd* function if $f(-z)=-f(z)$ for all z in its domain.

We next give the definition for perhaps the most familiar kinds of functions.

DEFINITION. A nonzero *polynomial* or *polynomial function* is a complex-valued function of a complex variable, $p:\mathbb{C}\to\mathbb{C}$, that is defined by a formula of the form

$$p(z)=\sum_{k=0}^{n}a_kz^k=a_0+a_1z+a_2z^2+\ldots+a_nz^n,$$

where the a_k's are complex numbers and $a_n\neq 0$. The integer n is called the *degree* of the polynomial p and is denoted by $\deg(p)$. The numbers $a_0,a_1,\ldots,a_n$ are called the *coefficients* of the polynomial. The domain of a polynomial function is all of $\mathbb{C}$; i.e., $p(z)$ is defined for every complex number z.
For technical reasons of consistency, the identically 0 function is called the *zero polynomial*. All of its coefficients are 0 and its degree is defined to be $-\infty$.
A *rational function* is a function r that is given by an equation of the form $r(z)=p(z)/q(z)$, where q is a nonzero polynomial and p is a (possibly zero) polynomial. The domain of a rational function is the set S of all $z\in\mathbb{C}$ for which $q(z)\neq 0$, i.e., for which $r(z)$ is defined.

Two other kinds of functions that are simple and important are step functions and polygonal functions.

DEFINITION. Let $[a,b]$ be a closed bounded interval of real numbers. By a *partition* of $[a,b]$ we mean a finite set $P=\{x_0<x_1<\ldots<x_n\}$ of $n+1$ points, where $x_0=a$ and $x_n=b$.
The n intervals $\{[x_{i-1},x_i]\}$, for $1\leq i\leq n$, are called the *closed subintervals* of the partition P, and the n intervals $\{(x_{i-1},x_i)\}$ are called the *open subintervals* of P.
We write $\|P\|$ for the maximum of the numbers (lengths of the subintervals) $\{x_i-x_{i-1}\}$, and call the number $\|P\|$ the *mesh size* of the partition P.

A function $h : [a,b] \to \mathbb{C}$ is called a *step function* if there exists a partition $P = \{x_0 < x_1 < \ldots < x_n\}$ of $[a,b]$ and n numbers $\{a_1, a_2, \ldots, a_n\}$ such that $h(x) = a_i$ if $x_{i-1} < x < x_i$. That is, h is a step function if it is a constant function on each of the (open) subintervals (x_{i-1}, x_i) determined by a partition P. Note that the values of a step function at the points $\{x_i\}$ of the partition are not restricted in any way.
A function $l : [a,b] \to \mathbb{R}$ is called a *polygonal function*, or a *piecewise linear function*, if there exists a partition $P = \{x_0 < x_1 < \ldots < x_n\}$ of $[a,b]$ and $n+1$ numbers $\{y_0, y_1, \ldots, y_n\}$ such that for each $x \in [x_{i-1}, x_i]$, $l(x)$ is given by the linear equation

$$l(x) = y_{i-1} + m_i(x - x_{i-1}),$$

where $m_i = (y_i - y_{i_1})/(x_i - x_{i-1})$. That is, l is a polygonal function if it is a linear function on each of the closed subintervals $[x_{i-1}, x_i]$ determined by a partition P. Note that the values of a piecewise linear function at the points $\{x_i\}$ of the partition P are the same, whether we think of x_i in the interval $[x_{i-1}, x_i]$ or $[x_i, x_{i+1}]$. (Check the two formulas for $l(x_i)$.)
The graph of a piecewise linear function is the polygonal line joining the $n+1$ points $\{(x_i, y_i)\}$.

There is a natural generalization of the notion of a step function that works for any domain S, e.g., a rectangle in the plane $\mathbb{C}$. Thus, if S is a set, we define a *partition* of S to be a finite collection $\{E_1, E_2, \ldots, E_n\}$ of subsets of S for which

(1) $\cup_{i=1}^n E_i = S$, and
(2) $E_i \cap E_j = \emptyset$ if $i \neq j$.

Then, a *step function* on S would be a function h that is constant on each subset E_i. We will encounter an even more elaborate generalized notion of a step function in Chapter V, but for now we will restrict our attention to step functions defined on intervals $[a,b]$.
The set of polynomials and the set of step functions are both closed under addition and multiplication, and the set of rational functions is closed under addition, multiplication, and division.

Exercise 3.1. (a) Prove that the sum and product of two polynomials is again a polynomial. Show that $\deg(p+q) \leq \max(\deg(p), \deg(q))$ and $\deg(pq) = \deg(p) + \deg(q)$. Show that a constant function is a polynomial, and that the degree of a nonzero constant function is 0.
(b) Show that the set of step functions is closed under addition and multiplication. Show also that the maximum and minimum of two step functions is again a step function. (Be careful to note that different step functions may be determined by different partitions. For instance, a partition determining the sum of two step functions may be different from the partitions determining the two individual step functions.) Note, in fact, that a step function can be determined by infinitely many different partitions. Prove that the sum, the maximum, and the minimum of two piecewise linear functions is again a piecewise linear function. Show by example that the product of two piecewise linear functions need not be piecewise linear.
(c) Prove that the sum, product, and quotient of two rational functions is again a rational function.
(d) Prove the **Root Theorem:** If $p(z) = \sum_{k=0}^n a_k z^k$ is a nonzero polynomial of degree n, and if c is a complex number for which $p(c) = 0$, then there exists a nonzero polynomial $q(z) = \sum_{j=0}^{n-1} b_j z^j$ of degree $n-1$ such that $p(z) = (z-c)q(z)$

for all z. That is, if c is a "root" of p, then $z-c$ is a factor of p. Show also that the leading coefficient b_{n-1} of q equals the leading coefficient a_n of p.
HINT: Write

$$p(z)=p(z)-p(c)=\sum_{k=0}^{n}a_k(z^k-c^k)=\dots.$$

(e) Let f be a function whose domain S equals $-S$. Define functions f_e and f_o by the formulas

$$f_e(z)=\frac{f(z)+f(-z)}{2}\text{ and }f_o(z)=\frac{f(z)-f(-z)}{2}.$$

Show that f_e is an even function, that f_o is an odd function, and that $f=f_e+f_o$. Show also that, if $f=g+h$, where g is an even function and h is an odd function, then $g=f_e$ and $h=f_o$. That is, there is only one way to write f as the sum of an even function and an odd function.
(f) Use part (e) to show that a polynomial p is an even function if and only if its only nonzero coefficients are even ones, i.e., the a_{2k}'s. Show also that a polynomial is an odd function if and only if its only nonzero coefficients are odd ones, i.e., the a_{2k+1}'s.
(g) Suppose $p(z)=\sum_{k=0}^{n}a_{2k}z^{2k}$ is a polynomial that is an even function. Show that

$$p(iz)=\sum_{k=0}^{n}(-1)^k a_{2k}z^{2k}=p^a(z),$$

where p^a is the polynomial obtained from p by alternating the signs of its nonzero coefficients.
(h) If $q(z)=\sum_{k=0}^{n}a_{2k+1}z^{2k+1}$ is a polynomial that is an odd function, show that

$$q(iz)=i\sum_{k=0}^{n}(-1)^k a_{2k+1}z^{2k+1}=iq^a(z),$$

where again q^a is the polynomial obtained from q by alternating the signs of its nonzero coefficients.
(i) If p is any polynomial, show that

$$p(iz)=p_e(iz)+p_o(iz)=p_e^a(z)+ip_o^a(z),$$

and hence that $p_e(iz)=p_e^a(z)$ and $p_o(iz)=ip_o^a(z)$.

POLYNOMIAL FUNCTIONS

If $p(z)=\sum_{k=0}^{n}a_kz^k$ and $q(z)=\sum_{j=0}^{m}b_jz^j$ are two polynomials, it certainly seems clear that they determine the same function only if they have identical coefficients. This is true, but by no means an obvious fact. Also, it seems clear that, as $|z|$ gets larger and larger, a polynomial function is more and more comparable to its leading term a_nz^n. We collect in the next theorem some elementary properties of polynomial functions, and in particular we verify the above "uniqueness of coefficients" result and the "behavior at infinity" result.

THEOREM 3.1.

(1) *Suppose $p(z) = \sum_{k=0}^n a_k z^k$ is a nonconstant polynomial of degree $n > 0$. Then $p(z) = 0$ for at most n distinct complex numbers.*

(2) *If r is a polynomial for which $r(z) = 0$ for an infinite number of distinct points, then r is the zero polynomial. That is, all of its coefficients are 0.*

(3) *Suppose p and q are nonzero polynomials, and assume that $p(z) = q(z)$ for an infinite number of distinct points. Then $p(z) = q(z)$ for all z, and p and q have the same coefficients. That is, they are the same polynomial.*

(4) *Let $p(z) = \sum_{j=0}^n c_j z^j$ be a polynomial of degree $n > 0$. Then there exist positive constants m and B such that*

$$\frac{|c_n|}{2}|z|^n \leq |p(z)| \leq M|z|^n$$

for all complex numbers z for which $|z| \geq B$. That is, For all complex numbers z with $|z| \geq B$, the numbers $|p(z)|$ and $|z|^n$ are "comparable."

(5) *If $f : [0, \infty) \to \mathbb{C}$ is defined by $f(x) = \sqrt{x}$, then there is no polynomial p for which $f(x) = p(x)$ for all $x \geq 0$. That is, the square root function does not agree with any polynomial function.*

PROOF. We prove part (1) using an argument by contradiction. Thus, suppose there does exist a counterexample to the claim, i.e., a nonzero polynomial p of degree n and $n+1$ distinct points $\{c_1, c_2, \dots, c_{n+1}\}$ for which $p(c_j) = 0$ for all $1 \leq j \leq n+1$. From the set of all such counterexamples, let p_0 be one with minimum degree n_0. That is, the claim in part (1) is true for any polynomial whose degree is smaller than n_0. We write

$$p_0(z) = \sum_{k=0}^{n_0} a_k z^k,$$

and we suppose that $p_0(c_j) = 0$ for $j = 1$ to $n_0 + 1$, where these c_k's are distinct complex numbers. We use next the Root Theorem (part (d) of Exercise 3.1) to write $p_0(z) = (z - c_{n_0+1})q(z)$, where $q(z) = \sum_{k=0}^{n_0-1} b_k z^k$. We have that q is a polynomial of degree $n_0 - 1$ and the leading coefficient a_{n_0} of p_0 equals the leading coefficient b_{n_0-1} of q. Note that for $1 \leq j \leq n_0$ we have

$$0 = p_0(c_j) = (c_j - c_{n_0+1})q(c_j),$$

which implies that $q(c_j) = 0$ for $1 \leq j \leq n_0$, since $c_j - c_{n_0+1} \neq 0$. But, since $\deg(q) < n_0$, the nonzero polynomial q can not be a counterexample to part (1), implying that $q(z) = 0$ for at most $n_0 - 1$ distinct points. We have arrived at a contradiction, and part (1) is proved.

Next, let r be a polynomial for which $r(z) = 0$ for an infinite number of distinct points. It follows from part (1) that r cannot be a nonzero polynomial, for in that case it would have a degree $n \geq 0$ and could be 0 for at most n distinct points. Hence, r is the zero polynomial, and part (2) is proved.

Now, to see part (3), set $r = p - q$. Then r is a polynomial for which $r(z) = 0$ for infinitely many z's. By part (2), it follows then that $r(z) = 0$ for all z, whence $p(z) = q(z)$ for all z. Moreover, $p - q$ is the zero polynomial, all of whose coefficients are 0, and this implies that the coefficients for p and q are identical.

To prove the first inequality in part (4), suppose that $|z|>1$, and from the backwards triangle inequality, note that

$$\begin{aligned}
|p(z)| &= |\sum_{k=0}^{n} c_k z^k| \\
&= |z|^n |\sum_{k=0}^{n} \frac{c_k}{z^{n-k}}| \\
&= |z|^n |(\sum_{k=0}^{n-1} \frac{c_k}{z^{n-k}}) + c_n| \\
&\geq |z|^n (|c_n| - |\sum_{k=0}^{n-1} \frac{c_k}{z^{n-k}}|) \\
&\geq |z|^n (|c_n| - \sum_{k=0}^{n-1} \frac{|c_k|}{|z|^{n-k}}) \\
&\geq |z|^n (|c_n| - \sum_{k=0}^{n-1} \frac{|c_k|}{|z|}) \\
&\geq |z|^n (|c_n| - \frac{1}{|z|} \sum_{k=0}^{n-1} |c_k|).
\end{aligned}$$

Set B equal to the constant $(2/|c_n|)\sum_{j=0}^{n-1}|c_j|$. Then, replacing the $1/|z|$ in the preceding calculation by $1/B$, we obtain

$$|p(z)| \geq m|z|^n$$

for every z for which $|z| \geq B$. This proves the first half of part (4).
To get the other half of part (4), suppose again that $|z|>1$. We have

$$|p(z)| \leq \sum_{k=0}^{n} |c_k||z|^k \leq \sum_{k=0}^{n} |c_k||z|^n,$$

so that we get the other half of part (4) by setting $M=\sum_{k=0}^{n}|c_k|$.
Finally, to see part (5), suppose that there does exist a polynomial p of degree n such that $\sqrt{x}=p(x)$ for all $x \geq 0$. Then $x=(p(x))^2$ for all $x \geq 0$. Now p^2 is a polynomial of degree $2n$. By part (2), the two polynomials $q(x)=x$ and $(p(x))^2$ must be the same, implying that they have the same degree. However, the degree of q is 1, which is odd, and the degree of p^2 is $2n$, which is even. Hence, we have arrived at a contradiction.

Exercise 3.2. (a) Let $r(z)=p(z)/q(z)$ and $r'(z)=p'(z)/q'(z)$ be two rational functions. Suppose $r(z)=r'(z)$ for infinitely many z's. Prove that $r(z)=r'(z)$ for all z in the intersection of their domains. Is it true that $p=p'$ and $q=q'$?
(b) Let p and q be polynomials of degree n and m respectively, and define a rational function r by $r=p/q$. Prove that there exist positive constants C and B such that $|r(z)|<C|z|^{n-m}$ for all complex numbers z for which $|z|>B$.

(c) Define $f : [0, \infty) \to \mathbb{R}$ by $f(x) = \sqrt{x}$. Show that there is no rational function r such that $f(x) = r(x)$ for all $x \geq 0$. That is, the square root function does not agree with a rational function.
(d) Define the real-valued function r on $\mathbb{R}$ by $r(x) = 1/(1 + x^2)$. Prove that there is no polynomial p such that $p(x) = r(x)$ for infinitely many real numbers x.
(e) If f is the real-valued function of a real variable given by $f(x) = |x|$, show that f is **not** a rational function.
HINT: Suppose $|x| = p(x)/q(x)$. Then $|x|q(x) = p(x)$ implying that $|x|q(x)$ is a polynomial $s(x)$. Now use Theorem 3.1 to conclude that $p(x) = xq(x)$ for all x and that $p(x) = -xq(x)$ for all x.
(f) Let f be any complex-valued function of a complex variable, and let $c_1, \ldots, c_n$ be n distinct complex numbers that belong to the domain of f. Show that there does exist a polynomial p of degree n such that $p(c_j) = f(c_j)$ for all $1 \leq j \leq n$.
HINT: Describe p in factored form.
(g) Give examples to show that the maximum and minimum of two polynomials need not be a polynomial or even a rational function.

Very important is the definition of the *composition* $g \circ f$ of two functions f and g.

DEFINITION. Let $f : S \to T$ and $g : T \to U$ be functions. We define a function $g \circ f$, with domain S and codomain U, by $(g \circ f)(x) = g(f(x))$.
If $f : S \to T$, $g : T \to S$, and $g \circ f(x) = x$ for all $x \in S$, then g is called a *left inverse* of f. If $f \circ g(y) = y$ for all $y \in T$, then g is called a *right inverse* for f. If g is both a left inverse and a right inverse, then g is called an *inverse* for f, f is called *invertible*, and we denote g by f^{-1}.

Exercise 3.3. (a) Suppose $f : S \to T$ has a left inverse. Prove that f is 1-1.
(b) Suppose $f : S \to T$ has a right inverse. Prove that f is onto.
(c) Show that the composition of two polynomials is a polynomial and that the composition of two rational functions is a rational function.
HINT: If p is a polynomial, show by induction that p^n is a polynomial. Now use Exercise 3.1.
(d) Find formulas for $g \circ f$ and $f \circ g$ for the following. What are the domains of these compositions?

(1) (i) $f(x) = 1 + x^2$ and $g(x) = 1/(1 + x)^{1/2}$.
(2) (ii) $f(x) = x/(x + 1)$ and $g(x) = x/(1 - x)$.
(3) (iii) $f(x) = ax + b$ and $g(x) = cx + d$.

CONTINUITY

Next, we come to the definition of continuity. Unlike the preceding discussion, which can be viewed as being related primarily to the algebraic properties of functions, this one is an analytic notion.

DEFINITION. Let S and T be sets of complex numbers, and let $f : S \to T$. Then f is said to be *continuous at a point* c of S if for every positive ϵ, there exists a positive δ such that if $x \in S$ satisfies $|x - c| < \delta$, then $|f(x) - f(c)| < \epsilon$. The function f is called *continuous on* S if it is continuous at every point c of S.
If the domain S of f consists of real numbers, then the function f is called *right continuous at* c if for every $\epsilon > 0$ there exists a $\delta > 0$ such that $|f(x) - f(c)| < \epsilon$

whenever $x \in S$ and $0 \leq x-c < \delta$, and is called *left continuous* at c if for every $\epsilon > 0$ there exists a $\delta > 0$ such that $|f(x)-f(c)| < \epsilon$ whenever $x \in S$ and $0 \geq x-c > -\delta$.

REMARK. If f is continuous at a point c, then the positive number δ of the preceding definition is not unique (any smaller number would work as well), but it does depend both on the number ϵ and on the point c. Sometimes we will write $\delta(\epsilon, c)$ to make this dependence explicit. Later, we will introduce a notion of uniform continuity in which δ only depends on the number ϵ and not on the particular point c.

The next theorem indicates the interaction between the algebraic properties of functions and continuity.

THEOREM 3.2. *Let S and T be subsets of $\mathbb{C}$, let f and g be functions from S into T, and suppose that f and g are both continuous at a point c of S. Then*

1. *There exists a $\delta > 0$ and a positive number M such that if $|y-c| < \delta$ and $y \in S$ then $|f(y)| \leq M$. That is, if f is continuous at c, then it is bounded near c.*
2. *$f+g$ is continuous at c.*
3. *fg is continuous at c.*
4. *$|f|$ is continuous at c.*
5. *If $g(c) \neq 0$, then f/g is continuous at c.*
6. *If f is a complex-valued function, and u and v are the real and imaginary parts of f, then f is continuous at c if and only if u and v are continuous at c.*

PROOF. We prove parts (1) and (5), and leave the remaining parts to the exercise that follows.

To see part (1), let $\epsilon = 1$. Then, since f is continuous at c, there exists a $\delta > 0$ such that if $|y-c| < \delta$ and $y \in S$ then $|f(y)-f(c)| < 1$. Since $|z-w| \geq ||z|-|w||$ for any two complex numbers z and w (backwards Triangle Inequality), it then follows that $||f(y)|-|f(c)|| < 1$, from which it follows that if $|y-c| < \delta$ then $|f(y)| < |f(c)|+1$. Hence, setting $M = |f(c)|+1$, we have that if $|y-c| < \delta$ and $y \in S$, then $|f(y)| \leq M$ as desired.

To prove part (5), we first make use of part 1. Let δ_1, M_1 and δ_2, M_2 be chosen so that if $|y-c| < \delta_1$ and $y \in S$ then

$$|f(y)| < M_1 \tag{3.1}$$

and if $|y-c| < \delta_2$ and $y \in S$ then

(3.2). $$|g(y)| < M_2$$

Next, let ϵ' be the positive number $|g(c)|/2$. Then, there exists a $\delta' > 0$ such that if $|y-c| < \delta'$ and $y \in S$ then $|g(y)-g(c)| < \epsilon' = |g(c)|/2$. It then follows from the backwards triangle inequality that

(3.3). $$|g(y)| > \epsilon' = |g(c)|/2 \,\text{so that}\,\ |1/g(y)| < 2/|g(c)|$$

Now, to finish the proof of part (5), let $\epsilon > 0$ be given. If $|y-c| < \min(\delta_1, \delta_2, \delta')$

and $y \in S$, then from Inequalities (3.1), (3.2), and (3.3) we obtain

$$\begin{aligned}|\frac{f(y)}{g(y)} - \frac{f(c)}{g(c)}| &= \frac{|f(y)g(c) - f(c)g(y)|}{|g(y)g(c)|} \\ &= \frac{|f(y)g(c) - f(c)g(c) + f(c)g(c) - f(c)g(y)|}{|g(y)||g(c)|} \\ &\leq \frac{|f(y) - f(c)||g(c)| + |f(c)||g(c) - g(y)|}{|g(y)||g(c)|} \\ &< (|f(y) - f(c)|M_2 + M_1|g(c) - g(y)|) \times \frac{2}{|g(c)|^2}.\end{aligned}$$

Finally, using the continuity of both f and g applied to the positive numbers $\epsilon_1 = \epsilon/(4M_2|g(c)|^2)$ and $\epsilon_2 = \epsilon/(4M_1|g(c)|^2)$, choose $\delta > 0$, with $\delta < \min(\delta_1, \delta_2, \delta')$, and such that if $|y - c| < \delta$ and $y \in S$ then $|f(y) - f(c)| < \frac{\epsilon}{4M_2/|g(c)|^2}$ and $|g(c) - g(y)| < \frac{\epsilon}{4M_1/|g(c)|^2}$. Then, if $|y - c| < \delta$ and $y \in S$ we have that

$$|\frac{f(y)}{g(y)} - \frac{f(c)}{g(c)}| < \epsilon$$

as desired.

Exercise 3.4. (a) Prove part (2) of the preceding theorem. (It's an $\epsilon/2$ argument.)
(b) Prove part (3) of the preceding theorem. (It's similar to the proof of part (5) only easier.)
(c) Prove part (4) of the preceding theorem.
(d) Prove part (6) of the preceding theorem.
(e) Suppose S is a subset of $\mathbb{R}$. Verify the above theorem replacing " continuity" with left continuity and right continuity.
(f) If S is a subset of $\mathbb{R}$, show that f is continuous at a point $c \in S$ if and only if it is both right continuous and left continuous at c.

THEOREM 3.3. (The composition of continuous functions is continuous.) Let S, T, and U be subsets of $\mathbb{C}$, and let $f : S \to T$ and $g : T \to U$ be functions. Suppose f is continuous at a point $c \in S$ and that g is continuous at the point $f(c) \in T$. Then the composition $g \circ f$ is continuous at c.

PROOF. Let $\epsilon > 0$ be given. Because g is continuous at the point $f(c)$, there exists an $\alpha > 0$ such that $|g(t) - g(f(c))| < \epsilon$ if $|t - f(c)| < \alpha$. Now, using this positive number α, and using the fact that f is continuous at the point c, there exists a $\delta > 0$ so that $|f(s) - f(c)| < \alpha$ if $|s - c| < \delta$. Therefore, if $|s - c| < \delta$, then $|f(s) - f(c)| < \alpha$, and hence $|g(f(s)) - g(f(c))| = |g \circ f(s) - g \circ f(c)| < \epsilon$, which completes the proof.

Exercise 3.5. (a) If $f : \mathbb{C} \to \mathbb{C}$ is the function defined by $f(z) = z$, prove that f is continuous at each point of $\mathbb{C}$.
(b) Use part (a) and Theorem 3.2 to conclude that every rational function is continuous on its domain.
(c) Prove that a step function $h : [a, b] \to \mathbb{C}$ is continuous everywhere on $[a, b]$ except possibly at the points of the partition P that determines h.

Exercise 3.6. (a) Let S be the set of nonnegative real numbers, and define $f : S \to S$ by $f(x) = \sqrt{x}$. Prove that f is continuous at each point of S.
HINT: For $c = 0$, use $\delta = \epsilon^2$. For $c \neq 0$, use the identity

$$\sqrt{y} - \sqrt{c} = (\sqrt{y} - \sqrt{c})\frac{\sqrt{y}+\sqrt{c}}{\sqrt{y}+\sqrt{c}} = \frac{y-c}{\sqrt{y}+\sqrt{c}} \leq \frac{y-c}{\sqrt{c}}.$$

(b) If $f : \mathbb{C} \to \mathbb{R}$ is the function defined by $f(z) = |z|$, show that f is continuous at every point of its domain.

Exercise 3.7. Using the previous theorems and exercises, explain why the following functions f are continuous on their domains. Describe the domains as well.
(a) $f(z) = (1 - z^2)/(1 + z^2)$.
(b) $f(z) = |1 + z + z^2 + z^3 - (1/z)|$.
(c) $f(z) = \sqrt{1 + \sqrt{1 - |z|^2}}$.

Exercise 3.8. (a) If c and d are real numbers, show that $\max(c, d) = (c + d)/2 + |c - d|/2$.
(b) If f and g are functions from S into $\mathbb{R}$, show that $\max(f, g) = (f+g)/2+|f-g|/2$.
(c) If f and g are real-valued functions that are both continuous at a point c, show that $\max(f, g)$ and $\min(f, g)$ are both continuous at c.

Exercise 3.9. Let $\mathbb{N}$ be the set of natural numbers, let P be the set of positive real numbers, and define $f : \mathbb{N} \to P$ by $f(n) = \sqrt{1 + n}$. Prove that f is continuous at each point of $\mathbb{N}$. Show in fact that every function $f : \mathbb{N} \to \mathbb{C}$ is continuous on this domain $\mathbb{N}$.
HINT: Show that for any $\epsilon > 0$, the choice of $\delta = 1$ will work.

Exercise 3.10. (Negations)
(a) Negate the statement: "For every $\epsilon > 0$, $|x| < \epsilon$.″
(b) Negate the statement: "For every $\epsilon > 0$, there exists an x for which $|x| < \epsilon$.″
(c) Negate the statement that " f is continuous at c.″

The next result establishes an equivalence between the basic ϵ, δ definition of continuity and a sequential formulation. In many cases, maybe most, this sequential version of continuity is easier to work with than the ϵ, δ version.

THEOREM 3.4. *Let $f : S \to \mathbb{C}$ be a complex-valued function on S, and let c be a point in S. Then f is continuous at c if and only if the following condition holds: For every sequence $\{x_n\}$ of elements of S that converges to c, the sequence $\{f(x_n)\}$ converges to $f(c)$. Or, said a different way, if $\{x_n\}$ converges to c, then $\{f(x_n)\}$ converges to $f(c)$. And, said yet a third (somewhat less precise) way, the function f converts convergent sequences to convergent sequences.*

PROOF. Suppose first that f is continuous at c, and let $\{x_n\}$ be a sequence of elements of S that converges to c. Let $\epsilon > 0$ be given. We must find a natural number N such that if $n \geq N$ then $|f(x_n) - f(c)| < \epsilon$. First, choose $\delta > 0$ so that $|f(y) - f(c)| < \epsilon$ whenever $y \in S$ and $|y - c| < \delta$. Now, choose N so that $|x_n - c| < \delta$ whenever $n \geq N$. Then if $n \geq N$, we have that $|x_n - c| < \delta$, whence $|f(x_n) - f(c)| < \epsilon$. This shows that the sequence $\{f(x_n)\}$ converges to $f(c)$, as desired.
We prove the converse by proving the contrapositive statement; i.e., we will show that if f is not continuous at c, then there does exist a sequence $\{x_n\}$ that converges

to c but for which the sequence $\{f(x_n)\}$ does not converge to $f(c)$. Thus, suppose f is **not** continuous at c. Then there exists an $\epsilon_0 > 0$ such that for every $\delta > 0$ there is a $y \in S$ such that $|y - c| < \delta$ but $|f(y) - f(c)| \geq \epsilon_0$. To obtain a sequence, we apply this statement to δ's of the form $\delta = 1/n$. Hence, for every natural number n there exists a point $x_n \in S$ such that $|x_n - c| < 1/n$ but $|f(x_n) - f(c)| \geq \epsilon_0$. Clearly, the sequence $\{x_n\}$ converges to c since $|x_n - c| < 1/n$. On the other hand, the sequence $\{f(x_n)\}$ cannot be converging to $f(c)$, because $|f(x_n) - f(c)|$ is always $\geq \epsilon_0$.

This completes the proof of the theorem.

CONTINUITY AND TOPOLOGY

Let $f : S \to T$ be a function, and let A be a subset of the codomain T. Recall that $f^{-1}(A)$ denotes the subset of the domain S consisting of all those $x \in S$ for which $f(x) \in A$.

Our original definition of continuity was in terms of ϵ's and δ's. Theorem 3.4 established an equivalent form of continuity, often called "sequential continuity," that involves convergence of sequences. The next result shows a connection between continuity and topology, i.e., open and closed sets.

THEOREM 3.5. *(1) Suppose S is a closed subset of $\mathbb{C}$ and that $f : S \to \mathbb{C}$ is a complex-valued function on S. Then f is continuous on S if and only if $f^{-1}(A)$ is a closed set whenever A is a closed subset of $\mathbb{C}$. That is, f is continuous on a closed set S if and only if the inverse image of every closed set is closed.*

(2) Suppose U is an open subset of $\mathbb{C}$ and that $f : U \to \mathbb{C}$ is a complex-valued function on U. Then f is continuous on U if and only if $f^{-1}(A)$ is an open set whenever A is an open subset of $\mathbb{C}$. That is, f is continuous on an open set U if and only if the inverse image of every open set is open.

PROOF. Suppose first that f is continuous on a closed set S and that A is a closed subset of $\mathbb{C}$. We wish to show that $f^{-1}(A)$ is closed. Thus, let $\{x_n\}$ be a sequence of points in $f^{-1}(A)$ that converges to a point c. Because S is a closed set, we know that $c \in S$, but in order to see that $f^{-1}(A)$ is closed, we need to show that $c \in f^{-1}(A)$. That is, we need to show that $f(c) \in A$. Now, $f(x_n) \in A$ for every n, and, because f is continuous at c, we have by Theorem 3.4 that $f(c) = \lim f(x_n)$. Hence, $f(c)$ is a limit point of A, and so $f(c) \in A$ because A is a closed set. Therefore, $c \in f^{-1}(A)$, and $f^{-1}(A)$ is closed.

Conversely, still supposing that S is a closed set, suppose f is not continuous on S, and let c be a point of S at which f fails to be continuous. Then, there exists an $\epsilon > 0$ and a sequence $\{x_n\}$ of elements of S such that $c = \lim x_n$ but such that $|f(c) - f(x_n)| \geq \epsilon$ for all n. (Why? See the proof of Theorem 3.4.) Let A be the complement of the open disk $B_\epsilon(f(c))$. Then A is a closed subset of $\mathbb{C}$. We have that $f(x_n) \in A$ for all n, but $f(c)$ is not in A. So, $x_n \in f^{-1}(A)$ for all n, but $c = \lim x_n$ is not in $f^{-1}(A)$. Hence, $f^{-1}(A)$ does not contain all of its limit points, and so $f^{-1}(A)$ is not closed. Hence, if f is not continuous on S, then there exists a closed set A such that $f^{-1}(A)$ is not closed. This completes the proof of the second half of part (1).

Next, suppose U is an open set, and assume that f is continuous on U. Let A be an open set in $\mathbb{C}$, and let c be an element of $f^{-1}(A)$. In order to prove that$f^{-1}(A)$ is open, we need to show that c belongs to the interior of $f^{-1}(A)$. Now, $f(c) \in A$, A

is open, and so there exists an $\epsilon > 0$ such that the entire disk $B_\epsilon(f(c)) \subseteq A$. Then, because f is continuous at the point c, there exists a $\delta > 0$ such that if $|x - c| < \delta$ then $|f(x) - f(c)| < \epsilon$. In other words, if $x \in B_\delta(c)$, then $f(x) \in B_\epsilon(f(c)) \subseteq A$. This means that $B_\delta(c)$ is contained in $f^{-1}(A)$, and hence c belongs to the interior of $f^{-1}(A)$. Hence, if f is continuous on an open set U, then $f^{-1}(A)$ is open whenever A is open. This proves half of part (2).

Finally, still assuming that U is open, suppose $f^{-1}(A)$ is open whenever A is open, let c be a point of S, and let us prove that f is continuous at c. Thus, let $\epsilon > 0$ be given, and let A be the open set $A = B_\epsilon(f(c))$. Then, by our assumption, $f^{-1}(A)$ is an open set. Also, c belongs to this open set $f^{-1}(A)$, and hence c belongs to the interior of $f^{-1}(A)$. Therefore, there exists a $\delta > 0$ such that the entire disk $b_\delta(c) \subseteq f^{-1}(A)$. But this means that if $\in S$ satisfies $|x - c| < \delta$, then $x \in B_\delta(c) \subseteq f^{-1}(A)$, and so $f(x) \in A = B_\epsilon(f(c))$. Therefore, if $|x - c| < \delta$, then $|f(x) - f(c)| < \epsilon$, which proves that f is continuous at c, and the theorem is completely proved.

DEEPER ANALYTIC PROPERTIES OF CONTINUOUS FUNCTIONS

We collect here some theorems that show some of the consequences of continuity. Some of the theorems apply to functions either of a real variable or of a complex variable, while others apply only to functions of a real variable. We begin with what may be the most famous such result, and this one is about functions of a real variable.

THEOREM 3.6. (Intermediate Value Theorem) If $f : [a, b] \to \mathbb{R}$ is a real-valued function that is continuous at each point of the closed interval $[a, b]$, and if v is a number (value) between the numbers $f(a)$ and $f(b)$, then there exists a point c between a and b such that $f(c) = v$.

PROOF. If $v = f(a)$ or $f(b)$, we are done. Suppose then, without loss of generality, that $f(a) < v < f(b)$. Let S be the set of all $x \in [a, b]$ such that $f(x) \leq v$, and note that S is nonempty and bounded above. ($a \in S$, and b is an upper bound for S.) Let $c = \sup S$. Then there exists a sequence $\{x_n\}$ of elements of S that converges to c. (See Exercise 2.20.) So, $f(c) = \lim f(x_n)$ by Theorem 3.4. Hence, $f(c) \leq v$. (Why?)

Now, arguing by contradiction, if $f(c) < v$, let ϵ be the positive number $v - f(c)$. Because f is continuous at c, there must exist a $\delta > 0$ such that $|f(y) - f(c)| < \epsilon$ whenever $|y - c| < \delta$ and $y \in [a, b]$. Since any smaller δ satisfies the same condition, we may also assume that $\delta < b - c$. Consider $y = c + \delta/2$. Then $y \in [a, b]$, $|y - c| < \delta$, and so $|f(y) - f(c)| < \epsilon$. Hence $f(y) < f(c) + \epsilon = v$, which implies that $y \in S$. But, since $c = \sup S$, c must satisfy $c \geq y = c + \delta/2$. This is a contradiction, so $f(c) = v$, and the theorem is proved.

The Intermediate Value Theorem tells us something qualitative about the range of a continuous function on an interval $[a, b]$. It tells us that the range is "connected;" i.e., if the range contains two points c and d, then the range contains all the points between c and d. It is difficult to think what the analogous assertion would be for functions of a complex variable, since "between" doesn't mean anything for complex numbers. We will eventually prove something called the Open Mapping Theorem in Chapter VII that could be regarded as the complex analog of the Intermediate Value Theorem.

The next theorem is about functions of either a real or a complex variable.

THEOREM 3.7. *Let $f : S \to \mathbb{C}$ be a continuous function, and let C be a compact (closed and bounded) subset of S. Then the image $f(C)$ of C is also compact. That is, the continuous image of a compact set is compact.*

PROOF. First, suppose $f(C)$ is not bounded. Thus, let $\{x_n\}$ be a sequence of elements of C such that, for each n, $|f(x_n)| > n$. By the Bolzano-Weierstrass Theorem, the sequence $\{x_n\}$ has a convergent subsequence $\{x_{n_k}\}$. Let $x = \lim x_{n_k}$. Then $x \in C$ because $C$is a closed subset of $\mathbb{C}$. Co, $f(x) = \lim f(x_{n_k})$ by Theorem 3.4. But since $|f(x_{n_k})| > n_k$, the sequence $\{f(x_{n_k})\}$ is not bounded, so cannot be convergent. Hence, we have arrived at a contradiction, and the set $f(C)$ must be bounded.

Now, we must show that the image $f(C)$ is closed. Thus, let y be a limit point of the image $f(C)$ of C, and let $y = \lim y_n$ where each $y_n \in f(C)$. For each n, let $x_n \in C$ satisfy $f(x_n) = y_n$. Again, using the Bolzano-Weierstrass Theorem, let $\{x_{n_k}\}$ be a convergent subsequence of the bounded sequence $\{x_n\}$, and write $x = \lim x_{n_k}$. Then $x \in C$, since C is closed, and from Theorem 3.4

$$y = \lim f(x_n) = \lim f(x_{n_k}) = f(x),$$

showing that $y \in f(C)$, implying that $f(C)$ is closed.

This theorem tells us something about the range of a continuous function of a real or complex variable. It says that if a subset of the domain is closed and bounded, so is the image of that subset.

The next theorem is about continuous real-valued functions of a complex variable, and it is one of the theorems to remember.

THEOREM 3.8. *Let f be a continuous real-valued function on a compact subset S of $\mathbb{C}$. Then f attains both a maximum and a minimum value on S. That is, there exist points z_1 and z_2 in S such that $f(z_1) \leq f(z) \leq f(z_2)$ for all $z \in S$.*

PROOF. We prove that f attains a maximum value, leaving the fact that f attains a minimum value to the exercise that follows. Let M_0 be the supremum of the set of all numbers $f(x)$ for $x \in S$. (How do we know that this supremum exists?) We will show that there exists an $z_2 \in S$ such that $f(z_2) = M_0$. This will finish the proof, since we would then have $f(z_2) = M_0 \geq f(z)$ for all $z \in S$. Thus, let $\{y_n\}$ be a sequence of elements in the range of f for which the sequence $\{y_n\}$ converges to M_0. (This is Exercise 2.20 again.) For each n, let x_n be an element of S such that $y_n = f(x_n)$. Then the sequence $\{f(x_n)\}$ converges to M_0. Let $\{x_{n_k}\}$ be a convergent subsequence of $\{x_n\}$. (How?) Let $z_2 = \lim x_{n_k}$. Then $z_2 \in S$, because S is closed, and $f(z_2) = \lim f(x_{n_k})$, because f is continuous. Hence, $f(z_2) = M_0$, as desired.

Exercise 3.11. (a) Prove that the f of the preceding theorem attains a minimum value on S.

(b) Give an alternate proof of Theorem 3.8 by using Theorem 3.7, and then proving that a closed and bounded subset of $\mathbb{R}$ contains both its supremum and its infimum.

(c) Let S be a compact subset of $\mathbb{C}$, and let c be a point of $\mathbb{C}$ that is not in S. Prove that there is a closest point to c in S. That is, show that there exists a point $w \in S$ such that $|w - c| \leq |z - c|$ for all points $z \in S$.

HINT: The function $z \rightarrow |z-c|$ is continuous on the set S.

Exercise 3.12. Let $f:[a,b] \rightarrow \mathbb{R}$ be a real-valued function that is continuous at each point of $[a,b]$.
(a) Prove that the range of f is a closed interval $[a',b']$. Show by example that the four numbers $f(a), f(b)$, a' and b' can be distinct.
(b) Suppose f is 1-1. Show that, if c is in the open interval (a,b), then $f(c)$ is in the open interval (a',b').

We introduce next a different kind of continuity called uniform continuity. The difference between regular continuity and uniform continuity is a bit subtle, and well worth some thought.

DEFINITION. A function $f:S \rightarrow \mathbb{C}$ is called *uniformly continuous* on S if for each positive number ϵ, there exists a positive number δ such that $|f(x)-f(y)|<\epsilon$ for all $x,y \in S$ satisfying $|x-y|<\delta$.

Basically, the difference between regular continuity and uniform conintuity is that the same δ works for all points in S.
Here is another theorem worth remembering.

THEOREM 3.9. *A continuous complex-valued function on a compact subset S of $\mathbb{C}$ is uniformly continuous.*

PROOF. We argue by contradiction. Thus, suppose f is continuous on S but not uniformly continuous. Then, there exists an $\epsilon>0$ for which no positive number δ satisfies the uniform continuity definition. Therefore, thinking of the δ's as ranging through the numbers $1/n$, we know that for each positive integer n, there exist two points x_n and y_n in S so that

(1) $|y_n-x_n|<1/n$, and
(2) $|f(y_n)-f(x_n)| \geq \epsilon$.

Otherwise, some $1/n$ would suffice for a δ. Let $\{x_{n_k}\}$ be a convergent subsequence of $\{x_n\}$ with limit x. By (1) and the triangle inequality, we deduce that x is also the limit of the corresponding subsequence $\{y_{n_k}\}$ of $\{y_n\}$. But then $f(x)=\lim f(x_{n_k})=\lim f(y_{n_k})$, implying that $0=\lim|f(y_{n_k})-f(x_{n_k})|$, which implies that $|f(y_{n_k})-f(x_{n_k})|<\epsilon$ for all large enough k. But that contradicts (2), and this completes the proof.

Continuous functions whose domains are not compact sets may or may not be uniformly continuous, as the next exercise shows.

Exercise 3.13. (a) Let $f:(0,1) \rightarrow \mathbb{R}$ be defined by $f(x)=1/x$. Prove that f is continuous at each x in its domain but that f is not uniformly continuous there.
HINT: Set $\epsilon=1$, and consider the pairs of points $x_n=1/n$ and $y_n=1/(n+1)$.
(b) Let $f:[1,\infty) \rightarrow [1,\infty)$ be defined by $f(x)=\sqrt{x}$. Prove that f is not bounded, but is nevertheless uniformly continuous on its domain.
HINT: Take $\delta=\epsilon$.

THEOREM 3.10. *Let $f:S \rightarrow T$ be a continuous 1-1 function from a compact (closed and bounded) subset of $\mathbb{C}$ onto the (compact) set T. Let $g:T \rightarrow S$ denote the inverse function f^{-1} of f. Then g is continuous. The inverse of a continuous function, that has a compact domain, is also continuous.*

PROOF. We prove that g is continuous by using Theorem 3.5; i.e., we will show that $g^{-1}(A)$ is closed whenever A is a closed subset of $\mathbb{C}$. But this is easy, since

$g^{-1}(A) = g^{-1}(A \cap S) = f(A \cap S)$, and this is a closed set by Theorem 3.7, because $A \cap S$ is compact. See part (e) of Exercise 2.24.

REMARK. Using the preceding theorem, and the exercise below, we will show that taking nth roots is a continuous function. that is, the function f defined by $f(x) = x^{1/n}$ is continuous.

Exercise 3.14. Use the preceding theorem to show the continuity of the following functions.
(a) Show that if n is an odd positive integer, then there exists a continuous function g defined on all of $\mathbb{R}$ such that $g(x)$ is an nth root of x for all real numbers x. That is, $(g(x))^n = x$ for all real x. (The function $f(x) = x^n$ is 1-1 and continuous.)
(b) Show that if n is any positive integer then there exists a unique continuous function g defined on $[0, \infty)$ such that $g(x)$ is an nth root of x for all nonnegative x.
(c) Let $r = p/q$ be a rational number. Prove that there exists a continuous function $g : [0, \infty) \to [0, \infty)$ such that $g(x)^q = x^p$ for all $x \geq 0$; i.e., $g(x) = x^r$ for all $x \geq 0$.

THEOREM 3.11. *Let f be a continuous 1-1 function from the interval $[a, b]$ onto the interval $[c, d]$. Then f must be strictly monotonic, i.e., strictly increasing everywhere or strictly decreasing everywhere.*

PROOF. Since f is 1-1, we clearly have that $f(a) \neq f(b)$, and, without loss of generality, let us assume that $c = f(a) < f(b) = d$. It will suffice to show that if α and β belong to the open interval (a, b), and $\alpha < \beta$, then $f(\alpha) \leq f(\beta)$. (Why will this suffice?) Suppose by way of contradiction that there exists $\alpha < \beta$ in (a, b) for which $f(\alpha) > f(\beta)$. We use the intermediate value theorem to derive a contradiction. Consider the four points $a < \alpha < \beta < b$. Either $f(a) < f(\alpha)$ or $f(\beta) < f(b)$. (Why?) In the first case $(f(a) < f(\alpha))$, $f([a, \alpha])$ contains every value between $f(a)$ and $f(\alpha)$. And, $f([\alpha, \beta])$ contains every value between $f(\alpha)$ and $f(\beta)$. So, let v be a number such that $f(a) < v$, $f(\beta) < v$, and $v < f(\alpha)$ (why does such a number v exist?). By the Intermediate Value Theorem, there exists $x_1 \in (a, \alpha)$ such that $v = f(x_1)$, and there exists an $x_2 \in (\alpha, \beta)$ such that $v = f(x_2)$. But this contradicts the hypothesis that f is 1-1, since $x_1 \neq x_2$. A similar argument leads to a contradiction in the second case $f(\beta) < f(b)$. (See the following exercise.) Hence, there can exist no such α and β, implying that f is strictly increasing on $[a, b]$.

Exercise 3.15. Derive a contradiction from the assumption that $f(\beta) < f(b)$ in the preceding proof.

POWER SERIES FUNCTIONS

The class of functions that we know are continuous includes, among others, the polynomials, the rational functions, and the nth root functions. We can combine these functions in various ways, e.g., sums, products, quotients, and so on. We also can combine continuous functions using composition, so that we know that nth roots of rational functions are also continuous. The set of all functions obtained in this manner is called the class of "algebraic functions." Now that we also have developed a notion of limit, or infinite sum, we can construct other continuous functions.

We introduce next a new kind of function. It is a natural generalization of a polynomial function. Among these will be the exponential function and the trigonometric functions. We begin by discussing functions of a complex varible, although totally analogous definitions and theorems hold for functions of a real variable.

DEFINITION. Let $\{a_n\}_0^\infty$ be a sequence of real or complex numbers. By the *power series function* $f(z) = \sum_{n=0}^\infty a_n z^n$ we mean the function $f : S \to \mathbb{C}$ where the domain S is the set of all $z \in \mathbb{C}$ for which the infinite series $\sum a_n z^n$ converges, and where f is the rule that assigns to such a $z \in S$ the sum of the series.
The numbers $\{a_n\}$ defining a power series function are called the *coefficients* of the function.
We associate to a power series function $f(z) = \sum_{n=0}^\infty a_n z^n$ its sequence $\{S_N\}$ of partial sums. We write

$$S_N(z) = \sum_{n=0}^N a_n z^n.$$

Notice that polynomial functions are very special cases of power series functions. They are the power series functions for which the coefficients $\{a_n\}$ are all 0 beyond some point. Note also that each partial sum S_N for any power series function is itself a polynomial function of degree less than or equal to N. Moreover, if f is a power series function, then for each z in its domain we have $f(z) = \lim_N S_N(z)$. Evidently, every power series function is a "limit" of a sequence of polynomials.
Obviously, the domain $S \equiv S_f$ of a power series function f depends on the coefficients $\{a_n\}$ determining the function. Our first goal is to describe this domain.

THEOREM 3.12. *Let f be a power series function: $f(z) = \sum_{n=0}^\infty a_n z^n$ with domain S. Then:*

(1) 0 *belongs to S.*
(2) *If a number t belongs to S, then every number u, for which $|u| < |t|$, also belongs to S.*
(3) *S is a disk of radius r around* 0 *in $\mathbb{C}$ (possibly open, possibly closed, possibly neither, possibly infinite). That is, S consists of the disk $B_r(0) = \{z : |z| < r\}$ possibly together with some of the points z for which $|z| = r$.*
(4) *The radius r of the disk in part (3) is given by the Cauchy-Hadamard formula:*

$$r = \frac{1}{\limsup |a_n|^{1/n}},$$

which we interpret to imply that $r = 0$ if and only if the limsup on the right is infinite, and $r = \infty$ if and only if that limsup is 0.

PROOF. Part (1) is clear.
To see part 2, assume that t belongs to S and that $|u| < |t|$. We wish to show that the infinites series $\sum a_n u^n$ converges. In fact, we will show that $\sum |a_n u^n|$ is convergent, i.e., that $\sum a_n u^n$ is absolutely convergent. We are given that the infinite series $\sum a_n t^n$ converges, which implies that the terms $a_n t^n$ tend to 0. Hence, let B be a number such that $|a_n z^n| \leq B$ for all n, and set $\alpha = |u|/|t|$. Then $\alpha < 1$, and therefore the infinite series $\sum B\alpha^n$ is convergent. Finally, $|a_n u^n| = |a_n t^n| \alpha^n \leq B\alpha^n$, which, by the Comparison Test, implies that $\sum |a_n u^n|$ is convergent, as desired.
Part (3) follows, with just a little thought, from part 2.
To prove part (4), note that $\limsup |a_n|^{1/n}$ either is finite or it is infinite. assume first that the sequence $\{|a_n|^{1/n}\}$ is not bounded; i.e., that $\limsup |a_n|^{1/n} = \infty$.

Then, given any number p, there are infinitely many terms $|a_n|^{1/n}$ that are larger than p. So, for any $z \neq 0$, there exist infinitely many terms $|a_n|^{1/n}$ that are larger than $1/|z|$. But then $|a_n z^n| > 1$ for all such terms. Therefore the infinite series $\sum a_n z^n$ is not convergent, since $\lim a_n z^n$ is not zero. So no such z is in the domain S. This shows that if $\limsup |a_n|^{1/n} = \infty$, then $r = 0 = 1/\limsup |a_n|^{1/n}$.

Now, suppose the sequence $\{|a_n|^{1/n}\}$ is bounded, and let L denote its limsup. We must show that $1/r = L$. We will show the following two claims: (a) if $1/|z| > L$, then $z \in S$, and (b) if $1/|z| < L$, then $z \notin S$. (Why will these two claims complete the proof?) Thus, suppose that $1/|z| > L$. Let β be a number satisfying $L < \beta < 1/|z|$, and let $\alpha = \beta|z|$. Then $0 < \alpha < 1$. Now there exists a natural number N so that $|a_n|^{1/n} < \beta$ for all $n \geq N$, or equivalently $|a_n| \leq \beta^n$ for all $n \geq N$. (See part (a) of Exercise 2.17.) This means that for all $n \geq N$ we have $|a_n z^n| = |a_n/\beta^n||\beta z|^n \leq \alpha^n$. This implies by the Comparison Test that the power series $\sum a_n z^n$ is absolutely convergent, whence convergent. Hence, $z \in S$, and this proves claim (a) above. Incidentally, note also that if $L = 0$, this argument shows that $r = \infty$, as desired.

To verify claim (b), suppose that $1/|z| < L$. Then there are infinitely many terms of the sequence $\{|a_n|^{1/n}\}$ that are greater than $1/|z|$. (Why?) For each such term, we would then have $|a_n z^n| \geq 1$. This means that the infinite series $\sum a_n z^n$ is not convergent and $z \notin S$, which shows claim b.

Hence, in all cases, we have that $r = 1/\limsup |a_n|^{1/n}$, as desired.

DEFINITION. If f is a power series function, the number r of the preceding theorem is called the *radius of convergence* of the power series. The disk S of radius r around 0, denoted by $B_r(0)$, is called the *disk of convergence.*

Exercise 3.16. Compute directly the radii of convergence for the following power series functions, i.e., without using the Cauchy-Hadamard formula. Then, when possible, verify that the Cauchy-Hadamard formula agrees with your computation.

(a) $f(z) = \sum z^n$.

(b) $f(z) = \sum n^2 z^n$.

(c) $f(z) = \sum(-1)^n (1/(n+1)) z^n$.

(d) $f(z) = \sum(1/(n+1)) z^{3n+1}$.

(e) $f(z) = \sum_{n=0}^{\infty} z^n/n!$.

Exercise 3.17. (a) Use part (e) of Exercise 3.1 to show that a power series function p is an even function if and only if its only nonzero coefficients are even ones, i.e., the a_{2k}'s. Show also that a power series function is an odd function if and only if its only nonzero coefficients are odd ones, i.e., the a_{2k+1}'s.

(b) Suppose $f(z) = \sum_{k=0}^{\infty} a_{2k} z^{2k}$ is a power series function that is an even function. Show that

$$f(iz) = \sum_{k=0}^{\infty} (-1)^k a_{2k} z^{2k} = f^a(z),$$

where f^a is the power series function obtained from f by alternating the signs of its coefficients. We call this function f^a the alternating version of f.

(c) If $g(z) = \sum_{k=0}^{\infty} a_{2k+1} z^{2k+1}$ is a power series function that is an odd function, show that

$$g(iz) = i \sum_{k=0}^{\infty} (-1)^k a_{2k+1} z^{2k+1} = i g^a(z),$$

where again g^a is the power series function obtained from g by alternating the signs of its coefficients.
(d) If f is any power series function, show that

$$f(iz) = f_e(iz) + f_o(iz) = f_e^a(z) + if_o^a(z),$$

and hence that $f_e(iz) = f_e^a(z)$ and $f_o(iz) = ip_o^a(z)$.

The next theorem will not come as a shock, but its proof is not so simple.

THEOREM 3.13. *Let $f(z) = \sum a_n z^n$ be a power series function with radius of convergence r. Then f is continuous at each point in the open disk $B_r(0)$, i.e., at each point z for which $|z| < r$.*

PROOF. Let $z \in B_r(0)$ be given. We must make some auxiliary constructions before we can show that f is continuous at z. First, choose a z' such that $|z| < |z'| < r$. Next, set $b_n = |na_n|$, and define $g(z) = \sum b_n z^n$. By the Cauchy-Hadamard formula, we see that the power series function g has the same radius of convergence as the power series function f. Indeed, $\limsup |b_n|^{1/n} = \limsup n^{1/n}|a_n| = \lim n^{1/n} \limsup |a_n|$. Therefore, z' belongs to the domain of g. Let M be a number such that each partial sum of the series $g(z') = \sum_{n=0}^N b_n z'^n$ is bounded by M.
Now, let $\epsilon > 0$ be given, and choose δ to be the minimum of the two positive numbers $\epsilon|z'|/M$ and $|z'| - |z|$. We consider any y for which $|y - z| < \delta$. Then $y \in B_r(0)$, $|y| < |z'|$, and

$$\begin{aligned}
|f(y) - f(z)| &= \lim |S_N(y) - S_N(z)| \\
&= \lim |\sum_{n=0}^N a_n(y^n - z^n)| \\
&\le \lim_N \sum_{n=0}^N |a_n||y^n - z^n| \\
&= \lim_N \sum_{n=1}^N |a_n||y - z| \sum_{j=0}^{n-1} |y^j||z^{n-1-j}| \\
&\le \lim_N \sum_{n=1}^N |a_n||y - z| \sum_{j=0}^{n-1} |z'|^{n-1} \\
&\le \lim_N |y - z|(1/|z'|) \sum_{n=0}^N n|a_n||z'|^n \\
&\le |y - z| \lim_N \frac{M}{|z'|} \\
&< \delta \lim_N \frac{M}{|z'|} \\
&\le \epsilon.
\end{aligned}$$

This completes the proof.

Exercise 3.18. (a) Let $f(z) = \sum_{n=0}^\infty a_n z^n$ be a power series function, and let $p(z) = \sum_{k=0}^m b_k z^k$ be a polynomial function. Prove that $f + p$ and fp are both

power series functions. Express the coefficients for $f+p$ and fp in terms of the a_n's and b_k's.
(b) Suppose f and g are power series functions. Prove that $f+g$ is a power series function. What is its radius of convergence? What about cf? What about fg? What about f/g? What about $|f|$?

Exercise 3.19. (a) Prove that every polynomial is a power series function with infinite radius of convergence.
(b) Prove that $1/z$ and $(1/(z-1)(z+2))$ are not power series functions. (Their domains aren't right.)
(c) Define $f(z)=\sum_{n=0}^{\infty}(-1)^n z^{2n+1}$. Prove that the radius of convergence of this power series function is 1, and that $f(z)=\frac{z}{1+z^2}$ for all $z\in B_1(0)$. Conclude that the rational function $z/(1+z^2)$ agrees with a power series function on the disk $B_1(0)$. But, they are **not** the same function.
HINT: Use the infinite geometric series.

Theorem 3.13 and Exercises 3.18 and 3.19 raise a very interesting and subtle point. Suppose $f(z)=\sum a_n z^n$ is a power series function having finite radius of convergence $r>0$. Theorem 3.13 says that f is continuous on the open disk, but it does not say anything about the continuity of f at points on the boundary of this disk that are in the domain of f, i,e., at points z_0 for which $|z_0|=r$. and $\sum a_n z_0^n$ converges. Suppose $g(z)$ is a continuous function whose domain contains the open disk $B_r(0)$ and also a point z_0, and assume that $f(z)=g(z)$ for all $z\in B_r(0)$. Does $f(z_0)$ have to agree with $g(z_0)$? It's worth some thought to understand just what this question means. It amounts to a question of the equality of two different kinds of limits. $f(z_0)$ is the sum of an infinite series, the limit of a sequence of partial sums, while, because g is continuous at z_0, $g(z_0=\lim_{z\to z_0} g(z)$. At the end of this chapter, we include a theorem of Abel that answers this question.
The next theorem is the analog for power series functions of part (2) of Theorem 3.1 for polynomials. We call it the "Identity Theorem," but it equally well could be known as the "Uniqueness of Coefficients Theorem," for it implies that different coefficients mean different functions.

THEOREM 3.14. (Identity Theorem) Let $f(z)=\sum a_n z^n$ be a power series function with positive radius of convergence r. Suppose $\{z_k\}$ is a sequence of nonzero distinct numbers in the domain of f such that:

(1) $\lim z_k=0$.
(2) $f(z_k)=0$ for all k.

Then f is identically 0 ($f(z)\equiv 0$ for all $z\in S$). Moreover, each coefficient a_n of f equals 0.

PROOF. Arguing by induction on n, let us prove that all the coefficients a_n are 0. First, since f is continuous at 0, and since $\lim z_k=0$, we have that a_0, which equals $f(0)$, $=\lim f(z_k)=0$.
Assume then that $a_0=a_1=\ldots=a_{n-1}=0$. Then

$$\begin{aligned} f(z) &= a_n z^n + a_{n+1}z^{n+1}+\ldots \\ &= z^n\sum_{j=0}^{\infty} b_j z^j, \end{aligned}$$

where $b_j = a_{n+j}$. If g is the power series function defined by $g(z) = \sum b_j z^j$, then, by the Cauchy-Hadamard Formula, we have that the radius of convergence for g is the same as that for f. (Why does $\limsup |b_j|^{1/j} = \limsup |a_k|^{1/k}$?) We have that $f(z) = z^n g(z)$ for all z in the common disk of convergence of these functions f and g. Since, for each k, $z_k \neq 0$ and $f(z_k) = z_k^n g(z_k) = 0$, it follows that $g(z_k) = 0$ for every k. Since g is continuous at 0, it then follows as in the argument above that $g(0) = 0$. But, $g(0) = b_0 = a_n$. Hence $a_n = 0$, and so by induction all the coefficients of the power series function f are 0. Clearly this implies that $f(z)$ is identically 0.

COROLLARY. *Suppose f and g are two power series functions, that $\{z_k\}$ is a sequence of nonzero points that converges to 0, and that $f(z_k) = g(z_k)$ for all k. Then f and g have the same coefficients, the same radius of convergence, and hence $f(z) = g(z)$ for all z in their common domain.*

Exercise 3.20. (a) Prove the preceding corollary. (Compare with the proof of Theorem 3.1.)
(b) Use the corollary, and the power series function $g(z) = z$, to prove that $f(z) = |z|$ is not a power series function.
(c) Show that there are power series functions that are not polynomial functions.
(d) Let $f(z) = \sum a_n z^n$ be a power series function with infinite radius of convergence, all of whose coefficients are positive. Show that there is no rational function $r = p/q$ for which $f(z) = r(z)$ for all complex numbers z. Conclude that the collection of power series functions provides some **new** functions.
HINT: Use the fact that for any n we have that $f(x) > a_n x^n$ for all positive x. Then, by choosing n appropriately, derive a contradiction to the resulting fact that $|p(x)/q(x)| > a_n x^n$ for all positive x. See part (b) of Exercise 3.2.

THE ELEMENTARY TRANSCENDENTAL FUNCTIONS

Having introduced a class of new functions (power series functions), we might well expect that some of these will have interesting and unexpected properties. So, which sets of coefficients might give us an exotic new function? Unfortunately, at this point in our development, we haven't much insight into this question. It is true, see Exercise 3.16, that most power series functions that we naturally write down have finite radii of convergence. Such functions may well be new and fascinating, but as a first example, we would prefer to consider a power series function that is defined everywhere, i.e., one with an infinite radius of convergence. Again revisiting Exercise 3.16, let us consider the coefficients $a_n = 1/n!$. This may seem a bit ad hoc, but let's have a look.

DEFINITION. Define a power series function, denoted by exp, as follows:

$$\exp(z) = \sum_{n=0}^{\infty} \frac{z^n}{n!}.$$

We will call this function, with 20-20 hindsight, the *exponential function.*

What do we know about this function, apart from the fact that it is defined for all complex numbers? We certainly do not know that it has anything to do with the function e^z; that will come in the next chapter. We do know what the number e is, but we do not know how to raise that number to a complex exponent.

All of the exponential function's coefficients are positive, and so by part (d) of Exercise 3.20 exp is not a rational function; it really is something new. It is natural to consider the even and odd parts $\exp_e$ and $\exp_o$ of this new function. And then, considering the constructions in Exercise 3.17, to introduce the alternating versions $\exp_e^a$ and $\exp_o^a$ of them.

DEFINITION. Define two power series functions cosh (hyperbolic cosine) and sinh (hyperbolic sine) by

$$\cosh(z) = \frac{\exp(z) + \exp(-z)}{2} \text{ and } \sinh(z) = \frac{\exp(z) - \exp(z)}{2},$$

and two other power series functions cos (cosine) and sin (sine) by

$$\cos(z) = \cosh(iz) = \frac{\exp(iz) + \exp(-iz)}{2}$$

and

$$\sin(z) = -i\sinh(iz) = \frac{\exp(iz) - \exp(-iz)}{2i}.$$

The five functions just defined are called the *elementary transcendental functions*, the sinh and cosh functions are called the basic *hyperbolic functions*, and the sine and cosine functions are called the basic *trigonometric* or *circular functions*. The connections between the hyperbolic functions and hyperbolic geometry, and the connection between the trigonometric functions and circles and triangles, will only emerge in the next chapter. From the very definitions, however, we can see a connection between the hyperbolic functions and the trigonometric functions. It's something like interchanging the roles of the real and imaginary axes. This is probably worth some more thought.

Exercise 3.21. (a) Verify the following equations:

$$\begin{aligned}\exp(z) &= \sum_{n=0}^{\infty} \frac{z^n}{n!} \\ &= 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \ldots + \frac{z^k}{k!} + \ldots, \\ &= \cosh(z) + \sinh(z).\end{aligned}$$

$$\begin{aligned}\sin(z) &= z - \frac{z^3}{3!} + \frac{z^5}{5!} - \frac{z^7}{7!} + \ldots + (-1)^k \frac{z^{2k+1}}{(2k+1)!} + \ldots \\ &= \sum_{k=0}^{\infty} (-1)^k \frac{z^{2k+1}}{(2k+1)!},\end{aligned}$$

$$\begin{aligned}\cos(z) &= 1 - \frac{z^2}{2!} + \frac{z^4}{4!} - \frac{z^6}{6!} + \ldots + (-1)^k \frac{z^{2k}}{(2k)!} + \ldots \\ &= \sum_{k=0}^{\infty} (-1)^k \frac{z^{2k}}{(2k)!},\end{aligned}$$

$$\sinh(z) = \sum_{k=0}^{\infty} \frac{z^{2k+1}}{(2k+1)!},$$

and

$$\cosh(z) = \sum_{k=0}^{\infty} \frac{z^{2k}}{(2k)!}.$$

(These expressions for the elementary transcendental functions are perhaps the more familiar ones from a calculus course.)
(b) Compute the radii of convergence for the elementary transcendental functions. HINT: Do not use the Cauchy-Hadamard formula. Just figure out for which z's the functions are defined.
(c) Verify that $\exp(0) = 1$, $\sin(0) = \sinh(0) = 0$, and $\cos(0) = \cosh(0) = 1$.
(d) Prove that all five of the elementary transcendental functions are not rational functions.
(e) Can you explain why $\sin^2(z)+\cos^2(z) \equiv 1$? What about the " Addition Formula"

$$\sin(z+w) = \sin(z)\cos(w) + \cos(z)\sin(w).$$

Exercise 3.22. (a) Show that the elementary transcendental functions map real numbers to real numbers. That is, as functions of a real variable, they are real-valued functions.
(b) Show that the exponential function exp is not bounded above. Show in fact that, for each nonnegative integer n, $\exp(x)/x^n$ is unbounded. Can you show that $\exp(x) = e^x$? What, in fact, does e^x mean if x is an irrational or complex number?

At this point, we probably need a little fanfare!

THEOREM 3.14159. (Definition of π) There exists a smallest positive number x for which $\sin(x) = 0$. We will denote this distinguished number x by the symbol π.

PROOF. First we observe that $\sin(1)$ is positive. Indeed, the infinite series for $\sin(1)$ is alternating. It follows from the alternating series test (Theorem 2.18) that $\sin(1) > 1 - 1/6 = 5/6$.
Next, again using the alternating series test, we observe that $\sin(4) < 0$. Indeed,

$$\sin(4) < 4 - \frac{4^3}{3!} + \frac{4^5}{5!} - \frac{4^7}{7!} + \frac{4^9}{9!} \approx -0.4553 < 0.$$

Hence, by the intermediate value theorem, there must exist a number c between 1 and 4 such that $\sin(c) = 0$. So, there is at least one positive number x such that $\sin(x) = 0$. However, we must show that there is a smallest positive number satisfying this equation.
Let A be the set of all $x > 0$ for which $\sin(x) = 0$. Then A is a nonempty set of real numbers that is bounded below. Define $\pi = \inf A$. We need to prove that $\sin(\pi) = 0$, and that $\pi > 0$. Clearly then it will be the smallest positive number x for which $\sin(x) = 0$.
By Exercise 2.20, there exists a sequence $\{x_k\}$ of elements of A such that $\pi = \lim x_k$. Since sin is continuous at π, it follows that $\sin(\pi) = \lim \sin(x_k) = \lim 0 = 0$. Finally,

if π were equal to 0, then by the Identity Theorem, Theorem 3.14, we would have that $\sin x = 0$ for all x. Since this is clearly not the case, we must have that $\pi > 0$. Hence, π is the smallest (minimum) positive number x for which $\sin(x) = 0$.

As hinted at earlier, the connection between this number π and circles is not at all evident at the moment. For instance, you probably will not be able to answer the questions in the next exercise.

Exercise 3.23. (a) Can you see why $\sin(x + 2\pi) \equiv \sin(x)$? That is, is it obvious that sin is a periodic function?
(b) Can you prove that $\cos(\pi) = -1$?

REMARK. Defining π to be the smallest positive zero of the sine function may strike many people as very much "out of the blue." However, the zeroes of a function are often important numbers. For instance, a zero of the function $x^2 - 2$ is a square root of 2, and that number we know was exztremely important to the Greeks as they began the study of what real numbers are. A zero of the function $z^2 + 1$ is something whose square is -1, i.e., negative. The idea of a square being negative was implausible at first, but is fundamental now, so that the zero of this particular function is critical for our understanding to numbers. Very likely, then, the zeroes of any "new" function will be worth studying. For instance, we will soon see that, perhaps disappointingly, there are no zeroes for the exponential function: $\exp(z)$ is never 0. Maybe it's even more interesting then that there are zeroes of the sine function.
The next theorem establishes some familiar facts about the trigonometric functions.

THEOREM 3.15.

(1) $\exp(iz) = \cos(z) + i\sin(z)$ *for all* $z \in \mathbb{C}$.

(2) *Let* $\{z_k\}$ *be a sequence of complex numbers that converges to 0. Then*

$$\lim \frac{\sin(z_k)}{z_k} = 0.$$

(3) *Let* $\{z_k\}$ *be a sequence of complex numbers that converges to 0. Then*

$$\lim \frac{1 - \cos(z_k)}{z_k^2} = \frac{1}{2}.$$

Exercise 3.24. Prove Theorem 3.15.
HINT: For parts (2) and (3), use Theorem 3.13.

ANALYTIC FUNCTIONS AND TAYLOR SERIES

DEFINITION. Let S be a subset of $\mathbb{C}$, let $f : S \to \mathbb{C}$ be a complex-valued function, and let c be a point of S. Then f is said to be *expandable in a Taylor series around* c with radius of convergence r if there exists an $r > 0$ such that $B_r(c) \subseteq S$, and $f(z)$ is given by the formula

$$f(z) = \sum_{n=0}^{\infty} a_n (z - c)^n$$

for all $z \in B_r(c)$.
Let S be a subset of $\mathbb{R}$, let $f : S \to \mathbb{R}$ be a real-valued function on S, and let c be a point of S. Then f is said to be *expandable in a Taylor series around* c with radius of convergence r if there exists an $r > 0$ such that the interval $(c - r, c + r) \subseteq S$, and $f(x)$ is given by the formula

$$f(x) = \sum_{n=0}^{\infty} a_n (x - c)^n$$

for all $x \in (c - r, c + r)$.
Suppose S is an open subset of $\mathbb{C}$. A function $f : S \to \mathbb{C}$ is called *analytic on* S if it is expandable in a Taylor series around every point c of S.
Suppose S is an open subset of $\mathbb{R}$. A function $f : S \to \mathbb{C}$ is called *real analytic on* S if it is expandable in a Taylor series around every point c of S.

THEOREM 3.16. *Suppose S is a subset of $\mathbb{C}$, that $f : S \to \mathbb{C}$ is a complex-valued function and that c belongs to S. Assume that f is expandable in a Taylor series around c with radius of convergence r. Then f is continuous at each $z \in B_r(c)$.*
Suppose S is a subset of $\mathbb{R}$, that $f : S \to \mathbb{R}$ is a real-valued function and that c belongs to S. Assume that f is expandable in a Taylor series around c with radius of convergence r. Then f is continuous at each $x \in (c - r, c + r)$.

PROOF. If we let g be the power series function given by $g(z) = \sum a_n z^n$, and T be the function defined by $T(z) = z - c$, then $f(z) = g(T(z))$, and this theorem is a consequence of Theorems 3.3 and 3.13.

Exercise 3.25. Prove that $f(z) = 1/z$ is analytic on its domain.
HINT: Use $r = |c|$, and then use the infinite geometric series.

Exercise 3.26. State and prove an Identity Theorem, analogous to Theorem 3.14, for functions that are expandable in a Taylor series around a point c.

Exercise 3.27. (a) Prove that every polynomial is expandable in a Taylor series around every point c.
HINT: Use the binomial theorem.
(b) Is the exponential function expandable in a Taylor series around the number -1?

UNIFORM CONVERGENCE

We introduce now two different notions of the limit of a sequence of functions. Let S be a set of complex numbers, and let $\{f_n\}$ be a sequence of complex-valued functions each having domain S.

DEFINITION. We say that the sequence $\{f_n\}$ *converges* or *converges pointwise* to a function $f : S \to \mathbb{C}$ if for every $x \in S$ and every $\epsilon > 0$ there exists a natural number N, depending on x and ϵ, such that for every $n \geq N$, $|f_n(x) - f(x)| < \epsilon$. That is, equivalently, $\{f_n\}$ converges pointwise to f if for every $x \in S$ the sequence $\{f_n(x)\}$ of numbers converges to the number $f(x)$.
We say that the sequence $\{f_n\}$ *converges uniformly* to a function f if for every $\epsilon > 0$, there exists an N, depending only on ϵ, such that for every $n \geq N$ and every $x \in S$, $|f_n(x) - f(x)| < \epsilon$.

If $\{u_n\}$ is a sequence of functions defined on S, we say that the infinite series $\sum u_n$ *converges uniformly* if the sequence $\{S_N = \sum_{n=0}^N u_n\}$ of partial sums converges uniformly.

These two definitions of convergence of a sequence of functions differ in subtle ways. Study the word order in the definitions.

Exercise 3.28. (a) Prove that if a sequence $\{f_n\}$ of functions converges uniformly on a set S to a function f then it converges pointwise to f.
(b) Let $S = (0,1)$, and for each n define $f_n(x) = x^n$. Prove that $\{f_n\}$ converges pointwise to the zero function, but that $\{f_n\}$ does not converge uniformly to the zero function. Conclude that pointwise convergence does **not** imply uniform convergence.
HINT: Suppose the sequence does converge uniformly. Take $\epsilon = 1/2$, let N be a corresponding integer, and consider x's of the form $x = 1 - h$ for tiny h's.
(c) Suppose the sequence $\{f_n\}$ converges uniformly to f on S, and the sequence $\{g_n\}$ converges uniformly to g on S. Prove that the sequence $\{f_n + g_n\}$ converges uniformly to $f + g$ on S.
(d) Suppose $\{f_n\}$ converges uniformly to f on S, and let c be a constant. Show that $\{cf_n\}$ converges uniformly to cf on S.
(e) Let $S = \mathbb{R}$, and set $f_n(x) = x + (1/n)$. Does $\{f_n\}$ converge uniformly on S? Does $\{f_n^2\}$ converge uniformly on S? What does this say about the limit of a product of uniformly convergent sequences versus the product of the limits?
(f) Suppose a and b are nonnegative real numbers and that $|a - b| < \epsilon^2$. Prove that $|\sqrt{a} - \sqrt{b}| < 2\epsilon$.
HINT: Break this into cases, the first one being when both $\sqrt{a}$ and $\sqrt{b}$ are less than ϵ.
(g) Suppose $\{f_n\}$ is a sequence of nonnegative real-valued functions that converges uniformly to f on S. Use part (f) to prove that the sequence $\{\sqrt{f_n}\}$ converges uniformly to $\sqrt{f}$.
(h) For each positive integer n, define f_n on $(-1,1)$ by $f_n(x) = |x|^{1+1/n}$. Prove that the sequence $\{f_n\}$ converges uniformly on $(-1,1)$ to the function $f(x) = |x|$.
HINT: Let $\epsilon > 0$ be given. Consider $|x|$'s that are $< \epsilon$ and $|x|$'s that are $\geq \epsilon$. For $|x| < \epsilon$, show that $|f_n(x) - f(x)| < \epsilon$ for all n. For $|x| \geq \epsilon$, choose N so that $|\epsilon^{1/n} - 1| < \epsilon$. How?

Exercise 3.29. Let $\{f_n\}$ be a sequence of functions on a set S, let f be a function on S, and suppose that for each n we have $|f(x) - f_n(x)| < 1/n$ for all $x \in S$. Prove that the sequence $\{f_n\}$ converges uniformly to f.

We give next four important theorems concerning uniform convergence. The first of these theorems is frequently used to prove that a given function is continuous. The theorem asserts that if f is the uniform limit of a sequence of continuous functions, then f is itself continuous.

THEOREM 3.17. (The uniform limit of continuous functions is continuous.) Suppose $\{f_n\}$ is a sequence of continuous functions on a set $S \subseteq \mathbb{C}$, and assume that the sequence $\{f_n\}$ converges uniformly to a function f. Then f is continuous on S.

PROOF. This proof is an example of what is called by mathematicians a "3ϵ argument."

Fix an $x \in S$ and an $\epsilon > 0$. We wish to find a $\delta > 0$ such that if $y \in S$ and $|y-x| < \delta$ then $|f(y) - f(x)| < \epsilon$.

We use first the hypothesis that the sequence converges uniformly. Thus, given this $\epsilon > 0$, there exists a natural number N such that if $n \geq N$ then $|f(z) - f_n(z)| < \epsilon/3$ for all $z \in S$. Now, because f_N is continuous at x, there exists a $\delta > 0$ such that if $y \in S$ and $|y-x| < \delta$ then $|f_N(y) - f_N(x)| < \epsilon/3$. So, if $y \in S$ and $|y-x| < \delta$, then

$$\begin{aligned}
|f(y) - f(x)| &= |f(y) - f_N(y) + f_N(y) - f_N(x) + f_N(x) - f(x)| \\
&\leq |f(y) - f_N(y)| + |f_N(y) - f_N(x)| + |f_N(x) - f(x)| \\
&< \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} \\
&= \epsilon.
\end{aligned}$$

This completes the proof.

REMARK. Many properties of functions are preserved under the taking of uniform limits, e.g., continuity, as we have just seen. However, not all properties are preserved under this limit process. Differentiability is not, integrability is sometimes, being a power series function is, and so on. We must be alert to be aware of when it works and when it does not.

THEOREM 3.18. (Weierstrass M-Test) Let $\{u_n\}$ be a sequence of complex-valued functions defined on a set $S \subseteq \mathbb{C}$. Write S_N for the partial sum $S_N(x) = \sum_{n=0}^{N} u_n(x)$. Suppose that, for each n, there exists an $M_n > 0$ for which $|u_n(x)| \leq M_n$ for all $x \in S$. Then

(1) If $\sum M_n$ converges, then the sequence $\{S_N\}$ converges uniformly to a function S. That is, the infinite series $\sum u_n$ converges uniformly.

(2) If each function u_n is continuous, and $\sum M_n$ converges, then the function S of part (1) is continuous.

PROOF. Because $\sum M_n$ is convergent, it follows from the Comparison Test that for each $x \in S$ the infinite series $\sum_{n=0}^{\infty} u_n(x)$ is absolutely convergent, hence convergent. Define a function S by $S(x) = \sum_{n=0}^{\infty} u_n(x) = \lim S_N(x)$.

To show that $\{S_N\}$ converges uniformly to S, let $\epsilon > 0$ be given, and choose a natural number N such that $\sum_{n=N+1}^{\infty} M_n < \epsilon$. This can be done because $\sum M_n$

converges. Now, for any $x \in S$ and any $m \geq N$, we have

$$\begin{aligned}
|S(x) - S_m(x)| &= |\lim_{k\to\infty} S_k(x) - S_m(x)| \\
&= |\lim_{k\to\infty} (S_k(x) - S_m(x))| \\
&= \lim_{k\to\infty} |S_k(x) - S_m(x)| \\
&= \lim_{k\to\infty} |\sum_{n=m+1}^{k} u_n(x)| \\
&\leq \lim_{k\to\infty} \sum_{n=m+1}^{k} |u_n(x)| \\
&\leq \lim_{k\to\infty} \sum_{n=m+1}^{k} M_n \\
&= \sum_{n=m+1}^{\infty} M_n \\
&\leq \sum_{n=N}^{\infty} M_n \\
&< \epsilon.
\end{aligned}$$

This proves part (1).
Part (2) now follows from part (1) and Theorem 3.17, since the S_N's are continuous.

THEOREM 3.19. *Let $f(z) = \sum_{n=0}^{\infty} a_n z^n$ be a power series function with radius of convergence $r > 0$, and let $\{S_N(z)\}$ denote the sequence of partial sums of this series:*

$$S_N(z) = \sum_{n=0}^{N} a_n z^n.$$

If $0 < r' < r$, then the sequence $\{S_N\}$ converges uniformly to f on the disk$B_{r'}(0)$.

PROOF. Define a power series function g by $g(z) = \sum_{n=0}^{\infty} |a_n| z^n$, and note that the radius of convergence for g is the same as that for f, i.e., r. Choose t so that $r' < t < r$. Then, since t belongs to the disk of convergence of the power series function g, we know that $\sum_{n=0}^{\infty} |a_n| t^n$ converges. Set $m_n = |a_n| t^n$, and note that $\sum m_n$ converges. Now, for each $z \in B_{r'}(0)$, we have that

$$|a_n z^n| \leq |a_n| {r'}^n \leq |a_n| t^n = m_n,$$

so that the infinite series $\sum a_n z^n$ converges uniformly on $B_{r'}(0)$ by the Weierstrass M-Test.

Exercise 3.30. Let $f(z) = \sum_{n=0}^{\infty} z^n$. Recall that the radius of convergence for f is 1. Verify that the sequence $\{S_N\}$ of partial sums of this power series function fails to converge uniformly on the full open disk of convergence $B_1(0)$, so that the requirement that $r' < r$ is necessary in the preceding theorem.

The next theorem shows that continuous, real-valued functions on closed bounded intervals are uniform limits of step functions. Step functions have not been mentioned lately, since they aren't continuous functions, but this next theorem will be crucial for us when we study integration in Chapter V.

THEOREM 3.20. *Let f be a continuous real-valued function on the closed and bounded interval $[a, b]$. Then there exists a sequence $\{h_n\}$ of step functions on $[a, b]$ that converges uniformly to f.*

PROOF. We use the fact that a continuous function on a compact set is uniformly continuous (Theorem 3.9).

For each positive integer n, let δ_n be a positive number satisfying $|f(x)-f(y)| < 1/n$ if $|x - y| < \delta_n$. Such a δ_n exists by the uniform continuity of f on $[a, b]$. Let $P_n = \{x_0 < x_1 < \ldots < x_{m_n}\}$ be a partition of $[a, b]$ for which $x_i - x_{i-1} < \delta_n$ for all $1 \leq i \leq m_n$. Define a step function h_n on $[a, b]$ as follows: If $x_{i-1} \leq x < x_i$, then $h_n(x) = f(x_{i-1})$. This defines $h_n(x)$ for every $x \in [a, b)$, and we complete the definition of h_n by setting $h_n(b) = f(b)$. It follows immediately that h_n is a step function.

Now, we claim that $|f(x) - h_n(x)| < 1/n$ for all $x \in [a, b]$. This is clearly the case for $x = b$, since $f(b) = h_n(b)$ for all n. For any other x, let i be the unique index such that $x_{i-1} \leq x < x_i$. Then

$$|f(x) - h_n(x)| = |f(x) - f(x_{i-1})| < 1/n$$

because $|x - x_{i-1}| < \delta_n$.

So, we have defined a sequence $\{h_n\}$ of step functions, and the sequence $\{h_n\}$ converges uniformly to f by Exercise 3.29.

We close this chapter with a famous theorem of Abel concerning the behavior of a power series function on the boundary of its disk of convergence. See the comments following Exercise 3.19.

THEOREM 3.21. (Abel) Suppose $f(z) = \sum_{n=0}^{\infty} a_n z^n$ is a power series function having finite radius of convergence $r > 0$, and suppose there exists a point z_0 on the boundary of $B_r(0)$ that is in the domain of f; i.e., $\sum a_n z_0^n$ converges to $f(z_0)$. Suppose g is a continuous function whose domain contains the open disk $B_r(0)$ as well as the point z_0, and assume that $f(z) = g(z)$ for all z in the open disk $B_r(0)$. Then $f(z_0)$ must equal $g(z_0)$.

PROOF. For simplicity, assume that $r = 1$ and that $z_0 = 1$. See the exercise that follows this proof. Write S_n for the partial sum of the a_n's: $S_n = \sum_{n=0}^{n} a_n$. In the following computation, we will use the Abel Summation Formula in the form

$$\sum_{n=0}^{N} a_n z^n = S_N z^N + \sum_{n=0}^{N-1} S_n (z^n - z^{n+1}).$$

See Exercise 2.30. Let ϵ be a positive number. Then, for any $0 < t < 1$ and any positive integer N, we have

$$\begin{aligned}
|g(1) - f(1)| &= |g(1) - f(t) + f(t) - \sum_{n=0}^{N} a_n t^n + \sum_{n=0}^{N} a_n t^n - f(1)| \\
&\leq |g(1) - g(t)| + |f(t) - \sum_{n=0}^{N} a_n t^n| + |\sum_{n=0}^{N} a_n t^n - f(1)| \\
&\leq |g(1) - g(t)| + |f(t) - \sum_{n=0}^{N} a_n t^n| \\
&\qquad + |S_N t^N + \sum_{n=0}^{N-1} S_n (t^n - t^{n+1}) - f(1)| \\
&= |g(1) - g(t)| + |f(t) - \sum_{n=0}^{N} a_n t^n \\
&\qquad + |S_N t^N + \sum_{n=0}^{N-1} (S_n - S_N)(t^n - t^{n+1}) + S_N \sum_{n=0}^{N-1} (t^n - t^{n+1}) - f(1)| \\
&= |g(1) - g(t)| + |f(t) - \sum_{n=0}^{N} a_n t^n \\
&\qquad + |\sum_{n=0}^{N-1} (S_n - S_N)(t^n - t^{n+1}) + S_N (t^N + \sum_{n=0}^{N-1} (t^n - t^{n+1})) - f(1)| \\
&\leq |g(1) - g(t)| + |f(t) - \sum_{n=0}^{N} a_n t^n| \\
&\qquad + |\sum_{n=0}^{N-1} (S_n - S_N)(t^n - t^{n+1})| + |S_N - f(1)| \\
&\leq |g(1) - g(t)| + |f(t) - \sum_{n=0}^{N} a_n t^n| \\
&\qquad + |\sum_{n=0}^{P} (S_n - S_N)(t^n - t^{n+1})| + |\sum_{n=P+1}^{N-1} (S_n - S_N)(t^n - t^{n+1})| + |S_N - f(1)| \\
&\leq |g(1) - g(t)| + |f(t) - \sum_{n=0}^{N} a_n t^n| \\
&\qquad + |\sum_{n=0}^{P} (S_n - S_N)(t^n - t^{n+1})| + \sum_{n=P+1}^{N-1} |S_n - S_N|(t^n - t^{n+1}) + |S_N - f(1)| \\
&= t_1 + t_2 + t_3 + t_4 + t_5.
\end{aligned}$$

First, choose an integer M_1 so that if P and N are both larger than M_1, then $t_4 < \epsilon$. (The sequence $\{S_k\}$ is a Cauchy sequence, and $\sum (t^k - t^{k+1}$ is telescoping.) Fix such a $P > M_1$. Then choose a $\delta > 0$ so that if $1 > t > 1 - \delta$, then both t_1 and $t_3 < \epsilon$. How?

Fix such a t. Finally, choose a N, greater than M_1, and also large enough so that both t_2 and t_5 are less than ϵ. (How?)
Now, $|g(1)-f(1)| < 5\epsilon$. Since this is true for every $\epsilon > 0$, it follows that $f(1) = g(1)$, and the theorem is proved.

Exercise 3.31. Let f, g, r, and z_0 be as in the statement of the preceding theorem. Define $\hat{f}(z) = f(z_0 z)$ and $\hat{g}(z) = g(z_0 z)$.
(a) Prove that $\hat{f}$ is a power series function $\hat{f}(z) = \sum_{n=0}^{\infty} b_n z^n$, with radius of convergence equal to 1, and such that $\sum_{n=0}^{\infty} b_n$ converges to $\hat{f}(1)$; i.e., 1 is in the domain of $\hat{f}$.
(b) Show that $\hat{g}$ is a continuous function whose domain contains the open disk $B_1(0)$ and the point $z = 1$.
(c) Show that, if $\hat{f}(1) = \hat{g}(1)$, then $f(z_0) = g(z_0)$. Deduce that the simplification in the preceding proof is justified.
(d) State and prove the generalization of Abel's Theorem to a function f that is expandable in a Taylor series around a point c.

CHAPTER IV
DIFFERENTIATION, LOCAL BEHAVIOR
$e^{i\pi} = -1.$

In this chapter we will finally see why $e^{i\pi}$ is -1. Along the way, we will give careful proofs of all the standard theorems of Differential Calculus, and in the process we will discover all the familiar facts about the trigonometric and exponential functions. At this point, we only know their definitions as power series functions. The fact that $\sin^2 + \cos^2 = 1$ or that $e^{x+y} = e^x e^y$ are not at all obvious. In fact, we haven't even yet defined what is meant by e^x for an arbitrary number x.
The main theorems of this chapter include:

(1) The **Chain Rule** (Theorem 4.7),
(2) The **Mean Value Theorem** (Theorem 4.9),
(3) The **Inverse Function Theorem** (Theorem 4.10),
(4) The **Laws of Exponents** (Corollary to Theorem 4.11 and Exercise 4.20), and
(5) **Taylor's Remainder Theorem** (Theorem 4.19).

THE LIMIT OF A FUNCTION

The concept of the derivative of a function is what most people think of as the beginning of calculus. However, before we can even define the derivative we must introduce a kind of generalization of the notion of continuity. That is, we must begin with the definition of the limit of a function.

DEFINITION. Let $f : S \to \mathbb{C}$ be a function, where $S \subseteq \mathbb{C}$, and let c be a limit point of S that is not necessarily an element of S. We say that f *has a limit* L *as* z *approaches* c, and we write

$$\lim_{z \to c} f(z) = L,$$

if for every $\epsilon > 0$ there exists a $\delta > 0$ such that if $z \in S$ and $0 < |z - c| < \delta$, then $|f(z) - L| < \epsilon$.
If the domain S is unbounded, we say that f *has a limit* L *as* z *approaches* ∞, and we write

$$L = \lim_{z \to \infty} f(z),$$

if for every $\epsilon > 0$ there exists a positive number B such that if $z \in S$ and $|z| \geq B$, then $|f(z) - L| < \epsilon$.
Analogously, if $S \subseteq \mathbb{R}$, we say $\lim_{x\to\infty} f(x) = L$ if for every $\epsilon > 0$ there exists a real number B such that if $x \in S$ and $x \geq B$, then $|f(x) - L| < \epsilon$. And we say that $\lim_{x\to-\infty} f(x) = L$ if for every $\epsilon > 0$ there exists a real number B such that if $x \in S$ and $x \leq B$, then $|f(x) - L| < \epsilon$.
Finally, for $f : (a, b) \to \mathbb{C}$ a function of a real variable, and for $c \in [a, b]$, we define the *one-sided* (left and right) limits of f at c. We say that f has a *left hand limit* of L at c, and we write $L = \lim_{x\to c-0} f(x)$, if for every $\epsilon > 0$ there exists a $\delta > 0$ such that if $x \in (a, b)$ and $0 < c - x < \delta$ then $|f(x) - L| < \epsilon$. We say that f has a *right hand limit* of L at c, and write $L = \lim_{x\to c+0} f(x)$, if for every $\epsilon > 0$ there exists a $\delta > 0$ such that if $x \in S$ and $0 < x - c < \delta$ then $|f(x) - L| < \epsilon$.

The first few results about limits of functions are not surprising. The analogy between functions having limits and functions being continuous is very close, so that for every elementary result about continuous functions there will be a companion result about limits of functions.

THEOREM 4.1. *Let c be a complex number. Let $f : S \to \mathbb{C}$ and $g : S \to \mathbb{C}$ be functions. Assume that both $f$and g have limits as x approaches c. Then:*

(1) *There exists a $\delta > 0$ and a positive number M such that if $z \in S$ and $0 < |z - c| < \delta$ then $|f(z)| < M$. That is, if f has a limit as z approaches c, then f is bounded near c.*

(2)
$$\lim_{z\to c}(f(z) + g(z)) = \lim_{z\to c} f(z) + \lim_{z\to c} g(z).$$

(3)
$$\lim_{z\to c}(f(z)g(z)) = \lim_{z\to c} f(z) \lim_{z\to c} g(z).$$

(4) *If $\lim_{z\to c} g(z) \neq 0$, then*
$$\lim_{z\to c} \frac{f(z)}{g(z)} = \frac{\lim_{z\to c} f(z)}{\lim_{z\to c} g(z)},$$

(5) *If u and v are the real and imaginary parts of a complex-valued function f, then u and v have limits as z approaches c if and only if f has a limit as z approaches c. And,*
$$\lim_{z\to c} f(z) = \lim_{z\to c} u(z) + i \lim_{z\to c} v(z).$$

Exercise 4.1. (a) Prove Theorem 4.1.
HINT: Compare with Theorem 3.2.
(b) Prove that $\lim_{x\to c} f(x) = L$ if and only if, for every sequence $\{x_n\}$ of elements of S that converges to c, we have $\lim f(x_n) = L$.
HINT: Compare with Theorem 3.4.
(c) Prove the analog of Theorem 4.1 replacing the limit as z approaches c by the limit as z approaches ∞.

Exercise 4.2. (a) Prove that a function $f : S \to \mathbb{C}$ is continuous at a point c of S if and only if $\lim_{x\to c} f(x) = f(c)$.
HINT: Carefully write down both definitions, and observe that they are verbetim the same.
(b) Let f be a function with domain S, and let c be a limit point of S that is not in S. Suppose g is a function with domain $S \cup \{c\}$, that $f(x) = g(x)$ for all $x \in S$, and that g is continuous at c. Prove that $\lim_{x\to c} f(x) = g(c)$.

Exercise 4.3. Prove that the following functions f have the specified limits L at the given points c.
(a) $f(x) = (x^3 - 8)/(x^2 - 4)$, $c = 2$, and $L = 3$.
(b) $f(x) = (x^2 + 1)/(x^3 + 1)$, $c = 1$, and $L = 1$.
(c) $f(x) = (x^8 - 1)/(x^6 + 1)$, $c = i$, and $L = -4/3$.
(d) $f(x) = (\sin(x) + \cos(x) - \exp(x))/(x^2)$, $c = 0$, and $L = -1$.

Exercise 4.4. Define f on the set S of all nonzero real numbers by $f(x) = c$ if $x < 0$ and $f(x) = d$ if $x > 0$. Show that $\lim_{x\to 0} f(x)$ exists if and only if $c = d$.

(b) Let $f : (a, b) \to \mathbb{C}$ be a complex-valued function on the open interval (a, b). Suppose c is a point of (a, b). Prove that $\lim_{x \to c} f(x)$ exists if and only if the two one-sided limits $\lim_{x \to c-0} f(x)$ and $\lim_{x \to c+0} f(x)$ exist and are equal.

Exercise 4.5. (Change of variable in a limit) Suppose $f : S \to \mathbb{C}$ is a function, and that $\lim_{x \to c} f(x) = L$. Define a function g by $g(y) = f(y + c)$.
(a) What is the domain of g?
(b) Show that 0 is a limit point of the domain of g and that $\lim_{y \to 0} g(y) = \lim_{x \to c} f(x)$.
(c) Suppose $T \subseteq \mathbb{C}$, that $h : T \to S$, and that $\lim_{y \to d} h(y) = c$. Prove that

$$\lim_{y \to d} f(h(y)) = \lim_{x \to c} f(x) = L.$$

REMARK. When we use the word " interior" in connection with a set S, it is obviously important to understand the context; i.e., is S being thought of as a set of real numbers or as a set of complex numbers. A point c is in the interior of a set S of complex numbers if the entire disk $B_\epsilon(c)$ of radius ϵ around c is contained in S. While, a point c belongs to the interior of a set S of real numbers if the entire interval $(c - \epsilon, c + \epsilon)$ is contained in S. Hence, in the following definition, we will be careful to distinguish between the cases that f is a function of a real variable or is a function of a complex variable.

THE DERIVATIVE OF A FUNCTION

Now begins what is ordinarily thought of as the first main subject of calculus, the derivative.

DEFINITION. Let S be a subset of $\mathbb{R}$, let $f : S \to \mathbb{C}$ be a complex-valued function (of a real variable), and let c be an element of the interior of S. We say that f is *differentiable at c* if

$$\lim_{h \to 0} \frac{f(c + h) - f(c)}{h}$$

exists. (Here, the number h is a real number.)
Analogously, let S be a subset of $\mathbb{C}$, let $f : S \to \mathbb{C}$ be a complex-valued function (of a complex variable), and let c be an element of the interior of S. We say that f is *differentiable at c* if

$$\lim_{h \to 0} \frac{f(c + h) - f(c)}{h}$$

exists. (Here, the number h is a complex number.)
If $f : S \to \mathbb{C}$ is a function either of a real variable or a complex variable, and if S' denotes the subset of S consisting of the points c where f is differentiable, we define a function $f' : S' \to \mathbb{C}$ by

$$f'(x) = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h}.$$

The function f' is called the *derivative* of f.
A continuous function $f : [a, b] \to \mathbb{C}$ that is differentiable at each point $x \in (a, b)$, and whose the derivative f' is continuous on (a, b), is called a *smooth* function on

$[a,b]$. If there exists a partition $\{a = x_0 < x_1 < \ldots < x_n = b\}$ of $[a,b]$ such that f is smooth on each subinterval $[x_{i-1}, x_i]$, then f is called *piecewise smooth* on $[a,b]$. Higher order derivatives are defined inductively. That is, f'' is the derivative of f', and so on. We use the symbol $f^{(n)}$ for the nth derivative of f.

REMARK. In the definition of the derivative of a function f, we are interested in the limit, as h approaches 0, not of f but of the quotient $q(h) = \frac{f(c+h)-f(c)}{h}$. Notice that 0 is not in the domain of the function q, but 0 is a limit point of that domain. This is the reason why we had to make such a big deal above out of the limit of a function. The function q is often called the *differential quotient.*

REMARK. As mentioned in Chapter III, we are often interested in solving for unknowns that are functions. The most common such problem is to solve a differential equation. In such a problem, there is an unknown function for which there is some kind of relationship between it and its derivatives. Differential equations can be extremely complicated, and many are unsolvable. However, we will have to consider certain relatively simple ones in this chapter, e.g., $f' = f$, $f' = -f$, and $f'' = \pm f$. There are various equivalent ways to formulate the definition of differentiable, and each of these ways has its advantages. The next theorem presents one of those alternative ways.

THEOREM 4.2. *Let c belong to the interior of a set S (either in $\mathbb{R}$ or in $\mathbb{C}$), and let $f : S \to \mathbb{C}$ be a function. Then the following are equivalent.*

(1) *f is differentiable at c. That is,*

$$\lim_{h\to 0} \frac{f(c+h)-f(c)}{h} \textit{ exists.}$$

(2)

$$\lim_{x\to c} \frac{f(x)-f(c)}{x-c} \textit{ exists.}$$

(3) *There exists a number L and a function θ such that the following two conditions hold:*

$$f(c+h) - f(c) = Lh + \theta(h) \tag{4.1}$$

and

$$\lim_{h\to 0} \frac{\theta(h)}{h} = 0. \tag{4.2}$$

In this case, L is unique and equals $f'(c)$, and the function θ is unique and equals $f(c+h) - f(c) - f'(c)h$.

PROOF. That (1) and (2) are equivalent follows from Exercise 4.5 by writing x as $c+h$.

Suppose next that f is differentiable at c, and define

$$L = f'(c) = \lim_{h\to 0} \frac{f(c+h)-f(c)}{h}.$$

Set

$$\theta(h) = f(c+h) - f(c) - f'(c)h.$$

Then clearly

$$f(c+h) - f(c) = Lh + \theta(h),$$

which is Equation (4.1). Also

$$\begin{aligned}|\frac{\theta(h)}{h}| &= |\frac{f(c+h) - f(c) - f'(c)h}{h}| \\ &= |\frac{f(c+h) - f(c)}{h} - f'(c)|,\end{aligned}$$

which tends to 0 as h approaches 0 because f is differentiable at c. Hence, we have established equations (4.1) and (4.2), showing that (1) implies (3).

Finally, suppose there is a number L and a function θ satisfying Equations (4.1) and (4.2). Then

$$\frac{f(c+h) - f(c)}{h} = L + \frac{\theta(h)}{h},$$

which converges to L as h approaches 0 by Equation (4.2) and part (2) of Theorem 4.1. Hence, $L = f'(c)$, and so $\theta(h) = f(c+h) - f(c) - f'(c)h$. Therefore, (3) implies (1), and the theorem is proved.

REMARK. Though it seems artificial and awkward, Condition (3) of this theorem is very convenient for many proofs. One should remember it.

Exercise 4.6. (a) What is the domain of the function θ of condition (3) in the preceding theorem? Is 0 in this domain? Are there any points in the interior of this domain?

(b) Let L and θ be as in part (3) of the preceding theorem. Prove that, given an $\epsilon > 0$ there exists a $\delta > 0$ such that if $|h| < \delta$ then $|\theta(h)| < \epsilon|h|$.

THEOREM 4.3. *If $f : S \to \mathbb{C}$ is a function, either of a real variable or a complex variable, and if f is differentiable at a point c of S, then f is continuous at c. That is, differentiability implies continuity.*

PROOF. We are assuming that $\lim_{h\to 0}(f(c+h) - f(c))/h = L$. Hence, there exists a positive number δ_0 such that $|\frac{f(c+h)-f(c)}{h} - L| < 1$ if $|h| < \delta_0$, implying that $|f(c+h) - f(c)| < |h|(|L| + 1)$whenever $|h| < \delta_0$. So, if $\epsilon > 0$ is given, let δ be the minimum of δ_0 and $\epsilon/(|L|+1)$. If $y \in S$ and $|y - c| < \delta$, then, thinking of y as being $c + h$,

$$|f(y) - f(c)| = |f(c+h) - f(c)| < |h|(|L| + 1) = |y - c|(|L| + 1) < \epsilon.$$

(Every y can be written as $c + h$ for some h, and $|y - c| = |h|$.)

Exercise 4.7. Define $f(z) = |z|$ for $z \in \mathbb{C}$.

(a) Prove that f is continuous at every point of $\mathbb{C}$.

(b) Show that, if f is differentiable at a point c, then $f'(c) = 0$.

HINT: Using part (b) of Exercise 4.1, evaluate $f'(c)$ in the following two ways.

$$f'(c) = \lim_{n\to\infty} \frac{|c + \frac{1}{n}| - |c|}{\frac{1}{n}}$$

and

$$f'(c) = \lim_{n\to\infty} \frac{|c + \frac{i}{n}| - |c|}{\frac{i}{n}}.$$

Show that the only way these two limits can be equal is for them to be 0.
(c) Conclude that f is not differentiable anywhere. Indeed, if it were, what would the function θ have to be, and why wouldn't it satisfy Equation 4.2?
(d) Suppose $f : \mathbb{R} \to \mathbb{R}$ is the function of a real variable that is defined by $f(x) = |x|$. Show that f is differentiable at every point $x \neq 0$. How does this result not contradict part (c)?

The following theorem generalizes the preceding exercise.

THEOREM 4.4. *Suppose $f : S \to \mathbb{R}$ is a real-valued function of a complex variable, and assume that f is differentiable at a point $c \in S$. Then $f'(c) = 0$. That is, every real-valued, differentiable function f of a complex variable satisfies $f'(c) = 0$ for all c in the domain of f'.*

PROOF. We compute $f'(c)$ in two ways.

$$f'(c) = \lim_n \frac{f(c + \frac{1}{n}) - f(c)}{\frac{1}{n}} \text{ is a real number..}$$

$$f'(c) = \lim_n \frac{f(c + \frac{i}{n}) - f(c)}{\frac{i}{n}} \text{ is a purely imaginary number.}$$

Hence, $f'(c)$ must be 0, as claimed.

REMARK. This theorem may come as a surprise, for it shows that there are very few real-valued differentiable functions of a complex variable. For this reason, whenever $f : S \to \mathbb{R}$ is a real-valued, differentiable function, we will presume that f is a function of a real variable; i.e., that the domain $S \subseteq \mathbb{R}$.
Evaluating $\lim_{h\to 0} q(h)$ in the two different ways, h real, and h pure imaginary, led to the proof of the last theorem. It also leads us to make definitions of what are called "partial derivatives" of real-valued functions whose domains are subsets of $\mathbb{C} \equiv \mathbb{R}^2$. As the next exercise will show, the theory of partial derivatives of real-valued functions is a much richer theory than that of standard derivatives of real-valued functions of a single complex variable.

DEFINITION. Let $f : S \to \mathbb{R}$ be defined on a set $S \subseteq \mathbb{C} \equiv \mathbb{R}^2$, and let $c = (a, b) = + + bi$ be a point in the interior of S. We define the *partial derivative of f with respect to x* at the point $c = (a, b)$ by the formula

$$\frac{\partial f}{\partial x}(a, b) = \lim_{h\to 0} \frac{f(a + h, b) - f(a, b)}{h},$$

and the *partial derivative of f with respect to y* at $c = (a, b)$ by the formula

$$\frac{\partial f}{\partial y}(a, b) = \lim_{h\to 0} \frac{f(a, b + h) - f(a, b)}{h},$$

whenever these limits exist. (In both these limits, the variable h is a real variable.)(

It is clear that the partial derivatives of a function arise when we fix either the real part of the variable or the imaginary part of the variable to be a constant, and then consider the resulting function of the other (real) variable. We will see in Exercise 4.8 that there is a definite difference between a function's being differentiable at a point $c = (a + bi)$ in the complex plane $\mathbb{C}$ versus its having partial derivatives at the point (a, b) in $\mathbb{R}^2$.

Exercise 4.8. (a) Suppose f is a complex-valued function of a complex variable, and assume that both the real and imaginary parts of f are differentiable at a point c. Show that f is differentiable at c and that $f'(c) = 0$.

(b) Let $f = u + iv$ be a complex-valued function of a complex variable that is differentiable at a point c. Prove that both partial derivatives of u and v exist at $c = (a, b)$, and in fact that

$$\frac{\partial u}{\partial x}(c) + i\frac{\partial v}{\partial x}(c) = f'(c)$$

and

$$\frac{\partial u}{\partial y}(c) + i\frac{\partial v}{\partial y}(c) = if'(c).$$

(c) Define a complex-valued function f on $\mathbb{C} \equiv \mathbb{R}^2$ by $f(z) = f(x + iy) = x - iy$. Write $f = u + iv$, and show that both partial derivatives of u and v exist at every point, but that f is not a differentiable function of the complex variable z.

The next theorem is, in part, what we call in calculus the "differentiation formulas."

THEOREM 4.5. *Let f and g be functions (either of a real variable or a complex variable), which are both differentiable at a point c. Let a and b be complex numbers. Then:*

(1) *$af + bg$ is differentiable at c, and $(af + bg)'(c) = af'(c) + bg'(c)$.*

(2) (Product Formula) fg is differentiable at c, and $(fg)'(c) = f'(c)g(c) + f(c)g'(c)$.

(3) (Quotient Formula) f/g is differentiable at c (providing that $g(c) \neq 0$), and
$$(\frac{f}{g})'(c) = \frac{g(c)f'(c) - f(c)g'(c)}{(g(c))^2}.$$

(4) If $f = u+iv$ is a complex-valued function, then f is differentiable at a point c if and only if u and v are differentiable at c, and $f'(c) = u'(c) + iv'(c)$.

PROOF. We prove part (2) and leave parts (1), (3), and (4) for the exercises. We have

$$\begin{aligned}\lim_{h\to 0}\frac{(fg)(c+h) - (fg)(c)}{h} &= \lim_{h\to 0}\frac{f(c+h)g(c+h) - f(c)g(c)}{h}\\ &= \lim_{h\to 0}\frac{f(c+h)g(c+h) - f(c)g(c+h)}{h}\\ &\quad + \lim_{h\to 0}\frac{f(c)g(c+h) - f(c)g(c)}{h}\\ &= \lim_{h\to 0}\frac{f(c+h) - f(c)}{h}\lim_{h\to 0}g(c+h)\\ &\quad + \lim_{h\to 0}f(c)\lim_{h\to 0}\frac{g(c+h) - g(c)}{h}\\ &= f'(c)g(c) + f(c)g'(c),\end{aligned}$$

where we have used Theorems 4.1, 4.2, and 4.3.

Exercise 4.9. (a) Prove parts (1), (3), and (4) of Theorem 4.5.

(b) If f and g are real-valued functions that are differentiable at a point c, what can be said about the differentiability of $\max(f, g)$?

(c) Let f be a constant function $f(z) \equiv k$. Prove that f is differentiable everywhere and that $f'(z) = 0$ for all z.

(d) Define a function f by $f(z) = z$. Prove that f is differentiable everywhere and that $f'(z) = 1$ for all z.

(e) Verify the usual derivative formulas for polynomial functions: If $p(z) = \sum_{k=0}^{n} a_k z^k$, then $p'(z) = \sum_{k=1}^{n} k a_k z^{k-1}$.

What about power series functions? Are they differentiable functions? If so, are their derivatives again power series functions? In fact, everything works as expected.

THEOREM 4.6. *Let f be a power series function $f(z) = \sum_{n=0}^{\infty} a_n z^n$ having radius of convergence $r > 0$. Then f is differentiable at each point z in its open disk $B_r(0)$ of convergence, and*

$$f'(z) = \sum_{n=0}^{\infty} n a_n z^{n-1} = \sum_{n=1}^{\infty} n a_n z^{n-1}.$$

PROOF. The proof will use part (3) of Theorem 4.2. Fix an z with $|z| < r$. Choose r' so that $|z| < r' < r$, and write α for $r' - |z|$, i.e., $|z| + \alpha = r'$. Note first that the infinite series $\sum_{n=0}^{\infty} |a_n| r'^n$ converges to a positive number we will call M. Also, from the Cauchy-Hadamard Formula, we know that the power series function $\sum n a_n w^n$ has the same radius of convergence as does f, and hence the infinite series $\sum n a_n z^{n-1}$ converges to a number we will denote by L. We define a function θ by $\theta(h) = f(z+h) - f(z) - Lh$ from which it follows immediately that

$$f(z+h) - f(z) = Lh + \theta(h),$$

which establishes Equation (4.1). To complete the proof that f is differentiable at z, it will suffice to establish Equation (4.2), i.e., to show that

$$\lim_{h \to 0} \frac{\theta(h)}{h} = 0.$$

That is, given $\epsilon > 0$ we must show that there exists a $\delta > 0$ such that if $0 < |h| < \delta$ then

$$|\theta(h)/h| = |\frac{f(z+h) - f(z)}{h} - L| < \epsilon.$$

Assuming, without loss of generality, that $|h| < \alpha$, we have that

$$
\begin{aligned}
|\frac{f(z+h)-f(z)}{h} - L| &= |\frac{\sum_{n=0}^{\infty} a_n(z+h)^n - \sum_{n=0}^{\infty} a_n z^n}{h} - L| \\
&= |\frac{\sum_{n=0}^{\infty} a_n(\sum_{k=0}^{n} \binom{n}{k} z^{n-k}h^k) - \sum_{n=0}^{\infty} a_n z^n}{h} - L| \\
&= |\frac{\sum_{n=0}^{\infty} a_n((\sum_{k=0}^{n} \binom{n}{k} z^{n-k}h^k) - z^n)}{h} - L| \\
&= |\frac{\sum_{n=1}^{\infty} a_n(\sum_{k=1}^{n} \binom{n}{k} z^{n-k}h^k)}{h} - L| \\
&= |\sum_{n=1}^{\infty} a_n(\sum_{k=1}^{n} \binom{n}{k} z^{n-k}h^{k-1}) - \sum_{n=1}^{\infty} na_n z^{n-1}| \\
&= |\sum_{n=1}^{\infty} a_n(\sum_{k=1}^{n} \binom{n}{k} z^{n-k}h^{k-1}) - \sum_{n=1}^{\infty} \binom{n}{1} a_n z^{n-1}| \\
&= |\sum_{n=2}^{\infty} a_n(\sum_{k=2}^{n} \binom{n}{k} z^{n-k}h^{k-1})| \\
&\leq \sum_{n=2}^{\infty}\sum_{k=2}^{n} |a_n|\binom{n}{k}|z|^{n-k}|h|^{k-1} \\
&\leq |h|\sum_{n=2}^{\infty}|a_n|\sum_{k=2}^{n}\binom{n}{k}|z|^{n-k}|h|^{k-2} \\
&\leq |h|\sum_{n=2}^{\infty}|a_n|\sum_{k=2}^{n}\binom{n}{k}|z|^{n-k}|\alpha|^{k-2} \\
&\leq |h|\frac{1}{\alpha^2}\sum_{n=0}^{\infty}|a_n|\sum_{k=0}^{n}\binom{n}{k}|z|^{n-k}\alpha^{k} \\
&= |h|\frac{1}{\alpha^2}\sum_{n=0}^{\infty}|a_n|(|z|+\alpha)^n \\
&= |h|\frac{1}{\alpha^2}\sum_{n=0}^{\infty}|a_n|r'^n \\
&= |h|\frac{M}{\alpha^2},
\end{aligned}
$$

so that if $\delta = \epsilon/\frac{M}{\alpha^2}$, then $|\theta(h)/h| < \epsilon$, whenever $|h| < \delta$, as desired.

REMARK. Theorem 4.6 shows that indeed power series functions are differentiable, and in fact their derivatives can be computed, just like polynomials, by differentiating term by term. This is certainly a result we would have hoped was true, but the proof is **not** trivial.

The next theorem, the Chain Rule, is another nontrivial one. It deals with the differentiability of the composition of two differentiable functions. Again, the result is what we would have wanted, the composition of two differentiable functions is itself differentiable, but the argument required to prove it is tricky.

THEOREM 4.7. (Chain Rule) Let $f : S \to \mathbb{C}$ be a function, and assume that f is differentiable at a point c. Suppose $g : T \to \mathbb{C}$ is a function, that $T \subseteq \mathbb{C}$, that the number $f(c) \in T$, and that g is differentiable at $f(c)$. Then the composition $g \circ f$ is differentiable at c and

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

PROOF. Using part (3) of Theorem 4.2, write

$$g(f(c)+k) - g(f(c)) = L_g k + \theta_g(k)$$

and

$$f(c+h) - f(c) = L_f h + \theta_f(h).$$

We know from that theorem that $L_g = g'(f(c))$ and $L_f = f'(c)$. And, we also know that

$$\lim_{k\to 0} \frac{\theta_g(k)}{k} = 0 \text{ and } \lim_{h\to 0} \frac{\theta_f(h)}{h} = 0.$$

Define a function $k(h) = f(c+h) - f(c)$. Then, by Theorem 4.3, we have that $\lim_{h\to 0} k(h) = 0$. We will show that $g \circ f$ is differentiable at c by showing that there exists a number L and a function θ satisfying the two conditions of part (3) of Theorem 4.2. Thus, we have that

$$\begin{aligned} g \circ f(c+h) - g \circ f(c) &= g(f(c+h)) - g(f(c)) \\ &= g(f(c) + k(h)) - g(f(c)) \\ &= L_g k(h) + \theta_g(k(h)) \\ &= L_g(f(c+h) - f(c)) + \theta_g(k(h)) \\ &= L_g(L_f h + \theta_f(h)) + \theta_g(k(h)) \\ &= L_g L_f h + L_g \theta_f(h) + \theta_g(k(h)). \end{aligned}$$

We define $L = L_g l_f = g'(f(c))f'(c)$, and we define the function θ by

$$\theta(h) = L_g \theta_f(h) + \theta_g(k(h)).$$

By our definitions, we have established Equation (4.1)

$$g \circ f(c+h) - g \circ f(c) = Lh + \theta(h),$$

so that it remains to verify Equation (4.2).
We must show that, given $\epsilon > 0$, there exists a $\delta > 0$ such that if $0 < |h| < \delta$ then $|\theta(h)/h| < \epsilon$. First, choose an $\epsilon' > 0$ so that

$$\text{(4.3).} \qquad |L_g|\epsilon' + |L_f|\epsilon' + {\epsilon'}^2 < \epsilon$$

Next, using part (b) of Exercise 4.6, choose a $\delta' > 0$ such that if $|k| < \delta'$ then $|\theta_g(k)| < \epsilon'|k|$. Finally, choose $\delta > 0$ so that if $0 < |h| < \delta$, then the following two inequalities hold. $|k(h)| < \delta'$ and $|\theta_f(h)| < \epsilon'|h|$. The first can be satisfied because

f is continuous at c, and the second is a consequence of part (b) of Exercise 4.6. Then: if $0 < |h| < \delta$,

$$\begin{aligned}|\theta(h)| &= |L_g\theta_f(h) + \theta_g(k(h))| \\ &\leq |L_g||\theta_f(h)| + |\theta_g(k(h))| \\ &< |L_g|\epsilon'|h| + \epsilon'|k(h)| \\ &= |L_g|\epsilon'|h| + \epsilon'|f(c+h) - f(c)| \\ &= |L_g|\epsilon'|h| + \epsilon'|L_f h + \theta_f(h)| \\ &\leq |L_g|\epsilon'|h| + \epsilon'|L_f||h| + \epsilon'|\theta_f(h)| \\ &< |L_g|\epsilon'|h| + \epsilon'|L_f||h| + \epsilon'\epsilon'|h| \\ &= (|L_g|\epsilon' + |L_f|\epsilon' + {\epsilon'}^2)|h|,\end{aligned}$$

whence

$$|\theta(h)/h| < (|L_g|\epsilon' + |L_f|\epsilon' + {\epsilon'}^2) < \epsilon,$$

as desired.

Exercise 4.10. (a) Derive the familiar formulas for the derivatives of the elementary transcendental functions:

$$\exp' = \exp,\ \sin' = \cos,\ , sinh' = cosh,\ \cosh' = \sinh \text{ and } \cos' = -\sin.$$

(b) Define a function f as follows. $f(z) = \cos^2(z) + \sin^2(z)$. Use part (a) and the Chain Rule to show that $f'(z) = 0$ for all $z \in \mathbb{C}$. Does this imply that $\cos^2(z) + \sin^2(z) = 1$ for all complex numbers z?
(c) Suppose f is expandable in a Taylor series around the point $c : f(z) = \sum_{n=0}^{\infty} a_n(z-c)^n$ for all $z \in B_r(c)$. Prove that f is differentiable at each point of the open disk $B_r(c)$, and show that

$$f'(z) = \sum_{n=1}^{\infty} na_n(z-c)^{n-1}.$$

HINT: Use Theorem 4.6 and the chain rule.

CONSEQUENCES OF DIFFERENTIABILITY, THE MEAN VALUE THEOREM

DEFINITION. Let $f : S \to \mathbb{R}$ be a real-valued function of a real variable, and let c be an element of the interior of S. Then f is said to *attain a local maximum* at c if there exists a $\delta > 0$ such that $(c-\delta, c+\delta) \subseteq S$ and $f(c) \geq f(x)$ for all $x \in (c-\delta, c+\delta)$.
The function f is said to *attain a local minimum* at c if there exists an interval $(c-\delta, c+\delta) \subseteq S$ such that $f(c) \leq f(x)$ for all $x \in (c-\delta, c+\delta)$.

The next theorem should be a familiar result from calculus.

THEOREM 4.8. (First Derivative Test for Extreme Values) Let $f : S \to \mathbb{R}$ be a real-valued function of a real variable, and let $c \in S$ be a point at which f attains a local maximum or a local minimum. If f is differentiable at c, then $f'(c)$ must be 0.

PROOF. We prove the theorem when f attains a local maximum at c. The proof for the case when f attains a local minimum is completely analogous.

Thus, let $\delta > 0$ be such that $f(c) \geq f(x)$ for all x such that $|x-c| < \delta$. Note that, if n is sufficiently large, then both $c+\frac{1}{n}$ and $c-\frac{1}{n}$ belong to the interval $(c-\delta, c+\delta)$. We evaluate $f'(c)$ in two ways. First,

$$f'(c) = \lim_n \frac{f(c+\frac{1}{n}) - f(c)}{\frac{1}{n}} \leq 0$$

because the numerator is always nonpositive and the denominator is always positive. On the other hand,

$$f'(c) = \lim_n \frac{f(c-\frac{1}{n}) - f(c)}{\frac{-1}{n}} \geq 0$$

since both numerator and denominator are nonpositive. Therefore, $f'(c)$ must be 0, as desired.

Of course we do not need a result like Theorem 4.8 for functions of a complex variable, since the derivative of every real-valued function of a complex variable necessarily is 0, independent of whether or not the function attains an extreme value.

REMARK. As mentioned earlier, the zeroes of a function are often important numbers. The preceding theorem shows that the zeroes of the derivative f' of a function f are intimately related to finding the extreme values of the function f. The zeroes of f' are often called the *critical points* for f. Part (a) of the next exercise establishes the familiar procedure from calculus for determining the maximum and minimum of a continuous real-valued function on a closed interval.

Exercise 4.11. (a) Let f be a continuous real-valued function on a closed interval $[a,b]$, and assume that f is differentiable at each point x in the open interval (a,b). Let M be the maximum value of f on this interval, and m be its minimum value on this interval. Write S for the set of all $x \in (a,b)$ for which $f'(x) = 0$. Suppose x is a point of $[a,b]$ for which $f(x)$ is either M or m. Prove that x either is an element of the set S, or x is one of the endpoints a or b.
(b) Let f be the function defined on $[0,1/2)$ by $f(t) = t/(1-t)$. Show that $f(t) < 1$ for all $t \in [0,1/2)$.
(c) Let $t \in (-1/2, 1/2)$ be given. Prove that there exists an $r < 1$, depending on t, such that $|t/(1+y)| < r$ for all y between 0 and t.
(d) Let t be a fixed number for which $0 < t < 1$. Show that, for all $0 \leq s \leq t$, $(t-s)/(1+s) \leq t$.

Probably the most powerful theorem about differentiation is the next one. It is stated as an equation, but its power is usually as an inequality; i.e., the absolute value of the left hand side is less than or equal to the absolute value of the right hand side.

THEOREM 4.9. (Mean Value Theorem) Let f be a real-valued continuous function on a closed bounded interval $[a,b]$, and assume that f is differentiable at each point x in the open interval (a,b). Then there exists a point $c \in (a,b)$ such that

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

PROOF. This proof is tricky. Define a function h on $[a,b]$ by

$$h(x) = x(f(b) - f(a)) - f(x)(b-a).$$

Clearly, h is continuous on $[a,b]$ and is differentiable at each point $x \in (a,b)$. Indeed,

$$h'(x) = f(b) - f(a) - f'(x)(b-a).$$

It follows from this equation that the theorem will be proved if we can show that there exists a point $c \in (a,b)$ for which $h'(c) = 0$. Note also that

$$h(a) = a(f(b) - f(a)) - f(a)(b-a) = af(b) - bf(a)$$

and

$$h(b) = b(f(b) - f(a)) - f(b)(b-a) = af(b) - bf(a),$$

showing that $h(a) = h(b)$.

Let m be the minimum value attained by the continuous function h on the compact interval $[a,b]$ and let M be the maximum value attained by h on $[a,b]$. If $m = M$, then h is a constant on $[a,b]$ and $h'(c) = 0$ for all $c \in (a,b)$. Hence, the theorem is true if $M = m$, and we could use any $c \in (a,b)$. If $m \neq M$, then at least one of these two extreme values is not equal to $h(a)$. Suppose $m \neq h(a)$. Of course, m is also not equal to $h(b)$. Let $c \in [a,b]$ be such that $h(c) = m$. Then, in fact, $c \in (a,b)$. By Theorem 4.8, $h'(c) = 0$.

We have then that in every case there exists a point $c \in (a,b)$ for which $h'(c) = 0$. This completes the proof.

REMARK. The Mean Value Theorem is a theorem about real-valued functions of a real variable, and we will see later that it fails for complex-valued functions of a complex variable. (See part (f) of Exercise 4.16.) In fact, it can fail for a complex-valued function of a real variable. Indeed, if $f(x) = u(x) + iv(x)$ is a continuous complex-valued function on the interval $[a,b]$, and differentiable on the open interval (a,b), then the Mean Value Theorem certainly holds for the two real-valued functions u and v, so that we would have

$$f(b) - f(a) = u(b) - u(a) + i(v(b) - v(a)) = u'(c_1)(b-a) + iv'(c_2)(b-a),$$

which is not $f'(c)(b-a)$ unless we can be sure that the two points c_1 and c_2 can be chosen to be equal. This simply is not always possible. Look at the function $f(x) = x^2 + ix^3$ on the interval $[0,1]$.

On the other hand, if f is a real-valued function of a complex variable (two real variables), then a generalized version of the Mean Value Theorem does hold. See part (c) of Exercise 4.35.

One of the first applications of the Mean Value Theorem is to show that a function whose derivative is identically 0 is necessarily a constant function. This seemingly obvious fact is just **not** obvious. The next exercise shows that this result holds for complex-valued functions of a complex variable, even though the Mean Value Theorem does not.

Exercise 4.12. (a) Suppose f is a continuous real-valued function on (a,b) and that $f'(x) = 0$ for all $x \in (a,b)$. Prove that f is a constant function on (a,b).

HINT: Show that $f(x) = f(a)$ for all $x \in [a,b]$ by using the Mean Value Theorem applied to the interval $[a,x]$.

(b) Let f be a complex-valued function of a real variable. Suppose f is differentiable at each point x in an open interval (a,b), and assume that $f'(x) = 0$ for all $x \in (a,b)$. Prove that f is a constant function.

HINT: Use the real and imaginary parts of f.
(c) Let f be a complex-valued function of a complex variable, and suppose that f is differentiable on a disk $B_r(c) \subseteq \mathbb{C}$, and that $f'(z) = 0$ for all $z \in B_r(c)$. Prove that $f(z)$ is constant on $B_r(c)$.
HINT: Let z be an arbitrary point in $B_r(c)$, and define a function $h : [0,1] \to \mathbb{C}$ by $h(t) = f((1-t)c + tz)$. Apply part (b) to h.

The next exercise establishes, at last, two important identities.

Exercise 4.13.) ($\cos^2 + \sin^2 = 1$ and $\exp(i\pi = -1$.)
(a) Prove that $\cos^2(z) + \sin^2(z) = 1$ for all complex numbers z.
(b) Prove that $\cos(\pi) = -1$.
HINT: We know from part (a) that $\cos(\pi) = \pm 1$. Using the Mean Value Theorem for the cosine function on the interval $[0, \pi]$, derive a contradiction from the assumption that $\cos(\pi) = 1$.
(c) Prove that $\exp(i\pi) = -1$.
HINT: Recall that $\exp(iz) = \cos(z) + i \sin(z)$ for all complex z. (Note that this does not yet tell us that $e^{i\pi} = -1$. We do not yet know that $\exp(z) = e^z$.)
(d) Prove that $\cosh^2 z - \sinh^2 z = 1$ for all complex numbers z.
(e) Compute the derivatives of the tangent and hyperbolic tangent functions $\tan = \sin / \cos$ and $\tanh = \sinh / \cosh$. Show in fact that

$$\tan' = \frac{1}{\cos^2} \text{ and } \tanh' = \frac{1}{\cosh^2}.$$

Here are two more elementary consequences of the Mean Value Theorem.

Exercise 4.14. (a) Suppose f and g are two complex-valued functions of a real (or complex) variable, and suppose that $f'(x) = g'(x)$ for all $x \in (a,b)$ (or $x \in B_r(c)$.) Prove that there exists a constant k such that $f(x) = g(x) + k$ for all $x \in (a,b)$ (or $x \in B_r(c)$.)
(b) Suppose $f'(z) = c \exp(az)$ for all z, where c and a are complex constants with $a \neq 0$. Prove that there exists a constant c' such that $f(z) = \frac{c}{a} \exp(az) + c'$. What if $a = 0$?
(c) (A generalization of part (a)) Suppose f and g are continuous real-valued functions on the closed interval $[a,b]$, and suppose there exists a partition $\{x_0 < x_1 < \ldots < x_n\}$ of $[a,b]$ such that both f and g are differentiable on each subinterval (x_{i-1}, x_i). (That is, we do not assume that f and g are differentiable at the endpoints.) Suppose that $f'(x) = g'(x)$ for every x in each open subinterval (x_{i-1}, x_i). Prove that there exists a constant k such that $f(x) = g(x) + k$ for all $x \in [a,b]$.
HINT: Use part (a) to conclude that $f = g + h$ where h is a step function, and then observe that h must be continuous and hence a constant.
(d) Suppose f is a differentiable real-valued function on (a,b) and assume that $f'(x) \neq 0$ for all $x \in (a,b)$. Prove that f is 1-1 on (a,b).

Exercise 4.15. Let $f : [a,b] \to \mathbb{R}$ be a function that is continuous on its domain $[a,b]$ and differentiable on (a,b). (We do not suppose that f' is continuous on (a,b).)
(a) Prove that f is nondecreasing on $[a,b]$ if and only if $f'(x) \geq 0$ for all $x \in (a,b)$. Show also that f is nonincreasing on $[a,b]$ if and only if $f'(x) \leq 0$ for all $x \in (a,b)$.
(b) Conclude that, if f' takes on both positive and negative values on (a,b), then f is **not** 1-1. (See the proof of Theorem 3.11.)

(c) Show that, if f' takes on both positive and negative values on (a, b), then there must exist a point $c \in (a, b)$ for which $f'(c) = 0$. (If f' were continuous, this would follow from the Intermediate Value Theorem. But, we are not assuming here that f' is continuous.)
(d) Prove the *Intermediate Value Theorem for Derivatives:* Suppose f is continuous on the closed bounded interval $[a, b]$ and differentiable on the open interval (a, b). If f' attains two distinct values $v_1 = f'(x_1) < v_2 = f'(x_2)$, then f' attains every value v between v_1 and v_2.
HINT: Suppose v is a value between v_1 and v_2. Define a function g on $[a, b]$ by $g(x) = f(x) - vx$. Now apply part (c) to g.

Here is another perfectly reasonable and expected theorem, but one whose proof is tough.

THEOREM 4.10. (Inverse Function Theorem) Suppose $f : (a, b) \to \mathbb{R}$ is a function that is continuous and 1-1 from (a, b) onto the interval (a', b'). Assume that f is differentiable at a point $c \in (a, b)$ and that $f'(c) \neq 0$. Then f^{-1} is differentiable at the point $f(c)$, and

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

PROOF. The formula $f^{-1\prime}(f(c)) = 1/f'(c)$ is no surprise. This follows directly from the Chain Rule. For, if $f^{-1}(f(x)) = x$, and f and f^{-1} are both differentiable, then $f^{-1\prime}(f(c))f'(c) = 1$, which gives the formula. The difficulty with this theorem is in proving that the inverse function f^{-1} of f is differentiable at $f(c)$. In fact, the first thing to check is that the point $f(c)$ belongs to the interior of the domain of f^{-1}, for that is essential if f^{-1} is to be differentiable there, and here is where the hypothesis that f is a real-valued function of a real variable is important. According to Exercise 3.12, the 1-1 continuous function f maps $[a, b]$ onto an interval $[a', b']$, and $f(c)$ is in the open interval (a', b'), i.e., is in the interior of the domain of f^{-1}. According to part (2) of Theorem 4.2, we can prove that f^{-1} is differentiable at $f(c)$ by showing that

$$\lim_{x \to f(c)} \frac{f^{-1}(x) - f^{-1}(f(c))}{x - f(c)} = \frac{1}{f'(c)}.$$

That is, we need to show that, given an $\epsilon > 0$, there exists a $\delta > 0$ such that if $0 < |x - f(c)| < \delta$ then

$$|\frac{f^{-1}(x) - f^{-1}(f(c))}{x - f(c)} - \frac{1}{f'(c)}| < \epsilon.$$

First of all, because the function $1/q$ is continuous at the point $f'(c)$, there exists an $\epsilon' > 0$ such that if $|q - f'(c)| < \epsilon'$, then

$$(4.4). \qquad |\frac{1}{q} - \frac{1}{f'(c)}| < \epsilon$$

Next, because f is differentiable at c, there exists a $\delta' > 0$ such that if $0 < |y - c| < \delta'$ then

$$(4.5). \qquad |\frac{f(y) - f(c)}{y - c} - f'(c)| < \epsilon'$$

Now, by Theorem 3.10, f^{-1} is continuous at the point $f(c)$, and therefore there exists a $\delta > 0$ such that if $|x - f(c)| < \delta$ then

$$(4.6). \qquad |f^{-1}(x) - f^{-1}(f(c)| < \delta'$$

So, if $|x - f(c)| < \delta$, then

$$|f^{-1}(x) - c| = |f^{-1}(x) - f^{-1}(f(c))| < \delta'.$$

But then, by Inequality 4.5,

$$|\frac{f(f^{-1}(x)) - f(c)}{f^{-1}(x) - c} - f'(c)| < \epsilon',$$

from which it follows, using Inequality 4.4, that

$$|\frac{f^{-1}(x) - f^{-1}(f(c))}{x - f(c)} - \frac{1}{f'(c)}| < \epsilon,$$

as desired.

REMARK. A result very like Theorem 4.10 is actually true for complex-valued functions of a complex variable. We will have to show that if c is in the interior of the domain S of a one-to-one, continuously differentiable, complex-valued function f of a complex variable, then $f(c)$ is in the interior of the domain $f(S)$ of f^{-1}. But, in the complex variable case, this requires a somewhat more difficult argument. Once that fact is established, the proof that f^{-1} is differentiable at $f(c)$ will be the same for complex-valued functions of complex variables as it is here for real-valued functions of a real variable. Though the proof of Theorem 4.10 is reasonably complicated for real-valued functions of a real variable, the corresponding result for complex functions is much more deep, and that proof will have to be postponed to a later chapter. See Theorem 7.10.

THE EXPONENTIAL AND LOGARITHM FUNCTIONS

We derive next the elementary properties of the exponential and logarithmic functions. Of course, by "exponential function," we mean the power series function exp. And, as yet, we have not even defined a logarithm function.

Exercise 4.16. (a) Define a complex-valued function $f : \mathbb{C} \to \mathbb{C}$ by $f(z) = \exp(z)\exp(-z)$. Prove that $f(z) = 1$ for all $z \in \mathbb{C}$.
(b) Conclude from part (a) that the exponential function is never 0, and that $\exp(-z) = 1/\exp(z)$.
(c) Show that the exponential function is always positive on $\mathbb{R}$, and
that $\lim_{x\to-\infty}\exp(x) = 0$.
(d) Prove that exp is continuous and 1-1 from $(-\infty, \infty)$ onto $(0, \infty)$.
(e) Show that the exponential function is **not** 1-1 on $\mathbb{C}$.
(f) Use parts b and e to show that the Mean Value Theorem is not in any way valid for complex-valued functions of a complex variable.

Using part (d) of the preceding exercise, we make the following important definition.

DEFINITION. We call the inverse $\exp^{-1}$ of the restriction of the exponential function to $\mathbb{R}$ the (natural) *logarithm function,* and we denote this function by $\ln$.

The properties of the exponential and logarithm functions are strongly tied to the simplest kinds of differential equations. The connection is suggested by the fact, we have already observed, that $\exp' = \exp$. The next theorem, corollary, and exercises make these remarks more precise.

THEOREM 4.11. *Suppose $f : \mathbb{C} \to \mathbb{C}$ is differentiable everywhere and satisfies the differential equation $f' = af$, where a is a complex number. Then $f(z) = c\exp(az)$, where $c = f(0)$.*

PROOF. Consider the function $h(z) = f(z)/\exp(az)$. Using the Quotient Formula, we have that

$$h'(z) = \frac{\exp(az)f'(z) - a\exp(az)f(z)}{[\exp(az)]^2} = \frac{\exp(az)(f'(z) - af(z))}{[\exp(z)]^2} = 0.$$

Hence, there exists a complex number c such that $h(z) = c$ for all z. Therefore, $f(z) = c\exp(az)$ for all z. Setting $z = 0$ gives $f(0) = c$, as desired.

COROLLARY. (Law of Exponents) For all complex numbers z and w, $\exp(z + w) = \exp(z)\exp(w)$.

PROOF OF THE COROLLARY. Fix w, define $f(z) = \exp(z + w)$, and apply the preceding theorem. We have $f'(z) = \exp(z + w) = f(z)$, so we get

$$\exp(z + w) = f(z) = f(0)\exp(z) = \exp(w)\exp(z).$$

Exercise 4.17. (a) If n is a positive integer and z is any complex number, show that $\exp(nz) = (\exp(z))^n$.
(b) If r is a rational number and x is any real number, show that $\exp(rx) = (\exp(x))^r$.

Exercise 4.18. (a) Show that ln is continuous and 1-1 from $(0, \infty)$ onto $\mathbb{R}$.
(b) Prove that the logarithm function ln is differentiable at each point $y \in (0, \infty)$ and that $\ln'(y) = 1/y$.
HINT: Write $y = \exp(c)$ and use Theorem 4.10.
(c) Derive the first law of logarithms: $\ln(xy) = \ln(x) + \ln(y)$.
(d) Derive the second law of logarithms: That is, if r is a rational number and x is a positive real number, show that $\ln(x^r) = r\ln(x)$.

We are about to make the connection between the number e and the exponential function. The next theorem is the first step.

THEOREM 4.12. $\ln(1) = 0$ *and* $\ln(e) = 1$.

PROOF. If we write $1 = \exp(t)$, then $t = \ln(1)$. But $\exp(0) = 1$, so that $\ln(1) = 0$, which establishes the first assertion.
Recall that

$$e = \lim_n (1 + \frac{1}{n})^n.$$

Therefore,

$$\begin{aligned}\ln(e) &= \ln(\lim_n (1+\frac{1}{n})^n)\\ &= \lim_n \ln((1+\frac{1}{n})^n)\\ &= \lim_n n\ln(1+\frac{1}{n})\\ &= \lim_n \frac{\ln(1+\frac{1}{n})}{\frac{1}{n}}\\ &= \lim_n \frac{\ln(1+\frac{1}{n})-\ln(1)}{\frac{1}{n}}\\ &= \ln'(1)\\ &= 1/1\\ &= 1.\end{aligned}$$

This establishes the second assertion of the theorem.

Exercise 4.19. (a) Prove that

$$e=\sum_{n=0}^{\infty}\frac{1}{n!}.$$

HINT: Use the fact that the logarithm function is 1-1.
(b) For r a rational number, show that $\exp(r)=e^r$.
(c) If a is a positive number and $r=p/q$ is a rational number, show that

$$a^r=\exp(r\ln(a)).$$

(d) Prove that e is irrational.
HINT: Let p_n/q_n be the nth partial sum of the series in part (a). Show that $q_n\le n!$, and that $\lim q_n(e-p_n/q_n)=0$. Then use Theorem 2.19.

We have finally reached a point in our development where we can make sense of raising any positive number to an arbitrary complex exponent. Of course this includes raising positive numbers to irrational powers. We make our general definition based on part (c) of the preceding exercise.

DEFINITION. For a a positive real number and z an arbitrary complex number, define a^z by

$$a^z=\exp(z\ln(a)).$$

REMARK. The point is that our old understanding of what a^r means, where $a>0$ and r is a rational number, coincides with the function $\exp(r\ln(a))$. So, this new definition of a^z coincides and is consistent with our old definition. And, it now allows us to raies a positive number a to an arbitrary complex exponent.

REMARK. Let the bugles sound!! Now, having made all the appropriate definitions and derived all the relevant theorems, we can finally prove that $e^{i\pi}=-1$. From the definition above, we see that if $a = e$, then we have $e^z = \exp(z)$. Then, from part (c) of Exercise 4.13, we have what we want:

$$e^{i\pi} = -1.$$

Exercise 4.20. (a) Prove that, for all complex numbers z and w, $e^{z+w} = e^z e^w$.
(b) If x is a real number and z is any complex number, show that

$$(e^x)^z = e^{xz}.$$

(c) Let a be a fixed positive number, and define a function $f : \mathbb{C} \to \mathbb{C}$ by $f(z) = a^z$. Show that f is differentiable at every $z \in \mathbb{C}$ and that $f'(z) = \ln(a)a^z$.
(d) Prove the general laws of exponents: If a and b are positive real numbers and z and w are complex numbers,

$$a^{z+w} = a^z a^w,$$

$$a^z b^z = (ab)^z,$$

and, if x is real,

$$a^{xw} = (a^x)^w.$$

(e) If y is a real number, show that $|e^{iy}| = 1$. If $z = x + iy$ is a complex number, show that $|e^z| = e^x$.
(f) Let $\alpha = a + bi$ be a complex number, and define a function $f : (0, \infty) \to \mathbb{C}$ by $f(x) = x^\alpha = e^{\alpha \ln(x)}$. Prove that f is differentiable at each point x of $(0, \infty)$ and that $f'(x) = \alpha x^{\alpha - 1}$.
(g) Let $\alpha = a + bi$ be as in part (f). For $x > 0$, show that $|x^\alpha| = x^a$.

THE TRIGONOMETRIC AND HYPERBOLIC FUNCTIONS

The laws of exponents and the algebraic connections between the exponential function and the trigonometric and hyperbolic functions, give the following "addition formulas:"

THEOREM 4.13. *The following identities hold for all complex numbers z and w.*

$$\sin(z + w) = \sin(z)\cos(w) + \cos(z)\sin(w).$$

$$\cos(z + w) = \cos(z)\cos(w) - \sin(z)\sin(w).$$

$$\sinh(z + w) = \sinh(z)\cosh(w) + \cosh(z)\sinh(w).$$

$$\cosh(z + w) = \cosh(z)\cosh(w) + \sinh(z)\sinh(w).$$

PROOF. We derive the first formula and leave the others to an exercise.
First, for any two real numbers x and y, we have

$$\begin{aligned}\cos(x + y) + i\sin(x + y) &= e^{i(x+y)} \\ &= e^{ix}e^{iy} \\ &= (\cos x + i\sin x) \times (\cos y + i\sin y) \\ &= \cos x\cos y - \sin x\sin y + i(\cos x\sin y + \sin x\cos y),\end{aligned}$$

which, equating real and imaginary parts, gives that

$$\cos(x+y) = \cos x \cos y - \sin x \sin y$$

and

$$\sin(x+y) = \sin x \cos y + \cos x \sin y.$$

The second of these equations is exactly what we want, but this calculation only shows that it holds for real numbers x and y. We can use the Identity Theorem to show that in fact this formula holds for all complex numbers z and w. Thus, fix a real number y. Let $f(z) = \sin z \cos y + \cos z \sin y$, and let

$$g(z) = \sin(z+y) = \frac{1}{2i}(e^{i(z+y)} - e^{-i(z+y)} = \frac{1}{2i}(e^{iz}e^{iy} - e^{-iz}e^{-iy}).$$

Then both f and g are power series functions of the variable z. Furthermore, by the previous calculation, $f(1/k) = g(1/k)$ for all positive integers k. Hence, by the Identity Theorem, $f(z) = g(z)$ for all complex z. Hence we have the formula we want for all complex numbers z and all real numbers y.

To finish the proof, we do the same trick one more time. Fix a complex number z. Let $f(w) = \sin z \cos w + \cos z \sin w$, and let

$$g(w) = \sin(z+w) = \frac{1}{2i}(e^{i(z+w)} - e^{-i(z+w)} = \frac{1}{2i}(e^{iz}e^{iw} - e^{-iz}e^{-iw}).$$

Again, both f and g are power series functions of the variable w, and they agree on the sequence $\{1/k\}$. Hence they agree everywhere, and this completes the proof of the first addition formula.

Exercise 4.21. (a) Derive the remaining three addition formulas of the preceding theorem.

(b) From the addition formulas, derive the two "half angle" formulas for the trigonometric functions:

$$\sin^2(z) = \frac{1-\cos(2z)}{2},$$

and

$$\cos^2(z) = \frac{1+\cos(2z)}{2}.$$

THEOREM 4.14. *The trigonometric functions* sin *and* cos *are periodic with period* 2π; *i.e.,* $\sin(z+2\pi) = \sin(z)$ *and* $\cos(z+2\pi) = \cos(z)$ *for all complex numbers* z.

PROOF. We have from the preceding exercise that $\sin(z+2\pi) = \sin(z)\cos(2\pi) + \cos(z)\sin(2\pi)$, so that the periodicity assertion for the sine function will follow if we show that $\cos(2\pi) = 1$ and $\sin(2\pi) = 0$. From part (b) of the preceding exercise, we have that

$$0 = \sin^2(\pi) = \frac{1-\cos(2\pi)}{2}$$

which shows that $\cos(2\pi) = 1$. Since $\cos^2 + \sin^2 = 1$, it then follows that $\sin(2\pi) = 0$.

The periodicity of the cosine function is proved similarly.

Exercise 4.22. (a) Prove that the hyperbolic functions sinh and cosh are periodic. What is the period?
(b) Prove that the hyperbolic cosine $\cosh(x)$ is never 0 for x a real number, that the hyperbolic tangent $\tanh(x) = \sinh(x)/\cosh(x)$ is bounded and increasing from $\mathbb{R}$ onto $(-1, 1)$, and that the inverse hyperbolic tangent has derivative given by ${\tanh^{-1}}'(y) = 1/(1 - y^2)$.
(c) Verify that for all $y \in (-1, 1)$

$$\tanh^{-1}(y) = \ln(\sqrt{\frac{1+y}{1-y}}).$$

Exercise 4.23. (Polar coordinates) Let z be a nonzero complex number. Prove that there exists a unique real number $0 \le \theta < 2\pi$ such that $z = re^{i\theta}$, where $r = |z|$.
HINT: If $z = a + bi$, then $z = r(\frac{a}{r} + \frac{b}{r}i$. Observe that $-1 \le \frac{a}{r} \le 1$, $-1 \le \frac{b}{r} \le 1$, and $(\frac{a}{r})^2 + (\frac{b}{r})^2 = 1$. Show that there exists a unique $0 \le \theta < 2\pi$ such that $\frac{a}{r} = \cos\theta$ and $\frac{b}{r} = \sin\theta$.

L'Hopital's Rule

Many limits of certain combinations of functions are difficult to evaluate because they lead to what's known as "indeterminate forms." These are expressions of the form $0/0$, ∞/∞, 0^0, $\infty - \infty$, 1^∞, and the like. They are precisely combinations of functions that are not covered by our limit theorems. See Theorem 4.1. The very definition of the derivative itself is such a case: $\lim_{h\to 0}(f(c + h) - f(c)) = 0$, $\lim_{h\to 0} h = 0$, and we are interested in the limit of the quotient of these two functions, which would lead us to the indeterminate form $0/0$. The definition of the number e is another example: $\lim(1 + 1/n) = 1$, $\lim n = \infty$, and we are interested in the limit of $(1 + 1/n)^n$, which leads to the indeterminate form 1^∞. L'Hopital's Rule, Theorem 4.16 below, is our strongest tool for handling such indeterminate forms.
To begin with, here is a useful generalization of the Mean Value Theorem.

THEOREM 4.15. (Cauchy Mean Value Theorem) Let f and g be continuous real-valued functions on a closed interval $[a, b]$, suppose $g(a) \ne g(b)$, and assume that both f and g are differentiable on the open interval (a, b). Then there exists a point $c \in (a, b)$ such that

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

Exercise 4.24. Prove the preceding theorem.
HINT: Define an auxiliary function h as was done in the proof of the original Mean Value Theorem.

The following theorem and exercise comprise what is called L'Hopital's Rule.

THEOREM 4.16. *Suppose f and g are differentiable real-valued functions on the bounded open interval (a, b) and assume that*

$$\lim_{x\to a+0} \frac{f'(x)}{g'(x)} = L,$$

where L is a real number. (Implicit in this hypothesis is that $g'(x) \neq 0$ for x in some interval $(a, a+\alpha)$.) Suppose further that either

$$\lim_{x\to a+0} f(x) = \lim_{x\to a+0} g(x) = 0$$

or

$$\lim_{x\to a+0} f(x) = \lim_{x\to a+0} g(x) = \infty.$$

then

$$\lim_{x\to a+0} \frac{f(x)}{g(x)} = L.$$

PROOF. Suppose first that

$$\lim_{x\to a+0} f(x) = \lim_{x\to a+0} g(x) = 0.$$

Observe first that, because $g'(x) \neq 0$ for all x in some interval $(a, a+\alpha)$, $g'(x)$ is either always positive or always negative on that interval. (This follows from part (d) of Exercise 4.15.) Therefore the function g must be strictly monotonic on the interval $(a, a+\alpha)$. Hence, since $\lim_{x\to a+0} g(x) = 0$, we must have that $g(x) \neq 0$ on the interval $(a, a+\alpha)$.

Now, given an $\epsilon > 0$, choose a positive $\delta < \alpha$ such that if $a < c < a+\delta$ then $|\frac{f'(c)}{g'(c)} - L| < \epsilon$. Then, for every natural number n for which $1/n < \delta$, and every $a < x < a+\delta$, we have by the Cauchy Mean Value Theorem that there exists a point c between $a + 1/n$ and x such that

$$|\frac{f(x) - f(a+1/n)}{g(x) - g(a+1/n)} - L| = |\frac{f'(c)}{g'(c)} - L| < \epsilon.$$

Therefore, taking the limit as n approaches ∞, we obtain

$$|\frac{f(x)}{g(x)} - L| = \lim_{n\to\infty} |\frac{f(x) - f(a+1/n)}{g(x) - g(a+1/n)} - L| \leq \epsilon$$

for all x for which $a < x < a+\delta$. This proves the theorem in this first case.

Next, suppose that

$$\lim_{x\to a+0} f(x) = \lim_{x\to a+0} g(x) = \infty.$$

This part of the theorem is a bit more complicated to prove. First, choose a positive α so that $f(x)$ and $g(x)$ are both positive on the interval $(a, a+\alpha)$. This is possible because both functions are tending to infinity as x approaches a. Now, given an $\epsilon > 0$, choose a positive number $\beta < \alpha$ such that

$$|\frac{f'(c)}{g'(c)} - L| < \frac{\epsilon}{2}$$

for all $a < c < a+\beta$. We express this absolute value inequality as the following pair of ordinary inequalities:

$$L - \frac{\epsilon}{2} < \frac{f'(c)}{g'(c)} < L + \frac{\epsilon}{2}.$$

Set $y = a + \beta$. Using the Cauchy Mean Value Theorem, and the preceding inequalities, we have that for all $a < x < y$

$$L - \frac{\epsilon}{2} < \frac{f(x) - f(y)}{g(x) - g(y)} < L + \frac{\epsilon}{2},$$

implying that

$$(L - \frac{\epsilon}{2})(g(x) - g(y)) + f(y) < f(x) < (L + \frac{\epsilon}{2})(g(x) - g(y)) + f(y).$$

Dividing through by $g(x)$ and simplifying we obtain

$$L - \frac{\epsilon}{2} - \frac{(L - \frac{\epsilon}{2})g(y)}{g(x)} + \frac{f(y)}{g(x)} < \frac{f(x)}{g(x)} < L + \frac{\epsilon}{2} - \frac{(L + \frac{\epsilon}{2})g(y)}{g(x)} + \frac{f(y)}{g(x)}.$$

Finally, using the hypothesis that $\lim_{x \to a+0} g(x) = \infty$, and the fact that $L, \epsilon, g(y)$, and $f(y)$ are all constants, choose a $\delta > 0$, with $\delta < \beta$, such that if $a < x < a + \delta$, then

$$|- \frac{(L - \frac{\epsilon}{2})g(y)}{g(x)} + \frac{f(y)}{g(x)}| < \frac{\epsilon}{2}$$

and

$$|- \frac{(L + \frac{\epsilon}{2})g(y)}{g(x)} + \frac{f(y)}{g(x)}| < \frac{\epsilon}{2}.$$

Then, for all $a < x < a + \delta$, we would have

$$L - \epsilon < \frac{f(x)}{g(x)} < L + \epsilon,$$

implying that

$$|\frac{f(x)}{g(x)} - L| < \epsilon,$$

and the theorem is proved.

Exercise 4.25. (a) Show that the conclusions of the preceding theorem also hold if we assume that

$$\lim_{x \to a+0} \frac{f'(x)}{g'(x)} = \infty.$$

HINT: Replace ϵ by a large real number B and show that $f(x)/g(x) > B$ if $0 < x - a < \delta$.

(b) Show that the preceding theorem, as well as part (a) of this exercise, also holds if we replace the (finite) endpoint a by $-\infty$.

HINT: Replace the δ's by negative numbers B.

(c) Show that the preceding theorem, as well as parts a and b of this exercise, hold if the limit as x approaches a from the right is replaced by the limit as x approaches b from the left.

HINT: Replace $f(x)$ by $f(-x)$ and $g(x)$ by $g(-x)$.

(d) Give an example to show that the converse of L'Hopital's Rule need not hold; i.e., find functions f and g for which $\lim_{x\to a+0} f(x) = \lim_{x\to a+0} g(x) = 0$,

$$\lim_{x\to a+0}\frac{f(x)}{g(x)} \text{ exists, but } \lim_{x\to a+0}\frac{f'(x)}{g'(x)} \text{ does not exist.}$$

(e) Deduce from the proof given above that if $\lim_{x\to a+0} f'(x)/g'(x) = L$ and $\lim_{x\to a+0} g(x) = \infty$, then $\lim_{x\to a+0} f(x)/g(x) = L$ independent of the behavior of f.
(f) Evaluate $\lim_{x\to\infty} x^{1/x}$, and $\lim_{x\to 0}(1-x)^{1/x}$.
HINT: Take logarithms.

HIGHER ORDER DERIVATIVES

DEFINITION. Let S be a subset of $\mathbb{R}$ (or $\mathbb{C}$), and Let $f : S \to \mathbb{C}$ be a function of a real (or complex) variable. We say that f is *continuously differentiable* on S^0 if f is differentiable at each point x of S^0 and the function f' is continuous on S^0. We say that $f \in C^1(S)$ if f is continuous on S and continuously differentiable on S^0. We say that f is *2-times continuously differentiable* on S^0 if the first derivative f' is itself continuously differentiable on S^0. And, inductively, we say that f is *k-times continuously differentiable* on S^0 if the $k-1$st derivative of f is itself continuously differentiable on S^0. We write $f^{(k)}$ for the kth derivative of f, and we write $f \in C^k(S)$ if f is continuous on S and is k times continuously differentiable on S^0. Of course, if $f \in C^k(S)$, then all the derivatives $f^{(j)}$, for $j \leq k$, exist nd are continuous on S^0. (Why?)
For completeness, we define $f^{(0)}$ to be f itself, and we say that $f \in C^\infty(S)$ if f is continuous on S and has infinitely many continuous derivatives on S^0; i.e., all of its derivatives exist and are continuous on S^0.
As in Chapter III, we say that f is *real-analytic* (or *complex-analytic*) on S if it is expandable in a Taylor series around each point $c \in S^0$

REMARK. Keep in mind that the definition above, as applied to functions whose domain S is a nontrivial subset of $\mathbb{C}$, has to do with functions of a complex variable that are continuously differentiable on the set S^0. We have seen that this is quite different from a function having continuous partial derivatives on S^0. We will return to partial derivatives at the end of this chapter.

THEOREM 4.17. *Let S be an open subset of $\mathbb{R}$ (or $\mathbb{C}$).*

1. *Suppose WS is a subset of $\mathbb{R}$. Then, for each $k \geq 1$, there exists a function in $C^k(S)$ that is not in $C^{k+1}(S)$. That is, $C^{k+1}(S)$ is a proper subset of $C^k(S)$.*
2. *If f is real-analytic (or complex-analytic) on S, then $f \in C^\infty(S)$.*
3. *There exists a function in $C^\infty(\mathbb{R})$ that is not real-analytic on $\mathbb{R}$. That is, the set of real-analytic functions on $\mathbb{R}$ is a proper subset of the set $C^\infty(\mathbb{R})$.*

REMARK. Suppose S is an open subset of $\mathbb{C}$. It is a famous result from the Theory of Complex Variables that if f is in $C^1(S)$, then f is necessarily complex analytic on S. We will prove this amazing result in Theorem 7.5. Part (3) of the theorem shows that the situation is quite different for real-valued functions of a real variable.

PROOF. For part (1), see the exercise below. Part (2) is immediate from part (c) of Exercise 4.10. Before finishing the proof of part (3), we present the following lemma:

LEMMA. *Let f be the function defined on all of $\mathbb{R}$ as follows.*

$$f(x) = \begin{cases} 0 & x \leq 0 \\ \frac{p(x)e^{-1/x}}{x^n} & x > 0 \end{cases}$$

where $p(x)$ is a fixed polynomial function and n is a fixed nonnegative integer. Then f is continuous at each point x of $\mathbb{R}$.

PROOF OF THE LEMMA. The assertion of the lemma is clear if $x \neq 0$. To see that f is continuous at 0, it will suffice to prove that

$$\lim_{x \to 0+0} \frac{p(x)e^{-1/x}}{x^n} = 0.$$

(Why?) But, for $x > 0$, we know from part (b) of Exercise 3.22 that $e^{1/x} > 1/(x^{n+1}(n+1)!)$, implying that $e^{-1/x} < x^{n+1}(n+1)!$. Hence, for $x > 0$,

$$|f(x)| = \frac{|p(x)|e^{-1/x}}{x^n} < (n+1)!x|p(x)|,$$

and this tends to 0 as x approaches 0 from the right, as desired.

. Returning to the proof of Theorem 4.17, we verify part (3) by observing that if f is as in the preceding lemma then f is actually differentiable, and its derivative f' is a function of the same sort. (Why?) It follows that any such function belongs to $C^\infty(\mathbb{R})$. On the other hand, a nontrivial such f cannot be expandable in a Taylor series around 0 because of the Identity Theorem. (Take $x_k = -1/k$.) This completes the proof.

Exercise 4.26. (a) Prove part (1) of Theorem 4.17. Use functions of the form $x^n \sin(1/x)$.

(b) Prove that any function of the form of the f in the lemma above is everywhere differentiable on $\mathbb{R}$, and its derivative has the same form. Conclude that any such function belongs to $C^\infty(\mathbb{R})$.

(c) For each positive integer n, define a function f_n on the interval $(-1, 1$ by $f_n(x) = |x|^{1+1/n}$. Prove that each f_n is differentiable at every point in $(-1, 1)$, including 0. Prove also that the sequence $\{f_n\}$ converges uniformly to the function $f(x) = |x|$. (See part (h) of Exercise 3.28.) Conclude that the uniform limit of differentiable functions of a real variable need not be differentiable. (Again, for functions of a complex variable, the situation is very different. In that case, the uniform limit of differentiable functions **is** differentiable. See Theorem 7.11.)

Exercise 4.27. (A smooth approximation to a step function.) Suppose $a < b < c < d$ are real numbers. Show that there exists a function χ in $C^\infty(\mathbb{R})$ such that $0 \leq \chi(x) \leq 1$ for all x, $\chi(x) \equiv 1$ for $x \in [b, c]$, and $\chi(x) \equiv 0$ for $x \notin (a, d)$. (If a is close to b and c is close to d, then this function is a C^∞ approximation to the step function that is 1 on the interval $[b, c]$ and 0 elsewhere.)

(a) Let f be a function like the one in the lemma. Think about the graphs of the functions $f(x-c)$ and $f(b-x)$. Construct a C^∞ function g that is 0 between b and c and positive everywhere else.

(b) Construct a C^∞ function h that is positive between a and d and 0 everywhere else.

(c) Let g and h be as in parts (a) and (b). If $j = g + h$, show that j is never 0, and write k for the C^∞ function $k = 1/j$.

(d) Examine the function hk, and show that it is the desired function χ.

THEOREM 4.18. (Formula for the coefficients of a Taylor Series function) Let f be expandable in a Taylor series around a point c :

$$f(x) = \sum a_n(x-c)^n.$$

Then for each n, $a_n = f^{(n)}(c)/n!$.

PROOF. Because each derivative of a Taylor series function is again a Taylor series function, and because the value of a Taylor series function at the point c is equal to its constant term a_0, we have that $a_1 = f'(c)$. Computing the derivative of the derivative, we see that $2a_2 = f''(c) = f^{(2)}(c)$. Continuing this, i.e., arguing by induction, we find that $n!a_n = f^{(n)}(c)$, which proves the theorem.

TAYLOR POLYNOMIALS AND TAYLOR'S REMAINDER THEOREM

DEFINITION. Let f be in $C^n(B_r(c))$ for c a fixed complex number, $r > 0$, and n a positive integer. Define the *Taylor polynomial* of degree n for f at c to be the polynomial $T^n \equiv T^n_{(f,c)}$ given by the formula:

$$(T^n_{(f,c)})(z) = \sum_{j=0}^{n} a_j(z-c)^j,$$

where $a_j = f^{(j)}(c)/j!$.

REMARK. If f is expandable in a Taylor series on $B_r(c)$, then the Taylor polynomial for f of degree n is nothing but the nth partial sum of the Taylor series for f on $B_r(c)$. However, any function that is n times differentiable at a point c has a Taylor polynomial of order n. Functions that are infinitely differentiable have Taylor polynomials of all orders, and we might suspect that these polynomials are some kind of good approximation to the function itself.

Exercise 4.28. Prove that f is expandable in a Taylor series function around a point c (with radius of convergence $r > 0$) if and only if the sequence $\{T^n_{(f,c)}\}$ of Taylor polynomials converges pointwise to f; i.e.,

$$f(z) = \lim(T^n_{(f,c)})(z)$$

for all z in $B_r(c)$.

Exercise 4.29. Let $f \in C^n(B_r(c))$. Prove that $f' \in C^{n-1}(B_r(c))$. Prove also that $(T^n_{(f,c)})' = T^{n-1}_{(f',c)}$.

The next theorem is, in many ways, the fundamental theorem of numerical analysis. It clearly has to do with approximating a general function by polynomials. It is a generalization of the Mean Value Theorem, and as in that case this theorem holds only for real-valued functions of a real variable.

THEOREM 4.19. (Taylor's Remainder Theorem) Let f be a real-valued function on an interval $(c-r, c+r)$, and assume that $f \in C^n((c-r, c+r))$, and that $f^{(n)}$ is differentiable on $(c-r, c+r)$. Then, for each x in $(c-r, c+r)$ there exists a y between c and x such that

$$f(x) - (T^n_{(f,c)})(x) = \frac{f^{(n+1)}(y)}{(n+1)!}(x-c)^{n+1}. \tag{4.7}$$

REMARK. If we write $f(x) = T^n_{f,c})(x) + R_{n+1}(x)$, where $R_{n+1}(x)$ is the error or remainder term, then this theorem gives a formula, and hence an estimate, for that remainder term. This is the evident connection with Numerical Analysis.

PROOF. We prove this theorem by induction on n. For $n = 0$, this is precisely the Mean Value Theorem. Thus,

$$f(x) - T^0_{f,c}(x) = f(x) - f(c) = f'(y)(x - c.$$

Now, assuming the theorem is true for all functions in $C^{n-1}((c-r, c+r))$, let us show it is true for the given function $f \in C^n((c - r, c + r))$. Set $g(x) = f(x) - (T^n_{(f,c)})(x)$ and let $h(x) = (x - c)^{n+1}$. Observe that both $g(c) = 0$ and $h(c) = 0$. Also, if $x \neq c$, then $h(x) \neq 0$. So, by the Cauchy Mean Value Theorem, we have that

$$\frac{g(x)}{h(x)} = \frac{g(x) - g(c)}{h(x) - h(c)} = \frac{g'(w)}{h'(w)}$$

for some w between c and x. Now

$$g'(w) = f'(w) - (t^n_{(f,c)})'(w) = f'(w) - (T^{n-1}_{(f',c)})(w)$$

(See the preceding exercise.), and $h'(w) = (n + 1)(w - c)^n$. Therefore,

$$\begin{aligned}\frac{f(x) - (T^n_{(f,c)})(x)}{(x - c)^{n+1}} &= \frac{g(x)}{h(x)} \\ &= \frac{g'(w)}{h'(w)} \\ &= \frac{f'(w) - (T^{n-1}_{(f',c)})(w)}{(n + 1)(w - c)^n}.\end{aligned}$$

We apply the inductive hypotheses to the function f' (which is in $C^{n-1}((c-r, c+r))$) and obtain

$$\begin{aligned}\frac{f(x) - (T^n_{(f,c)})(x)}{(x - c)^{n+1}} &= \frac{f'(w) - (T^{n-1}_{(f',c)})(w)}{(n + 1)(w - c)^n} \\ &= \frac{\frac{f'^{(n)}(y)}{n!}(w - c)^n}{(n + 1)(w - c)^n} \\ &= \frac{f'^{(n)}(y)}{(n + 1)!} \\ &= \frac{f^{(n+1)}(y)}{(n + 1)!}\end{aligned}$$

for some y between c and w. But this implies that

$$f(x) - (T^n_{(f,c)})(x) = \frac{f^{(n+1)}(y)(x - c)^{n+1}}{(n + 1)!},$$

for some y between c and x, which finishes the proof of the theorem.

Exercise 4.30. Define $f(x) = 0$ for $x \leq 0$ and $f(x) = e^{-1/x}$ for $x > 0$. Verify that $f \in C^\infty(\mathbb{R})$, that $f^{(n)}(0) = 0$ for all n, and yet f is not expandable in a Taylor series around 0. Interpret Taylor's Remainder Theorem for this function. That is, describe the remainder $R_{n+1}(x)$.

As a first application of Taylor's Remainder Theorem we give the following result, which should be familiar from calculus. It is the generalized version of what's ordinarily called the "second derivative test."

THEOREM 4.20. (Test for Local Maxima and Minima) Let f be a real-valued function in $C^n(c-r, c+r)$, suppose that the $n+1$st derivative $f^{(n+1)}$ of f exists everywhere on $(c-r, c+r)$ and is continuous at c, and suppose that $f^{(k)}(c) = 0$ for all $1 \leq k \leq n$ and that $f^{(n+1)}(c) \neq 0$. Then:

(1) If n is even, f attains neither a local maximum nor a local minimum at c. In this case, c is called an *inflection point.*
(2) If n is odd and $f^{(n+1)}(c) < 0$, then f attains a local maximum at c.
(3) If n is odd and $f^{(n+1)}(c) > 0$, then f attains a local minimum at c.

PROOF. Since $f^{(n+1)}$ is continuous at c, there exists a $\delta > 0$ such that $f^{(n+1)}(y)$ has the same sign as $f^{(n+1)}(c)$ for all $y \in (c-\delta, c+\delta)$. We have by Taylor's Theorem that if $x \in (c-\delta, c+\delta)$ then there exists a y between x and c such that

$$f(x) = (T^n_{(f,c)})(x) + \frac{f^{(n+1)}(y)}{(n+1)!}(x-c)^{n+1},$$

from which it follows that

$$\begin{aligned} f(x) - f(c) &= \sum_{k=1}^{n} f^{(k)}(c)k!(x-c)^k + \frac{f^{(n+1)}(y)}{(n+1)!}(x-c)^{n+1} \\ &= \frac{f^{(n+1)}(y)}{(n+1)!}(x-c)^{n+1}. \end{aligned}$$

Suppose n is even. It follows then that if $x < c$, the sign of $(x-c)^{n+1}$ is negative, so that the sign of $f(x) - f(c)$ is the opposite of the sign of $f^{(n+1)}(c)$. On the other hand, if $x > c$, then $(x-c)^{n+1} > 0$, so that the sign of $f(x) - f(c)$ is the same as the sign of $f^{(n+1)}(c)$. So, $f(x) > f(c)$ for all nearby x on one side of c, while $f(x) < f(c)$ for all nearby x on the other side of c. Therefore, f attains neither a local maximum nor a local minimum at c. This proves part (1).
Now, if n is odd, the sign of $f(x) - f(c)$ is the same as the sign of $f^{(n+1)}(y)$, which is the same as the sign of $f^{(n+1)}(c)$, for all $x \in (c-\delta, c+\delta)$. Hence, if $f^{(n+1)}(c) < 0$, then $f(x) - f(c) < 0$ for all $x \in (c-\delta, c+\delta)$, showing that f attains a local maximum at c. And, if $f^{(n+1)}(c) > 0$, then the sign of $f(x) - f(c)$ is positive for all $x \in (c-\delta, c+\delta)$, showing that f attains a local minimum at c. This proves parts (2) and (3).

The General Binomial Theorem

We use Taylor's Remainder Theorem to derive a generalization of the Binomial Theorem to nonintegral exponents. First we must generalize the definition of binomial coefficient.

DEFINITION. Let α be a complex number, and let k be a nonnegative integer. We define the general *binomial coefficient* $\binom{\alpha}{k}$ by

$$\binom{\alpha}{k} = \frac{\alpha(\alpha-1)\ldots(\alpha-k+1)}{k!}.$$

If α is itself a positive integer and $k \leq \alpha$, then $\binom{\alpha}{k}$ agrees with the earlier definition of the binomial coefficient, and $\binom{\alpha}{k} = 0$ when $k > \alpha$. However, if α is not an integer, but just an arbitrary complex number, then every $\binom{\alpha}{k} \neq 0$.

Exercise 4.31. Estimates for the size of binomial coefficients. Let α be a fixed complex number.
(a) Show that

$$|\binom{\alpha}{k}| \leq \prod_{j=1}^{k}(1 + \frac{|\alpha|}{j})$$

for all nonnegative integers k.
HINT: Note that

$$|\binom{\alpha}{k}| \leq \frac{|\alpha|(|alpha| + 1)(|alpha| + 2) \dots (|\alpha| + k - 1)}{k!}.$$

(b) Use part (a) to prove that there exists a constant C such that

$$|\binom{\alpha}{k}| \leq C2^k$$

for all nonnegative integers k.
HINT: Note that $(1 + |\alpha|/j) < 2$ for all $j > |\alpha|$.
(c) Show in fact that for each $\epsilon > 0$ there exists a constant C_ϵ such that

$$|\binom{\alpha}{k}| \leq C_\epsilon(1 + \epsilon)^k$$

for all nonnegative integers k.
(d) Let $h(t)$ be the power series function given by $h(t) = \sum_{k=0}^{\infty} \binom{\alpha}{k} t^k$. Use the ratio test to show that the radius of convergence for h equals 1.

REMARK. The general Binomial Theorem, if there is one, should be something like the following:

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

The problem is to determine when this infinite series converges, i.e., for what values of the three variables x, y, and α does it converge. It certainly is correct if $x = 0$, so we may as well assume that $x \neq 0$, in which case we are considering the validity of the formula

$$(x + y)^\alpha = x^\alpha (1 + t)^\alpha = x^\alpha \sum_{k=0}^{\infty} \binom{\alpha}{k} t^k,$$

where $t = y/x$. Therefore, it will suffice to determine for what values of t and α does the infinite series

$$\sum_{k=0}^{\infty} \binom{\alpha}{k} t^k$$

equal

$$(1 + t)^\alpha.$$

The answer is that, for n arbitrary complex number α, this series converges to the correct value for all $t \in (-1,1)$. (Of course, t must be larger than -1 for the expression $(1+t)^\alpha$ even to be defined.) However, the next theorem only establishes this equality for t's in the subinterval $(-1/2,1/2)$. As mentioned earlier, its proof is based on Taylor's Remainder Theorem. We must postpone the complete proof to the next chapter, where we will have a better version of Taylor's Theorem.

THEOREM 4.21. *Let $\alpha = a + bi$ be a fixed complex number. Then*

$$(1+t)^\alpha = \sum_{k=0}^{\infty} \binom{\alpha}{k} t^k$$

for all $t \in (-1/2, 1/2)$.

PROOF. Of course, this theorem is true if α is a nonnegative integer, for it is then just the original Binomial Theorem, and in fact in that case it holds for every complex number t. For a general complex number α, we have only defined x^α for positive x's, so that $(1+t)^\alpha$ is not even defined for $t < -1$.

Now, for a general $\alpha = a + bi$, consider the function $g : (-1/2,1/2) \to \mathbb{C}$ defined by $g(t) = (1+t)^\alpha$. Observe that the nth derivative of g is given by

$$g^{(n)}(t) = \frac{\alpha(\alpha-1)\dots(\alpha-n+1)}{(1+t)^{n-\alpha}}.$$

Then $g \in C^\infty((-1/2,1/2))$. (Of course, g is actually in $C^\infty(-1,1)$, but the present theorem is only concerned with t's in $(-1/2,1/2)$.)

For each nonnegative integer k define

$$a_k = g^{(k)}(0)/k! = \frac{\alpha(\alpha-1)\dots(\alpha-k+1)}{k!} = \binom{\alpha}{k},$$

and set h equal to the power series function given by $h(t) = \sum_{k=0}^\infty a_k t^k$. According to part (d) of the preceding exercise, the radius of convergence for the power series$\sum a_k t^k$ is 1. The aim of this theorem is to show that $g(t) = h(t)$ for all $-1/2 < t < 1/2$. In other words, we wish to show that g agrees with this power series function at least on the interval $(-1/2,1/2)$. It will suffice to show that the sequence $\{S_n\}$ of partial sums of the power series function h converges to the function g, at least on $(-1/2,1/2)$. We note also that the nth partial sum of this power series is just the nth Taylor polynomial T_g^n for g.

$$S_n(t) = \sum_{k=0}^{n} \binom{\alpha}{k} t^k = \sum_{k=0}^{n} \frac{g^{(k)}(0)}{k!} t^k.$$

Now, fix a t strictly between $-1/2$ and $1/2$, and let $r < 1$ be as in part (c) of Exercise 4.11. That is, $|t/(1+y)| < r$ for every y between 0 and t. (This is an important inequality for our proof, and this is one place where the hypothesis that $t \in (-1/2,1/2)$ is necessary.) Note also that, for any $y \in (-1/2,1/2)$, we have $|(1+y)^\alpha| = (1+y)^a$, and this is trapped between $(1/2)^a$ and $(3/2)^a$. Hence, there exists a number M such that $|(1+y)^\alpha| \leq M$ for all $y \in (-1/2,1/2)$.

Next, choose an $\epsilon > 0$ for which $\beta = (1+\epsilon)r < 1$. We let C_ϵ be a constant satisfying the inequality in Part (c) of Exercise 4.31. So, using Taylor's Remainder Theorem, we have that there exists a y between 0 and t for which

$$\begin{aligned}|g(t) - \sum_{k=0}^{n} a_k t^k| &= |g(t) - (T^n_{(g,0)}(t)| \\ &= |\frac{g^{(n+1)}(y)}{(n+1)!}t^{n+1}| \\ &= |\frac{\alpha(\alpha-1)\dots(\alpha-n)}{(n+1)!(1+y)^{n+1-\alpha}}t^{n+1}| \\ &\leq |\binom{\alpha}{n+1}||(1+y)^\alpha||\frac{t}{1+y}|^{n+1} \\ &\leq C_\epsilon(1+\epsilon)^{n+1}M|\frac{t}{1+y}|^{n+1} \\ &\leq C_\epsilon(1+\epsilon)^{n+1}Mr^{n+1} \\ &\leq C_\epsilon M\beta^{n+1},.\end{aligned}$$

Taking the limit as n tends to ∞, and recalling that $\beta < 1$, shows that $g(t) = h(t)$ for all $-1/2 < t < 1/2$, which completes the proof.

MORE ON PARTIAL DERIVATIVES

We close the chapter with a little more concerning partial derivatives. Thus far, we have discussed functions of a single variable, either real or complex. However, it is difficult not to think of a function of one complex variable $z = x+iy$ as equally well being a function of the two real variables x and y. We will write (a, b) and $a+bi$ to mean the same point in $\mathbb{C} \equiv \mathbb{R}^2$, and we will write $|(a, b)|$ and $|a + bi|$ to indicate the same quantity, i.e., the absolute value of the complex number $a + bi \equiv (a, b)$. We have seen in Theorem 4.4 that the only real-valued, differentiable functions of a complex variable are the constant functions. However, this is far from the case if we consider real-valued functions of two real variables, as is indicated in Exercise 4.8. Consequently, we make the following definition of differentiability of a real-valued function of two real variables. Note that it is clearly different from the definition of differentiability of a function of a single complex variable, and though the various notations for these two kinds of differentiability are clearly ambiguous, we will leave it to the context to indicate which kind we are using.

DEFINITION. Let $f : S \to \mathbb{R}$ be a function whose domain is a subset S of $\mathbb{R}^2$, and let $c = (a, b)$ be a point in the interior S^0 of S. We say that f is *differentiable, as a function of two real variables,* at the point (a, b) if there exists a pair of real numbers L_1 and L_2 and a function θ such that

$$f(a + h_1, b + h_2) - f(a, b) = L_1h_1 + L_2h_2 + \theta(h_1, h_2) \tag{4.8}$$

and

$$\lim_{|(h_1,h_2)|\to 0} \frac{\theta(h_1, h_2)}{|(h_1, h_2)|} = 0. \tag{4.9}$$

One should compare this definition with part (3) of Theorem 4.2.
Each partial derivative of a function f is again a real-valued function of two real variables, and so it can have partial derivatives of its own. We use simplifying notation like f_{xyxx} and $f_{yyyxyy\ldots}$ to indicate "higher order" mixed partial derivatives. For instance, f_{xxyx} denotes the fourth partial derivative of f, first with respect to x, second with respect to x again, third with respect to y, and finally fourth with respect to x. These higher order partial derivatives are called *mixed partial derivatives.*

DEFINITION. Suppose S is a subset of $\mathbb{R}^2$, and that f is a continuous real-valued function on S. If both partial derivatives of f exist at each point of the interior S^0 of S, and both are continuous on S^0, then f is said to belong to $C^1(S)$. If all kth order mixed partial derivatives exist at each point of S^0, and all of them are continuous on S^0, then f is said to belong to $C^k(S)$. Finally, if all mixed partial derivatives, of arbitrary orders, exist and are continuous on S^0, then f is said to belong to $C^\infty(S)$.

Exercise 4.32. (a) Suppose f is a real-valued function of two real variables and that it is differentiable, as a function of two real variables, at the point (a, b). Show that the numbers L_1 and L_2 in the definition are exactly the partial derivatives of f at (a, b). That is,

$$L_1 = \frac{\partial f}{\partial x}(a,b) = \lim_{h\to 0} \frac{f(a+h,b) - f(a,b)}{h}$$

and

$$L_2 = \frac{\partial f}{\partial y}(a,b) = \lim_{h\to 0} \frac{f(a,b+h) - f(a,b)}{h}.$$

(b) Define f on $\mathbb{R}^2$ as follows: $f(0,0) = 0$, and if $(x,y) \neq (0,0)$, then $f(x,y) = xy/(x^2+y^2)$. Show that both partial derivatives of f at $(0,0)$ exist and are 0. Show also that f is **not**, as a function of two real variables, differentiable at $(0,0)$.
HINT: Let h and k run through the numbers $1/n$.
(c) What do parts (a) and (b) tell about the relationship between a function of two real variables being differentiable at a point (a,b) and its having both partial derivatives exist at (a,b)?
(d) Suppose $f = u + iv$ is a complex-valued function of a complex variable, and assume that f is differentiable, as a function of a complex variable, at a point $c = a + bi \equiv (a,b)$. Prove that the real and imaginary parts u and v of f are differentiable, as functions of two real variables. Relate the five quantities

$$\frac{\partial u}{\partial x}(a,b), \frac{\partial u}{\partial y}(a,b), \frac{\partial v}{\partial x}(a,b), \frac{\partial v}{\partial y}(a,b), \text{ and } f'(c).$$

Perhaps the most interesting theorem about partial derivatives is the "mixed partials are equal" theorem. That is, $f_{xy} = f_{yx}$. The point is that this is**not** always the case. An extra hypothesis is necessary.

THEOREM 4.22. (Theorem on mixed partials) Let $f : S \to \mathbb{R}$ be such that both second order partials derivatives f_{xy} and f_{yx} exist at a point (a,b) of the interior of S, and assume in addition that one of these second order partials exists at every point in a disk $B_r(a,b)$ around (a,b) and that it is continuous at the point (a,b). Then $f_{xy}(a,b) = f_{yx}(a,b)$.

PROOF. Suppose that it is f_{yx} that is continuous at (a,b). Let $\epsilon > 0$ be given, and let $\delta_1 > 0$ be such that if $|(c,d)-(a,b)| < \delta_1$ then $|f_{yx}(c,d) - f_{yx}(a,b)| < \epsilon$. Next, choose a δ_2 such that if $0 < |k| < \delta_2$, then

$$|f_{xy}(a,b) - \frac{f_x(a,b+k) - f_x(a,b)}{k}| < \epsilon,$$

and fix such a k. We may also assume that $|k| < \delta_1/2$. Finally, choose a $\delta_3 > 0$ such that if $0 < |h| < \delta_3$, then

$$|f_x(a,b+k) - \frac{f(a+h,b+k) - f(a,b+k)}{h}| < |k|\epsilon,$$

and

$$|f_x(a,b) - \frac{f(a+h,b) - f(a,b)}{h}| < |k|\epsilon,$$

and fix such an h. Again, we may also assume that $|h| < \delta_1/2$.
In the following calculation we will use the Mean Value Theorem twice.

$$\begin{aligned}
0 &\le |f_{xy}(a,b) - f_{yx}(a,b)| \\
&\le |f_{xy}(a,b) - \frac{f_x(a,b+k) - f_x(a,b)}{k}| \\
&\quad + |\frac{f_x(a,b+k) - f_x(a,b)}{k} - f_{yx}(a,b)| \\
&\le \epsilon + |\frac{f_x(a,b+k) - \frac{f(a+h,b+k) - f(a,b+k)}{h}}{k}| \\
&\quad + |\frac{\frac{f(a+h,b) - f(a,b)}{h} - f_x(a,b)}{k}| \\
&\quad + |\frac{f(a+h,b+k) - f(a,b+k) + (f(a+h,b) - f(a,b))}{hk} - f_{yx}(a,b)| \\
&< 3\epsilon + |\frac{f(a+h,b+k) - f(a,b+k) + (f(a+h,b) - f(a,b))}{hk} - f_{yx}(a,b)| \\
&= 3\epsilon + |\frac{f_y(a+h,b') - f_y(a,b')}{h} - f_{yx}(a,b)| \\
&= 3\epsilon + |f_{yx}(a',b') - f_{yx}(a,b)| \\
&< 4\epsilon,
\end{aligned}$$

because b' is between b and $b+k$, and a' is between a and $a+h$, so that $|(a',b') - (a,b)| < \delta_1/\sqrt{2} < \delta_1$. Hence, $|f_{xy}(a,b) - f_{yx}(a,b) < 4\epsilon$, for an arbitrary ϵ, and so the theorem is proved.

Exercise 4.33. Let f be defined on $\mathbb{R}^2$ by $f(0,0) = 0$ and, for $(x,y) \neq (0,0)$, $f(x,y) = x^3y/(x^2+y^2)$.
(a) Prove that both partial derivatives f_x and f_y exist at each point in the plane.
(b) Show that $f_{yx}(0,0) = 1$ and $f_{xy}(0,0) = 0$.

(c) Show that f_{xy} exists at each point in the plane, but that it is not continuous at $(0,0)$.

The following exercise is an obvious generalization of the First Derivative Test for Extreme Values, Theorem 4.8, to real-valued functions of two real variables.

Exercise 4.34. Let $f : S \to \mathbb{R}$ be a real-valued function of two real variables, and let $c = (a,b) \in S^0$ be a point at which f attains a local maximum or a local minimum. Show that if either of the partial derivatives $\partial f/\partial x$ or $\partial f/\partial y$ exists at c, then it must be equal to 0.
HINT: Just consider real-valued functions of a real variable like $x \to f(x,b)$ or $y \to f(a,y)$, and use Theorem 4.8.

Whenever we make a new definition about functions, the question arises of how the definition fits with algebraic combinations of functions and how it fits with the operation of composition. In that light, the next theorem is an expected one.

THEOREM 4.23. (Chain Rule again) Suppose S is a subset of $\mathbb{R}^2$, that (a,b) is a point in the interior of S, and that $f : S \to \mathbb{R}$ is a real-valued function that is differentiable, as a function of two real variables, at the point (a,b). Suppose that T is a subset of $\mathbb{R}$, that c belongs to the interior of T, and that $\phi : T \to \mathbb{R}^2$ is differentiable at the point c and $\phi(c) = (a,b)$. Write $\phi(t) = (x(t), y(t))$. Then the composition $f \circ \phi$ is differentiable at c and

$$f \circ \phi'(c) = \frac{\partial f}{\partial x}(a,b)x'(c) + \frac{\partial f}{\partial y}(a,b)y'(c) = \frac{\partial f}{\partial x}(\phi(c))x'(c) + \frac{\partial f}{\partial y}(\phi(c))y'(c).$$

PROOF. From the definition of differentiability of a real-valued function of two real variables, write

$$f(a+h_1, b+h_2) - f(a,b) = L_1h_1 + L_2h_2 + \theta_f(H_1, h_2).$$

and from part (3) of Theorem 4.2, write

$$\phi(c+h) - \phi(c) = \phi'(c)h + \theta_\phi(h),$$

or, in component form,

$$x(c+h) - x(c) = x(c+h) - a = x'(c)h + \theta_x(h)$$

and

$$y(c+h) - y(c) = y(c+h) - b = y'(c)h + \theta_y(h).$$

We also have that

$$\lim_{|(h_1,h_2)|\to 0} \frac{\theta_f((h_1,h_2))}{|(h_1,h_2)|} = 0,$$

$$\lim_{h\to 0} \frac{\theta_x(h)}{h} = 0,$$

and

$$\lim_{h\to 0} \frac{\theta_y(h)}{h} = 0.$$

We will show that $f\circ\phi$ is differentiable at c by showing that there exists a number L and a function θ satisfying the two conditions of part (3) of Theorem 4.2. Define

$$k_1(h),k_2(h)) = \phi(c+h)-\phi(c) = (x(c+h)-x(c), y(c+h)-y(c)).$$

Thus, we have that

$$\begin{aligned} f\circ\phi(c+h)-f\circ\phi(c) &= f(\phi(c+h))-f(\phi(c)) \\ &= f(x(c+h),y(c+h))-f(x(c),y(c)) \\ &= f(a+k_1(h),b+k_2(h))-f(a,b) \\ &= L_1k_1(h)+L_2k_2(h)+\theta_f(k_1(h),k_2(h)) \\ &= l_1(x(c+h)-x(c))+L_2(y(c+h)-y(c)) \\ &\qquad +\theta_f(k_1(h),k_2(h)) \\ &= L_1(x'(c)h+\theta_x(h))+L_2(y'(c)h+\theta_y(h)) \\ &\qquad +\theta_f(k_1(h),k_2(h)) \\ &= (L_1x'(c)+L_2y'(c))h \\ &\qquad +L_1\theta_x(h)+L_2\theta_y(h)+\theta_f(k_1(h),k_2(h)). \end{aligned}$$

We define $L=(L_1x'(c)+L_2y'(c))$ and $\theta(h)=l_1\theta_x(h)+L_2\theta_y(h)+\theta_f(k_1(h),k_2(h))$. By these definitions and the calculation above we have Equation (4.1)

$$f\circ\phi(c+h)-f\circ\phi(c)=Lh+\theta(h),$$

so that it only remains to verify Equation (4.2) for the function θ. We have seen above that the first two parts of θ satisfy the desired limit condition, so that it is just the third part of θ that requires some proof. The required argument is analogous to the last part of the proof of the Chain Rule (Theorem 4.7), and we leave it as an exercise.

Exercise 4.35. (a) Finish the proof to the preceding theorem by showing that

$$\lim_{h\to 0}\frac{\theta_f(k_1(h),k_2(h))}{h}=0.$$

HINT: Review the corresponding part of the proof to Theorem 4.7.
(b) Suppose $f:S\to\mathbb{R}$ is as in the preceding theorem and that ϕ is a real-valued function of a real variable. Suppose f is differentiable, as a function of two real variables, at the point (a,b), and that ϕ is differentiable at the point $c=f(a,b)$. Let $g=\phi\circ f$. Find a formula for the partial derivatives of the real-valued function g of two real variables.
(c) (A generalized Mean Value Theorem) Suppose u is a real-valued function of two real variables, both of whose partial derivatives exist at each point in a disk $B_r(a,b)$. Show that, for any two points (x,y) and (x',y') in $B_r(a,b)$, there exists a point $(\hat{x},\hat{y})$ on the line segment joining (x,y) to (x',y') such that

$$u(x,y)-u(x',y')=\frac{\partial u}{\partial x}(\hat{x},\hat{y})(x-x')+\frac{\partial u}{\partial y}(\hat{x},\hat{y})(y-y').$$

HINT: Let $\phi : [0,1] \to \mathbb{R}^2$ be defined by $\phi(t) = (1-t)(x', y') + t(x, y)$. Now use the preceding theorem.
(d) Verify that the assignment $f \to \partial f/\partial x$ is linear; i.e., that

$$\frac{\partial(f+g)}{\partial x} = \frac{\partial f}{\partial x} + \frac{\partial g}{\partial x}.$$

Check that the same is true for partial derivatives with respect to y.

CHAPTER V
INTEGRATION, AVERAGE BEHAVIOR
$A = \pi r^2$.

In this chapter we will derive the formula $A = \pi r^2$ for the area of a circle of radius r. As a matter of fact, we will first have to settle on exactly what is the definition of the area of a region in the plane, and more subtle than that, we must decide what kinds of regions in the plane "have" areas. Before we consider the problem of area, we will develop the notion of the integral (or average value) of a function defined on an interval $[a, b]$, which notion we will use later to compute areas, once they have been defined.

The main results of this chapter include:

(1) The definition of **integrability** of a function, and the definition of the **integral** of an integrable function,
(2) The **Fundamental Theorem of Calculus** (Theorem 5.9),
(3) The **Integral Form of Taylor's Remainder Theorem** (Theorem 5.12),
(4) The **General Binomial Theorem** (Theorem 5.13),
(5) The definition of the **area** of a **geometric set,**
(6) $A = \pi r^2$ (Theorem 5.15), and
(7) The **Integral Test** (Theorem 5.17).

INTEGRALS OF STEP FUNCTIONS

We begin by defining the integral of certain (but not all) bounded, real-valued functions whose domains are closed bounded intervals. Later, we will extend the definition of integral to certain kinds of unbounded complex-valued functions whose domains are still intervals, but which need not be either closed or bounded. First, we recall from Chapter III the following definitions.

DEFINITION. Let $[a, b]$ be a closed bounded interval of real numbers. By a *partition* of $[a, b]$ we mean a finite set $P = \{x_0 < x_1 < \ldots < x_n\}$ of $n + 1$ points, where $x_0 = a$ and $x_n = b$.

The n intervals $\{[x_{i-1}, x_i]\}$ are called the *closed subintervals* of the partition P, and the n intervals $\{(x_{i-1}, x_i)\}$ are called the *open subintervals* or *elements* of P.

We write $\|P\|$ for the maximum of the numbers (lengths of the subintervals) $\{x_i - x_{i-1}\}$, and call $\|P\|$ the *mesh size* of the partition P.

If a partition $P = \{x_i\}$ is contained in another partition $Q = \{y_j\}$, i.e., each x_i equals some y_j, then we say that Q is *finer* than P.

Let f be a function on an interval $[a, b]$, and let $P = \{x_0 < \ldots < x_n\}$ be a partition of $[a, b]$. Physicists often consider sums of the form

$$S_{P,\{y_i\}} = \sum_{i=1}^{n} f(y_i)(x_i - x_{i-1}),$$

where y_i is a point in the subinterval (x_{i-1}, x_i). These sums (called Riemann sums) are approximations of physical quantities, and the limit of these sums, as the mesh of the partition becomes smaller and smaller, should represent a precise value of the physical quantity. What precisely is meant by the " limit" of such sums is already a subtle question, but even having decided on what that definition should be, it is as important and difficult to determine whether or not such a limit exists for many (or even any) functions f. We approach this question from a slightly different point of view, but we will revisit Riemann sums in the end.

Again we recall from Chapter III the following.

DEFINITION. Let $[a,b]$ be a closed bounded interval in $\mathbb{R}$. A real-valued function $h : [a,b] \to \mathbb{R}$ is called a *step function* if there exists a partition $P = \{x_0 < x_1 < \ldots < x_n\}$ of $[a,b]$ such that for each $1 \leq i \leq n$ there exists a number a_i such that $h(x) = a_i$ for all $x \in (x_{i-1}, x_i)$.

REMARK. A step function h is constant on the open subintervals (or elements) of a certain partition. Of course, the partition is not unique. Indeed, if P is such a partition, we may add more points to it, making a larger partition having more subintervals, and the function h will still be constant on these new open subintervals. That is, a given step function can be described using various distinct partitions.

Also, the values of a step function at the partition points themselves is irrelevant. We only require that it be constant on the open subintervals.

Exercise 5.1. Let h be a step function on $[a,b]$, and let $P = \{x_0 < x_1 < \ldots < x_n\}$ be a partition of $[a,b]$ such that $h(x) = a_i$ on the subinterval (x_{i-1}, x_i) determined by P.

(a) Prove that the range of h is a finite set. What is an upper bound on the cardinality of this range?

(b) Prove that h is differentiable at all but a finite number of points in $[a,b]$. What is the value of h' at such a point?

(c) Let f be a function on $[a,b]$. Prove that f is a step function if and only if $f'(x)$ exists and $= 0$ for every $x \in (a,b)$ except possibly for a finite number of points.

(d) What can be said about the values of h at the endpoints $\{x_i\}$ of the subintervals of P?

(e) Let h be a step function on $[a,b]$, and let j be a function on $[a,b]$ for which $h(x) = j(x)$ for all $x \in [a,b]$ except for one point c. Show that j is also a step function.

(f) If k is a function on $[a,b]$ that agrees with a step function h except at a finite number of points $c_1, c_2, \ldots, c_N$, show that k is also a step function.

Exercise 5.2. Let $[a,b]$ be a fixed closed bounded interval in $\mathbb{R}$, and let $H([a,b])$ denote the set of all step functions on $[a,b]$.

(a) Using Part (c) of Exercise 5.1, prove that the set $H([a,b])$ is a vector space of functions; i.e., it is closed under addition and scalar multiplication.

(b) Show that $H([a,b])$ is closed under multiplication; i.e., if $h_1, h_2 \in H([a,b])$, then $h_1h_2 \in H([a,b])$.

(c) Show that $H([a,b])$ is closed under taking maximum and minimum and that it contains all the real-valued constant functions.

(d) We call a function χ an *indicator function* if it equals 1 on an interval (c,d) and is 0 outside $[c,d]$. To be precise, we will denote this indicator function by $\chi_{(c,d)}$. Prove that every indicator function is a step function, and show also that every step function h is a linear combination of indicator functions:

$$h = \sum_{j=1}^{n} a_j \chi_{(c_j,d_j)}.$$

(e) Define a function k on $[0,1]$ by setting $k(x)=0$ if x is a rational number and $k(x)=1$ if x is an irrational number. Prove that the range of k is a finite set, but that k is **not** a step function.

Our first theorem in this chapter is a fundamental consistency result about the "area under the graph" of a step function. Of course, the graph of a step function looks like a collection of horizontal line segments, and the region under this graph is just a collection of rectangles. Actually, in this remark, we are implicitly thinking that the values $\{a_i\}$ of the step function are positive. If some of these values are negative, then we must re-think what we mean by the area under the graph. We first introduce the following bit of notation.

DEFINITION. Let h be a step function on the closed interval $[a,b]$. Suppose $P=\{x_0<x_1<\ldots<x_n\}$ is a partition of $[a,b]$ such that $h(x)=a_i$ on the interval (x_{i-1},x_i). Define the *weighted average* of h *relative to* P to be the number $S_P(h)$ defined by

$$S_P(h)=\sum_{i=1}^n a_i(x_i-x_{i-1}).$$

REMARK. Notice the similarity between the formula for a weighted average and the formula for a Riemann sum. Note also that if the interval is a single point, i.e., $a=b$, then the only partition P of the interval consists of the single point $x_0=a$, and every weighted average $S_P(h)=0$.

The next theorem is not a surprise, although its proof takes some careful thinking. It is simply the assertion that the weighted averages are independent of the choice of partition.

THEOREM 5.1. *Let h be a step function on the closed interval $[a,b]$. Suppose $P=\{x_0<x_1<\ldots<x_n\}$ is a partition of $[a,b]$ such that $h(x)=a_i$ on the interval (x_{i-1},x_i), and suppose $Q=\{y_0<y_1<\ldots<y_m\}$ is another partition of $[a,b]$ such that $h(x)=b_j$ on the interval (y_{j-1},y_j). Then the weighted average of h relative to P is the same as the weighted average of h relative to Q. That is, $S_P(h)=S_Q(h)$.*

PROOF. Suppose first that the partition Q is obtained from the partition P by adding one additional point. Then $m=n+1$, and there exists an i_0 between 1 and $n-1$ such that

(1) for $0\leq i\leq i_0$ we have $y_i=x_i$.
(2) $x_{i_0}<y_{i_0+1}<x_{i_0+1}$.
(3) For $i_0<i\leq n$ we have $x_i=y_{i+1}$.

In other words, y_{i_0+1} is the only point of Q that is not a point of P, and y_{i_0+1} lies strictly between x_{i_0} and x_{i_0+1}.

Because h is constant on the interval $(x_{i_0},x_{i_0+1})=(y_{i_0},y_{i_0+2})$, it follows that

(1) For $1\leq i\leq i_0$, $a_i=b_i$.
(2) $b_{i_0+1}=b_{i_0+2}=a_{i_0+1}$.
(3) For $i_0+1\leq i\leq n$, $a_i=b_{i+1}$.

So,

$$\begin{aligned}
S_P(h) &= \sum_{i=1}^{n} a_i(x_i - x_{i-1}) \\
&= \sum_{i=1}^{i_0} a_i(x_i - x_{i-1}) + a_{i_0+1}(x_{i_0+1} - x_{i_0}) \\
&\qquad + \sum_{i=i_0+2}^{n} a_i(x_i - x_{i-1}) \\
&= \sum_{i=1}^{i_0} b_i(y_i - y_{i-1}) + a_{i_0+1}(y_{i_0+2} - y_{i_0}) \\
&\qquad + \sum_{i=i_0+2}^{n} b_{i+1}(y_{i+1} - y_i) \\
&= \sum_{i=1}^{i_0} b_i(y_i - y_{i-1}) + a_{i_0+1}(y_{i_0+2} - y_{i_0+1} + y_{i_0+1} - y_{i_0}) \\
&\qquad + \sum_{i=i_0+3}^{n+1} b_i(y_i - y_{i-1}) \\
&= \sum_{i=1}^{i_0} b_i(y_i - y_{i-1}) + b_{i_0+1}(y_{i_0+1} - y_{i_0}) + b_{i_0+2}(y_{i_0+2} - y_{i_0+1}) \\
&\qquad + \sum_{i=i_0+3}^{m} b_i(y_i - y_{i-1}) \\
&= \sum_{i=1}^{m} b_i(y_i - y_{i-1}) \\
&= S_Q(h),
\end{aligned}$$

which proves the theorem in this special case where Q is obtained from P by adding just one more point.

It follows easily now by induction that if Q is obtained from P by adding any finite number of additional points, then h is constant on each of the open subintervals determined by Q, and $S_Q(h) = S_P(h)$.

Finally, let $Q = \{y_0 < y_1 < \ldots < y_m\}$ be an arbitrary partition of $[a, b]$, for which h is constant on each of the open subintervals (y_{j-1}, y_j) determined by Q. Define R to be the partition of $[a, b]$ obtained by taking the union of the partition points $\{x_i\}$ and $\{y_j\}$. Then R is a partition of $[a, b]$ that is obtained by adding a finite number of points to the partition P, whence $S_R(h) = S_P(h)$. Likewise, R is obtained from the partition Q by adding a finite number of points, whence $S_R(h) = S_Q(h)$, and this proves that $S_Q(h) = S_P(h)$, as desired.

DEFINITION. Let $[a, b]$ be a fixed closed bounded interval in $\mathbb{R}$. We define the *integral* of a step function h on $[a, b]$, and denote it by $\int h$, as follows: If $P = \{x_0 < x_1 < \ldots < x_n\}$ is a partition of $[a, b]$, for which $h(x) = a_i$ for all $x \in (x_{i-1}, x_i)$, then

$$\int h = S_P(h) = \sum_{i=1}^{n} a_i(x_i - x_{i-1}).$$

REMARK. The integral of a step function h is defined to be the weighted average of h relative to a partition P of $[a,b]$. Notice that the preceding theorem is crucial in order that this definition of $\int h$ be unambiguously defined. The integral of a step function should **not** depend on which partition is used. Theorem 5.1 asserts precisely this fact.

Note also that if the interval is a single point, i.e., $a = b$, then the integral of every step function h is 0.

We use a variety of notations for the integral of h :

$$\int h = \int_a^b h = \int_a^b h(t)\,dt.$$

The following exercise provides a very useful way of describing the integral of a step function. Not only does it show that the integral of a step function looks like a Riemann sum, but it provides a description of the integral that makes certain calculations easier. See, for example, the proof of the next theorem.

Exercise 5.3. Suppose h is a step function on $[a,b]$ and that $R = \{z_0 < z_1 < \ldots < z_n\}$ is a partition of $[a,b]$ for which h is constant on each subinterval (z_{i-1}, z_i) of R.

(a) Prove that

$$\int h = S_R(h) = \sum_{i=1}^{n} h(w_i)(z_i - z_{i-1}),$$

where, for each $1 \leq i \leq n$, w_i is any point in (z_{i-1}, z_i). (Note then that the integral of a step function takes the form of a Riemann sum.)

(b) Show that $\int h$ is independent of the values of h at the points $\{z_i\}$ of the partition R.

Exercise 5.4. Let h_1 and h_2 be two step functions on $[a,b]$.

(a) Suppose that $h_1(x) = h_2(x)$ for all $x \in [a,b]$ except for one point c. Prove that $\int h_1 = \int h_2$.

HINT: Let P be a partition of $[a,b]$, for which both h_1 and h_2 are constant on its open subintervals, and for which c is one of the points of P. Now use the preceding exercise to calculate the two integrals.

(b) Suppose $h_1(x) = h_2(x)$ for all but a finite number of points $c_1, \ldots, c_N \in [a,b]$. Prove that $\int h_1 = \int h_2$.

We have used the terminology "weighted average" of a step function relative to a partition P. The next exercise shows how the integral of a step function can be related to an actual average value of the function.

Exercise 5.5. Let h be a step function on the closed interval $[a,b]$, and let $P = \{x_0 < x_1 < \ldots < x_n\}$ be a partition of $[a,b]$ for which $h(x) = a_i$ on the interval (x_{i-1}, x_i). Let us think of the interval $[a,b]$ as an interval of time, and suppose that the function h assumes the value a_i for the interval of time between x_{i-1} and

x_i. Show that the average value $A(h)$ taken on by h throughout the entire interval ($[a,b]$) of time is given by

$$A(h) = \frac{\int h}{b-a}.$$

THEOREM 5.2. *Let $H([a,b])$ denote the vector space of all step functions on the closed interval $[a,b]$. Then the assignment $h \to \int h$ of $H([a,b])$ into $\mathbb{R}$ has the following properties:*

(1) (Linearity) $H([a,b])$ is a vector space. Furthermore, $\int (h_1 + h_2) = \int h_1 + \int h_2$, and $\int ch = c \int h$ for all $h_1, h_2, h \in H([a,b])$, and for all real numbers c.

(2) If $h = \sum_{i=1}^n a_i \chi_{(c_i,d_i)}$ is a linear combination of indicator functions (See part (d) of Exercise 5.2), then $\int h = \sum_{i=1}^n a_i(d_i - c_i)$.

(3) (Positivity) If $h(x) \geq 0$ for all $x \in [a,b]$, then $\int h \geq 0$.

(4) (Order-preserving) If h_1 and h_2 are step functions for which $h_1(x) \leq h_2(x)$ for all $x \in [a,b]$, then $\int h_1 \leq \int h_2$.

PROOF. That $H([a,b])$ is a vector space was proved in part (a) of Exercise 5.2. Suppose $P = \{x_0 < x_1 < \ldots < x_n\}$ is a partition of $[a,b]$ such that $h_1(x)$ is constant for all $x \in (x_{i-1}, x_i)$, and suppose $Q = \{y_0 < y_1 < \ldots < y_m\}$ is a partition of $[a,b]$ such that $h_2(x)$ is constant for all $x \in (y_{j-1}, y_j)$. Let $R = \{z_0 < z_1 < \ldots < z_r\}$ be the partition of $[a,b]$ obtained by taking the union of the x_i's and the y_j's. Then h_1 and h_2 are both constant on each open subinterval of R, since each such subinterval is contained in some open subinterval of P and also is contained in some open subinterval of Q. Therefore, $h_1 + h_2$ is constant on each open subinterval of R. Now, using Exercise 5.3, we have that

$$\begin{aligned}\int (h_1 + h_2) &= \sum_{k=1}^r ((h_1 + h_2)(w_k))(z_k - z_{k-1}) \\ &= \sum_{k=1}^r h_1(w_k)(z_k - z_{k-1}) + \sum_{k=1}^r h_2(w_k)(z_k - z_{k-1}) \\ &= \int h_1 + \int h_2.\end{aligned}$$

This proves the first assertion of part (1).

Next, let $P = \{x_0 < x_1 < \ldots < x_n\}$ be a partition of $[a,b]$ such that $h(x)$ is constant on each open subinterval of P. Then $ch(x)$ is constant on each open subinterval of P, and using Exercise 5.3 again, we have that

$$\begin{aligned}\int (ch) &= \sum_{i=1}^n ch(w_i)(x_i - x_{i-1}) \\ &= c \sum_{i=1}^n h(w_i)(x_i - x_{i-1}) \\ &= c \int h,\end{aligned}$$

which completes the proof of the other half of part (1).

To see part (2), we need only verify that $\int \chi_{(c_i,d_i)} = d_i - c_i$, for then part (2) will follow from part (1). But $\chi_{(c_i,d_i)}$ is just a step function determined by the four point partition $\{a, c_i, d_i, b\}$ and values 0 on (a, c_i) and (d_i, b) and 1 on (c_i, d_i). Therefore, we have that $\int \chi_{(c_i,d_i)} = d_i - c_i$.

If $h(x) \geq 0$ for all x, and $P = \{x_0 < x_1 < \ldots < x_n\}$ is as above, then

$$\int h = \sum_{i=1}^{n} h(w_i)(x_i - x_{i-1}) \geq 0,$$

and this proves part (3).

Finally, suppose $h_1(x) \leq h_2(x)$ for all $x \in [a, b]$. By Exercise 5.2, we know that the function $h_3 = h_2 - h_1$ is a step function on $[a, b]$. Also, $h_3(x) \geq 0$ for all $x \in [a, b]$. So, by part (3), $\int h_3 \geq 0$. Then, by part (1),

$$0 \leq \int h_3 = \int (h_2 - h_1) = \int h_2 - \int h_1,$$

which implies that $\int h_1 \leq \int h_2$, as desired.

Exercise 5.6. (a) Let h be the constant function c on $[a, b]$. Show that $\int h = c(b - a)$.

(b) Let $a < c < d < b$ be real numbers, and let h be the step function on $[a, b]$ that equals r for $c < x < d$ and 0 otherwise. Prove that $\int_a^b h(t)\, dt = r(d - c)$.

(c) Let h be a step function on $[a, b]$. Prove that $|h|$ is a step function, and that $|\int h| \leq \int |h|$.

HINT: Note that $-|h|(x) \leq h(x) \leq |h|(x)$. Now use the preceding theorem.

(d) Suppose h is a step function on $[a, b]$ and that c is a constant for which $|h(x)| \leq c$ for all $x \in [a, b]$. Prove that $|\int h| \leq c(b - a)$.

INTEGRABLE FUNCTIONS

We now wish to extend the definition of the integral to a wider class of functions. This class will consist of those functions that are **uniform limits** of step functions. The requirement that these limits be uniform is crucial. Pointwise limits of step functions doesn't work, as we will see in Exercise 5.7 below. The initial step in carrying out this generalization is the following.

THEOREM 5.3. *Let $[a, b]$ be a closed bounded interval, and let $\{h_n\}$ be a sequence of step functions that converges uniformly to a function f on $[a, b]$. Then the sequence $\{\int h_n\}$ is a convergent sequence of real numbers.*

PROOF. We will show that $\{\int h_n\}$ is a Cauchy sequence in $\mathbb{R}$. Thus, given an $\epsilon > 0$, choose an N such that for any $n \geq N$ and any $x \in [a, b]$, we have

$$|f(x) - h_n(x)| < \frac{\epsilon}{2(b - a)}.$$

Then, for any m and n both $\geq N$ and any $x \in [a, b]$, we have

$$|h_n(x) - h_m(x)| \leq |h_n(x) - f(x)| + |f(x) - h_m(x)| < \frac{\epsilon}{b - a}.$$

Therefore,

$$|\int h_n - \int h_m| = |\int (h_n - h_m)| \leq \int |h_n - h_m| \leq \int \frac{\epsilon}{b-a} = \epsilon,$$

as desired.

The preceding theorem provides us with a perfectly good idea of how to define the integral of a function f that is the uniform limit of a sequence of step functions. However, we first need to establish another kind of consistency result.

THEOREM 5.4. *If $\{h_n\}$ and $\{k_n\}$ are two sequences of step functions on $[a, b]$, each converging uniformly to the same function f, then*

$$\lim \int h_n = \lim \int k_n.$$

PROOF. Given $\epsilon > 0$, choose N so that if $n \geq N$, then $|h_n(x) - f(x)| < \epsilon/(2(b-a))$ for all $x \in [a, b]$, and such that $|f(x) - k_n(x)| < \epsilon/(2(b-a))$ for all $x \in [a, b]$. Then, $|h_n(x) - k_n(x)| < \epsilon/(b-a)$ for all $x \in [a, b]$ if $n \geq N$. So,

$$|\int h_n - \int k_n| \leq \int |h_n - k_n| \leq \int \frac{\epsilon}{b-a} = \epsilon$$

if $n \geq N$. Taking limits gives

$$|\lim \int h_n - \lim \int k_n| \leq \epsilon.$$

Since this is true for arbitrary $\epsilon > 0$, it follows that $\lim \int h_n = \lim \int k_n$, as desired.

DEFINITION. Let $[a, b]$ be a closed bounded interval of real numbers. A function $f : [a, b] \to \mathbb{R}$ is called *integrable* on $[a, b]$ if it is the uniform limit of a sequence $\{h_n\}$ of step functions.
Let $I([a, b])$ denote the set of all functions that are integrable on $[a, b]$. If $f \in I([a, b])$, define the *integral* of f, denoted $\int f$, by

$$\int f = \lim \int h_n,$$

where $\{h_n\}$ is some (any) sequence of step functions that converges uniformly to f on $[a, b]$.

As in the case of step functions, we use the following notations:

$$\int f = \int_a^b f = \int_a^b f(t)\,dt.$$

REMARK. Note that Theorem 5.4 is crucial in order that this definition be unambiguous. Indeed, we will see below that this critical consistency result is one place where uniform limits of step functions works while pointwise limits do not. See parts (c) and (d) of Exercise 5.7. Note also that it follows from this definition that $\int_a^a f = 0$, because $\int_a^a h = 0$ for any step function. In fact, we will derive almost everything about the integral of a general integrable function from the corresponding results about the integral of a step function. No surprise. This is the essence of mathematical analysis, approximation.

Exercise 5.7. Define a function f on the closed interval $[0,1]$ by $f(x)=1$ if x is a rational number and $f(x)=0$ if x is an irrational number.
(a) Suppose h is a step function on $[0,1]$. Prove that there must exist an $x\in[0,1]$ such that $|f(x)-h(x)|\geq 1/2$.
HINT: Let (x_{i-1},x_i) be an interval on which h is a constant c. Now use the fact that there are both rationals and irrationals in this interval.
(b) Prove that f is not the uniform limit of a sequence of step functions. That is, f is **not** an integrable function.
(c) Consider the two sequences $\{h_n\}$ and $\{k_n\}$ of step functions defined on the interval $[0,1]$ by $h_n=\chi_{(0,1/n)}$, and $k_n=n\chi_{(0,1/n)}$. Show that both sequences $\{h_n\}$ and $\{k_n\}$ converge pointwise to the 0 function on $[0,1]$.
HINT: All functions are 0 at $x=0$. For $x>0$, choose N so that $1/N<x$. Then, for any $n\geq N$, $h_n(x)=k_n(x)=0$.
(d) Let h_n and k_n be as in part (c). Show that $\lim\int h_n=0$, but $\lim\int k_n=1$. Conclude that the consistency result in Theorem 5.4 does not hold for pointwise limits of step functions.

Exercise 5.8. Define a function f on the closed interval $[0,1]$ by $f(x)=x$.
(a) For each positive integer n, let P_n be the partition of $[0,1]$ given by the points $\{0<1/n<2/n<3/n<\ldots<(n-1)/n<1\}$. Define a step function h_n on $[0,1]$ by setting $h_n(x)=i/n$ if $\frac{i-1}{n}<x<\frac{i}{n}$, and $h_n(i/n)=i/n$ for all $0\leq i\leq n$. Prove that $|f(x)-h_n(x)|<1/n$ for all $x\in[0,1]$, and then conclude that f is the uniform limit of the h_n's whence $f\in I([0,1])$.
(b) Show that

$$\int h_n=\sum_{i=1}^{n}\frac{i}{n^2}=\frac{n(n+1)}{2n^2}.$$

(c) Show that $\int_0^1 f(t)\,dt=1/2$.

The next exercise establishes some additional properties of integrable functions on an interval $[a,b]$.

Exercise 5.9. Let $[a,b]$ be a closed and bounded interval, and let f be an element of $I([a,b])$.
(a) Show that, for each $\epsilon>0$ there exists a step function h on $[a,b]$ such that $|f(x)-h(x)|<\epsilon$ for all $x\in[a,b]$.
(b) For each positive integer n let h_n be a step function satisfying the conclusion of part (a) for $\epsilon=1/n$. Define $k_n=h_n-1/n$ and $l_n=h_n+1/n$. Show that k_n and l_n are step functions, that $k_n(x)<f(x)<l_n(x)$ for all $x\in[a,b]$, and that $|l_n(x)-k_n(x)|=l_n(x)-k_n(x)=2/n$ for all x. Hence, $\int_a^b(l_n-k_n)=\frac{2}{n}(b-a)$.
(c) Conclude from part (b) that, given any $\epsilon>0$, there exist step functions k and l such that $k(x)\leq f(x)\leq l(x)$ for which $\int(l(x)-k(x))<\epsilon$.
(d) Prove that there exists a sequence $\{j_n\}$ of step functions on $[a,b]$, for which $j_n(x)\leq j_{n+1}(x)\leq f(x)$ for all x, that converges uniformly to f. Show also that there exists a sequence $\{j'_n\}$ of step functions on $[a,b]$, for which $j'_n(x)\geq j'_{n+1}(x)\geq f(x)$ for all x, that converges uniformly to f. That is, if $f\in I([a,b])$, then f is the

uniform limit of a nondecreasing sequence of step functions and also is the uniform limit of a nonincreasing sequence of step functions.
HINT: To construct the j_n's and j'_n's, use the step functions k_n and l_n of part (b), and recall that the maximum and minimum of step functions is again a step function.
(e) Show that if $f(x) \geq 0$ for all $x \in [a,b]$, and g is defined by $g(x) = \sqrt{f(x)}$, then $g \in I([a,b])$.
HINT: Write $f = \lim h_n$ where $h_n(x) \geq 0$ for all x and n. Then use part (g) of Exercise 3.28.
(f) (Riemann sums again.) Show that, given an $\epsilon > 0$, there exists a partition P such that if $Q = \{x_0 < x_1 < \ldots < x_n\}$ is any partition finer than P, and $\{w_i\}$ are any points for which $w_i \in (x_{i-1}, x_i)$, then

$$|\int_a^b f(t)\,dt - \sum_{i=1}^n f(w_i)(x_i - x_{i-1})| < \epsilon.$$

HINT: Let P be a partition for which both the step functions k and l of part (c) are constant on the open subintervals of P. Verify that for any finer partition Q, $l(w_i) \geq f(w_i) \geq k(w_i)$, and hence

$$\sum_i l(w_i)(x_i - x_{i-1}) \geq \sum_i f(w_i)(x_i - x_{i-1}) \geq \sum_i k(w_i)(x_i - x_{i-1}).$$

DEFINITION. A bounded real-valued function f on a closed bounded interval $[a,b]$ is called *Riemann-integrable* if, given any $\epsilon > 0$, there exist step functions k and l, on $[a,b]$ for which $k(x) \leq f(x) \leq l(x)$ for all x, such that $\int (l-k) < \epsilon$. We denote the set of all functions on $[a,b]$ that are Riemann-integrable by $I_R([a,b])$.

REMARK. The notion of Riemann-integrability was introduced by Riemann in the mid nineteenth century and was the first formal definition of integrability. Since then several other definitions have been given for an integral, culminating in the theory of Lebesgue integration. The definition of integrability that we are using in this book is slightly different and less general from that of Riemann, and both of these are very different and less general from the definition given by Lebesgue in the early twentieth century. Part (c) of Exercise 5.9 above shows that the functions we are calling integrable are necessarily Riemann-integrable. We will see in Exercise 5.10 that there are Riemann-integrable functions that are **not** integrable in our sense. In both cases, Riemann's and ours, an integrable function f must be trapped between two step functions k and l. In our definition, we must have $l(x) - k(x) < \epsilon$ for all $x \in [a,b]$, while in Riemann's definition, we only need that $\int l - k < \epsilon$. The distinction is that a small step function must have a small integral, but it isn't necessary for a step function to be (uniformly) small in order for it to have a small integral. It only has to be small on most of the interval $[a,b]$.
On the other hand, all the definitions of integrability on $[a,b]$ include among the integrable functions the continuous ones. And, all the different definitions of integral give the same value to a continuous function. The differences then in these definitions shows up at the point of saying exactly which functions are integrable. Perhaps the most enlightening thing to say in this connection is that it is impossible to make a "good" definition of integrability in such a way that every function is

integrable. Subtle points in set theory arise in such attempts, and many fascinating and deep mathematical ideas have come from them. However, we will stick with our definition, since it is simpler than Riemann's and is completely sufficient for our purposes.

THEOREM 5.5. *Let $[a, b]$ be a fixed closed and bounded interval, and let $I([a, b])$ denote the set of integrable functions on $[a, b]$. Then:*

1. *Every element of $I([a, b])$ is a bounded function. That is, integrable functions are necessarily bounded functions.*
2. *$I([a, b])$ is a vector space of functions.*
3. *$I([a, b])$ is closed under multiplication; i.e., if f and $g \in I([a, b])$, then $fg \in I([a, b])$.*
4. *Every step function is in $I([a, b])$.*
5. *If f is a continuous real-valued function on $[a, b]$, then f is in $I([a, b])$. That is, every continuous real-valued function on $[a, b]$ is integrable on $[a, b]$.*

PROOF. Let $f \in I([a, b])$, and write $f = \lim h_n$, where $\{h_n\}$ is a sequence of step functions that converges uniformly to f. Given the positive number $\epsilon = 1$, choose N so that $|f(x) - h_N(x)| < 1$ for all $x \in [a, b]$. Then $|f(x)| \leq |h_N(x)| + 1$ for all $x \in [a, b]$. Because h_N is a step function, its range is a finite set, so that there exists a number M for which $|h_N(x)| \leq M$ for all $x \in [a, b]$. Hence, $|f(x)| \leq M + 1$ for all $x \in [a, b]$, and this proves part (1).

Next, let f and g be integrable, and write $f = \lim h_n$ and $g = \lim k_n$, where $\{h_n\}$ and $\{k_n\}$ are sequences of step functions that converge uniformly to f and g respectively. If s and t are real numbers, then the sequence $\{sh_n + tk_n\}$ converges uniformly to the function $sf + tg$. See parts (c) and (d) of Exercise 3.28. Therefore, $sf + tg \in I([a, b])$, and $I([a, b])$ is a vector space, proving part (2).

Note that part (3) does not follow immediately from Exercise 3.28; the product of uniformly convergent sequences may not be uniformly convergent. To see it for this case, let $f = \lim h_n$ and $g = \lim k_n$ be elements of $I([a, b])$. By part (1), both f and g are bounded, and we write M_f and M_g for numbers that satisfy $|f(x)| \leq M_f$ and $|g(x)| \leq M_g$ for all $x \in [a, b]$. Because the sequence $\{k_n\}$ converges uniformly to g, there exists an N such that if $n \geq N$ we have $|g(x) - k_n(x)| < 1$ for all $x \in [a, b]$. This implies that, if $n \geq N$, then $|k_n(x)| \leq M_g + 1$ for all $x \in [a, b]$.

Now we show that fg is the uniform limit of the sequence $h_n k_n$. For, if $n \geq N$, then

$$\begin{aligned} |f(x)g(x) - h_n(x)k_n(x)| &= |f(x)g(x) - f(x)k_n(x) + f(x)k_n(x) - h_n(x)k_n(x)| \\ &\leq |f(x)||g(x) - k_n(x)| + |k_n(x)||f(x) - h_n(x)| \\ &\leq M_f|g(x) - k_n(x)| + (M_g + 1)|f(x) - h_n(x)|, \end{aligned}$$

which implies that $fg = \lim(h_n k_n)$.

If h is itself a step function, then it is obviously the uniform limit of the constant sequence $\{h\}$, which implies that h is integrable.

Finally, if f is continuous on $[a, b]$, it follows from Theorem 3.20 that f is the uniform limit of a sequence of step functions, whence $f \in I([a, b])$.

Exercise 5.10. Let f be the function defined on $[0, 1]$ by $f(x) = \sin(1/x)$ if $x \neq 0$ and $f(0) = 0$.

(a) Show that f is continuous at every nonzero x and discontinuous at 0.

HINT: Observe that, on any interval $(0, \delta)$, the function $\sin(1/x)$ attains both the values 1 and -1.

(b) Show that f is not integrable on $[0, 1]$.

HINT: Suppose $f = \lim h_n$. Choose N so that $|f(x) - h_N(x)| < 1/2$ for all $x \in [0, 1]$. Let P be a partition for which h_N is constant on its open subintervals, and examine the situation for x's in the interval (x_0, x_1).

(c) Show that f **is** Riemann-integrable on $[0, 1]$. Conclude that $I([a, b])$ is a proper subset of $I_R([a, b])$.

Exercise 5.11. (a) Let f be an integrable function on $[a, b]$. Suppose g is a function for which $g(x) = f(x)$ for all $x \in [a, b]$ except for one point c. Prove that g is integrable and that $\int g = \int f$.

HINT: If $f = \lim h_n$, define $k_n(x) = h_n(x)$ for all $x \neq c$ and $k_n(c) = g(c)$. Then use Exercise 5.4.

(b) Again, let f be an integrable function on $[a, b]$. Suppose g is a function for which $g(x) = f(x)$ for all but a finite number of points $c_1, \ldots, c_N \in [a, b]$. Prove that $g \in I([a, b])$, and that $\int g = \int f$.

(c) Suppose f is a function on the closed interval $[a, b]$, that is uniformly continuous on the open interval (a, b). Prove that f is integrable on $[a, b]$.

HINT: Just reproduce the proof to Theorem 3.20.

REMARK. In view of part (b) of the preceding exercise, we see that whether a function f is integrable or not is totally independent of the values of the function at a fixed finite set of points. Indeed, the function needn't even be defined at a fixed finite set of points, and still it can be integrable. This observation is helpful in many instances, e.g., in parts (d) and (e) of Exercise 5.21.

THEOREM 5.6. *The assignment $f \rightarrow \int f$ on $I([a, b])$ satisfies the following properties.*

1. (Linearity) $I([a, b])$ is a vector space, and $\int(\alpha f + \beta g) = \alpha \int f + \beta \int g$ for all $f, g \in I([a, b])$and $\alpha, \beta \in \mathbb{R}$.
2. (Positivity) If $f(x) \geq 0$ for all $x \in [a, b]$, then $\int f \geq 0$.
3. (Order-preserving) If $f, g \in I([a, b])$ and $f(x) \leq g(x)$ for all $x \in [a, b]$, then $\int f \leq \int g$.
4. If $f \in I([a, b])$, then so is $|f|$, and $|\int f| \leq \int |f|$.
5. If f is the uniform limit of functions f_n, each of which is in $I([a, b])$, then $f \in I([a, b])$ and $\int f = \lim \int f_n$.
6. Let $\{u_n\}$ be a sequence of functions in $I([a, b])$. Suppose that for each n there is a number m_n, for which $|u_n(x)| \leq m_n$ for all $x \in [a, b]$, and such that the infinite series $\sum m_n$ converges. Then the infinite series $\sum u_n$ converges uniformly to an integrable function, and $\int \sum u_n = \sum \int u_n$.

PROOF. That $I([a, b])$ is a vector space was proved in part (2) of Theorem 5.5. Let f and g be in $I([a, b])$, and write $f = \lim h_n$ and $g = \lim k_n$, where the h_n's and the k_n's are step functions. Then $\alpha f + \beta g = \lim(\alpha h_n + \beta k_n)$, so that, by Theorem

5.2 and the definition of the integral, we have

$$\begin{aligned}\int(\alpha f+\beta g) &= \lim\int(\alpha h_n+\beta k_n)\\ &= \lim(\alpha\int h_n+\beta\int k_n)\\ &= \alpha\lim\int h_n+\beta\lim\int k_n\\ &= \alpha\int f+\beta\int g,\end{aligned}$$

which proves part (1).

Next, if $f\in I([a,b])$ satisfies $f(x)\geq 0$ for all $x\in[a,b]$, let $\{l_n\}$ be a nonincreasing sequence of step functions that converges uniformly to f. See part (d) of Exercise 5.9. Then $l_n(x)\geq f(x)\geq 0$ for all x and all n. So, again by Theorem 5.2, we have that

$$\int f=\lim\int l_n\geq 0.$$

This proves part (2).

Part (3) now follows by combining parts (1) and (2) just as in the proof of Theorem 5.2.

To see part (4), let $f\in I([a,b])$ be given. Write $f=\lim h_n$. Then $|f|=\lim|h_n|$. For

$$||f(x)|-|h_n(x)||\leq|f(x)-h_n(x)|.$$

Therefore, $|f|$ is integrable. Also,

$$\int|f|=\lim\int|h_n|\geq\lim|\int h_n|=|\lim\int h_n|=|\int f|.$$

To see part (5), let $\{f_n\}$ be a sequence of elements of $I([a,b])$, and suppose that $f=\lim f_n$. For each n, let h_n be a step function on $[a,b]$ such that $|f_n(x)-h_n(x)|<1/n$ for all $x\in[a,b]$. Note also that it follows from parts (3) and (4) that

$$|\int f_n-\int h_n|<\frac{b-a}{n}.$$

Now $\{h_n\}$ converges uniformly to f. For,

$$\begin{aligned}|f(x)-h_n(x)| &\leq|f(x)-f_n(x)|+|f_n(x)-h_n(x)|\\ &<|f(x)-f_n(x)|+\frac{1}{n},\end{aligned}$$

showing that $f=\lim h_n$. Therefore, $f\in I([a,b])$. Moreover, $\int f=\lim\int h_n$. Finally, $\int f=\lim\int f_n$, for

$$\begin{aligned}|\int f-\int f_n| &\leq|\int f-\int h_n|+|\int h_n-\int f_n|\\ &\leq|\int f-\int h_n|+\frac{b-a}{n}.\end{aligned}$$

This completes the proof of part (5).

Part (6) follows directly from part (5) and the Weierstrass M Test (Theorem 3.18). For, part (1) of that theorem implies that the infinite series $\sum u_n$ converges uniformly, and then $\int \sum u_n = \sum \int u_n$ follows from part (5) of this theorem.

As a final extension of our notion of integral, we define the integral of certain complex-valued functions.

DEFINITION. Let $[a,b]$ be a fixed bounded and closed interval. A complex-valued function $f = u + iv$ is called *integrable* if its real and imaginary parts u and v are integrable. In this case, we define

$$\int_a^b f = \int_a^b (u + iv) = \int_a^b u + i \int_a^b v.$$

THEOREM 5.7.

(1) *The set of all integrable complex-valued functions on $[a,b]$ is a vector space over the field of complex numbers, and*

$$\int_a^b (\alpha f + \beta g) = \alpha \int_a^b f + \beta \int_a^b g$$

for all integrable complex-valued functions f and g and all complex numbers α and β.

(2) *If f is an integrable complex-valued function on $[a,b]$, then so is $|f|$, and $|\int_a^b f| \leq \int_a^b |f|$.*

PROOF. We leave the verification of part (1) to the exercise that follows.

To see part (2), suppose that f is integrable, and write $f = u + iv$. Then $|f| = \sqrt{u^2 + v^2}$, so that $|f|$ is integrable by Theorem 5.5 and part (e) of Exercise 5.9. Now write $z = \int_a^b f$, and write z in polar coordinates as $z = re^{i\theta}$, where $r = |z| = |\int_a^b f|$. (See Exercise 4.23.) Define a function g by $g(x) = e^{-i\theta} f(x)$ and notice that $|g| = |f|$. Then $\int_a^b g = e^{-i\theta} \int_a^b f = r$, which is a real number. Writing $g = \widehat{u} + i\widehat{v}$, we then have that $r = \int \widehat{u} + i \int \widehat{v}$, implying that $\int \widehat{v} = 0$. So,

$$\begin{aligned} |\int_a^b f| &= r \\ &= \int_a^b g \\ &= \int_a^b \widehat{u} + i \int_a^b \widehat{v} \\ &= \int_a^b \widehat{u} \\ &= |\int_a^b \widehat{u}| \\ &\leq \int_a^b |\widehat{u}| \\ &\leq \int_a^b |g| \\ &= \int_a^b |f|, \end{aligned}$$

as desired.

Exercise 5.12. Prove part (1) of the preceding theorem.
HINT: Break $\alpha, \beta, \int f$, and $\int g$ into real and imaginary parts.

THE FUNDAMENTAL THEOREM OF CALCULUS

We begin this section with a result that is certainly not a surprise, but we will need it at various places in later proofs, so it's good to state it precisely now.

THEOREM 5.8. *Suppose $f \in I([a,b])$, and suppose $a < c < b$. Then $f \in I([a,c])$, $f \in I([c,b])$, and*

$$\int_a^b f = \int_a^c f + \int_c^b f.$$

PROOF. Suppose first that h is a step function on $[a,b]$, and let $P = \{x_0 < x_1 < \ldots < x_n\}$ be a partition of $[a,b]$ such that $h(x) = a_i$ on the subinterval (x_{i-1}, x_i) of P. Of course, we may assume without loss of generality that c is one of the points of P, say $c = x_k$. Clearly h is a step function on both intervals $[a,c]$ and $[c,b]$. Now, let $Q_1 = \{a = x_0 < x_1 < \ldots < c = x_k\}$ be the partition of $[a,c]$ obtained by intersecting P with $[a,c]$, and let $Q_2 = \{c = x_k < x_{k+1} < \ldots < x_n = b\}$ be the partition of $[c, b]$ obtained by intersecting P with $[c, b]$. We have that

$$\begin{aligned}\int_a^b h &= S_P(h)\\ &= \sum_{i=1}^n a_i(x_i - x_{i-1})\\ &= \sum_{i=1}^k a_i(x_i - x_{i-1}) + \sum_{i=k+1}^n a_i(x_i - x_{i-1})\\ &= S_{Q_1}(h) + S_{Q_2}(h)\\ &= \int_a^c h + \int_c^b h,\end{aligned}$$

which proves the theorem for step functions.
Now, write $f = \lim h_n$, where each h_n is a step function on $[a, b]$. Then clearly $f = \lim h_n$ on $[a, c]$, which shows that $f \in I([a, c])$, and

$$\int_a^c f = \lim \int_a^c h_n.$$

Similarly, $f = \lim h_n$ on $[c, b]$, showing that $f \in I([c, b])$, and

$$\int_c^b f = \lim \int_c^b h_n.$$

Finally,

$$\begin{aligned}\int_a^b f &= \lim \int_a^b h_n\\ &= \lim(\int_a^c h_n + \int_c^b h_n)\\ &= \lim \int_a^c h_n + \lim \int_c^b h_n\\ &= \int_a^c f + \int_c^b f,\end{aligned}$$

as desired.

I's time for the trumpets again! What we call the Fundamental Theorem of Calculus was discovered by Newton and Leibniz more or less simultaneously in the seventeenth century, and it is without doubt the cornerstone of all we call mathematical analysis today. Perhaps the main theoretical consequence of this theorem is that it provides a procedure for inventing "new" functions. Polynomials are rather natural functions, power series are a simple generalization of polynomials, and then what? It all came down to thinking of a function of a variable x as being the area beneath a curve between a fixed point a and the varying point x. By now, we have polished and massaged these ideas into a careful, detailed development of the subject, which has substantially obscured the original ingenious insights of Newton and Leibniz. On the other hand, our development and proofs are complete, while theirs were based heavily on their intuition. So, here it is.

THEOREM 5.9. (Fundamental Theorem of Calculus) Suppose f is an arbitrary element of $I([a,b])$. Define a function F on $[a,b]$ by $F(x)=\int_a^x f$. Then:

(1) F is continuous on $[a,b]$, and $F(a)=0$.
(2) If f is continuous at a point $c\in(a,b)$, then F is differentiable at c and $F'(c)=f(c)$.
(3) Suppose that f is continuous on $[a,b]$. If G is any continuous function on $[a,b]$ that is differentiable on (a,b) and satisfies $G'(x)=f(x)$ for all $x\in(a,b)$, then

$$\int_a^b f(t)\,dt = G(b)-G(a).$$

REMARK. Part (2) of this theorem is the heart of it, the great discovery of Newton and Leibniz, although most beginning calculus students often think of part (3) as the main statement. Of course it is that third part that enables us to actually compute integrals.

PROOF. Because $f\in I([a,b])$, we know that $f\in I([a,x])$ for every $x\in[a,b]$, so that $F(x)$ at least is defined.

Also, we know that f is bounded; i.e., there exists an M such that $|f(t)|\le M$ for all $t\in[a,b]$. Then, if $x,y\in[a,b]$ with $x\ge y$, we have that

$$\begin{aligned}|F(x)-F(y)| &= |\int_a^x f - \int_a^y f| \\ &= |\int_a^y f + \int_y^x f - \int_a^y f| \\ &= |\int_y^x f| \\ &\le \int_y^x |f| \\ &\le \int_y^x M \\ &= M(x-y),\end{aligned}$$

so that $|F(x)-F(y)|\le M|x-y|<\epsilon$ if $|x-y|<\delta=\epsilon/M$. This shows that F is (uniformly) continuous on $[a,b]$. Obviously, $F(a)=\int_a^a f=0$, and part (1) is proved.

Next, suppose that f is continuous at $c\in(a,b)$, and write $L=f(c)$. Let $\epsilon>0$ be given. To show that F is differentiable at c and that $F'(c)=f(c)$, we must find a $\delta>0$ such that if $0<|h|<\delta$ then

$$|\frac{F(c+h)-F(c)}{h}-L|<\epsilon.$$

Since f is continuous at c, choose $\delta>0$ so that $|f(t)-f(c)|<\epsilon$ if $|t-c|<\delta$. Now, assuming that $h > 0$ for the moment, we have that

$$\begin{aligned} F(c+h) - F(c) &= \int_a^{c+h} f - \int_a^c f \\ &= \int_a^c f + \int_c^{c+h} f - \int_a^c f \\ &= \int_c^{c+h} f, \end{aligned}$$

and

$$L = \frac{\int_c^{c+h} L}{h}.$$

So, if $0 < h < \delta$, then

$$\begin{aligned} |\frac{F(c+h) - F(c)}{h} - L| &= |\frac{\int_c^{c+h} f(t)\,dt}{h} - \frac{\int_c^{c+h} L}{h}| \\ &= |\frac{\int_c^{c+h} (f(t) - L)\,dt}{h}| \\ &\leq \frac{\int_c^{c+h} |f(t) - L|\,dt}{h} \\ &= \frac{\int_c^{c+h} |f(t) - f(c)|\,dt}{h} \\ &\leq \frac{\int_c^{c+h} \epsilon}{h} \\ &= \epsilon, \end{aligned}$$

where the last inequality follows because for $t \in [c, c+h]$, we have that $|t-c| \leq h < \delta$. A similar argument holds if $h < 0$. (See the following exercise.) This proves part (2).

Suppose finally that G is continuous on $[a, b]$, differentiable on (a, b), and that $G'(x) = f(x)$ for all $x \in (a, b)$. Then, $F - G$ is continuous on $[a, b]$, differentiable on (a, b), and by part (2) $(F - G)'(x) = F'(x) - G'(x) = f(x) - f(x) = 0$ for all $x \in (a, b)$. It then follows from Exercise 4.12 that $F - G$ is a constant function C, whence,

$$G(b) - G(a) = F(b) + C - F(a) - C = F(b) = \int_a^b f(t)\,dt,$$

and the theorem is proved.

Exercise 5.13. (a) Complete the proof of part (2) of the preceding theorem; i.e., take care of the case when $h < 0$.

HINT: In this case, $a < c + h < c$. Then, write $\int_a^c f = \int_a^{c+h} f + \int_{c+h}^c f$.

(b) Suppose f is a continuous function on the closed interval $[a, b]$, and that f' exists and is continuous on the open interval (a, b). Assume further that f' is integrable on the closed interval $[a, b]$. Prove that $f(x) - f(a) = \int_a^x f'$ for all $x \in [a, b]$. Be careful to understand how this is different from the Fundamental Theorem.

(c) Use the Fundamental Theorem to prove that for $x \geq 1$ we have

$$\ln(x) = F(x) \equiv \int_1^x \frac{1}{t}\, dt,$$

and for $0 < x < 1$ we have

$$\ln(x) = F(x) \equiv -\int_x^1 \frac{1}{t}\, dt.$$

HINT: Show that these two functions have the same derivative and agree at $x = 1$.

CONSEQUENCES OF THE FUNDAMENTAL THEOREM

The first two theorems of this section constitute the basic "techniques of integration" taught in a calculus course. However, the careful formulations of these standard methods of evaluating integrals have some subtle points, i.e., some hypotheses. Calculus students are rarely told about these details.

THEOREM 5.10. (Integration by Parts Formula) Let f and g be integrable functions on $[a, b]$, and as usual let F and G denote the functions defined by

$$F(x) = \int_a^x f, \text{ and } G(x) = \int_a^x g.$$

Then

$$\int_a^b fG = [F(b)G(b) - F(a)G(a)] - \int_a^b Fg.$$

Or, recalling that $f = F'$ and $g = G'$,

$$\int_a^b F'G = [F(b)G(b) - F(a)G(a)] - \int_a^b FG'.$$

Exercise 5.14. (a) Prove the preceding theorem.
HINT: Replace the upper limit b by a variable x, and differentiate both sides. By the way, how do we know that the functions Fg and fG are integrable?
(b) Suppose f and g are integrable functions on $[a, b]$ and that both f' and g' are continuous on (a, b) and integrable on $[a, b]$. (Of course f' and g' are not even defined at the endpoints a and b, but they can still be integrable on $[a, b]$. See the remark following Exercise 5.11.) Prove that

$$\int_a^b fg' = [f(b)g(b) - f(a)g(a)] - \int_a^b f'g.$$

THEOREM 5.11. (Integration by Substitution) Let f be a continuous function on $[a, b]$, and suppose g is a continuous, one-to-one function from $[c, d]$ onto $[a, b]$ such that g is continuously differentiable on (c, d), and such that $a = g(c)$ and $b = g(d)$. Assume finally that g' is integrable on $[c, d]$. Then

$$\int_a^b f(t)\, dt = \int_c^d f(g(s))g'(s)\, ds.$$

PROOF. It follows from our assumptions that the function $f(g(s))g'(s)$ is continuous on (a,b) and integrable on $[c,d]$. It also follows from our assumptions that g maps the open interval (c,d) onto the open interval (a,b). As usual, let F denote the function on $[a,b]$ defined by $F(x)=\int_a^x f(t)\,dt$. Then, by part (2) of the Fundamental Theorem, F is differentiable on (a,b), and $F'=f$. Then, by the chain rule, $F\circ g$ is continuous and differentiable on (c,d) and

$$(F\circ g)'(s)=F'(g(s))g'(s)=f(g(s))g'(s).$$

So, by part (3) of the Fundamental Theorem, we have that

$$\begin{aligned}\int_c^d f(g(s))g'(s)\,ds &= \int_c^d (F\circ g)'(s)\,ds\\ &=(F\circ g)(d)-(F\circ g)(c)\\ &=F(g(d))-F(g(c))\\ &=F(b)-F(a)\\ &=\int_a^b f(t)\,dt,\end{aligned}$$

which finishes the proof.

Exercise 5.15. (a) Prove the "Mean Value Theorem" for integrals: If f is continuous on $[a,b]$, then there exists a $c\in(a,b)$ such that

$$\int_a^b f(t)\,dt=f(c)(b-a).$$

(b) (Uniform limits of differentiable functions. Compare with Exercise 4.26.) Suppose $\{f_n\}$ is a sequence of continuous functions on a closed interval $[a,b]$ that converges pointwise to a function f. Suppose that each derivative f_n' is continuous on the open interval (a,b), is integrable on the closed interval $[a,b]$, and that the sequence $\{f_n'\}$ converges uniformly to a function g on (a,b). Prove that f is differentiable on (a,b), and $f'=g$.
HINT: Let x be in (a,b), and let c be in the interval (a,x). Justify the following equalities, and use them together with the Fundamental Theorem to make the proof.

$$f(x)-f(c)=\lim(f_n(x)-f_n(c))=\lim\int_c^x f_n'=\int_c^x g.$$

We revisit now the Remainder Theorem of Taylor, which we first presented in Theorem 4.19. The point is that there is another form of this theorem, the integral form, and this version is more powerful in some instances than the original one, e.g., in the general Binomial Theorem below.

THEOREM 5.12. (Integral Form of Taylor's Remainder Theorem) Let c be a real number, and let f have $n+1$ derivatives on $(c-r,c+r)$, and suppose that $f^{(n+1)}\in I([c-r,c+r])$. Then for each $c<x<c+r$,

$$f(x)-T^n_{(f,c)}(x)=\int_c^x f^{(n+1)}(t)\frac{(x-t)^n}{n!}\,dt,$$

where T_f^n denotes the nth Taylor polynomial for f.
Similarly, for $c - r < x < c$,

$$f(x) - T^n_{(f,c)}(x) = \int_x^c f^{(n+1)}(t)\frac{(x-t)^n}{n!}\,dt.$$

Exercise 5.16. Prove the preceding theorem.
HINT: Argue by induction on n, and integrate by parts.

REMARK. We return now to the general Binomial Theorem, first studied in Theorem 4.21. The proof given there used the derivative form of Taylor's remainder Theorem, but we were only able to prove the Binomial Theorem for $|t| < 1/2$. The theorem below uses the integral form of Taylor's Remainder Theorem in its proof, and it gives the full binomial theorem, i.e., for all t for which $|t| < 1$.

THEOREM 5.13. (General Binomial Theorem) Let $\alpha = a+bi$ be a fixed complex number. Then

$$(1+t)^\alpha = \sum_{k=0}^\infty \binom{\alpha}{k} t^k$$

for all $t \in (-1, 1)$.

PROOF. For clarity, we repeat some of the proof of Theorem 4.21. Given a general $\alpha = a+bi$, consider the function $g : (-1, 1) \to \mathbb{C}$ defined by $g(t) = (1+t)^\alpha$. Observe that the nth derivative of g is given by

$$g^{(n)}(t) = \frac{\alpha(\alpha-1)\dots(\alpha-n+1)}{(1+t)^{n-\alpha}}.$$

Then $g \in C^\infty((-1,1))$.
For each nonnegative integer k define

$$a_k = g^{(k)}(0)/k! = \frac{\alpha(\alpha-1)\dots(\alpha-k+1)}{k!} = \binom{\alpha}{k},$$

and set $h(t) = \sum_{k=0}^\infty a_k t^k$. The radius of convergence for the power series function h is 1, as was shown in Exercise 4.31. We wish to show that $g(t) = h(t)$ for all $-1 < t < 1$. That is, we wish to show that g is a Taylor series function around 0. It will suffice to show that the sequence $\{S_n\}$ of partial sums of the power series function h converges to the function g. We note also that the nth partial sum is just the nth Taylor polynomial T_g^n for g.
Now, fix a t strictly between 0 and 1. The argument for t's between -1 and 0 is completely analogous.. Choose an $\epsilon > 0$ for which $\beta = (1+\epsilon)t < 1$. We let C_ϵ be a numbers such that $|\binom{\alpha}{n}| \le C_\epsilon(1+\epsilon)^n$ for all nonnegative integers n. See Exercise 4.31. We will also need the following estimate, which can be easily deduced as a calculus exercise (See part (d) of Exercise 4.11.). For all s between 0 and t, we have $(t-s)/(1+s) \le t$. Note also that, for any $s \in (0,t)$, we have $|(1+s)^\alpha| = (1+s)^a$, and this is trapped between 1 and $(1+t)^a$. Hence, there exists a number M_t such that $|(1+s)^{\alpha-1}| \le M_t$ for all $s \in (-0, t)$. We will need this estimate in the calculation that follows.

Then, by the integral form of Taylor's Remainder Theorem, we have:

$$\begin{aligned}
|g(t)-\sum_{k=0}^{n}a_kt^k| &= |g(t)-T_g^n(t)|\\
&= |\int_0^t g^{(n+1)}(s)\frac{(t-s)^n}{n!}\,ds|\\
&= |\int_0^t \binom{(n+1)\times\alpha}{n+1}(1+s)^{\alpha-n-1}(t-s)^n\,ds|\\
&\leq \int_0^t |\binom{\alpha}{n+1}||(1+s)^{\alpha-1}|(n+1)|(\frac{t-s}{1+s}|^n\,ds\\
&\leq \int_0^t |\binom{\alpha}{n+1}|M_t(n+1)t^n\,ds\\
&\leq C_\epsilon M_t(n+1)\int_0^t(1+\epsilon)^{n+1}t^n\,ds\\
&= C_\epsilon M_t(n+1)(1+\epsilon)^{n+1}t^{n+1}\\
&= C_\epsilon M_t(n+1)\beta^{n+1},
\end{aligned}$$

which tends to 0 as n goes to ∞, because $\beta < 1$. This completes the proof for $0 < t < 1$.

AREA OF REGIONS IN THE PLANE

It would be desirable to be able to assign to each subset S of the Cartesian plane $\mathbb{R}^2$ a nonnegative real number $A(S)$ called its area. We would insist based on our intuition that (i) if S is a rectangle with sides of length L and W then the number $A(S)$ should be LW, so that this abstract notion of area would generalize our intuitively fundamental one. We would also insist that (ii) if S were the union of two disjoint parts, $S = S_1 \cup S_2$, then $A(S)$ should be $A(S_1) + A(S_2)$. (We were taught in high school plane geometry that the whole is the sum of its parts.) In fact, even if S were the union of an infinite number of disjoint parts, $S = \cup_{n=1}^{\infty} S_n$ with $S_i \cap S_j = \emptyset$ if $i \neq j$, we would insist that (iii) $A(S) = \sum_{n=1}^{\infty} A(S_n)$.

The search for such a definition of area for every subset of $\mathbb{R}^2$ motivated much of modern mathematics. Whether or not such an assignment exists is intimately related to subtle questions in basic set theory, e.g., the *Axiom of Choice* and the *Continuum Hypothesis*. Most mathematical analysts assume that the Axiom of Choice holds, and as a result of that assumption, it has been shown that there can be no assignment $S \to A(S)$ satisfying the above three requirements. Conversely, if one does not assume that the Axiom of Choice holds, then it has also been shown that it is perfectly consistent to assume as a basic axiom that such an assignment $S \to A(S)$ does exist. We will not pursue these subtle points here, leaving them to a course in Set Theory or Measure Theory. However, Here's a statement of the Axiom of Choice, and we invite the reader to think about how reasonable it seems.

AXIOM OF CHOICE. *Let $\mathcal{S}$ be a collection of sets. Then there exists a set A that contains exactly one element out of each of the sets S in $\mathcal{S}$.*

The difficulty mathematicians encountered in trying to define area turned out to be involved with defining $A(S)$ for **every** subset $S \in \mathbb{R}^2$. To avoid this difficulty, we will restrict our attention here to certain " reasonable" subsets S. Of course, we certainly want these sets to include the rectangles and all other common geometric sets.

DEFINITION. By a (open) *rectangle* we will mean a set $R = (a,b) \times (c,d)$ in $\mathbb{R}^2$. That is, $R = \{(x,y) : a < x < b \text{ and } c < y < d\}$. The analogous definition of a *closed rectangle* $[a,b] \times [c,d]$ should be clear: $[a,b] \times [c,d] = \{(x,y) : a \leq x \leq b,\ c \leq y \leq d\}$.
By the *area* of a (open or closed) rectangle $R = (a,b) \times (c,d)$ or $[a,b] \times [c,d]$ we mean the number $A(R) = (b-a)(d-c)$.

. The fundamental notion behind our definition of the area of a set S is this. If an open rectangle $R = (a,b) \times (c,d)$ is a subset of S, then the area $A(S)$ surely should be greater than or equal to $A(R) = (b-a)(d-c)$. And, if S contains the disjoint union of several open rectangles, then the area of S should be greater than or equal to the sum of their areas.
We now specify precisely for which sets we will define the area. Let $[a,b]$ be a fixed closed bounded interval in $\mathbb{R}$ and let l and u be two continuous real-valued functions on $[a,b]$ for which $l(x) < u(x)$ for all $x \in (a,b)$.

DEFINITION. Given $[a,b], l$, and u as in the above, let S be the set of all pairs $(x,y) \in \mathbb{R}^2$, for which $a < x < b$ and $l(x) < y < u(x)$. Then S is called an *open geometric set.* If we replace the $<$ signs with $\leq$ signs, i.e., if S is the set of all (x,y) such that $a \leq x \leq b$, and $l(x) \leq y \leq u(x)$, then S is called a *closed geometric set.* In either case, we say that S is bounded on the left and right by the vertical line segments $\{(a,y) : l(a) \leq y \leq u(a)\}$ and $\{(b,y) : l(b) \leq y \leq u(b)\}$, and it is bounded below by the graph of the function l and bounded above by the graph of the function u. We call the union of these four bounding curves the *boundary* of S, and denote it by C_S.
If the bounding functions u and l of a geometric set S are smooth or piecewise smooth functions, we will call S a *smooth* or *piecewise smooth* geometric set.
If S is a closed geometric set, we will indicate the corresponding open geometric set by the symbol S^0.

The symbol S^0 we have introduced for the open geometric set corresponding to a closed one is the same symbol that we have used previously for the interior of a set. Study the exercise that follows to see that the two uses of this notation agree.

Exercise 5.17. (a) Show that rectangles, triangles, and circles are geometric sets. What in fact is the definition of a circle?
(b) Find some examples of sets that are **not** geometric sets. Think about a horseshoe on its side, or a heart on its side.
(c) Let f be a continuous, nonnegative function on $[a,b]$. Show that the "region" under the graph of f is a geometric set.
(d) Show that the intersection of two geometric sets is a geometric set. Describe the left, right, upper, and lower boundaries of the intersection. Prove that the interior $(S_1 \cap S_2)^0$ of the intersection of two geometric sets S_1 and S_2 coincides with the intersection $S_1^0 \cap S_2^0$ of their two interiors.
(e) Give an example to show that the union of two geometric sets need not be a geometric set.
(f) Show that every closed geometric set is compact.

(g) Let S be a closed geometric set. Show that the corresponding open geometric set S^0 coincides with the interior of S, i.e., the set of all points in the interior of S. HINT: Suppose $a < x < b$ and $l(x) < y < u(x)$. Begin by showing that, because both l and u are continuous, there must exist an $\epsilon > 0$ and a $\delta > 0$ such that $a < x - \delta < x + \delta < b$ and $l(x) < y - \epsilon < y + \epsilon < u(x)$.

Now, given a geometric set S (either open or closed), that is determined by an interval $[a, b]$ and two bounding functions u and l, let $P = \{x_0 < x_1 < \ldots < x_n\}$ be a partition of $[a, b]$. For each $1 \leq i \leq n$, define numbers c_i and d_i as follows:

$$c_i = \sup_{x_{i-1}<x<x_i} l(x), \text{ and } d_i = \inf_{x_{i-1}<x<x_i} u(x).$$

Because the functions l and u are continuous, they are necessarily bounded, so that the supremum and infimum above are real numbers. For each $1 \leq i \leq n$ define R_i to be the open rectangle $(x_{i-1}, x_i) \times (c_i, d_i)$. Of course, d_i may be $< c_i$, in which case the rectangle R_i is the empty set. In any event, we see that the partition P determines a finite set of (possibly empty) rectangles $\{R_i\}$, and we denote the union of these rectangles by the symbol $\mathcal{C}_P. = \cup_{i=1}^n (x_{i-1}, x_i) \times (c_i, d_i)$.
The area of the rectangle R_i is $(x_i - x_{i-1})(d_i - c_i)$ if $c_i < d_i$ and 0 otherwise. We may write in general that $A(R_i) = (x_i - x_{i-1}) \max((d_i - c_i), 0)$. Define the number A_P by

$$A_P = \sum_{i=1}^n (x_i - x_{i-1})(d_i - c_i).$$

Note that A_P is not exactly the sum of the areas of the rectangles determined by P because it may happen that $d_i < c_i$ for some i's, so that those terms in the sum would be negative. In any case, it is clear that A_P is less than or equal to the sum of the areas of the rectangles, and this notation simplifies matters later.
For any partition P, we have $S \supseteq \mathcal{C}_P$, so that, if $A(S)$ is to denote the area of S, we want to have

$$\begin{aligned} A(S) &\geq \sum_{i=1}^n A(R_i) \\ &= \sum_{i=1}^n (x_i - x_{i-1}) \max((d_i - c_i), 0) \\ &\geq \sum_{i=1}^n (x_i - x_{i-1})(d_i - c_i) \\ &= A_P. \end{aligned}$$

DEFINITION. Let S be a geometric set (either open or closed), bounded on the left by $x = a$, on the right by $x = b$, below by the graph of l, and above by the graph of u. Define the *area* $A(S)$ of S by

$$A(S) = \sup_P A_P = \sup_{P=\{x_0<x_1<\ldots<x_n\}} \sum_{i=1}^n (x_i - x_{i-1})(d_i - c_i),$$

where the supremum is taken over all partitions P of $[a, b]$, and where the numbers c_i and d_i are as defined above.

Exercise 5.18. (a) Using the notation of the preceding paragraphs, show that each rectangle R_i is a subset of the set S and that $R_i \cap R_j = \emptyset$ if $i \neq j$. It may help to draw a picture of the set S and the rectangles $\{R_i\}$. Can you draw one so that $d_i < c_i$?
(b) Suppose S_1 is a geometric set and that S_2 is another geometric set that is contained in S_1. Prove that $A(S_2) \leq A(S_1)$.
HINT: For each partition P, compare the two A_P's.

Exercise 5.19. Let T be the triangle in the plane with vertices at the three points $(0,0), (0,H)$, and $(B,0)$. Show that the area $A(T)$, as defined above, agrees with the formula $A = (1/2)BH$, where B is the base and H is the height.

The next theorem gives the connection between area (geometry) and integration (analysis). In fact, this theorem is what most calculus students think integration is all about.

THEOREM 5.14. *Let S be a geometric set, i.e., a subset of $\mathbb{R}^2$ that is determined in the above manner by a closed bounded interval $[a,b]$ and two bounding functions l and u. Then*

$$A(S) = \int_a^b (u(x) - l(x))\,dx.$$

PROOF. Let $P = \{x_0 < x_1 < \ldots < x_n\}$ be a partition of $[a,b]$, and let c_i and d_i be defined as above. Let h be a step function that equals d_i on the open interval (x_{i-1}, x_i), and let k be a step function that equals c_i on the open interval (x_{i-1}, x_i). Then on each open interval (x_{i-1}, x_i) we have $h(x) \leq u(x)$ and $k(x) \geq l(x)$. Complete the definitions of h and k by defining them at the partition points so that $h(x_i) = k(x_i)$ for all i. Then we have that $h(x) - k(x) \leq u(x) - l(x)$ for all $x \in [a,b]$. Hence,

$$A_P = \sum_{i=1}^n (x_i - x_{i-1})(d_i - c_i) = \int_a^b (h - k) \leq \int_a^b (u - l).$$

Since this is true for every partition P of $[a,b]$, it follows by taking the supremum over all partitions P that

$$A(S) = \sup_P A_P \leq \int_a^b (u(x) - l(x))\,dx,$$

which proves half of the theorem; i.e., that $A(S) \leq \int_a^b u - l$.
To see the other inequality, let h be any step function on $[a,b]$ for which $h(x) \leq u(x)$ for all x, and let k be any step function for which $k(x) \geq l(x)$ for all x. Let $P = \{x_0 < x_1 < \ldots < x_n\}$ be a partition of $[a,b]$ for which both h and k are constant on the open subintervals (x_{i-1}, x_i) of P. Let $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ be the numbers such that $h(x) = a_i$ on (x_{i-1}, x_i) and $k(x) = b_i$ on (x_{i-1}, x_i). It follows, since $h(x) \leq u(x)$ for all x, that $a_i \leq d_i$. Also, it follows that $b_i \geq c_i$. Therefore,

$$\int_a^b (h - k) = \sum_{i=1}^n (a_i - b_i)(x_i - x_{i-1}) \leq \sum_{i=1}^n (x_i - x_{i-1})(d_i - c_i) = A_P \leq A(S).$$

Finally, let $\{h_m\}$ be a nondecreasing sequence of step functions that converges uniformly to u, and let $\{k_m\}$ be a nonincreasing sequence of step functions that converges uniformly to l. See part (d) of Exercise 5.9. Then

$$\int_a^b (u-l) = \lim_m \int_a^b (h_m - k_m) \leq A(S),$$

which proves the other half of the theorem.

OK! Trumpet fanfares, please!

THEOREM 5.15. ($A = \pi r^2$.) If S is a circle in the plane having radius r, then the area $A(S)$ of S is πr^2.

PROOF. Suppose the center of the circle S is the point (h,k). This circle is a geometric set. In fact, we may describe the circle with center (h,k) and radius r as the subset S of $\mathbb{R}^2$ determined by the closed bounded interval $[h-r, h+r]$ and the functions

$$u(x) = k + \sqrt{r^2 - (x-h)^2}$$

and

$$l(x) = k - \sqrt{r^2 - (x-h)^2}.$$

By the preceding theorem, we then have that

$$A(S) = \int_{h-r}^{h+r} 2\sqrt{r^2 - (x-h)^2}\, dx = \pi r^2.$$

We leave the verification of the last equality to the following exercise.

Exercise 5.20. Evaluate the integral in the above proof:

$$\int_{h-r}^{h+r} 2\sqrt{r^2 - (x-h)^2}\, dx.$$

Be careful to explain each step by referring to theorems and exercises in this book. It may seem like an elementary calculus exercise, but we are justifying each step here.

REMARK. There is another formula for the area of a geometric set that is sometimes very useful. This formula gives the area in terms of a "double integral." There is really nothing new to this formula; it simply makes use of the fact that the number (length) $u(x) - l(x)$ can be represented as the integral from $l(x)$ to $u(x)$ of the constant 1. Here's the formula:

$$A(S) = \int_a^b (\int_{l(x)}^{u(x)} 1\, dy)\, dx.$$

The next theorem is a result that justifies our definition of area by verifying that the whole is equal to the sum of its parts, something that any good definition of area should satisfy.

THEOREM 5.16. *Let S be a closed geometric set, and suppose $S = \cup_{i=1}^n S_i,$ where the sets $\{S_i\}$ are closed geometric sets for which $S_i^0 \cap S_j^0 = \emptyset$ if $i \neq j.$ Then*

$$A(S) = \sum_{i=1}^n A(S_i).$$

PROOF. Suppose S is determined by the interval $[a, b]$ and the two bounding functions l and u, and suppose S_i is determined by the interval $[a_i, b_i]$ and the two bounding functions l_i and u_i. Because $S_i \subseteq S$, it must be that the interval $[a_i, b_i]$ is contained in the interval $[a, b]$. Initially, the bounding functions l_i and u_i are defined and continuous on $[a_i, b_i]$, and we extend their domain to all of $[a, b]$ by defining $l_i(x) = u_i(x) = 0$ for all $x \in [a, b]$ that are not in $[a_i, b_i]$. The extended functions l_i and u_i may not be continuous on all of $[a, b]$, but they are still integrable on $[a, b]$. (Why?) Notice that we now have the formula

$$A(S_i) = \int_{a_i}^{b_i} (u_i(x) - l_i(x))\, dx = \int_a^b (u_i(x) - l_i(x))\, dx.$$

Next, fix an x in the open interval (a, b). We must have that the vertical intervals $(l_i(x), u_i(x))$ and $(l_j(x), u_j(x))$ are disjoint if $i \neq j$. Otherwise, there would exist a point y in both intervals, and this would mean that the point (x, y) would belong to both S_i^0 and S_j^0, which is impossible by hypothesis. Therefore, for each $x \in (a, b)$, the intervals $\{(l_i()x), u_i(x))\}$ are pairwise disjoint open intervals, and they are all contained in the interval $(l(x), u(x))$, because the S_i's are subsets of S. Hence, the sum of the lengths of the open intervals $\{(l_i(x), u_i(x))\}$ is less than or equal to the length of $(l(x), u(x))$. Also, for any point y in the closed interval $[l(x), u(x)]$, the point (x, y) must belong to one of the S_i's, implying that y is in the closed interval $[l_i(x), u_i(x)]$ for some i. But this means that the sum of the lengths of the closed intervals $[l_i(x), u_i(x)]$ is greater than or equal to the length of the interval $[l(x), u(x)]$. Since open intervals and closed intervals have the same length, we then see that $(u(x) - l(x) = \sum_{i=1}^n (u_i(x) - l_i(x))$.
We now have the following calculation:

$$\begin{aligned}\sum_{i=1}^n A(S_i) &= \sum_{i=1}^n \int_{a_i}^{b_i} (u_i(x) - l_i(x))\, dx \\ &= \sum_{i=1}^n \int_a^b (u_i(x) - l_i(x))\, dx \\ &= \int_a^b \sum_{i=1}^n (u_i(x) - l_i(x))\, dx \\ &= \int_a^b (u(x) - l(x))\, dx \\ &= A(S),\end{aligned}$$

which completes the proof.

EXTENDING THE DEFINITION OF INTEGRABILITY

We now wish to extend the definition of the integral to a wider class of functions, namely to some that are unbounded and Others whose domains are not closed and bounded intervals. This extended definition is somewhat ad hoc, and these integrals are sometimes called "improper integrals."

DEFINITION. Let f be a real or complex-valued function on the open interval (a, b) where a is possibly $-\infty$ and b is possibly $+\infty$. We say that f is *improperly-integrable* on (a, b) if it is integrable on each closed and bounded subinterval $[a', b'] \subset (a, b)$, and for each point $c \in (a, b)$ we have that the two limits $\lim b' \to b - 0 \int_c^{b'} f$ and $\lim_{a' \to a+0} \int_{a'}^c f$ exist.
More generally, We say that a real or complex-valued function f, not necessarily defined on all of the open interval (a, b), is *improperly-integrable* on (a, b) if there exists a partition $\{x_i\}$ of $[a, b]$ such that f is defined and improperly-integrable on each open interval (x_{i-1}, x_i).
We denote the set of all functions f that are improperly-integrable on an open interval (a, b) by $I_i((a, b))$.
Analogous definitions are made for a function's being integrable on half-open intervals $[a, b)$ and $(a, b]$.

Note that, in order for f to be improperly-integrable on an open interval, we only require f to be defined at almost all the points of the interval, i.e., at every point except the endpoints of some partition.

Exercise 5.21. (a) Let f be defined and improperly-integrable on the open interval (a, b). Show that $\lim_{a' \to a+0} \int_{a'}^c f + \lim_{b' \to b-0} \int_c^{b'} f$ is the same for all $c \in (a, b)$.
(b) Define a function f on $(0, 1)$ by $f(x) = (1 - x)^{-1/2}$. Show that f is improperly-integrable on $(0, 1)$ and that f is **not** bounded. (Compare this with part (1) of Theorem 5.5.)
(c) Define a function g on $(0, 1)$ by $g(x) = (1 - x)^{-1}$. Show that g is **not** improperly-integrable on $(0, 1)$, and, using part (b), conclude that the product of improperly-integrable functions on $(0, 1)$ need not itself be improperly-integrable. (Compare this with part (3) of Theorem 5.5.)
(d) Define h to be the function on $(0, \infty)$ given by $h(x) = 1$ for all x. Show that h is not improperly-integrable on $(0, \infty)$. (Compare this with parts (4) and (5) of Theorem 5.5.)

Part (a) of the preceding exercise is just the consistency condition we need in order to make a definition of the integral of an improperly-integrable function over an open interval.

DEFINITION. Let f be defined and improperly-integrable on an open interval (a, b). We define the *integral* of f over the interval (a, b), and denote it by $\int_a^b f$, by

$$\int_a^b f = \lim_{a' \to a+0} \int_{a'}^c f + \lim_{b' \to b-0} \int_c^{b'} f.$$

In general, if f is improperly-integrable over an open interval, i.e., f is defined and improperly-integrable over each subinterval of (a, b) determined by a partition $\{x_i\}$, then we define the *integral* of f over the interval (a, b) by

$$\int_a^b f = \sum_{i=1}^n \int_{x_{i-1}}^{x_i} f.$$

THEOREM 5.17. *Let (a,b) be a fixed open interval (with a possibly equal to $-\infty$ and b possibly equal to $+\infty$), and let $I_i((a,b))$ denote the set of improperly-integrable functions on (a,b). Then:*

(1) $I_i((a,b))$ *is a vector space of functions.*
(2) (Linearity) $\int_a^b(\alpha f+\beta g)=\alpha\int_a^b f+\beta\int_a^b g$ for all $f,g\in I_i((a,b))$and $\alpha,\beta\in\mathbb{C}$.
(3) (Positivity) If $f(x)\geq 0$ for all $x\in(a,b)$, then $\int_a^b f\geq 0$.
(4) (Order-preserving) If $f,g\in I_i((a,b))$ and $f(x)\leq g(x)$ for all $x\in(a,b)$, then $\int_a^b f\leq\int_a^b g$.

Exercise 5.22. (a) Use Theorems 5.5, 5.6, 5.7, and properties of limits to prove the preceding theorem.
(b) Let f be defined and improperly-integrable on (a,b). Show that, given an $\epsilon>0$, there exists a $\delta>0$ such that for any $a<a'<a+\delta$ and any $b-\delta<b'<b$ we have $|\int_a^{a'}f|+|\int_{b'}^b f|<\epsilon$.
(c) Let f be improperly-integrable on an open interval (a,b). Show that, given an $\epsilon>0$, there exists a $\delta>0$ such that if (c,d) is any open subinterval of (a,b) for which $d-c<\delta$, then $|\int_c^d f|<\epsilon$.
HINT: Let $\{x_i\}$ be a partition of $[a,b]$ such that f is defined and improperly-integrable on each subinterval (x_{i-1},x_i). For each i, choose a δ_i using part (b). Now f is bounded by M on all the intervals $[x_{i-1}+\delta_i,x_i-\delta_i]$, so $\delta=\epsilon/M$ should work there.
(d) Suppose f is a continuous function on a closed bounded interval $[a,b]$ and is continuously differentiable on the open interval (a,b). Prove that f' is improperly-integrable on (a,b), and evaluate $\int_a^b f'$.
HINT: Fix a point $c\in(a,b)$, and use the Fundamental Theorem of Calculus to show that the two limits exist.
(e) (Integration by substitution again.) Let $g:[c,d]\to[a,b]$ be continuous on $[c,d]$ and satisfy $g(c)=a$ and $g(d)=b$. Suppose there exists a partition $\{x_0<x_1<\ldots<x_n\}$ of the interval $[c,d]$ such that g is continuously differentiable on each subinterval (x_{i-1},x_i). Prove that g' is improperly-integrable on the open interval (c,d). Show also that if f is continuous on $[a,b]$, we have that

$$\int_a^b f(t)\,dt=\int_c^d f(g(s))g'(s)\,ds.$$

HINT: Integrate over the subintervals (x_{i-1},x_i), and use part (d).

REMARK. Note that there are parts of Theorems 5.5 and 5.6 that are not asserted in Theorem 5.17. The point is that these other properties do not hold for improperly-integrable functions on open intervals. See the following exercise.

Exercise 5.23. (a) Define f to be the function on $[1,\infty)$ given by $f(x)=(-1)^{n-1}/n$ if $n-1\leq x<n$. Show that f is improperly-integrable on $(1,\infty)$, but that $|f|$ is not improperly-integrable on $(1,\infty)$. (Compare this with part (4) of Theorem 5.6.)
HINT: Verify that $\int_1^N f$ is a partial sum of a convergent infinite series, and then verify that $\int_1^N|f|$ is a partial sum of a divergent infinite series.

(b) Define the function f on $(1,\infty)$ by $f(x)=1/x$. For each positive integer n, define the function f_n on $(1,\infty)$ by $f_n(x)=1/x$ if $1<x<n$ and $f_n(x)=0$ otherwise. Show that each f_n is improperly-integrable on $(1,\infty)$, that f is the uniform limit of the sequence $\{f_n\}$, but that f is not improperly-integrable on $(1,\infty)$. (Compare this with part (5) of Theorem 5.6.)
(c) Suppose f is a nonnegative real-valued function on the half-open interval (a,∞) that is integrable on every closed bounded subinterval $[a,b']$. For each positive integer $n\geq a$, define $y_n=\int_a^n f(x)\,dx$. Prove that f is improperly-integrable on $[a,\infty)$ if and only if the sequence $\{y_n\}$ is convergent. In that case, show that $\int_a^\infty f=\lim y_n$.

We are now able to prove an important result relating integrals over infinite intervals and convergence of infinite series.

THEOREM 5.18. *Let f be a positive function on $[1,\infty)$, assume that f is integrable on every closed bounded interval $[1,b]$, and suppose that f is nonincreasing; i.e., if $x<y$ then $f(x)\geq f(y)$. For each positive integer i, set $a_i=f(i)$, and let S_N denote the Nth partial sum of the infinite series $\sum a_i : S_N=\sum_{i=1}^N a_i$. Then:*

(1) *For each N, we have*

$$S_N-a_1\leq\int_1^N f(x)\,dx\leq S_{N-1}.$$

(2) *For each N, we have that*

$$S_{N-1}-\int_1^N f(x)\,dx\leq a_1-a_N\leq a_1;$$

i.e., the sequence $\{S_{N-1}-\int_1^N f\}$ is bounded above.

(3) *The sequence $\{S_{N-1}-\int_1^N f\}$ is nondecreasing.*

(4) (Integral Test) The infinite series $\sum a_i$ converges if and only if the function f is improperly-integrable on $(1,\infty)$.

PROOF. For each positive integer N, define a step function k_N on the interval $[1,N]$ as follows. Let $P=\{x_0<x_1<\ldots<x_{N-1}\}$ be the partition of $[1,N]$ given by the points $\{1<2<3<\ldots<N\}$, i.e., $x_i=i+1$. Define $k_N(x)$ to be the constant $c_i=f(i+1)$ on the interval $[x_{i-1},x_i)=[i,i+1)$. Complete the definition of k_N by setting $k_N(N)=f(N)$. Then, because f is nonincreasing, we have that $k_N(x)\leq f(x)$ for all $x\in[1,N]$. Also,

$$\begin{aligned}\int_1^N k_N &= \sum_{i=1}^{N-1} c_i(x_i-x_{i-1})\\ &= \sum_{i=1}^{N-1} f(i+1)\\ &= \sum_{i=2}^{N} f(i)\\ &= \sum_{i=2}^{N} a_i\\ &= S_N-a_1,\end{aligned}$$

which then implies that

$$S_N - a_1 = \int_1^N k_N(x)\,dx \le \int_1^N f(x)\,dx.$$

This proves half of part (1).
For each positive integer $N > 1$ define another step function l_N, using the same partition P as above, by setting $l_N(x) = f(i)$ if $i \le x < i+1$ for $1 \le i < N$, and complete the definition of l_N by setting $l_N(N) = f(N)$. Again, because f is nonincreasing, we have that $f(x) \le l_N(x)$ for all $x \in [1, N]$. Also

$$\begin{aligned}\int_1^N l_N &= \sum_{i=1}^{N-1} f(i)\\ &= \sum_{i=1}^{N-1} a_i\\ &= S_{N-1},\end{aligned}$$

which then implies that

$$\int_1^N f(x)\,dx \le \int_1^N l_N(x)\,dx = S_{N-1},$$

and this proves the other half of part (1).
It follows from part (1) that

$$S_{N-1} - \int_1^N f(x)\,dx \le S_{N-1} - S_N + a_1 = a_1 - a_N,$$

and this proves part (2).
We see that the sequence $\{S_{N-1} - \int_1^N f\}$ is nondecreasing by observing that

$$\begin{aligned}S_N - \int_1^{N+1} f - S_{N-1} + \int_1^N f &= a_N - \int_N^{N+1} f\\ &= f(N) - \int_N^{N+1} f\\ &\ge 0,\end{aligned}$$

because f is nonincreasing.
Finally, to prove part (4), note that both of the sequences $\{S_N\}$ and $\{\int_1^N f\}$ are nondecreasing. If f is improperly-integrable on $[1, \infty)$, then $\lim_N \int_1^N f$ exists, and $S_N \le a_1 + \int_1^\infty f(x)\,dx$ for all N, which implies that $\sum a_i$ converges by Theorem 2.14. Conversely, if $\sum a_i$ converges, then $\lim S_N$ exists. Since $\int_1^N f(x)\,dx \le S_{N-1}$, it then follows, again from Theorem 2.14, that $\lim_N \int_1^N f(x)\,dx$ exists. So, by the preceding exercise, f is improperly-integrable on $[1, \infty)$.

We may now resolve a question first raised in Exercise 2.32. That is, for $1 < s < 2$, is the infinite series $\sum 1/n^s$ convergent or divergent? We saw in that exercise that this series is convergent if s is a rational number.

Exercise 5.24. (a) Let s be a real number. Use the Integral Test to prove that the infinite series $\sum 1/n^s$ is convergent if and only if $s > 1$.
(b) Let s be a complex number $s = a + bi$. Prove that the infinite series $\sum 1/n^s$ is absolutely convergent if and only if $a > 1$.

Exercise 5.25. Let f be the function on $[1, \infty)$ defined by $f(x) = 1/x$.
(a) Use Theorem 5.18 to prove that the sequence $\{\sum_{i=1}^{N} \frac{1}{i} - \ln N\}$ converges to a positive number $\gamma \leq 1$. (This number γ is called Euler's constant.)
HINT: Show that this sequence is bounded above and nondecreasing.
(b) Prove that

$$\sum_{i=1}^{\infty} \frac{(-1)^{i+1}}{i} = \ln 2.$$

HINT: Write S_{2N} for the $2N$th partial sum of the series. Use the fact that

$$S_{2N} = \sum_{i=1}^{2N} \frac{1}{i} - 2\sum_{i=1}^{N} \frac{1}{2i}.$$

Now add and subtract $\ln(2N)$ and use part (a).

INTEGRATION IN THE PLANE

Let S be a closed geometric set in the plane. If f is a real-valued function on S, we would like to define what it means for f to be "integrable" and then what the "integral" of f is. To do this, we will simply mimic our development for integration of functions on a closed interval $[a, b]$.
So, what should be a "step function" in this context? That is, what should is a "partition" of S be in this context? Presumably a step function is going to be a function that is constant on the "elements" of a partition. Our idea is to replace the subintervals determined by a partition of the interval $[a, b]$ by geometric subsets of the geometric set S.

DEFINITION. The *overlap* of two geometric sets S_1 and S_2 is defined to be the interior $(S_1 \cap S_2)^0$ of their intersection. S_1 and S_2 are called *nonoverlapping* if this overlap $(S_1 \cap S_2)^0$ is the empty set.

DEFINITION. A *partition* of a closed geometric set S in $\mathbb{R}^2$ is a finite collection $\{S_1, S_2, \ldots, S_n\}$ of nonoverlapping closed geometric sets for which $\cup_{i=1}^{n} S_i = S$; i.e., the union of the S_i's is all of the geometric set S.
The open subsets $\{S_i^0\}$ are called the *elements* of the partition.
A *step function* on the closed geometric set S is a real-valued function h on S for which there exists a partition $P = \{S_i\}$ of S such that $h(z) = a_i$ for all $z \in S_i^0$; i.e., h is constant on each element of the partition P.

REMARK. One example of a partition of a geometric set, though not at all the most general kind, is the following. Suppose the geometric set S is determined by the interval $[a, b]$ and the two bounding functions u and l. Let $\{x_0 < x_1 < \ldots < x_n\}$ be a partition of the interval $[a, b]$. We make a partition $\{S_i\}$ of S by constructing vertical lines at the points x_i from $l(x_i)$ to $u(x_i)$. Then S_i is the geometric set determined by the interval $[x_{i-1}, x_i]$ and the two bounding functions u_i and l_i that are the restrictions of u and l to the interval $[x_{i-1}, x_i]$.

A step function is constant on the open geometric sets that form the elements of some partition. We say nothing about the values of h on the "boundaries" of these geometric sets. For a step function h on an interval $[a, b]$, we do not worry about the finitely many values of h at the endpoints of the subintervals. However, in the plane, we are ignoring the values on the boundaries, which are infinite sets. As a consequence, a step function on a geometric set may very well have an infinite range, and may not even be a bounded function, unlike the case for a step function on an interval. The idea is that the boundaries of geometric sets are "negligible" sets as far as area is concerned, so that the values of a function on these boundaries shouldn't affect the integral (average value) of the function.

Before continuing our development of the integral of functions in the plane, we digress to present an analog of Theorem 3.20 to functions that are continuous on a closed geometric set.

THEOREM 5.19. *Let f be a continuous real-valued function whose domain is a closed geometric set S. Then there exists a sequence $\{h_n\}$ of step functions on S that converges uniformly to f.*

PROOF. As in the proof of Theorem 3.20, we use the fact that a continuous function on a compact set is uniformly continuous.

For each positive integer n, let δ_n be a positive number satisfying $|f(z)-f(w)| < 1/n$ if $|z - w| < \delta_n$. Such a δ_n exists by the uniform continuity of f on S. Because S is compact, it is bounded, and we let $R = [a, b] \times [c, d]$ be a closed rectangle that contains S. We construct a partition $\{S_i^n\}$ of S as follows. In a checkerboard fashion, we write R as the union $\cup R_i^n$ of small, closed rectangles satisfying

(1) If z and w are in R_i^n, then $|z - w| < \delta_n$. (The rectangles are that small.)
(2) $R_i^{n\,0} \cap R_j^{n\,0} = \emptyset$. (The interiors of these small rectangles are disjoint.)

Now define $S_i^n = S \cap R_i^n$. Then $S_i^{n\,0} \cap S_j^{n\,0} = \emptyset$, and $S = \cup S_i^n$. Hence, $\{S_i^n\}$ is a partition of S.

For each i, choose a point z_i^n in S_i^n, and set $a_i^n = f(z_i^n)$. We define a step function h_n as follows: If z belongs to one (and of course only one) of the open geometric sets $S_i^{n\,0}$, set $h_n(z) = a_i^n$. And, if z does not belong to any of the open geometric sets $S_i^{n\,0}$, set $h_n(z) = f(z)$. It follows immediately that h_n is a step function.

Now, we claim that $|f(z)-h_n(z)| < 1/n$ for all $z \in S$. For any z in one of the $S_i^{n\,0}$'s, we have

$$|f(z) - h_n(z)| = |f_z) - a_i^n| = |f(z) - f(z_I^n)| < 1/n$$

because $|z - z_i^n| < \delta_n$. And, for any z not in any of the $S_i^{n\,0}$'s, $f(z) - h_n(z) = 0$. So, we have defined a sequence $\{h_n\}$ of step functions on S, and the sequence $\{h_n\}$ converges uniformly to f by Exercise 3.29.

What follows now should be expected. We will define the integral of a step function h over a geometric set S by

$$\int_S h = \sum_{i=1}^{n} a_i \times A(S_i).$$

We will define a function f on S to be integrable if it is the uniform limit of a sequence $\{h_n\}$ of step functions, and we will then define the integral of f by

$$\int_S f = \lim \int_S h_n.$$

Everything should work out nicely. Of course, we have to check the same two consistency questions we had for the definition of the integral on $[a, b]$, i.e., the analogs of Theorems 5.1 and 5.4.

THEOREM 5.20. *Let S be a closed geometric set, and let h be a step function on S. Suppose $P = \{S_1, \dots, S_n\}$ and $Q = \{T_1, \dots, T_m\}$ are two partitions of S for which $h(z)$ is the constant a_i on S_i^0 and $h(z)$ is the constant b_j on T_j^0. Then*

$$\sum_{i=1}^{n} a_i A(S_i) = \sum_{j=1}^{m} b_j A(T_j).$$

PROOF. We know by part (d) of Exercise 5.17 that the intersection of two geometric sets is itself a geometric set. Also, for each fixed index j, we know that the sets $\{T_j \cap S_i^0\}$ are pairwise disjoint. Then, by Theorem 5.16, we have that $A(T_j) = \sum_{i=1}^{n} A(T_j \cap S_i)$. Similarly, for each fixed i, we have that $A(S_i = \sum_{j=1}^{m} A(T_j \cap S_i)$. Finally, for each pair i and j, for which the set $T_j^0 \cap S_i^0$ is not empty, choose a point $z_{i,j} \in T_j^0 \cap S_i^0$, and note that $a_i = h(z_{i,j}) = b_j$, because $z_{i,j}$ belongs to both S_i^0 and T_j^0.
With these observations, we then have that

$$\begin{aligned}\sum_{i=1}^{n} a_i A(S_i) &= \sum_{i=1}^{n} a_i \sum_{j=1}^{m} A(T_j \cap S_i)\\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} a_i A(T_j \cap S_i)\\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} h(z_{i,j}) A(T_j \cap S_i)\\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} b_j A(T_j \cap S_i)\\ &= \sum_{j=1}^{m} \sum_{i=1}^{n} b_j A(T_j \cap S_i)\\ &= \sum_{j=1}^{m} b_j \sum_{i=1}^{n} A(T_j \cap S_i)\\ &= \sum_{j=1}^{m} b_j A(T_j),\end{aligned}$$

which completes the proof.

OK, the first consistency condition is satisfied. Moving right along:

DEFINITION. Let h be a step function on a closed geometric set S. Define the *integral* of h over the geometric set S by the formula

$$\int_S h = \int_S H(z)\,dz = \sum_{i=1}^{n} a_i A(S_i),$$

where $S_1, \ldots, S_n$ is a partition of S for which h is the constant a_i on the interior S_i^0 of the set S_i.

Just as in the case of integration on an interval, before checking the second consistency result, we need to establish the following properties of the integral of step functions.

THEOREM 5.21. *Let $H(S)$ denote the vector space of all step functions on the closed geometric set S. Then the assignment $h \to \int h$ of $H(S)$ into $\mathbb{R}$ has the following properties:*

(1) (Linearity) $H(S)$ is a vector space, and $\int_S (h_1 + h_2) = \int_S h_1 + \int_S h_2$, and $\int_S ch = c \int_S h$ for all $h_1, h_2, h \in H(S)$, and for all real numbers c.

(2) If $h = \sum_{i=1}^n c_i \chi_{S_i}$ is a linear combination of indicator functions of geometric sets that are subsets of S, then $\int h = \sum_{i=1}^n c_i A(S_i)$.

(3) (Positivity) If $h(z) \geq 0$ for all $z \in S$, then $\int_S h \geq 0$.

(4) (Order-preserving) If h_1 and h_2 are step functions on S for which $h_1(z) \leq h_2(z)$ for all $z \in S$, then $\int_S h_1 \leq \int_S h_2$.

PROOF. Suppose h_1 is constant on the elements of a partition $P = \{S_i\}$ and h_2 is constant on the elements of a partition $Q = \{T_j\}$. Let V be the partition of the geometric set S whose elements are the sets $\{U_k\} = \{S_i^0 \cap T_j^0\}$. Then both h_1 and h_2 are constant on the elements U_k of V, so that $h_1 + h_2$ is also constant on these elements. Therefore, $h_1 + h_2$ is a step function, and

$$\int (h_1 + h_2) = \sum_k (a_k + b_k) A(U_k) = \sum_k a_k A(U_k) + \sum_k b_k A(U_k) = \int h_1 + \int h_2,$$

and this proves the first assertion of part (1).
The proof of the other half of part (1), as well as parts (2), (3), and (4), are totally analogous to the proofs of the corresponding parts of Theorem 5.2, and we omit the arguments here.

Now for the other necessary consistency condition:

THEOREM 5.22. *let S be a closed geometric set in the plane.*

(1) *If $\{h_n\}$ is a sequence of step functions that converges uniformly to a function f on S, then the sequence $\{\int_S h_n\}$ is a convergent sequence of real numbers.*

(2) *If $\{h_n\}$ and $\{k_n\}$ are two sequences of step functions on S that converge uniformly to the same function f, then*

$$\lim \int_S h_n = \lim \int_S k_n.$$

Exercise 5.26. Prove Theorem 5.22. Mimic the proofs of Theorems 5.3 and 5.4.

DEFINITION. If f is a real-valued function on a closed geometric set S in the plane, then f is *integrable on* S if it is the uniform limit of a sequence $\{h_n\}$ of step functions on S.
We define the *integral* of an integrable function f on S by

$$\int_S f \equiv \int_S f(z)\,dz = \lim \int_S h_n,$$

where $\{h_n\}$ is a sequence of step functions on S that converges uniformly to f.

THEOREM 5.23. *Let S be a closed geometric set in the plane, and let $I(S)$ denote the set of integrable functions on S. Then:*

(1) *$I(S)$ is a vector space of functions.*
(2) *If f and $g \in I(S)$, and one of them is bounded, then $fg \in I(S)$.*
(3) *Every step function is in $I(S)$.*
(4) *If f is a continuous real-valued function on S, then f is in $I(S)$. That is, every continuous real-valued function on S is integrable on S.*

Exercise 5.27. (a) Prove Theorem 5.23. Note that this theorem is the analog of Theorem 5.5, but that some things are missing.
(b) Show that integrable functions on S are not necessarily bounded; not even step functions have to be bounded.
(c) Show that, if $f \in I(S)$, and g is a function on S for which $f(x,y) = g(x,y)$ for all (x,y) in the interior S^0 of S, then $g \in I(S)$. That is, integrable functions on S can do whatever they like on the boundary.

THEOREM 5.24. *Let S be a closed geometric set. The assignment $f \to \int f$ on $I(S)$ satisfies the following properties.*

(1) (Linearity) $I(S)$ is a vector space, and $\int_S(\alpha f + \beta g) = \alpha \int_S f + \beta \int_S g$ for all $f, g \in I(S)$and $\alpha, \beta \in \mathbb{R}$.
(2) (Positivity) If $f(z) \geq 0$ for all $z \in S$, then $\int_S f \geq 0$.
(3) (Order-preserving) If $f, g \in I(S)$ and $f(z) \leq g(z)$ for all $z \in S$, then $\int_S f \leq \int_S g$.
(4) If $f \in I(S)$, then so is $|f|$, and $|\int_S f| \leq \int_S |f|$.
(5) If f is the uniform limit of functions f_n, each of which is in $I(S)$, then $f \in I(S)$ and $\int_S f = \lim \int_S f_n$.
(6) Let $\{u_n\}$ be a sequence of functions in $I(S)$, and suppose that for each n there is a number m_n, for which $|u_n(z)| \leq m_n$ for all $z \in S$, and such that the infinite series $\sum m_n$ converges. Then the infinite series $\sum u_n$ converges uniformly to an integrable function, and $\int_S \sum u_n = \sum \int_S u_n$.
(7) If $f \in I(S)$, and $\{S_1, \ldots, S_n\}$ is a partition of S, then $f \in I(S_i)$ for all i, and

$$\int_S = \sum_{i=1}^{n} \int_{S_i} f.$$

Exercise 5.28. Prove Theorem 5.24. It is mostly the analog to Theorem 5.6. To see the last part, let h_i be the step function that is identically 1 on S_i; check that $h_i f \in I(S_i)$; then examine $\sum_i \int_S f h_i$.

Of course, we could now extend the notion of integrability over a geometric set S to include complex-valued functions just as we did for integrability over an interval $[a,b]$. However, real-valued functions on geometric sets will suffice for the purposes of this book.

We include here, to be used later in Chapter VII, a somewhat technical theorem about constructing partitions of a geometric set.

THEOREM 5.25. *Let $S_1, \ldots, S_n$ be closed, nonoverlapping, geometric sets, all contained in a geometric set S. Then there exists a partition $\widehat{S}_1, \ldots, \widehat{S}_M$ of S such* *that for $1 \leq i \leq n$ we have $S_i = \widehat{S}_i$. In other words, the s_i's are the first n elements of a partition of S.*

PROOF. Suppose S is determined by the interval $[a, b]$ and the two bounding functions u and l. We prove this theorem by induction on n.

If $n = 1$, let S_1 be determined by the interval $[a_1, b_1]$ and the two bounding functions u_1 and l_1. Set $\widehat{S}_1 = S_1$, and define four more geometric sets $\widehat{S}_2, \dots, \widehat{S}_5$ as follows:

(1) $\widehat{S}_2$ is determined by the interval $[a, a_1]$ and the two bounding functions u and l restricted to that interval.

(2) S_3 is determined by the interval $[a_1, b_1]$ and the two bounding functions u and u_1 restricted to that interval.

(3) S_4 is determined by the interval $[a_1, b_1]$ and the two bounding functions l and l_1 restricted to that interval.

(4) $\widehat{S}_5$ is determined by the interval $[b_1, b]$ and the two bounding functions u and l restricted to that interval.

Observe that the five sets $\widehat{S}_1, \widehat{S}_2, \dots, \widehat{S}_5$ constitute a partition of the geometric set S, proving the theorem in the case $n = 1$.

Suppose next that the theorem is true for any collection of n sets satisfying the hypotheses. Then, given $S_1, \dots, S_{n+1}$ as in the hypothesis of the theorem, apply the inductive hypothesis to the n sets $S_1, \dots, S_n$ to obtain a partition $T_1, \dots, T_m$ of S for which $T_i = S_i$ for all $1 \leq i \leq n$. For each $n+1 \leq i \leq m$, consider the geometric set $S'_i = S_{n+1} \cap T_i$ of the geometric set T_i. We may apply the case $n = 1$ of this theorem to this geometric set to conclude that S'_i is the first element $S'_{i,1}$ of a partition $\{S'_{i,1}, S'_{i,2}, \dots, S'_{i,m_i}\}$ of the geometric set T_i.

Define a partition $\{\widehat{S}_k\}$ of S as follows: For $1 \leq k \leq n$, set $\widehat{S}_k = T_k$. Set $\widehat{S}_{n+1} = \cup_{i=n+1}^{m} S'_{i,1} = S_{n+1}$. And define the rest of the partition $\{\widehat{S}_k\}$ to be made up of the remaining sets $S'_{i,j}$ for $n+1 \leq i \leq m$ and $2 \leq j \leq m_i$. It follows directly that this partition $\{\widehat{S}_k\}$ satisfies the requirements of the theorem.

Exercise 5.29. Let $S_1, \dots, S_n$ be as in the preceding theorem. Suppose S_k is determined by the interval $[a_k, b_k]$ and the two bounding functions u_k and l_k. We will say that S_k is "below" S_j, equivalently S_j is "above" S_k, if there exists a point x such that $u_k(x) < l_j(x)$. Note that this implies that $x \in [a_k, b_k] \cap [a_j, b_j]$.

(a) Suppose S_k is below S_j, and suppose $(z, y_k) \in S_k$ and $(z, y_j) \in S_j$. Show that $y_j > y_k$. That is, if S_k is below S_j, then no part of S_k can be above S_j.

(b) Suppose S_2 is below S_1 and S_3 is below S_2. Show that no part of S_3 can be above S_1.

HINT: By way of contradiction, let $x_1 \in [a_1, b_1]$ be such that $u_2(x_1) < l_1(x_1)$; let $x_2 \in [a_2, b_2]$ be such that $u_3(x_2) < l_2(x_2)$; and suppose $x_3 \in [a_3, b_3]$ is such that $u_1(x_3) < l_3(x_3)$. Derive contradictions for all possible arrangements of the three points x_1, x_2, and x_3.

(c) Prove that there exists an index k_0 such that S_{k_0} is minimal in the sense that there is no other S_j that is below S_{k_0}.

HINT: Argue by induction on n. Thus, let $\{T_l\}$ be the collection of all S_k's that are below S_1, and note that there are at most $n-1$ elements of $\{T_l\}$. By induction, there is one of the T_l's, i.e., an S_{k_0} that is minimal for that collection. Now, using part (b), show that this S_{k_0} must be minimal for the original collection.

There is one more concept about integrating over geometric sets that we will need in later chapters. We have only considered sets that are bounded on the left and right by straight vertical lines and along the top and bottom by graphs of continuous functions $y = u(x)$ and $y = l(x)$. We equally well could have discussed sets that are bounded above and below by straight horizontal lines and bounded on the left and right by graphs of continuous functions $x = l(y)$ and $x = r(y)$. These additional sets do not provide anything particularly important, so we do not discuss them. However, there are times when it is helpful to work with geometric sets with the roles of horizontal and vertical reversed. We accomplish this with the following definition.

DEFINITION. Let S be a subset of $\mathbb{R}^2$. By the *symmetric image* of S we mean the set $\widehat{S}$ of all points $(x, y) \in \mathbb{R}^2$ for which the point $(y, x) \in S$.

The symmetric image of a set is just the reflection of the set across the $y = x$ line in the plane. Note that the symmetric image of the rectangle $[a, b] \times [c, d]$ is again a rectangle, $[c, d] \times [a, b]$, and therefore the area of a rectangle is equal to the area of its symmetric image. This has the implication that if the symmetric image of a geometric set is also a geometric set, then they both have the same area. The symmetric image of a geometric set doesn't have to be a geometric set itself. For instance, consider the examples suggested in part (b) of Exercise 5.17. But clearly rectangles, triangles, and circles have this property, for their symmetric images are again rectangles, triangles, and circles. For a geometric set, whose symmetric image is again a geometric set, there are some additional computational properties of the area of S as well as the integral of functions over S, and we present them in the following exercises.

Exercise 5.30. Suppose S is a closed geometric set, which is determined by a closed interval $[a, b]$ and two bounding functions $u(x)$ and $l(x)$. Suppose the symmetric image $\widehat{S}$ of S is also a closed geometric set, determined by an interval $[\widehat{a}, \widehat{b}]$ and two bounding functions $\widehat{u}(x)$ and $\widehat{l}(x)$.

(a) Make up an example to show that the numbers $\widehat{a}$ and $\widehat{b}$ need not have anything to do with the numbers a and b, and that the functions $\widehat{u}$ and $\widehat{l}$ need not have anything to do with the functions u and l.

(b) Prove that S and $\widehat{S}$ have the same area.

HINT: use the fact that the area of a geometric set is approximately equal to the sum of the areas of certain rectangles, and then use the fact that the area of the symmetric image of a rectangle is the same as the area of the rectangle.

(c) Show that for every point $(x, y) \in S$, we must have $\widehat{a} \leq y \leq \widehat{b}$, and for every such y, we must have $\widehat{l}(y) \leq x \leq \widehat{u}(y)$.

HINT: If $(x, y) \in S$, then $(y, x) \in \widehat{S}$.

(d) Prove that the area $A(S)$ of S is given by the formula

$$A(S) = \int_a^b \int_{l(x)}^{u(x)} 1\, dydx = \int_{\widehat{a}}^{\widehat{b}} \int_{\widehat{l}(y)}^{\widehat{u}(y)} 1\, dxdy.$$

(See the remark preceding Theorem 5.16.)

(e) Let S be the right triangle having vertices $(a, c), (b, c)$, and (b, d), where $d > c$. Describe the symmetric image of S; i.e., find the corresponding $\widehat{a}, \widehat{b}, \widehat{u}$, and $\widehat{l}$. Use part (d) to obtain the following formulas for the area of S :

$$A(S)=\int_a^b\int_c^{d+\frac{t-b}{a-b}(c-d)}1\,dsdt=\int_c^d\int_{b+\frac{t-d}{c-d}(a-b)}1\,dsdt.$$

Exercise 5.31. (a) Prove that if S_1 and S_2 are geometric sets whose symmetric images are again geometric sets, then the symmetric image of the geometric set $S_1\cap S_2$ is also a geometric set.

(b) Suppose T is a closed geometric set that is contained in a closed geometric set S. Assume that both the symmetric images $\widehat{T}$ and $\widehat{S}$ are also geometric sets. If $\widehat{S}$ is determined by an interval $[\widehat{a},\widehat{b}]$ and two bounding functions $\widehat{u}$ and $\widehat{l}$, prove that

$$A(T)=\int_{\widehat{a}}^{\widehat{b}}\int_{\widehat{l}(s)}^{\widehat{u}(s)}\chi_T(t,s)\,dtds,$$

where χ_T is the indicator function of the set T; i.e., $\chi_T(t,s)=1$ if $(t,s)\in T$, and $\chi_T(t,s)=0$ if $(t,s)\notin T$.

HINT: See the proof of Theorem 5.16, give names to all the intervals and bounding functions, and in the end use part (d) of the preceding exercise.

(c) Suppose $\{S_i\}$ is a partition of a geometric set S, and suppose the symmetric images of S and all the S_i's are also geometric sets. Suppose h is a step function that is the constant a_i on the element S_i^0 of the partition $\{S_i\}$. Prove that $\int_S h=\sum_{i=1}^n a_i\int_S\chi_{S_i^0}$, and therefore that

$$\int_S h=\int_a^b\int_{l(t)}^{u(t)}h(t,s)\,dsdt=\int_{\widehat{a}}^{\widehat{b}}\int_{\widehat{l}(s)}^{\widehat{u}(s)}h(t,s)\,dtds.$$

HINT: Use part (b).

(d) Let S be a geometric set whose symmetric image $\widehat{S}$ is also a geometric set, and suppose f is a continuous function on S. Show that

$$\int_S f=\int_a^b\int_{l(t)}^{u(t)}f(t,s)\,dsdt=\int_{\widehat{a}}^{\widehat{b}}\int_{\widehat{l}(s)}^{\widehat{u}(s)}f(t,s)\,dtds.$$

HINT: Make use of the fact that the step functions constructed in Theorem 5.19 satisfy the assumptions of part (c). Then take limits.

(e) Let S be the triangle in part (e) of the preceding exercise. If f is a continuous function on S, show that the integral of f over S is given by the formulas

$$\int_S f=\int_a^b\int_c^{d+\frac{t-b}{a-b}(c-d)}f(t,s)\,dsdt=\int_c^d\int_{b+\frac{s-d}{c-d}(a-b)}f(t,s)\,dtds.$$

CHAPTER VI
INTEGRATION OVER SMOOTH CURVES IN THE PLANE
$C = 2\pi r$

In this chapter we will define what we mean by a smooth curve in the plane and what is meant by its arc length. These definitions are a good bit more tricky than one might imagine. Indeed, it is the subtlety of the definition of arc length that prevented us from defining the trigonometric functions in terms of wrapping the real line around the circle, a definition frequently used in high school trigonometry courses. Having made a proper definition of arc length, we will then be able to establish the formula $C = 2\pi r$ for the circumference of a circle of radius r.

By the "plane," we will mean $\mathbb{R}^2 \equiv \mathbb{C}$, and we will on occasion want to carefully distinguish between these two notions of the plane, i.e., two real variables x and y as opposed to one complex variable $z = x + iy$. In various instances, for clarity, we will use notations like $x + iy$ and (x, y), remembering that both of these represent the same point in the plane. As $x + iy$, it is a single complex number, while as (x, y) we may think of it as a vector in $\mathbb{R}^2$ having a magnitude and, if nonzero, a direction.

We also will define in this chapter three different kinds of integrals over such curves. The first kind, called "integration with respect to arc length," will be completely analogous to the integral defined in Chapter V for functions on a closed and bounded interval, and it will only deal with functions whose domain is the set consisting of the points on the curve. The second kind of integral, called a "contour integral," is similar to the first one, but it emphasizes in a critical way that we are integrating a complex-valued function over a curve in the complex plane $\mathbb{C}$ and not simply over a subset of $\mathbb{R}^2$. The applications of contour integrals is usually to functions whose domains are open subsets of the plane that contain the curve as a proper subset, i.e., whose domains are larger than just the curve. The third kind of integral over a curve, called a "line integral," is conceptually very different from the first two. In fact, we won't be integrating functions at all but rather a new notion that we call "differential forms." This is actually the beginnings of the subject called differential geometry, whose intricacies and power are much more evident in higher dimensions than 2.

The main points of this chapter include:

(1) The definition of a **smooth curve**, and the definition of its **arc length**,
(2) the derivation of the formula $C = 2\pi r$ for the circumference of a circle of radius r (Theorem 6.5),
(3) the definition of the **integral with respect to arc length**,
(4) the definition of a **contour integral**,
(5) the definition of a **line integral**, and
(6) **Green's Theorem** (Theorem 6.14).

SMOOTH CURVES IN THE PLANE

Our first project is to make a satisfactory definition of a smooth curve in the plane, for there is a good bit of subtlety to such a definition. In fact, the material in this chapter is all surprisingly tricky, and the proofs are good solid analytical arguments, with lots of ϵ's and references to earlier theorems.

Whatever definition we adopt for a curve, we certainly want straight lines, circles, and other natural geometric objects to be covered by our definition. Our intuition is that a curve in the plane should be a "1-dimensional" subset, whatever that may mean. At this point, we have no definition of the dimension of a general set, so this is probably not the way to think about curves. On the other hand, from the point of view of a physicist, we might well define a curve as the trajectory followed by a particle moving in the plane, whatever that may be. As it happens, we do have some notion of how to describe mathematically the trajectory of a moving particle. We suppose that a particle moving in the plane proceeds in a continuous manner relative to time. That is, the position of the particle at time t is given by a continuous function $f(t) = x(t) + iy(t) \equiv (x(t), y(t))$, as t ranges from time a to time b. A good first guess at a definition of a curve joining two points z_1 and z_2 might well be that it is the range C of a continuous function f that is defined on some closed bounded interval $[a, b]$. This would be a curve that joins the two points $z_1 = f(a)$ and $z_2 = f(b)$ in the plane. Unfortunately, this is also not a satisfactory definition of a curve, because of the following surprising and bizarre mathematical example, first discovered by Guiseppe Peano in 1890.

THE PEANO CURVE. *The so-called "Peano curve" is a continuous function f defined on the interval $[0, 1]$, whose range is the entire unit square $[0, 1] \times [0, 1]$ in $\mathbb{R}^2$.*

Be careful to realize that we're talking about the "range" of f and not its graph. The graph of a real-valued function could never be the entire square. This Peano function is a complex-valued function of a real variable. Anyway, whatever definition we settle on for a curve, we do not want the entire unit square to be a curve, so this first attempt at a definition is obviously not going to work.

Let's go back to the particle tracing out a trajectory. The physicist would probably agree that the particle should have a continuously varying velocity at all times, or at nearly all times, i.e., the function f should be continuously differentiable. Recall that the velocity of the particle is defined to be the rate of change of the position of the particle, and that's just the derivative f' of f. We might also assume that the particle is never at rest as it traces out the curve, i.e., the derivative $f'(t)$ is never 0. As a final simplification, we could suppose that the curve never crosses itself, i.e., the particle is never at the same position more than once during the time interval from $t = a$ to $t = b$. In fact, these considerations inspire the formal definition of a curve that we will adopt below.

Recall that a function f that is continuous on a closed interval $[a, b]$ and continuously differentiable on the open interval (a, b) is called a smooth function on $[a, b]$. And, if there exists a partition $\{t_0 < t_1 < \ldots < t_n\}$ of $[a, b]$ such that f is smooth on each subinterval $[t_{i-1}, t_i]$, then f is called piecewise smooth on $[a, b]$. Although the derivative of a smooth function is only defined and continuous on the open interval (a, b), and hence possibly is unbounded, it follows from part (d) of Exercise 5.22 that this derivative is improperly-integrable on that open interval. We recall also that just because a function is improperly-integrable on an open interval, its absolute value may not be improperly-integrable. Before giving the formal definition of a smooth curve, which apparently will be related to smooth or piecewise smooth functions, it is prudent to present an approximation theorem about smooth functions. Theorem 3.20 asserts that every continuous function on a closed bounded interval is the uniform limit of a sequence of step functions. We give next a similar, but stronger, result about smooth functions. It asserts that a smooth function can be approximated "almost uniformly" by piecewise linear functions.

THEOREM 6.1. *Let f be a smooth function on a closed and bounded interval $[a,b]$, and assume that $|f'|$ is improperly-integrable on the open interval (a,b). Given an $\epsilon > 0$, there exists a piecewise linear function p for which*

(1) $|f(x) - p(x)| < \epsilon$ *for all* $x \in [a,b]$.

(2) $\int_a^b |f'(x) - p'(x)|\,dx < \epsilon$.

That is, the functions f and p are close everywhere, and their derivatives are close on average in the sense that the integral of the absolute value of the difference of the derivatives is small.

PROOF. Because f is continuous on the compact set $[a,b]$, it is uniformly continuous. Hence, let $\delta > 0$ be such that if $x, y \in [a,b]$, and $|x - y| < \delta$, then $|f(x) - f(y)| < \epsilon/2$.

Because $|f'|$ is improperly-integrable on the open interval (a,b), we may use part (b) of Exercise 5.22 to find a $\delta' > 0$, which may also be chosen to be $< \delta$, such that $\int_a^{a+\delta'} |f'| + \int_{b-\delta'}^b |f'| < \epsilon/2$, and we fix such a δ'.

Now, because f' is uniformly continuous on the compact set $[a+\delta', b-\delta']$, there exists an $\alpha > 0$ such that $|f'(x) - f'(y)| < \epsilon/4(b-a)$ if x and y belong to $[a+\delta', b-\delta']$ and $|x - y| < \alpha$. Choose a partition $\{x_0 < x_1 < \ldots < x_n\}$ of $[a,b]$ such that $x_0 = a, x_1 = a+\delta', x_{n-1} = b-\delta', x_n = b$, and $x_i - x_{i-1} < \min(\delta, \alpha)$ for $2 \leq i \leq n-1$. Define p to be the piecewise linear function on $[a,b]$ whose graph is the polygonal line joining the $n+1$ points $(a, f(x_1))$, $\{(x_i, f(x_i))\}$ for $1 \leq i \leq n-1$, and $(b, f(x_{n-1}))$. That is, p is constant on the outer subintervals $[a, x_1]$ and $[x_{n-1}, b]$ determined by the partition, and its graph between x_1 and x_{n-1} is the polygonal line joining the points $\{(x_1, f(x_1)), \ldots, (x_{n-1}, f(x_{n-1}))\}$. For example, for $2 \leq i \leq n-1$, the function p has the form

$$p(x) = f(x_{i-1}) + \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}}(x - x_{i-1})$$

on the interval $[x_{i-1}, x_i]$. So, $p(x)$ lies between the numbers $f(x_{i-1})$ and $f(x_i)$ for all i. Therefore,

$$|f(x) - p(x)| \leq |f(x) - f(x_i)| + |f(x_i) - l(x)| \leq |f(x) - f(x_i)| + |f(x_i) - f(x_{i-1})| < \epsilon.$$

Since this inequality holds for all i, part (1) is proved.

Next, for $2 \leq i \leq n-1$, and for each $x \in (x_{i-1}, x_i)$, we have $p'(x) = (f(x_i) - f(x_{i-1}))/(x_i - x_{i-1})$, which, by the Mean Value Theorem, is equal to $f'(y_i)$ for some $y_i \in (x_{i-1}, x_i)$. So, for each such $x \in (x_{i-1}, x_i)$, we have $|f'(x) - p'(x)| = |f'(x) - f'(y_i)|$, and this is less than $\epsilon/4(b-a)$, because $|x - y_i| < \alpha$. On the two outer intervals, $p(x)$ is a constant, so that $p'(x) = 0$. Hence,

$$\begin{aligned}\int_a^b |f' - p'| &= \sum_{i=1}^n \int_{x_{i-1}}^{x_i} |f' - p'| \\ &= \int_a^{x_1} |f'| + \sum_{i=2}^{n-1} |f' - p'| + \int_{x_{n-1}}^b |f'| \\ &\leq \int_a^{a+\delta'} |f'| + \int_{b-\delta'}^b |f'| + \frac{\epsilon}{4(b-a)} \int_{x_1}^{x_{n-1}} 1 \\ &< \epsilon.\end{aligned}$$

The proof is now complete.

REMARK. It should be evident that the preceding theorem can easily be generalized to a piecewise smooth function f, i.e., a function that is continuous on $[a,b]$, continuously differentiable on each subinterval (t_{i-1},t_i) of a partition $\{t_0 < t_1 < \ldots < t_n\}$, and whose derivative f' is absolutely integrable on (a,b). Indeed, just apply the theorem to each of the subintervals (t_{i-1},t_i), and then carefully piece together the piecewise linear functions on those subintervals.
Now we are ready to define what a smooth curve is.

DEFINITION. By a *smooth curve* from a point z_1 to a different point z_2 in the plane, we mean a set $C \subseteq \mathbb{C}$ that is the range of a 1-1, smooth, function $\phi : [a,b] \to \mathbb{C}$, where $[a,b]$ is a bounded closed interval in $\mathbb{R}$, where $z_1 = \phi(a)$ and $z_2 = \phi(b)$, and satisfying $\phi'(t) \neq 0$ for all $t \in (a,b)$.
More generally, if $\phi : [a,b] \to \mathbb{R}^2$ is 1-1 and piecewise smooth on $[a,b]$, and if $\{t_0 < t_1 < \ldots < t_n\}$ is a partition of $[a,b]$ such that $\phi'(t) \neq 0$ for all $t \in (t_{i-1},t_i)$, then the range C of ϕ is called a *piecewise smooth curve* from $z_1 = \phi(a)$ to $z_2 = \phi(b)$. In either of these cases, ϕ is called a *parameterization* of the curve C.

Note that we do not assume that $|\phi'|$ is improperly-integrable, though the preceding theorem might have made you think we would.

REMARK. Throughout this chapter we will be continually faced with the fact that a given curve can have many different parameterizations. Indeed, if $\phi_1 : [a,b] \to C$ is a parameterization, and if $g : [c,d] \to [a,b]$ is a smooth function having a nonzero derivative, then $\phi_2(s) = \phi_1(g(s))$ is another parameterization of C. Since our definitions and proofs about curves often involve a parametrization, we will frequently need to prove that the results we obtain are independent of the parameterization. The next theorem will help; it shows that any two parameterizations of C are connected exactly as above, i.e., there always is such a function g relating ϕ_1 and ϕ_2.

THEOREM 6.2. *Let $\phi_1 : [a,b] \to C$ and $\phi_2 : [c,d] \to C$ be two parameterizations of a piecewise smooth curve C joining z_1 to z_2. Then there exists a piecewise smooth function $g : [c,d] \to [a,b]$ such that $\phi_2(s) = \phi_1(g(s))$ for all $s \in [c,d]$. Moreover, the derivative g' of g is nonzero for all but a finite number of points in $[c,d]$.*

PROOF. Because both ϕ_1 and ϕ_2 are continuous and 1-1, it follows from Theorem 3.10 that the function $g = \phi_1^{-1} \circ \phi_2$ is continuous and 1-1 from $[c,d]$ onto $[a,b]$. Moreover, from Theorem 3.11, it must also be that g is strictly increasing or strictly decreasing. Write $\phi_1(t) = u_1(t) + iv_1(t) \equiv (u_1(t),v_1(t))$, and $\phi_2(s) = u_2(s) + iv_2(s) \equiv (u_2(s),v_2(s))$. Let $\{x_0 < x_1 < \ldots < x_p\}$ be a partition of $[a,b]$ for which ϕ_1' is continuous and nonzero on the subintervals (x_{j-1},x_j), and let $\{y_0 < y_1 < \ldots < y_q\}$ be a partition of $[c,d]$ for which ϕ_2' is continuous and nonzero on the subintervals (y_{k-1},y_k). Then let $\{s_0 < s_1 < \ldots < s_n\}$ be the partition of $[c,d]$ determined by the finitely many points $\{y_k\} \cup \{g^{-1}(x_j)\}$. We will show that g is continuously differentiable at each point s in the subintervals (s_{i-1},s_i).
Fix an s in one of the intervals (s_{i-1},s_i), and let $t = \phi_1^{-1}(\phi_2(s)) = g(s)$. Of course this means that $\phi_1(t) = \phi_2(s)$, or $u_1(t) = u_2(s)$ and $v_1(t) = v_2(s)$. Then t is in some one of the intervals (x_{j-1},x_j), so that we know that $\phi_1'(t) \neq 0$. Therefore, we must have that at least one of $u_1'(t)$ or $v_1'(t)$ is nonzero. Suppose it is $v_1'(t)$ that is nonzero. The argument, in case it is $u_1'(t)$ that is nonzero, is completely

analogous. Now, because v_1' is continuous at t and $v_1'(t) \neq 0$, it follows that v_1 is strictly monotonic in some neighborhood $(t-\delta, t+\delta)$ of t and therefore is 1-1 on that interval. Then v_1^{-1} is continuous by Theorem 3.10, and is differentiable at the point $v_1(t)$ by the Inverse Function Theorem. We will show that on this small interval $g = v_1^{-1} \circ v_2$, and this will prove that g is continuously differentiable at s. Note first that if $\phi_2(\sigma) = x+iy$ is a point on the curve C, then $v_2(\phi_2^{-1}(x+iy)) = y$. Then, for any $\tau \in [a,b]$, we have

$$\begin{aligned} v_1^{-1}(v_2(g^{-1}(\tau))) &= v_1^{-1}(v_2(\phi_2^{-1}(\phi_1(\tau)))) \\ &= v_1^{-1}(v_2(\phi_2^{-1}(u_1(\tau)+iv_1(\tau)))) \\ &= v_1^{-1}(v_1(\tau)) \\ &= \tau, \end{aligned}$$

showing that $v_1^{-1} \circ v_2 = g^{-1^{-1}} = g$. Hence g is continuously differentiable at every point s in the subintervals (s_{i-1}, s_i). Indeed $g'(\sigma) = v_1^{-1'}(v_2(\sigma))v_2'(\sigma)$ for all σ near s, and hence g is piecewise smooth.

Obviously, $\phi_2(s) = \phi_1(g(s))$ for all s, implying that $\phi_2'(s) = \phi_1'(g(s))g'(s)$. Since $\phi_2'(s) \neq 0$ for all but a finite number of points s, it follows that $g'(s) \neq 0$ for all but a finite number of points, and the theorem is proved.

COROLLARY. *Let ϕ_1 and ϕ_2 be as in the theorem. Then, for all but a finite number of points $z = \phi_1(t) = \phi_2(s)$ on the curve C, we have*

$$\frac{\phi_1'(t)}{|\phi_1'(t)|} = \frac{\phi_2'(s)}{|\phi_2'(s)|}.$$

PROOF OF THE COROLLARY. From the theorem we have that

$$\phi_2'(s) = \phi_1'(g(s))g'(s) = \phi_1'(t)g'(s)$$

for all but a finite number of points $s \in (c,d)$. Also, g is strictly increasing, so that $g'(s) \geq 0$ for all points s where g is differentiable. And in fact, $g'(s) \neq 0$ for all but a finite number of s's, because $g'(s)$ is either $(v_1^{-1} \circ v_2)'(s)$ or $(u_1^{-1} \circ u_2)'(s)$, and these are nonzero except for a finite number of points. Now the corollary follows by direct substitution.

REMARK. If we think of $\phi'(t) = (x'(t), y'(t))$ as a vector in the plane $\mathbb{R}^2$, then the corollary asserts that the direction of this vector is independent of the parameterization, at least at all but a finite number of points. This direction vector will come up again as the unit tangent of the curve.

The adjective "smooth" is meant to suggest that the curve is bending in some reasonable way, and specifically it should mean that the curve has a tangent, or tangential direction, at each point. We give the definition of tangential direction below, but we note that in the context of a moving particle, the tangential direction is that direction in which the particle would continue to move if the force that is keeping it on the curve were totally removed. If the derivative $\phi'(t) \neq 0$, then this vector is the velocity vector, and its direction is exactly what we should mean by the tangential direction.

The adjective "piecewise" will allow us to consider curves that have a finite number of points where there is no tangential direction, e.g., where there are "corners."
We are carefully orienting our curves at the moment. A curve C from z_1 to z_2 is being distinguished from the same curve from z_2 to z_1, even though the set C is the same in both instances. Which way we traverse a curve will be of great importance at the end of this chapter, when we come to Green's Theorem.

DEFINITION. Let C, the range of $\phi : [a, b] \to C$, be a piecewise smooth curve, and let $z = (x, y) = \phi(c)$ be a point on the curve. We say that the curve C has a tangential direction at z, relative to the parameterization ϕ, if the following limit exists:

$$\lim_{t\to c} \frac{\phi(t) - z}{|\phi(t) - z|} = \lim_{t\to c} \frac{\phi(t) - \phi(c)}{|\phi(t) - \phi(c)|}.$$

If this limit exists, it is a vector of length 1 in $\mathbb{R}^2$, and this unit vector is called the unit tangent (relative to the parameterization ϕ) to C at z.
The curve C has a *unit tangent* at the point z if there exists a parameterization ϕ for which the unit tangent at z relative to ϕ exists.

Exercise 6.1. (a) Restate the definition of tangential direction and unit tangent using the $\mathbb{R}^2$ version of the plane instead of the $\mathbb{C}$ version. That is, restate the definition in terms of pairs (x, y) of real numbers instead of a complex number z.
(b) Suppose $\phi : [a, b] \to C$ is a parameterization of a piecewise smooth curve C, and that $t \in (a, b)$ is a point where ϕ is differentiable with $\phi'(t) \neq 0$. Show that the unit tangent (relative to the parameterization ϕ) to C at $z = \phi(t)$ exists and equals $\phi'(t)/|\phi'(t)|$. Conclude that, except possibly for a finite number of points, the unit tangent to C at z is independent of the parameterization.
(c) Let C be the graph of the function $f(t) = |t|$ for $t \in [-1, 1]$. Is C a smooth curve? Is it a piecewise smooth curve? Does C have a unit tangent at every point?
(d) Let C be the graph of the function $f(t) = t^{2/3} = (t^{1/3})^2$ for $t \in [-1, 1]$. Is C a smooth curve? Is it a piecewise smooth curve? Does C have a unit tangent at every point?
(e) Consider the set C that is the right half of the unit circle in the plane. Let $\phi_1 : [-1, 1] \to C$ be defined by

$$\phi_1(t) = (\cos(t\frac{\pi}{2}), \sin(t\frac{\pi}{2})),$$

and let $\phi_2 : [-1, 1] \to C$ be defined by

$$\phi_2(t) = (\cos(t^3\frac{\pi}{2}), \sin(t^3\frac{\pi}{2})).$$

Prove that ϕ_1 and ϕ_2 are both parameterizations of C. Discuss the existence of a unit tangent at the point $(1, 0) = \phi_1(0) = \phi_2(0)$ relative to these two parameterizations.
(f) Suppose $\phi : [a, b] \to C$ is a parameterization of a curve C from z_1 to z_2. Define ψ on $[a, b]$ by $\psi(t) = \phi(a + b - t)$. Show that ψ is a parameterization of a curve from z_2 to z_1.

Exercise 6.2. (a) Suppose f is a smooth, real-valued function defined on the closed interval $[a, b]$, and let $C \subseteq \mathbb{R}^2$ be the graph of f. Show that C is a smooth curve, and find a "natural" parameterization $\phi : [a, b] \to C$ of C. What is the unit tangent to C at the point $(t, f(t))$?

(b) Let z_1 and z_2 be two distinct points in $\mathbb{C}$, and define $\phi : [0,1] \to$ by $\phi(t) = (1-t)z_1 + tz_2$. Show that ϕ is a parameterization of the straight line from the point z_1 to the point z_2. Consequently, a straight line is a smooth curve. (Indeed, what is the definition of a straight line?)
(c) Define a function $\phi : [-r,r] \to \mathbb{R}^2$ by $\phi(t) = (t, \sqrt{r^2 - t^2})$. Show that the range C of ϕ is a smooth curve, and that ϕ is a parameterization of C.
(d) Define ϕ on $[0, \pi/2)$ by $\phi(t) = e^{it}$. For what curve is ϕ a parametrization?
(e) Let $z_1, z_2, \dots, z_n$ be n distinct points in the plane, and suppose that the polygonal line joing these points in order never crosses itself. Construct a parameterization of that polygonal line.
(f) Let S be a piecewise smooth geometric set determined by the interval $[a,b]$ and the two piecewise smooth bounding functions u and l. Suppose z_1 and z_2 are two points in the interior S^0 of S. Show that there exists a piecewise smooth curve C joining z_1 to z_2, i.e., a piecewise smooth function $\phi : [\widehat{a}, \widehat{b}] \to C$ with $\phi(\widehat{a}) = z_1$ and $\phi(\widehat{b}) = z_2$, that lies entirely in S^0.
(g) Let C be a piecewise smooth curve, and suppose $\phi : [a,b] \to \mathbb{C}$ is a parameterization of C. Let $[c,d]$ be a subinterval of $[a,b]$. Show that the range of the restriction of ϕ to $[c,d]$ is a smooth curve.

Exercise 6.3. Suppose C is a smooth curve, parameterized by $\phi = u + iv : [a,b] \to C$.
(a) Suppose that $u'(t) \neq 0$ for all $t \in (a,b)$. Prove that there exists a smooth, real-valued function f on some closed interval $[a', b']$ such that C coincides with the graph of f.
HINT: f should be something like $v \circ u^{-1}$.
(b) What if $v'(t) \neq 0$ for all $t \in (a,b)$?

Exercise 6.4. Let C be the curve that is the range of the function $\phi : [-1,1] \to \mathbb{C}$, where $\phi(t) = t^3 + t^6 i$).
(a) Is C a piecewise smooth curve? Is it a smooth curve? What points z_1 and z_2 does it join?
(b) Is ϕ a parameterization of C?
(c) Find a parameterization for C by a function $\psi : [3,4] \to \mathbb{C}$.
(d) Find the unit tangent to C and the point $0 + 0i$.

Exercise 6.5. Let C be the curve parameterized by $\phi : [-\pi, \pi - \epsilon] \to C$ defined by $\phi(t) = e^{it} = \cos(t) + i\sin(t)$.
(a) What curve does ϕ parameterize?
(b) Find another parameterization of this curve, but base on the interval $[0, 1-\epsilon]$.

ARC LENGTH

Suppose C is a piecewise smooth curve, parameterized by a function ϕ. Continuing to think like a physicist, we might guess that the length of this curve could be computed as follows. The particle is moving with velocity $\phi'(t)$. This velocity is thought of as a vector in $\mathbb{R}^2$, and as such it has a direction and a magnitude or speed. The speed is just the absolute value $|\phi'(t)|$ of the velocity vector $\phi'(t)$. Now distance is speed multiplied by time, and so a good guess for the formula for the length L of the curve C would be

$$L = \int_a^b |\phi'(t)|\, dt. \tag{6.1}$$

Two questions immediately present themselves. First, and of primary interest, is whether the function $|\phi'|$ is improperly-integrable on (a,b)? We know by Exercise 5.22 that ϕ' itself is improperly-integrable, but we also know from Exercise 5.23 that a function can be improperly-integrable on an open interval and yet its absolute value is not. In fact, the answer to this first question is no (See Exercise 6.6.). We know only that $|\phi'|$ exists and is continuous on the open subintervals of a partition of $[a,b]$.

The second question is more subtle. What if we parameterize a curve in two different ways, i.e., with two different functions ϕ_1 and ϕ_2? How do we know that the two integral formulas for the length have to agree? Of course, maybe most important of all to us, we also must justify the physicist's intuition. That is, we must give a rigorous mathematical definition of the length of a smooth curve and show that Formula (6.1) above does in fact give the length of the curve. First we deal with the independence of parameterization question.

THEOREM 6.3. *Let C be a smooth curve joining (distinct) points z_1 to z_2 in $\mathbb{C}$, and let $\phi_1 : [a,b] \to C$ and $\phi_2 : [c,d] \to C$ be two parameterizations of C. Suppose $|\phi_2'|$ is improperly-integrable on (c,d). Then $|\phi_1'|$is improperly-integrable on (a,b), and*

$$\int_a^b \|\phi_1'(t)\|\,dt = \int_c^d \|\phi_2'(s)\|\,ds.$$

PROOF. We will use Theorem 6.2. Thus, let $g = \phi_1^{-1} \circ \phi_2$, and recall that g is continuous on $[c,d]$ and continuously differentiable on each open subinterval of a certain partition of $[c,d]$. Therefore, by part (d) of Exercise 5.22, g' is improperly-integrable on (c,d).

Let $\{x_0 < x_1 < \ldots < x_p\}$ be a partition of $[a,b]$ for which ϕ_1' is continuous and nonzero on the subintervals (x_{j-1},x_j). To show that $|\phi_1'|$ is improperly-integrable on (a,b), it will suffice to show this integrability on each subinterval (x_{j-1},x_j). Thus, fix a closed interval $[a',b'] \subset (x_{j-1},x_j)$, and let $[c',d']$ be the closed subinterval of $[c,d]$ such that g maps $[c',d']$ 1-1 and onto $[a',b']$. Hence, by part (e) of Exercise 5.22, we have

$$\begin{aligned}\int_{a'}^{b'} |\phi_1'(t)|\,dt &= \int_{c'}^{d'} |\phi_1'(g(s))|g'(s)\,ds\\ &= \int_{c'}^{d'} |\phi_1'(g(s))||g'()s)|\,ds\\ &= \int_{c'}^{d'} |\phi_1'(g(s))g'(s)|\,ds\\ &= \int_{c'}^{d'} |(\phi_1 \circ g)'(s)|\,ds\\ &= \int_{c'}^{d'} |\phi_2'(s)|\,ds\\ &\le \int_c^d |\phi_2'(s)|\,ds,\end{aligned}$$

which, by taking limits as a' goes to x_{j-1} and b' goes to x_j, shows that $|\phi_1'|$ is improperly-integrable over (x_{j-1},x_j) for every j, and hence integrable over all of

(a, b). Using part (e) of Exercise 5.22 again, and a calculation similar to the one above, we deduce the equality

$$\int_a^b |\phi_1'| = \int_c^d |\phi_2'|,$$

and the theorem is proved.

Exercise 6.6. (A curve of infinite length) Let $\phi : [0, 1] : \mathbb{R}^2$ be defined by $\phi(0) = (0, 0)$, and for $t > 0$, $\phi(t) = (t, t\sin(1/t))$. Let C be the smooth curve that is the range of ϕ.
(a) Graph this curve.
(b) Show that

$$\begin{aligned}|\phi'(t)| &= \sqrt{1 + \sin^2(1/t) - \frac{\sin(2/t)}{t} + \frac{\cos^2(1/t)}{t^2}} \\ &= \frac{1}{t}\sqrt{t^2 + t^2\sin^2(1/t) - t\sin(2/t) + \cos^2(1/t)}.\end{aligned}$$

(c) Show that

$$\int_\delta^1 |\phi'(t)|\,dt = \int_1^{1/\delta} \frac{1}{t}\sqrt{\frac{1}{t^2} + \frac{\sin^2(t)}{t^2} - \frac{\sin(2t)}{t} + \cos^2(t)}\,dt.$$

(d) Show that there exists an $\epsilon > 0$ so that for each positive integer n we have $\cos^2(t) - \sin(2t)/t > 1/2$ for all t such that $|t - n\pi| < \epsilon$.
(e) Conclude that $|\phi'|$ is not improperly-integrable on $(0, 1)$. Deduce that, if Formula (6.1) is correct for the length of a curve, then this curve has infinite length.

Next we develop a definition of the length of a parameterized curve from a purely mathematical or geometric point of view. Happily, it will turn out to coincide with the physically intuitive definition discussed above.
Let C be a piecewise smooth curve joining the points z_1 and z_2, and let $\phi : [a, b] \to C$ be a parameterization of C. Let $P = \{a = t_0 < t_1 < \ldots < t_n = b\}$ be a partition of the interval $[a, b]$. For each $0 \leq j \leq n$ write $z_j = \phi(t_j)$, and think about the polygonal trajectory joining these points $\{z_j\}$ in order. The length L_P^ϕ of this polygonal trajectory is given by the formula

$$L_P^\phi = \sum_{j=1}^n |z_j - z_{j-1}|,$$

and this length is evidently an approximation to the length of the curve C. Indeed, since the straight line joining two points is the shortest curve joining those points, these polygonal trajectories all should have a length smaller than or equal to the length of the curve. These remarks motivate the following definition.

DEFINITION. Let $\phi : [a, b] \to C$ be a parameterization of a piecewise smooth curve $C \subset \mathbb{C}$. By the *length* L^ϕ of C, relative to the parameterization ϕ, we mean the number $L^\phi = \sup_P L_P^\phi$, where the supremum is taken over all partitions P of $[a, b]$.

REMARK. Of course, the supremum in the definition above could well equal infinity in some cases. Though it is possible for a curve to have an infinite length, the ones we will study here will have finite lengths. This is another subtlety of this subject. After all, every smooth curve is a compact subset of $\mathbb{R}^2$, since it is the continuous image of a closed and bounded interval, and we think of compact sets as being "finite" in various ways. However, this finiteness does not necessarily extend to the length of a curve.

Exercise 6.7. Let $\phi : [a,b] \to \mathbb{R}^2$ be a parameterization of a piecewise smooth curve C, and let P and Q be two partitions of $[a,b]$.
(a) If P is finer than Q, i.e., $Q \subseteq P$, show that $L^\phi_Q \leq L^\phi_P$.
(b) If $\phi(t) = u(t) + iv(t)$, express L^ϕ_P in terms of the numbers $u(t_j)$ and $v(t_j)$.

Of course, we again face the annoying possibility that the definition of length of a curve will depend on the parameterization we are using. However, the next theorem, taken together with Theorem 6.3, will show that this is not the case.

THEOREM 6.4. *If C is a piecewise smooth curve parameterized by $\phi : [a,b] \to C$, then*

$$L^\phi = \int_a^b |\phi'(t)|\, dt,$$

specifically meaning that one of these quantities is infinite if and only if the other one is infinite.

PROOF. We prove this theorem for the case when C is a smooth curve, leaving the general argument for a piecewise smooth curve to the exercises. We also only treat here the case when L^ϕ is finite, also leaving the argument for the infinite case to the exercises. Hence, assume that $\phi = u + iv$ is a smooth function on $[a,b]$ and that $L^\phi < \infty$.
Let $\epsilon > 0$ be given. Choose a partition $P = \{t_0 < t_1 < \ldots < t_n\}$ of $[a,b]$ for which

$$L^\phi - L^\phi_P = L^\phi - \sum_{j=1}^n |\phi(t_j) - \phi(t_{j-1})| < \epsilon.$$

Because ϕ is continuous, we may assume by making a finer partition if necessary that the t_j's are such that $|\phi(t_1) - \phi(t_0)| < \epsilon$ and $|\phi(t_n) - \phi(t_{n-1})| < \epsilon$. This means that

$$L^\phi - \sum_{j=2}^{n-1} |\phi(t_j) - \phi(t_{j-1})| < 3\epsilon.$$

The point of this step (trick) is that we know that ϕ' is continuous on the open interval (a,b), but we will use that it is uniformly continuous on the compact set $[t_1, t_{n-1}]$. Of course that means that $|\phi'|$ is integrable on that closed interval, and in fact one of the things we need to prove is that $|\phi'|$ is improperly-integrable on the open interval (a,b).
Now, because ϕ' is uniformly continuous on the closed interval $[t_1, t_{n-1}]$, there exists a $\delta > 0$ such that $|\phi'(t) - \phi'(s)| < \epsilon$ if $|t - s| < \delta$ and t and s are in the interval $[t_1, t_{n-1}]$. We may assume, again by taking a finer partition if necessary, that the

mesh size of $P$is less than this δ. Then, using part (f) of Exercise 5.9, we may also assume that the partition P is such that

$$|\int_{t_1}^{t_{n-1}} |\phi'(t)|\,dt - \sum_{j=2}^{n-1} |\phi'(s_j)|(t_j - t_{j-1})| < \epsilon$$

no matter what points s_j in the interval (t_{j-1}, t_j) are chosen. So, we have the following calculation, in the middle of which we use the Mean Value Theorem on the two functions u and v.

$$\begin{aligned}
0 &\leq |L^\phi - \int_{t_1}^{t_{n-1}} |\phi'(t)|\,dt| \\
&\leq |L^\phi - \sum_{j=2}^{n-1} |\phi(t_j) - \phi(t_{j-1})| \\
&\qquad + |\sum_{j=2}^{n-1} |\phi(t_j) - \phi(t_{j-1})| - \int_{t_1}^{t_{n-1}} |\phi'(t)|\,dt| \\
&\leq 3\epsilon + |\sum_{j=2}^{n-1} |\phi(t_j) - \phi(t_{j-1})| - \int_{t_1}^{g_{n-1}} |\phi'(t)|\,dt| \\
&= 3\epsilon + |\sum_{j=2}^{n-1} |u(t_j) - u(t_{j-1}) + i(v(t_j) - v(t_{j-1})| - \int_{t_1}^{t_{n-1}} |\phi'(t)|\,dt| \\
&= 3\epsilon + |\sum_{j=2}^{n-1} \sqrt{(u(t_j) - u(t_{j-1}))^2 + (v(t_j) - v(t_{j-1}))^2} \\
&\qquad - \int_{t_1}^{g_{n-1}} |\phi'(t)|\,dt| \\
&= 3\epsilon + |\sum_{j=2}^{n-1} \sqrt{(u'(s_j))^2 + (v'(r_j))^2}(t_j - t_{j-1}) \\
&\qquad - \int_{t_1}^{t_{n-1}} |\phi'(t)|\,dt| \\
&\leq 3\epsilon + |\sum_{j=2}^{n-1} \sqrt{(u'(s_j))^2 + (v'(s_j))^2}(t_j - t_{j-1}) \\
&\qquad - \int_{t_1}^{t_{n-1}} |\phi'(t)|\,dt| \\
&\qquad + \sum_{j=2}^{n-1} |\sqrt{(u(s_j))^2 + (v'(r_j))^2} - \sqrt{(u(s_j))^2 + (v'(s_j))^2}|(t_j - t_{j-1}) \\
&= 3\epsilon + |\sum_{j=2}^{n-1} |\phi'(s_j)|(t_j - t_{j-1}) - \int_{t_1}^{t_{n-1}} |\phi'(t)|\,dt| \\
&\qquad + \sum_{j=2}^{n-1} |\sqrt{(u(s_j))^2 + (v'(r_j))^2} - \sqrt{(u(s_j))^2 + (v'(s_j))^2}|(t_j - t_{j-1}) \\
&\leq 4\epsilon + \sum_{j=2}^{n-1} \frac{|(v'(r_j))^2 - (v'(s_j))^2|}{\sqrt{(u'(s_j))^2 + (v'(r_j))^2} + \sqrt{(u'(s_j))^2 + (v'(s_j))^2}}(t_j - t_{j-1}) \\
&\leq 4\epsilon + \sum_{j=2}^{n-1} \frac{|v'(r_j) - v'(s_j)||v'(r_j) + v'(s_j)|}{|v'(r_j)| + |v'(s_j)|}(t_j - t_{j-1}) \\
&\leq 4\epsilon + \sum_{j=2}^{n-1} |v'(r_j) - v'(s_j)|(t_j - t_{j-1}) \\
&\leq 4\epsilon + \sum_{j=2}^{n-1} |\phi'(r_j) - \phi'(s_j)|(t_j - t_{j-1}) \\
&\leq 4\epsilon + \sum_{j=2}^{n-1} \epsilon(t_j - t_{j-1}) \\
&= 4\epsilon + \epsilon(t_{n-1} - t_1) \\
&< \epsilon(4 + b - a).
\end{aligned}$$

This implies that

$$L^{\phi} - \epsilon(4+b-a) \leq \int_{t_1}^{t_{n-1}} |\phi'| \leq L^{\phi} + \epsilon(4+b-a).$$

If we now let t_1 approach a and t_{n-1} approach b, we get

$$L^{\phi} - \epsilon(4+b-a) \leq \int_a^b |\phi'| \leq L^{\phi} + \epsilon(4+b-a),$$

which completes the proof, since ϵ is arbitrary.

Exercise 6.8. (a) Take care of the piecewise case in the preceding theorem.
(b) Take care of the case when L^{ϕ} is infinite in the preceding theorem.

We now have all the ingredients necessary to define the length of a smooth curve.

DEFINITION. Let C be a piecewise smooth curve in the plane. The *length* or *arc length* $L \equiv L(C)$ of C is defined by the formula

$$L(C) = L^{\phi} = \sup_P L_P^{\phi},$$

where ϕ is any parameterization of C.
If z and w are two points on a piecewise smooth curve C, we will denote by $L(z,w)$ the arc length of the portion of the curve between z and w.

REMARK. According to Theorems 6.3 and 6.4, we have the following formula for the length of a piecewise smooth curve:

$$L = \int_a^b |\phi'(t)|\, dt,$$

where ϕ is any parameterization of C.
It should come as no surprise that the length of a curve C from z_1 to z_2 is the same as the length of that same curve C, but thought of as joining z_2 to z_1. Nevertheless, let us make the calculation to verify this. If $\phi : [a,b] \to C$ is a parameterization of this curve from z_1 to z_2, then we have seen in part (f) of exercise 6.1 that $\psi : [a,b] \to C$, defined by $\psi(t) = \phi(a+b-t)$, is a parameterization of C from z_2 to z_1. We just need to check that the two integrals giving the lengths are equal. Thus,

$$\int_a^b |\psi'(t)|\, dt = \int_a^b |\phi'(a+b-t)(-1)|\, dt = \int_a^b |\phi'(a+b-t)|\, dt = \int_a^b |\phi'(s)|\, ds,$$

where the last equality follows by changing variables, i.e., setting $t = a+b-s$.
We can now derive the formula for the circumference of a circle, which was one of our main goals. TRUMPETS?

THEOREM 6.5. *Let C be a circle of radius r in the plane. Then the length of C is $2\pi r$.*

PROOF. Let the center of the circle be denoted by (h,k). We can parameterize the top half of the circle by the function ϕ on the interval $[0,\pi]$ by $\phi(t) = h + r\cos(t) + i(k + r\sin(t))$. So, the length of this half circle is given by

$$L = \int_0^{\pi} |\phi'(t)|\, dt = \int_0^{\pi} |-r\sin(t) + ir\cos(t)|\, dt = \int_0^{\pi} r\, dt = \pi r.$$

The same kind of calculation would show that the lower half of the circle has length πr, and hence the total length is $2\pi r$.

The integral formula for the length of a curve is frequently not much help, especially if you really want to know how long a curve is. The integrals that show up are frequently not easy to work out.

Exercise 6.9. (a) Let C be the portion of the graph of the function $y = x^2$ between $x = 0$ and $x = 1$. Let $\phi : [0,1] \to C$ be the parameterization of this curve given by $\phi(t) = t + t^2 i$. Find the length of this curve.
(b) Define $\phi : [-0,\pi] \to \mathbb{C}$ by $\phi(t) = a\cos(t) + ib\sin(t)$. What curve does ϕ parameterize, and can you find its length?

INTEGRATION WITH RESPECT TO ARC LENGTH

We introduce next what would appear to be the best parameterization of a piecewise smooth curve, i.e., a parameterization by arc length. We will then use this parameterization to define the integral of a function whose domain is the curve.

THEOREM 6.6. *Let C be a piecewise smooth curve of finite length L joining two distinct points z_1 to z_2. Then there exists a parameterization $\gamma : [0,L] \to C$ for which the arc length of the curve joining $\gamma(t)$ to $\gamma(u)$ is equal to $|u - t|$ for all $t < u \in [0,L]$.*

PROOF. Let $\phi : [a,b] \to C$ be a parameterization of C. Define a function $F : [a,b] \to [0,L]$ by

$$F(t) = \int_a^t |\phi'(s)|\, ds.$$

In other words, $F(t)$ is the length of the portion of C that joins the points $z_1 = \phi(a)$ and $\phi(t)$. By the Fundamental Theorem of Calculus, we know that the function F is continuous on the entire interval $[a,b]$ and is continuously differentiable on every subinterval (t_{i-1}, t_i) of the partition P determined by the piecewise smooth parameterization ϕ. Moreover, $F'(t) = |\phi'(t)| > 0$ for all $t \in (t_{i-1}, t_i)$, implying that F is strictly increasing on these subintervals. Therefore, if we write $s_i = F(t_i)$, then the s_i's form a partition of the interval $[0,L]$, and the function $F : (t_{i-1}, t_i) \to (s_{i-1}, s_i)$ is invertible, and its inverse F^{-1} is continuously differentiable. It follows then that $\gamma = \phi \circ F^{-1} : [0,L] \to C$ is a parameterization of C. The arc length between the points $\gamma(t)$ and $\gamma(u)$ is the arc length between $\phi(F^{-1}(t))$ and $\phi(F^{-1}(u))$, and this is given by the formula

$$\begin{aligned}\int_{F^{-1}(t)}^{F^{-1}(u)} |\phi'(s)|\, ds &= \int_a^{F^{-1}(u)} |\phi'(s)|\, ds - \int_a^{F^{-1}(t)} |\phi'(s)|\, ds \\ &= F(F^{-1}(u)) - F(F^{-1}(t)) \\ &= u - t,\end{aligned}$$

which completes the proof.

COROLLARY. *If γ is the parameterization by arc length of the preceding theorem, then, for all $t \in (s_{i-1}, s_i)$, we have $|\gamma'(s)| = 1$.*

PROOF OF THE COROLLARY. We just compute

$$\begin{aligned}|\gamma'(s)| &= |(\phi\circ F^{-1})'(s)|\\ &= |\phi'(F^{-1}(s))(F^{-1})'(s)|\\ &= |\phi'(F^{-1}(s)||\frac{1}{F'(F^{-1}(s))}|\\ &= |\phi'(f^{-1}(s))|\frac{1}{|\phi'(f^{-1}(s))|}\\ &= 1,\end{aligned}$$

as desired.

We are now ready to make the first of our three definitions of integral over a curve. This first one is pretty easy.
Suppose C is a piecewise smooth curve joining z_1 to z_2 of finite length L, parameterized by arc length. Recall that this means that there is a 1-1 function γ from the interval $[0,L]$ onto C that satisfies the condidition that the arc length betweenthe two points $\gamma(t)$ and $\gamma(s)$ is exactly the distance between the points t and s. We can just identify the curve C with the interval $[0,L]$, and relative distances will correspond perfectly. A partition of the curve C will correspond naturally to a partition of the interval $[0,L]$. A step function on the dcurve will correspond in an obvious way to a step function on the interval $[0,L]$, and the formula for the integral of a step function on the curve is analogous to what it is on the interval. Here are the formal definitions:

DEFINITION. Let C be a piecewise smooth curve of finite length L joining distinct points, and let $\gamma:[0,L]\to C$ be a parameterization of C by arc length. By a *partition* of C we mean a set $\{z_0,z_1,\ldots,z_n\}$ of points on C such that $z_j=\gamma(t_j)$ for all j, where the points $\{t_0<t_1<\ldots<t_n\}$ form a partition of the interval $[0,L]$. The portions of the curve between the points z_{j-1} and z_j, i.e., the set $\gamma(t_{j-1},t_j)$, are called the *elements* of the partition.
A *step fucntion* on C is a real-valued function h on C for which there exists a partition $\{z_0,z_1,\ldots,z_n\}$ of C such that $h(z)$ is a constant a_j on the portion of the curve between z_{j-1} and z_j.

Before defining the integral of a step function on a curve, we need to establish the usual consistency result, encountered in the previous cases of integration on intervals and integration over geometric sets, the proof of which this time we put in an exercise.

Exercise 6.10. Suppose h is a function on a piecewise smooth curve of finite length L, and assume that there exist two partitions $\{z_0,z_1,\ldots,z_n\}$ and $\{w_0,w_1,\ldots,w_m\}$ of C such that $h(z)$ is a constant a_k on the portion of the curve between z_{k-1} and z_k, and $h(z)$ is a constant b_j on the portion of the curve between w_{j-1} and w_j. Show that

$$\sum_{k=1}^n a_kL(z_{k-1},z_k)=\sum_{j=1}^m b_jL(w_{j-1},w_j).$$

HINT: Make use of the fact that $h\circ\gamma$ is a step function on the interval $[0,L]$.

Now we can make the definition of the integral of a step function on a curve.

DEFINITION. Let h be a step function on a piecewise smooth curve C of finite length L. The *integral, with respect to arc length* of h over C is denoted by $\int_C h(s)\,ds$, and is defined by

$$\int_C h(s)\,ds = \sum_{j=1}^{n} a_j L(z_{j-1}, z_j),$$

where $\{z_0, z_1, \dots, z_n\}$ is a partition of C for which $h(z)$ is the constant a_j on the portion of C between z_{j-1} and z_j.

Of course, integrable functions on C with respect to arc length will be defined to be functions that are uniform limits of step functions. Again, there is the consistency issue in the definition of the integral of an integrable function.

Exercise 6.11. (a) Suppose $\{h_n\}$ is a sequence of step functiohns on a piecewise smooth curve C of finite length, and assume that the sequence $\{h_n\}$ converges uniformly to a function f. Prove that the sequence $\{\int_C h_n(s)\,ds\}$ is a convergent sequence of real numbers.
(b) Suppose $\{h_n\}$ and $\{k_n\}$ are two sequences of step functions on a piecewise smooth curve C of finite length l, and that both sequences converge uniformly to the same function f. Prove that

$$\lim \int_C h_n(s)\,ds = \lim \int_C k_n(s)\,ds.$$

DEFINITION. Let C be a piecewise smooth curve of finite length L. A function f with domain C is called *integrable with respect to arc length* on C if it is the uniform limit of step functions on C.
The *integral with respect to arc length* of an integrable function f on C is again denoted by $\int_C f(s)\,ds$, and is defined by

$$\int_C f(s)\,ds = \lim \int_C h_n(s)\,ds,$$

where $\{h_n\}$ is a sequence of step functions that converges uniformly to f on C.

In a sense, we are simply identifying the curve C with the interval $[0, L]$ by means of the 1-1 parameterizing function γ. The next theorem makes this quite plain.

THEOREM 6.7. *Let C be a piecewise smooth curve of finite length L, and let γ be a parameterization of C by arc length. If f is an integrable function on C, then*

$$\int_C f(s)\,ds = \int_0^L f(\gamma(t))\,dt.$$

PROOF. First, if h is a step function on C, let $\{z_j\}$ be a partition of C for which $h(z)$ is a constant a_j on the portion of the curve between z_{j-1} and z_j. Let $\{t_j\}$ be the partition of $[0, L]$ for which $z_j = \gamma(t_j)$ for every j. Note that $h \circ \gamma$ is a step function on $[0, L]$, and that $h \circ \gamma(t) = a_j$ for all $t \in (t_{j-1}, t_j)$. Then,

$$\begin{aligned}\int_C h(s)\,ds &= \sum_{j=1}^{N} a_j L(z_{j-1}, z_j)\\ &= \sum_{j=1}^{n} a_j L(\gamma(t_{j-1}), \gamma(t_j))\\ &= \sum_{j=1}^{n} a_j (t_j - t_{j-1})\\ &= \int_0^L h \circ \gamma(t)\,dt,\end{aligned}$$

which proves the theorem for step functions.

Finally, if $f = \lim h_n$ is an integrable function on C, then the sequence $\{h_n \circ \gamma\}$ converges uniformly to $f \circ \gamma$ on $[0, L]$, and so

$$\begin{aligned}\int_C f(s)\,ds &= \lim \int_C h_n(s)\,ds\\ &= \lim \int_0^L h_n(\gamma(t))\,dt\\ &= \int_0^l f(\gamma(t))\,dt,\end{aligned}$$

where the final equality follows from Theorem 5.6. Hence, Theorem 6.7 is proved.

Although the basic definitions of integrable and integral, with respect to arc length, are made in terms of the particular parameterization γ of the curve, for computational purposes we need to know how to evaluate these integrals using different parameterizations. Here is the result:

THEOREM 6.8. *Let C be a piecewise smooth curve of finite length L, and let $\phi : [a, b] \to C$ be a parameterization of C. If f is an integrable function on C. Then*

$$\int_C f(s)\,ds = \int_a^b f(\phi(t))|\phi'(t)|\,dt.$$

PROOF. Write $\gamma : [0, L] \to C$ for a parameterization of C by arc length. As in the proof to Theorem 6.3, we write $g : [a, b] \to [0, L]$ for $\gamma^{-1} \circ \phi$. Just as in that proof, we know that g is a piecewise smooth function on the interval $[a, b]$. Hence, recalling that $|\gamma'(t)| = 1$ and $g'(t) > 0$ for all but a finite number of points, the following calculation is justified:

$$\begin{aligned}\int_C f(s)\,ds &= \int_0^L f(\gamma(t))\,dt\\ &= \int_0^L f(\gamma(t))|\gamma'(t)|\,dt\\ &= \int_a^b f(\gamma(g(u)))|\gamma'(g(u))|g'(u)\,du\\ &= \int_a^b f(\gamma(g(u)))|\gamma'(g(u))||g'(u)|\,du\\ &= \int_a^b f(\phi(u))|\gamma'(g(u))g'(u)|\,du\\ &= \int_a^b f(\phi(u))|('gamma\circ g)'(u)|\,du\\ &= \int_a^b f(\phi(u))|\phi'(u)|\,du,\end{aligned}$$

as desired.

Exercise 6.12. Let C be the straight line joining the points $(0,1)$ and $(1,2)$.
(a) Find the arc length parameterization $\gamma : [0,\sqrt{2}] \to C$.
(b) Let f be the function on this curve given by $f(x,y) = x^2y$. Compute $\int_C f(s)\,ds$.
(c) Let f be the function on this curve that is defined by $f(x,y)$ is the distance from (x,y) to the point $(0,3)$. Compute $\int_c f(s)\,ds$.

The final theorem of this section sums up the properties of integrals with respect to arc length. There are no surprises here.

THEOREM 6.9. *Let C be a piecewise smooth curve of finite length L, and write $I(C)$ for the set of all functions that are integrable with respect to arc length on C. Then:*

(1) *$I(C)$ is a vector space ovr the real numbers, and*

$$\int_C (af(s)+bg(s))\,ds = a\int_C f(s)\,ds + b\int_C g(s)\,ds$$

for all $f,g \in I(C)$ and all $a,b \in \mathbb{R}$.

(2) (Positivity) If $f(z) \geq 0$ for all $z \in C$, then $\int_C f(s)\,ds \geq 0$.

(3) If $f \in I(C)$, then so is $|f|$, and $|\int_C f(s)\,ds| \leq \int_C |f(s)|\,ds$.

(4) If f is the uniform limit of functions f_n, each of which is in $I(C)$, then $f \in I(C)$ and $\int_C f(s)\,ds = \lim \int_C f_n(s)\,ds$.

(5) Let $\{u_n\}$ be a sequence of functions in $I(C)$, and suppose that for each n there is a number m_n, for which $|u_n(z)| \leq m_n$ for all $z \in C$, and such that the infinite series $\sum m_n$ converges. Then the infinite series $\sum u_n$ converges uniformly to an integrable function, and $\int_C \sum u_n(s)\,ds = \sum \int_C u_n(s)\,ds$.

Exercise 6.13. (a) Prove the preceding theorem. Everything is easy if we compose all functions on C with the parameterization γ, obtaining functions on $[0,L]$, and then use Theorem 5.6.

(b) Suppose C is a piecewise smooth curve of finite length joining z_1 and z_2. Show that the integral with respect to arc length of a function f over C is the same whether we think of C as being a curve from z_1 to z_2 or, the other way around, a curve from z_2 to z_1.

REMARK. Because of the result in part (b) of the preceding exercise, we speak of "integrating over C" when we are integrating with respect to arc length. We do not speak of "integrating from z_1 to z_2," since the direction doesn't matter. This is in marked contrast to the next two kinds of integrals over curves that we will discuss.

here is one final bit of notation. Often, the curves of interest to us are graphs of real-valued functions. If $g : [a,b] \to \mathbb{R}$ is a piecewise smooth function, then its graph C is a piecewise smooth curve, and we write $\int_{\mathrm{graph}(g)} f(s)\,ds$ for the integral with respect to arc length of f over $C = \mathrm{graph}(g)$.

CONTOUR INTEGRALS

We discuss next what appears to be a simpler notion of integral over a curve. In this one, we really do regard the curve C as a subset of the complex plane as opposed to two-dimensional real space; we will be integrating complex-valued functions; and we explicitly think of the parameterizations of the curve as complex-valued functions on an interval $[a,b]$. Also, in this definition, a curve C from z_1 to z_2 will be distinguished from its reverse, i.e., the same set C thought of as a curve from z_2 to z_1.

DEFINITION. Let C be a piecewise smooth curve from z_1 to z_2 in the plane $\mathbb{C}$, parameterized by a (complex-valued) function $\phi : [a,b] \to C$. If f is a continuous, complex-valued function on C, The *contour integral of f from z_1 to z_2 along C* will be denoted by $\int_C f(\zeta)\,d\zeta$ or more precisely by $\int_{C\,z_1}^{z_2} f(\zeta)\,d\zeta$, and is defindd by

$$\int_{C\,z_1}^{z_2} f(\zeta)\,d\zeta = \int_a^b f(\phi(t))\phi'(t)\,dt.$$

REMARK. There is, as usual, the question about whether this definition depends on the parameterization. Again, it does not. See the next exercise.

The definition of a contour integral looks very like a change of variables formula for integrals. See Theorem 5.11 and part (e) of Exercise 5.22. This is an example of how mathematicians often use a true formula from one context to make a new definition in another context.

Notice that the only difference between the computation of a contour integral and an integral with respect to arc length on the curve is the absence of the absolute value bars around the factor $\phi'(t)$. This will make contour integrals more subtle than integrals with respect to arc length, just as conditionally convergent infinite series are more subtle than absolutely convergent ones.

Note also that there is no question about the integrability of $f(\phi(t))\phi'(t)$, because of Exercise 5.22. f is bounded, ϕ' is improperly-integrable on (a,b), and therefore so is their product.

Exercise 6.14. (a) State and prove the "independence of parameterization" result for contour integrals.

(b) Prove that

$$\int_{C\,z_1}^{z_2} f(\zeta)\,d\zeta = -\int_{C\,z_2}^{z_1} f(\zeta)\,d\zeta.$$

Just remember how to parameterize the curve in the opposite direction.
(c) Establish the following relation between the absolute value of a contour integral and a corresponding integral with respect to arc length.

$$|\int_C f(\zeta)\,d\zeta| \leq \int_C |f(s)|\,ds.$$

Not all the usual properties hold for contour integrals, e.g., like those in Theorem 6.9 above. The functions here, and the values of their contour integrals, are complex numbers, so all the properties of integrals having to do with positivity and inequalities, except for the one in part (c) of Exercise 6.14, no longer make any sense. However, we do have the following results for contour integrals, the verification of which is just as it was for Theorem 6.9.

THEOREM 6.10. *Let C be a piecewise smooth curve of finite length joining z_1 to z_2. Then the contour integrals of continuous functions on C have the following properties.*

(1) *If f and g are any two continuous functions on C, and a and b are any two complex numbers, then*

$$\int_C (af(\zeta) + bg(\zeta))\,d\zeta = a\int_C f(\zeta)\,d\zeta + b\int_C g(\zeta)\,d\zeta.$$

(2) *If f is the uniform limit on C of a sequence $\{f_n\}$ of continuous functions, then $\int_C f(\zeta)\,d\zeta = \lim \int_C f_n(\zeta)\,d\zeta$.*
(3) *Let $\{u_n\}$ be a sequence of continuous functions on C, and suppose that for each n there is a number m_n, for which $|u_n(z)| \leq m_n$ for all $z \in C$, and such that the infinite series $\sum m_n$ converges. Then the infinite series $\sum u_n$ converges uniformly to a continuous function, and $\int_C \sum u_n(\zeta)\,d\zeta = \sum \int_C u_n(\zeta)\,d\zeta$.*

In the next exercise, we give some important contour integrals, which will be referred to several times in the sequel. Make sure you understand them.

Exercise 6.15. Let c be a point in the complex plane, and let r be a positive number. Let C be the curve parameterized by $\phi : [-\pi, \pi - \epsilon] : C$ defined by $\phi(t) = c + re^{it} = c + r\cos(t) + ir\sin(t)$. For each integer $n \in \mathbb{Z}$, define $f_n(z) = (z-c)^n$.
(a) What two points z_1 and z_2 does C join, and what happens to z_2 as ϵ approaches 0?
(b) Compute $\int_C f_n(\zeta)\,d\zeta$ for all integers n, positive and negative.
(c) What happens to the integrals computed in part (b) when ϵ approaches 0?
(d) Set $\epsilon = \pi$, and compute $\int_C f_n(\zeta)\,d\zeta$ for all integers n.
(e) Again, set $\epsilon = \pi$. Evaluate

$$\int_C \frac{\cos(\zeta - c)}{\zeta - c}\,d\zeta \text{ and } \int_C \frac{\sin(\zeta - c)}{\zeta - c}\,d\zeta.$$

HINT: Make use of the infinite series representations of the trigonometric functions.

VECTOR FIELDS, DIFFERENTIAL FORMS, AND LINE INTEGRALS

We motivate our third definition of an integral over a curve by returning to physics. This definition is very much a real variable one, so that we think of the plane as $\mathbb{R}^2$ instead of $\mathbb{C}$. A connection between this real variable definition and the complex variable definition of a contour integral will emerge later.

DEFINITION. By a *vector field* on an open subset U of $\mathbb{R}^2$, we mean nothing more than a continuous function $\vec{V}(x,y) \equiv (P(x,y), Q(x,y))$ from U into $\mathbb{R}^2$. The functions P and Q are called the *components* of the vector field $\vec{V}$.
We will also speak of *smooth* vector fields, by which we will mean vector fields $\vec{V}$ both of whose component functions P and Q have continuous partial derivatives

$$\frac{\partial P}{\partial x}, \frac{\partial P}{\partial y}, \frac{\partial Q}{\partial x} textand \frac{\partial Q}{\partial y}$$

on U.

REMARK. The idea from physics is to think of a vector field as a force field, i.e., something that exerts a force at the point (x,y) with magnitude $|\vec{V}(x,y)|$ and acting in the direction of the vector $\vec{V}(x,y)$. For a particle to move within a force field, "work" must be done, that is energy must be provided to move the particle against the force, or energy is given to the particle as it moves under the influence of the force field. In either case, the basic definition of work is the product of force and distance traveled. More precisely, if a particle is moving in a direction $\vec{u}$ within a force field, then the work done on the particle is the product of the component of the force field in the direction of $\vec{u}$ and the distance traveled by the particle in that direction. That is, we must compute dot products of the vectors $\vec{V}(x,y)$ and $\vec{u}(x,y)$. Therefore, if a particle is moving along a curve C, parameterized with respect to arc length by $\gamma : [0,L] \to C$, and we write $\gamma(t) = (x(t), y(t))$, then the work $W(z_1, z_2)$ done on the particle as it moves from $z_1 = \gamma(0)$ to $z_2 = \gamma(L)$ within the force field $\vec{V}$, should intuitively be given by the formula

$$\begin{aligned} W(z_1,z_2) &= \int_0^L \langle \vec{V}(\gamma(t)) \mid \gamma'(t) \rangle \, dt \\ &= \int_0^L P(x(t),y(t))x'(t) + Q(x(t),y(t))y'(t) \, dt \\ &\equiv \int_C P\,dx + Q\,dy, \end{aligned}$$

where the last expression is explicitly defining the shorthand notation we will be using.
The preceding discussion leads us to a new notion of what kind of object should be "integrated" over a curve.

DEFINITION. A *differential form* on a subset U of $\mathbb{R}^2$ is denoted by $\omega = Pdx + Qdy$, and is determined by two continuous real-valued functions P and Q on U. We say that ω is *bounded* or *uniformly continuous* if the functions P and Q are bounded or uniformly continuous functions on U. We say that the differential form ω is *smooth of order k* if the set U is open, and the functions P and Q have continuous mixed partial derivatives of order k.

If $\omega = Pdx + Qdy$ is a differential form on a set U, and if C is any piecewise smooth curve of finite length contained in U, then we define the *line integral* $\int_C \omega$ of ω over C by

$$\int_C \omega = \int_C P\,dx + Q\,dy = \int_0^L P(\gamma(t))x'(t) + Q(\gamma(t))y'(t)\,dt,$$

where $\gamma(t) = (x(t), y(t))$ is a parameterization of C by arc length.

REMARK. There is no doubt that the integral in this definition exists, because P and Q are continuous functions on the compact set C, hence bounded, and γ' is integrable, implying that both x' and y' are integrable. Therefore $P(\gamma(t))x'(t) + Q(\gamma(t))y'(t)$ is integrable on $(0, L)$.
These differential forms ω really should be called "differential 1-forms." For instance, an example of a differential 2-form would look like $R\,dxdy$, and in higher dimensions, we could introduce notions of differential forms of higher and higher orders, e.g., in 3 dimension things like $P\,dxdy + Q\,dzdy + R\,dxdz$. Because we will always be dealing with $\mathbb{R}^2$, we will have no need for higher order differential forms, but the study of such things is wonderful. Take a course in Differential Geometry!
Again, we must see how this quantity $\int_C \omega$ depends, if it does, on different parameterizations. As usual, it does not.

Exercise 6.16. Suppose $\omega = Pdx + Qdy$ is a differential form on a subset U of $\mathbb{R}^2$.
(a) Let C be a piecewise smooth curve of finite length contained in U that joins z_1 to z_2. Prove that

$$\int_C \omega = \int_C P\,dx + Q\,dy = \int_a^b P(\phi(t))x'(t) + Q(\phi(t))y'(t)\,dt$$

for any parameterization $\phi : [a, b] \to C$ having components $x(t)$ and $y(t)$.
(b) Let C be as in part (a), and let $\widehat{C}$ denote the reverse of C, i.e., the same set C but thought of as a curve joining z_2 to z_1. Show that $\int_{\widehat{C}} \omega = -\int_C \omega$.
(c) Let C be as in part (a). Prove that

$$|\int_C P\,dx + Q\,dy| \leq (M_P + M_Q)L,$$

where M_P and M_Q are bounds for the continuous functions $|P|$ and $|Q|$ on the compact set C, and where L is the length of C.

EXAMPLE. The simplest interesting example of a differential form is constructed as follows. Suppose U is an open subset of $\mathbb{R}^2$, and let $f : U \to \mathbb{R}$ be a differentiable real-valued function of two real variables; i.e., both of its partial derivatives exist at every point $(x, y) \in U$. (See the last section of Chapter IV.) Define a differential form $\omega = df$, called the *differential* of f, by

$$df = \frac{\partial f}{\partial x}\,dx + \frac{\partial f}{\partial y}\,dy,$$

i.e., $P = \partial f/\partial x$ and $Q = \partial f/\partial y$. These differential forms df are called *exact differential forms.*

REMARK. Not every differential form ω is exact, i.e., of the form df. Indeed, determining which ω's are df's boils down to what may be the simplest possible partial differential equation problem. If ω is given by two functions P and Q, then saying that $\omega = df$ amounts to saying that f is a solution of the pair of simultaneous partial differential equations

$$\frac{\partial f}{\partial x} = P \text{ and } \frac{\partial f}{\partial y} = Q.$$

See part (b) of the exercise below for an example of a nonexact differential form. Of course if a real-valued function f has continuous partial derivatives of the second order, then Theorem 4.22 tells us that the mixed partials f_{xy} and f_{yx} must be equal. So, if $\omega = Pdx + Qdy = df$ for some such f, Then P and Q would have to satisfy $\partial P/\partial y = \partial Q/\partial x$. Certainly not every P and Q would satisfy this equation, so it is in fact trivial to find examples of differential forms that are not differentials of functions. A good bit more subtle is the question of whether every differential form $Pdx + Qdy$, for which $\partial P/\partial y = \partial Q/\partial x$, is equal to some df. Even this is not true in general, as part (c) of the exercise below shows. The open subset U on which the differential form is defined plays a significant role, and, in fact, differential forms provide a way of studying topologically different kinds of open sets.

In fact, although it may seem as if a differential form is really nothing more than a pair of functions, the concept of a differential form is in part a way of organizing our thoughts about partial differential equation problems into an abstract mathematical context. This abstraction is a good bit more enlightening in higher dimensional spaces, i.e., in connection with functions of more than two variables. Take a course in Multivariable Analysis!

Exercise 6.17. (a) Solve the pair of simultaneous partial differential equations

$$\frac{\partial f}{\partial x} = x + y \text{ and } \frac{\partial f}{\partial y} = x - y.$$

(b) Show that it is impossible to solve the pair of simultaneous partial differential equations

$$\frac{\partial f}{\partial x} = x + y \text{ and } \frac{\partial f}{\partial y} = y^3.$$

Hence, conclude that the differential form $\omega = (x+y)dx + y^3dy$ is not the differential df of any real-valued function f.

(c) Let U be the open subset of $\mathbb{R}^2$ that is the complement of the single point $(0,0)$. Let $P(x,y) = -y/(x^2+y^2)$ and $Q(x,y) = x/(x^2+y^2)$. Show that $\partial P/\partial y = \partial Q/\partial x$ at every point of U, but that $\omega = Pdx + Qdy$ is not the differential df of any smooth function f on U.

HINT: If P were f_x, then f would have to be of the form $f(x,y) = -\tan^{-1}(x/y) + g(y)$, where g is some differentiable function of y. Show that if $Q = f_y$ then $g(y)$ is a constant c. Hence, $f(x,y)$ must be $-\tan^{-1}(x/y) + c$. But this function f is not continuous, let alone differentiable, at the point $(1,0)$. Consider $\lim f(1,1/n)$ and $\lim f(1,-1/n)$.

The next thing we wish to investigate is the continuity of $\int_C \omega$ as a function of the curve C. This brings out a significant difference in the concepts of line integrals versis integrals with respect to arc length. For the latter, we typically think of a fixed curve and varying functions, whereas with line integrals, we typically think of a fixed differential form and variable curves. This is not universally true, but should be kept in mind.

THEOREM 6.11. *Let $\omega = Pdx + Qdy$ be a fixed, bounded, uniformly continuous differential form on a set U in $\mathbb{R}^2$, and let C be a fixed piecewise smooth curve of finite length L, parameterized by $\phi : [a,b] \to C$, that is contained in U. Then, given an $\epsilon > 0$ there exists a $\delta > 0$ such that, for any curve $\widehat{C}$ contained in U, $|\int_C \omega - \int_{\widehat{C}} \omega| < \epsilon$ whenever the following conditions on the curve $\widehat{C}$ hold:*

1. *$\widehat{C}$ is a piecewise smooth curve of finite length $\widehat{L}$ contained in U, parameterized by $\widehat{\phi} : [a,b] \to \widehat{C}$.*
2. *$|\phi(t) - \widehat{\phi}(t)| < \delta$ for all $t \in [a,b]$.*
3. *$\int_a^b |\phi'(t) - \widehat{\phi}'(t)|\, dt < \delta$.*

PROOF. Let $\epsilon > 0$ be given. Because both P and Q are bounded on U, let M_P and M_Q be upper bounds for the functions $|P|$ and $|Q|$ respectively. Also, since both P and Q are uniformly continuous on U, there exists a $\delta > 0$ such that if $|(c,d) - (c',d')| < \delta$, then $|P(c,d) - P(c',d')| < \epsilon/4L$ and $|Q(c,d) - Q(c',d')| < \epsilon/4L$. We may also choose this δ to be less than both $\epsilon/4M_P$ and $\epsilon/4M_Q$. Now, suppose $\widehat{C}$ is a curve of finite length $\widehat{L}$, parameterized by $\widehat{\phi} : [a,b] \to \widehat{C}$, and that $|\phi(t) - \widehat{\phi}(t)| < \delta$ for all $t \in [a,b]$, and that $\int_a^b |\phi'(t) - \widehat{\phi}'(t)| < \delta$. Writing $\phi(t) = (x(t), y(t))$ and $\widehat{\phi}(t) = (\widehat{x}(t), \widehat{y}(t))$, we have

$$\begin{aligned}
0 &\leq |\int_C P\,dx + Q\,dy - \int_{\widehat{C}} P\,dx + Q\,dy| \\
&= |\int_a^b P(\phi(t))x'(t) - P(\widehat{\phi}(t))\widehat{x}'(t) + Q(\phi(t))y'(t) - Q(\widehat{\phi}(t))\widehat{y}'(t)\,dt| \\
&\leq \int_a^b |P(\phi(t))x'(t) - P(\widehat{\phi}(t))\widehat{x}'(t)|\,dt + \int_a^b |Q(\phi(t))y'(t) - Q(\widehat{\phi}(t))\widehat{y}'(t)|\,dt \\
&\leq \int_a^b |P(\phi(t)) - P(\widehat{\phi}(t))||x'(t)|\,dt + \int_a^b |P(\widehat{\phi}(t))||x'(t) - \widehat{x}'(t)|\,dt \\
&\qquad + \int_a^b |Q(\phi(t)) - Q(\widehat{\phi}(t))||y'(t)|\,dt + \int_a^b |Q(\widehat{\phi}(t))||y'(t) - \widehat{y}'(t)|\,dt \\
&\leq \frac{\epsilon}{4L}\int_a^b |x'(t)|\,dt + M_P\int_a^b |x'(t) - \widehat{x}'(t)|\,dt \\
&\qquad + \frac{\epsilon}{4L}\int_a^b |y'(t)|\,dt + M_Q\int_a^b |y'(t) - \widehat{y}'(t)|\,dt \\
&\leq \frac{\epsilon}{4L}\int_a^b |\phi'(t)|\,dt + M_P\int_a^b |\phi'(t) - \widehat{\phi}'(t)|\,dt \\
&\qquad + \frac{\epsilon}{4L}\int_a^b |\phi'(t)|\,dt + M_Q\int_a^b |\phi'(t) - \widehat{\phi}'(t)|\,dt \\
&< \frac{\epsilon}{4} + \frac{\epsilon}{4} + M_P\delta + M_Q\delta \\
&< \epsilon,
\end{aligned}$$

as desired.

Again, we have a special notation when the curve C is a graph. If $g : [a,b] \to \mathbb{R}$ is a piecewise smooth function, then its graph C is a piecewise smooth curve, and we write $\int_{\text{graph}(g)} P\,dx + Q\,dy$ for the line integral of the differential form $Pdx + Qdy$ over the curve $C = \text{graph}(g)$.
As alluded to earlier, there is a connection between contour integrals and line integrals. It is that a single contour integral can often be expressed in terms of two line integrals. Here is the precise statement.

THEOREM 6.12. *Suppose C is a piecewise curve of finite length, and that $f = u + iv$ is a complex-valued, continuous function on C. Let $\phi : [a,b] \to C$ be a parameterization of C, and write $\phi(t) = x(t) + iy(t)$. Then*

$$\int_C f(\zeta)\,d\zeta = \int_C (U\,dx - v\,dy) + \int_C (v\,dx + u\,dy).$$

PROOF. We just compute:

$$\begin{aligned}\int_C f(\zeta)\,d\zeta &= \int_a^b f(\phi(t))\phi'(t)\,dt\\ &= \int_a^b (u(\phi(t)) + iv(\phi(t)))(x'(t) + iy'(t))\,dt\\ &= \int_a^b (u(\phi(t))x'(t) - v(\phi(t))y'(t))\\ &\qquad + i(v(\phi(t))x'(t) + u(\phi(t))y'(t))\,dt\\ &= \int_a^b (u(\phi(t))x'(t) - v(\phi(t))y'(t))\,dt\\ &\qquad + i\int_a^b (v(\phi(t))x'(t) + u(\phi(t))y'(t))\,dt\\ &= \int_C u\,dx - v\,dy + i\int_C v\,dx + u\,dy,\end{aligned}$$

as asserted.

INTEGRATION AROUND CLOSED CURVES, AND GREEN'S THEOREM

Thus far, we have discussed integration over curves joining two distinct points z_1 and z_2. Very important in analysis is the concept of integrating around a closed curve, i.e., one that starts and ends at the same point. There is nothing really new here; the formulas for all three kinds of integrals we have defined will look the same, in the sense that they all are described interms of some parameterization ϕ. A parameterization $\phi : [a,b] \to C$ of a closed curve C is just like the parameterization for a curve joining two points, except that the two points $\phi(a)$ and $\phi(b)$ are equal. Two problems are immediately apparent concerning integrating around a closed curve. First, where do we start on the curve, which point is the initial point? And second, which way to we go around the curve? Recall tha if $\phi : [a,b] \to C$ is a parameterization of C, then $\psi : [a,b] \to C$, defined by $\psi(t) = \phi(a + b - t)$, is a parameterization of C that is the reverse of ϕ, i.e., it goes around the curve in the other direction. If we are integrating with respect to arc length, this reverse direction won't make a difference, but, for contour integrals and line integrals, integrating in the reverse direction will introduce a minus sign.

The first question mentioned above is not so difficult to handle. It doesn't really matter where we start on a closed curve; the parameterization can easily be shifted.

Exercise 6.18. Let $\phi[a,b] \to \mathbb{R}^2$ be a piecewise smooth function that is 1-1 except that $\phi(a) = \phi(b)$. For each $0 < c < b-a$, define $\widehat{\phi} : [a+c, b+c] : \mathbb{R}^2$ by $\widehat{\phi}(t) = \phi(t)$ for $a+c \leq t \leq b$, and $\widehat{\phi}(t) = \phi(t-b+a$ for $b \leq t \leq b+c$.

(a) Show that $\widehat{\phi}$ is a piecewise smooth function, and that the range C of ϕ coincides with the range of $\widehat{\phi}$.

(b) Let f be an integrable (with respect to arc length) function on C. Show that

$$\int_a^b f(\phi(t))|\phi'(t)|\,dt = \int_{a+c}^{b+c} f(\widehat{\phi}(t))|\widehat{\phi}'(t)|\,dt.$$

That is, the integral $\int_C f(s)\,ds$ of f with respect to arc length around the closed curve C is independent of where we start.

(c) Let f be a continuous complex-valued function on C. Show that

$$\int_a^b f(\phi(t))\phi'(t)\,dt = \int_{a+c}^{b+c} f(\widehat{\phi}(t))\widehat{\phi}'(t)\,dt.$$

That is, the contour integral $\int_C f(\zeta)\,d\zeta$ of f around the closed curve C is independent of where we start.

(d) Let $\omega = Pdx + Qdy$ be a differential form on C. Prove that

$$\int_a^b P(\phi(t))x'(t) + Q(\phi(t))y'(t)\,dt = \int_{a+c}^{b+c} P(\widehat{\phi}(t))\widehat{x}'(t) + Q(\widehat{\phi}(t))\widehat{y}'(t)\,dt.$$

That is, the line integral $\int_C \omega$ of ω around C is independent of where we start.

The question of which way we proceed around a closed curve is one that leads to quite intricate and difficult mathematics, at least when we consider totaly general smooth curves. For our purposes it wil, suffice to study a special kind of closed curve, i.e., curves that are the boundaries of piecewise smooth geometric sets. Indeed, the intricate part of the general situation has a lot to do with determining which is the "inside" of the closed curve and which is the "outside," a question that is easily settled in the case of a geometric set. Simple pictures make this general question seem silly, but precise proofs that there is a definite inside and a definite outside are difficult, and eluded mathematicians for centuries, culminating in the famous Jordan Curve Theorem, which asserts exactly what our intuition predicts:

JORDAN CURVE THEOREM. *The complement of a closed curve is the union of two disjoint components, one bounded and one unbounded.*

We define the bounded component to be the inside of the curve and the unbounded component to be the outside.

We adopt the following convention for how we integrate around the boundary of a piecewise smooth geometric set S. That is, the curve C_S will consist of four parts: the lower boundary (graph of the lower bounding function l), the righthand boundary (a portion of the vertical line $x = b$), the upper boundary (the graph of the upper bounding function u), and finally the lefthand side (a portion of the vertical line $x = a$). By *integrating around* such a curve C_S, we will always mean proceeding counterclockwise around the curves. Specifically, we move from left to right along the lower boundary, from bottom to top along the righthand boundary, from right to left across the upper boundary, and from top to bottom along the lefthand boundary. Of course, as shown in the exercise above, it doesn't matter where we start.

Exercise 6.19. Let S be the closed piecewise smooth geometric set that is determined by the interval $[a, b]$ and the two piecewise smooth bounding functions u and l. Assume that the boundary C_S of S has finite length. Suppose the graph of u intersects the lines $x = a$ and $x = b$ at the points (a, c) and (b, d), and suppose that the graph of l intersects those lines at the points (a, e) and (b, f). Find a parameterization $\phi : [a', b'] \to C_S$ of the curve C_S.
HINT: Try using the interval $[a, b + d - f + b - a + c - e]$ as the domain $[a', b']$ of ϕ.

The next theorem, though simple to state and use, contains in its proof a combinatorial idea that is truly central to all that follows in this chapter. In its simplest form, it is just the realization that the line integral in one direction along a curve is the negative of the line integral in the opposite direction.

THEOREM 6.13. *Let $S_1, \ldots, S_n$ be a collection of closed geometric sets that constitute a partition of a geometric set S, and assume that the boundaries of all the S_i's, as well as the boundary of S, have finite length. Suppose ω is a continuous differential form on all the boundaries $\{C_{S_k}\}$. Then*

$$\int_{C_S} \omega = \sum_{k=1}^{n} \int_{C_{S_k}} \omega.$$

PROOF. We give a careful proof for a special case, and then outline the general argument. Suppose then that S is a piecewise smooth geometric set, determined by the interval $[a, b]$ and the two bounding functions u and l, and assume that the boundary C_S has finite length. Suppose $m(x)$ is a piecewise smooth function on $[a, b]$, satisfying $\int_a^b |m'| < \infty$, and assume that $l(x) < m(x) < u(x)$ for all $x \in (a, b)$. Let S_1 be the geometric set determined by the interval $[a, b]$ and the two bounding functions m and l, and let S_2 be the geometric set determined by the interval $[a, b]$ and the two bounding functions u and m. We note first that the two geometric sets S_1 and S_2 comprise a partition of the geometric set S, so that this is indeed a pspecial case of the theorem.

Next, consider the following eight line integrals: First, integrate from left to write along the graph of m, second, up the line $x = b$ from $(b, m(b))$ to $(b, u(b))$, third, integrate from right to left across the graph of u, fourth, integrate down the line $x = a$ from $(a, u(a))$ to $(a, m(a))$, fifth, continue down the line $x = a$ from $(a, m(a))$ to $(a, l(a))$, sixth, integrate from left to right across the graph of l, seventh, integrate up the line $x = b$ from $(b, l(b))$ to $(b, m(b))$, and finally, integfrate from right to left across the graph of m.

The first four line integrals comprise the line integral around the geometric set S_2, and the last four comprise the line integral around the geometric set S_1. On the other hand, the first and eighth line integrals here cancel out, for one is just the reverse of the other. Hence, the sum total of these eight line integrals, integrals 2–7, is just the line integral around the boundary C_S of S. Therefore

$$\int_{C_S} \omega = \int_{C_{S_1}} \omega + \int_{C_{S_2}} \omega$$

as desired.
We give next an outline of the proof for a general partition $S_1, \dots, S_n$ of S. Let S_k be determined by the interval $[a_k, b_k]$ and the two bounding functions u_k and l_k. Observe that, if the boundary C_{S_k} of S_k intersects the boundary C_{S_j} of S_j in a curve C, then the line integral of ω along C, when it is computed as part of integrating counterclockwise around S_k, is the negative of the line integral along C, when it is computed as part of the line integral counterclockwise around S_j. Indeed, the first line integral is the reverse of the second one. (A picture could be helpful.) Consequently, when we compute the sum of the line integrals of ω around the C_{S_k}'s, All terms cancel out except those line integrals that ar computed along parts of the boundaries of the S_k's that intersect no other S_j. But such parts of the boundaries of the S_k's must coincide with parts of the boundary of S. Therefore, the sum of the line integrals of ω around the boundaries of the S_k's equals the line integral of ω around the boundary of S, and this is precisely what the theorem asserts.

Exercise 6.20. Prove the analog of Theorem 6.13 for contour integrals: Let $S_1, \dots, S_n$ be a collection of closed geometric sets that constitute a partition of a geometric set S, and assume that the boundaries of all the S_i's, as well as the boundary of S, have finite length. Suppose f is a continuous complex-valued function on all the boundaries $\{C_{S_k}\}$ as well as on the boundary C_S. Then

$$\int_{C_S} f(\zeta)\, d\zeta = \sum_{k=1}^{n} \int_{C_{S_k}} f(\zeta)\, d\zeta.$$

We come now to the most remarkable theorem in the subject of integration over curves, Green's Theorem. Another fanfare, please!

THEOREM 6.14. (Green) Let S be a piecewise smooth, closed, geometric set, let C_S denote the closed curve that is the boundary of S, and assume that C_S is of finite length. Suppose $\omega = Pdx + Qdy$ is a continuous differential form on S that is smooth on the interior S^0 of S. Then

$$\int_{C_S} \omega = \int_{C_S} P\,dx + Q\,dy = \int_S \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}.$$

REMARK. The first thing to notice about this theorem is that it connects an integral around a (1-dimensional) curve with an integral over a (2-dimensional) set, suggesting a kind of connection between a 1-dimensional process and a 2-dimensional one. Such a connection seems to be unexpected, and it should therefore have some important implications, as indeed Green's Theorem does.

The second thing to think about is the case when ω is an exact differential *df* of a smooth function f of two real variables. In that case, Green's Theorem says

$$\int_{C_S} \frac{\partial f}{\partial x}\,dx + \frac{\partial f}{\partial y}\,dy = \int_S (f_{yx} - f_{xy}),$$

which would be equal to 0 if $f \in C^2(S)$, by Theorem 4.22. Hence, the integral of *df* around any such curve would be 0. If U is an open subset of $\mathbb{R}^2$, there may or may not be some other ω's, called *closed differential forms,* having the property that their integral around every piecewise smooth curve of finite length in U is 0, and the study of these closed differential forms ω that are not exact differential forms *df* has led to much interesting mathematics. It turns out that the structure of the open set U, e.g., how many "holes" there are in it, is what's important. Take a course in Algebraic Topology!

The proof of Green's Theorem is tough, and we break it into several steps.

LEMMA 1. *Suppose S is the rectangle $[a,b] \times [c,d]$. Then Green's Theorem is true.*

PROOF OF LEMMA 1. We think of the closed curve C_S bounding the rectangle as the union of four straight lines, C_1, C_2, C_3 and C_4, and we parameterize them as follows: Let $\phi : [a,b] \to C_1$ be defined by $\phi(t) = (t,c)$; let $\phi : [b, b+d-c] \to C_2$ be defined by $\phi(t) = (b, t-b+c)$; let $\phi : [b+d-c, b+d-c+b-a] \to C_3$ be defined by $\phi(t) = (b+d-c+b-t, d)$; and let $\phi : [b+d-c+b-a, b+d-c+b-a+d-c] \to C_4$ be defined by $\phi(t) = (a, b+d-c+b-a+d-t)$. One can check directly to see that this ϕ parameterizes the boundary of the rectangle $S = [a,b] \times [c,d]$.

As usual, we write $\phi(t) = (x(t), y(t))$. Now, we just compute, use the Fundamental

Theorem of Calculus in the middle, and use part (d) of Exercise 5.30 at the end.

$$\begin{aligned}
\int_{C_S}\omega &= \int_{C_1}\omega+\int_{C_2}\omega \\
&\quad+\int_{C_3}\omega+\int_{C_4}\omega \\
&= \int_{C_1}P\,dx+Q\,dy+\int_{C_2}P\,dx+Q\,dy \\
&\quad+\int_{C_3}P\,dx+Q\,dy+\int_{C_4}P\,dx+Q\,dy \\
&= \int_a^b P(\phi(t))x'(t)+Q(\phi(t))y'(t)\,dt \\
&\quad+\int_b^{b+d-c} P(\phi(t))x'(t)+Q(\phi(t))y'(t)\,dt \\
&\quad+\int_{b+d-c}^{b+d-c+b-a} P(\phi(t))x'(t)+Q(\phi(t))y'(t)\,dt \\
&\quad+\int_{b+d-c+b-a}^{b+d-c+b-a+d-c} P(\phi(t))x'(t)+Q(\phi(t))y'(t)\,dt \\
&= \int_a^b P(t,c)\,dt+\int_b^{b+d-c} Q(b,t-b+c)\,dt \\
&\quad+\int_{b+d-c}^{b+d-c+b-a} P(b+d-c+b-t,d)(-1)\,dt \\
&\quad+\int_{b+d-c+b-a}^{b+d-c+b-a+d-c} Q(a,b+d-c+b-a+d-t)(-1)\,dt \\
&= \int_a^b P(t,c)\,dt+\int_c^d Q(b,t)\,dt \\
&\quad-\int_a^b P(t,d)\,dt-\int_c^d Q(a,t)\,dt \\
&= \int_c^d (Q(b,t)-Q(a,t))\,dt-\int_a^b (P(t,d)-P(t,c))\,dt \\
&= \int_c^d\int_a^b \frac{\partial Q}{\partial x}(s,t)\,dsdt \\
&\quad-\int_a^b\int_c^d \frac{\partial P}{\partial y}(t,s)\,dsdt \\
&= \int_S(\frac{\partial Q}{\partial x}-\frac{\partial P}{\partial y},
\end{aligned}$$

proving the lemma.

LEMMA 2. *Suppose S is a right triangle whose vertices are of the form $(a,c),(b,c)$ and (b,d). Then Green's Theorem is true.*

PROOF OF LEMMA 2. We parameterize the boundary C_S of this triangle as follows: For $t\in[a,b]$, set $\phi(t)=(t,c)$; for $t\in[b,b+d-c]$, set $\phi(t)=(b,t+c-b)$;

and for $t \in [b+d-c, b+d-c+b-a]$, set $\phi(t) = (b+d-c+b-t, b+d-c+d-t)$. Again, one can check that this ϕ parameterizes the boundary of the triangle S. Write $\phi(t) = (x(t), y(t))$. Again, using the Fundamental Theorem and Exercise 5.30, we have

$$\begin{aligned}
\int_{C_S} \omega &= \int_{C_S} P\,dx + Q\,dy \\
&= \int_a^b P(\phi(t))x'(t) + Q(\phi(t))y'(t)\,dt \\
&\quad + \int_b^{b+d-c} P(\phi(t))x'(t) + Q(\phi(t))y'(t)\,dt \\
&\quad + \int_{b+d-c}^{b+d-c+b-a} P(\phi(t))x'(t) + Q(\phi(t))y'(t)\,dt \\
&= \int_a^b P(t,c)\,dt + \int_b^{b+d-c} Q(b, t+c-b)\,dt \\
&\quad + \int_{b+d-c}^{b+d-c+b-a} P(b+d-c+b-t, b+d-c+d-t)(-1)\,dt \\
&\quad + \int_{b+d-c}^{b+d-c+b-a} Q(b+d-c+b-t, b+d-c+d-t)(-1)\,dt \\
&= \int_a^b P(t,c)\,dt + \int_c^d Q(b,t)\,dt \\
&\quad - \int_a^b P(s, (d + \frac{s-b}{a-b}(c-d)))\,ds \\
&\quad - \int_c^d Q(b + \frac{s-d}{c-d}(a-b)), s)\,ds \\
&= \int_c^d (Q(b,s) - Q((b + \frac{s-d}{c-d}(a-b)), s))\,ds \\
&\quad - \int_a^b (P(s, (d + \frac{s-b}{a-b}(c-d))) - P(s,c))\,ds \\
&= \int_c^d \int_{b+\frac{s-d}{c-d}(a-b)}^b \frac{\partial Q}{\partial x}(t,s)\,dtds \\
&\quad - \int_a^b \int_c^{d+\frac{s-b}{a-b}(c-d)} \frac{\partial P}{\partial y}(s,t)\,dtds \\
&= \int_S (\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y},
\end{aligned}$$

which proves Lemma 2.

LEMMA 3. *Suppose $S_1, \dots, S_n$ is a partition of the geometric set S, and that the boundary C_{S_k} has finite length for all $1 \leq k \leq n$. If Green's Theorem holds for each geometric set S_k, then it holds for S.*

PROOF OF LEMMA 3. From Theorem 6.13 we have

$$\int_{C_S} \omega = \sum_{k=1}^n \int_{C_{S_k}} \omega,$$

and from Theorem 5.24 we have

$$\int_S Q_x - P_y = \sum_{k=1}^n \int_{S_k} Q_x - P_y.$$

Since Green's Theorem holds for each k, we have that

$$\int_{C_{S_k}} \omega = \int_{S_k} Q_x - P_y,$$

and therefore

$$\int_{C_S} \omega = \int_S Q_x - P_y,$$

as desired.

Exercise 6.21. (a) Prove Green's Theorem for a right triangle with vertices of the form $(a, c), (b, c)$, and (a, d).
(b) Prove Green's Theorem for a trapezoid having vertices of the form $(a, c), (b, c)$, (b, d), and (a, e), where both d and e are greater than c.
HINT: Represent this trapezoid as the union of a rectangle and a right triangle that share a border. Then use Lemma 3.
(c) Prove Green's Theorem for S any quadrilateral that has two vertical sides.
(d) Prove Green's Theorem for any geometric set S whose upper and lower bounding functions are piecewise linear functions.
HINT: Show that S can be thought of as a finite union of quadrilaterals, like those in part (c), each one sharing a vertical boundary with the next. Then, using induction and the previous exercise finish the argument.

We need one final lemma before we can complete the general proof of Green's Theorem. This one is where the analysis shows up; there are carefully chosen ϵ's and δ's.

LEMMA 4. *Suppose S is contained in an open set U and that ω is smooth on all of U. Then Green's Theorem is true.*

PROOF OF LEMMA 4. Let the piecewise smooth geometric set S be determined by the interval $[a, b]$ and the two bounding functions u and l. Using Theorem 2.11, choose an $r > 0$ such that the neighborhood $N_r(S) \subseteq U$. Now let $\epsilon > 0$ be given, and choose *delta* to satisfy the following conditions:

(1) (a) $\delta < r/2$, from which it follows that the open neighborhood $N_\delta(S)$ is a subset of the compact set $\overline{N}_{r/2}(S)$. (See part (f) of Exercise 2.24.)
(2) (b) $\delta < \epsilon/4M$, where M is a common bound for all four continuous functions $|P|, |Q|, |P_y|$, and $|Q_x|$ on the compact set $\overline{N}_{r/2}(S)$.
(3) (c) $\delta < \epsilon/4M(b - a)$.
(4) (d) δ satisfies the conditions of Theorem 6.11.

Next, using Theorem 6.1, choose two piecewise linear functions p_u and p_l so that

(1) $|u(x) - p_u(x)| < \delta/2$ for all $x \in [a, b]$.
(2) $|l(x) - p_l(x)| < \delta/2$ for all $x \in [a, b]$.
(3) $\int_a^b |u'(x) - p_u'(x)|\, dx < \delta$.
(4) $\int_a^b |l'(x) - p_l'(x)|\, dx < \delta$.

Let $\widehat{S}$ be the geometric set determined by the interval $[a, b]$ and the two bounding functions $\widehat{u}$ and $\widehat{l}$, where $\widehat{u} = p_u + \delta/2$ and $\widehat{l} = p_l - \delta/2$. We know that both $\widehat{u}$ and $\widehat{l}$ are piecewise linear functions. We have to be a bit careful here, since for some x's it could be that $p_u(x) < p_l(x)$. Hence, we could not simply use p_u and p_l themselves as bounding functions for $\widehat{S}$. We do know from (1) and (2) that $u(x) < \widehat{u}(x)$ and $l(x) > \widehat{l}(x)$, which implies that the geometric set S is contained in the geometric set $\widehat{S}$. Also $\widehat{S}$ is a subset of the neighborhood $N_\delta(s)$, which in turn is a subset of the compact set $\overline{N}_{r/2}(S)$.

Now, by part (d) of the preceding exercise, we know that Green's Theorem holds for $\widehat{S}$. That is

$$\int_{C_{\widehat{S}}} \omega = \int_{\widehat{S}} (Q_x - P_y).$$

We will show that Green's Theorem holds for S by showing two things: (i) $|\int_{C_S} \omega - \int_{C_{\widehat{S}}} \omega| < 4\epsilon$, and (ii) $|\int_S (Q_x) - P_y) - \int_{\widehat{S}} (Q_x - P_y)| < \epsilon$. We would then have, by the usual adding and subtracting business, that

$$|\int_{C_S} \omega - \int_S (Q_x - P_y)| < 5\epsilon,$$

and, since ϵ is an arbitrary positive number, we would obtain

$$\int_{C_S} \omega = \int_S (Q_x - P_y).$$

Let us estabish (i) first. We have from (1) above that $|u(x) - \widehat{u}(x)| < \delta$ for all $x \in [a, b]$, and from (3) that

$$\int_a^b |u'(x) - \widehat{u}'(x)|\, dx = \int_a^b |u'(x) - p_u'(x)|\, dx < \delta.$$

Hence, by Theorem 6.11,

$$\int_{\mathrm{graph}(u)} \omega - \int_{\mathrm{graph}(\widehat{u})} \omega| < \epsilon.$$

Similarly, using (2) and (4) above, we have that

$$|\int_{\mathrm{graph}(l)} \omega - \int_{\mathrm{graph}(\widehat{l})} \omega| < \epsilon.$$

Also, the difference of the line integrals of ω along the righthand boundaries of S and $\widehat{S}$ is less than ϵ. Thus

$$\begin{aligned}
|\int_{C(b,l(b))}^{(b,u(b))} \omega - \int_{C(b,\widehat{l}(b))}^{(b,\widehat{u}(b))} \omega| &= |\int_{l(b)}^{u(b)} Q(b,t)\, dt - \int_{\widehat{l}(b)}^{\widehat{u}(b)} Q(b,t)\, dt| \\
&\le |\int_{u(b)}^{\widehat{u}(b)} Q(b,t)\, dt| + |\int_{\widehat{l}(b)}^{l(b)} Q(b,t)\, dt| \\
&\le M(|l(b) - \widehat{l}(b)| + |u(b) - \widehat{u}(b)|) \\
&\le M(\delta + \delta) \\
&= 2M\delta \\
&< \epsilon.
\end{aligned}$$

Of course, a similar calculation shows that

$$|\int_{C(a,u(a))}^{(a,l(a))}\omega - \int_{C(a,\widehat{u}(a))}^{(a,\widehat{l}(a))}\omega| < \epsilon.$$

These four line integral inequalities combine to give us that

$$|\int_{C_S}\omega - \int_{C_{\widehat{S}}}\omega| < 4\epsilon,$$

establishing (i).

Finally, to see (ii), we just compute

$$\begin{aligned}
0 &\leq |\int_{\widehat{S}}(Q_y - P_x) - \int_S (Q_y - P_x)| \\
&= |\int_a^b \int_{\widehat{l}(t)}^{\widehat{u}(t)} (Q_x((t,s) - P_y(t,s))\,dsdt - \int_a^b \int_{l(t)}^{u(t)} (Q_x(t,s) - P_y(t,s))\,dsdt| \\
&= |\int_a^b \int_{\widehat{l}(t)}^{l(t)} (Q_x(t,s) - P_y(t,s))\,dsdt + \int_a^b \int_{u(t)}^{\widehat{u}(t)} (Q_x(t,s) - P_y(t,s))\,dsdt| \\
&\leq 2M(\int_a^b |l(t) - \widehat{l}(t)| + |\widehat{u}(t) - u(t)|\,dt \\
&\leq 4M\delta(b-a) \\
&< \epsilon.
\end{aligned}$$

This establishes (ii), and the proof is complete.

At last, we can finish the proof of this remarkable result.

PROOF OF GREEN'S THEOREM. As usual, let S be determined by the interval $[a,b]$ and the two bounding functions u and l. Recall that $u(x) - l(x) > 0$ for all $x \in (a,b)$. For each natural number $n > 2$, let S_n be the geometric set that is determined by the interval $[a + 1/n, b - 1/n]$ and the two bounding functions u_n and l_n, where $u_n = u - (u-l)/n$ restricted to the interval $[a + 1/n, b - 1/n]$, and $l_n = l + (u-l)/n$ restricted to $[a + 1/n, b - 1/n]$. Then each S_n is a piecewise smooth geometric set, whose boundary has finite length, and each S_n is contained in the open set S^0 where by hypothesis ω is smooth. Hence, by Lemma 4, Green's Theorem holds for each S_n. Now it should follow directly, by taking limits, that Green's Theorem holds for S. In fact, this is the case, and we leave the details to the exercise that follows.

Exercise 6.22. Let S, ω, and the S_n's be as in the preceding proof.

(a) Using Theorem 6.11, show that

$$\int_{C_S}\omega = \lim \int_{C_{S_n}}\omega.$$

(b) Let f be a bounded integrable function on the geometric set S. Prove that

$$\int_S f = \lim \int_{S_n} f.$$

(c) Complete the proof to Green's Theorem; i.e., take limits.

REMARK. Green's Theorem is primarily a theoretical result. It is rarely used to "compute" a line integral around a curve or an integral of a function over a geometric set. However, there is one amusing exception to this, and that is when the differential form $\omega = x\,dy$. For that kind of ω, Green's Theorem says that the area of the geometric set S can be computed as follows:

$$A(S) = \int_S 1 = \int_S \frac{\partial Q}{\partial x} = \int_{C_S} x\,dy.$$

This is certainly a different way of computing areas of sets from the methods we developed earlier. Try this way out on circles, ellipses, and the like.

CHAPTER VII
THE FUNDAMENTAL THEOREM OF ALGEBRA, AND THE FUNDAMENTAL THEOREM OF ANALYSIS

In this chapter we will discover the incredible difference between the analysis of functions of a single complex variable as opposed to functions of a single real variable. Up to this point, in some sense, we have treated them as being quite similar subjects, whereas in fact they are extremely different in character. Indeed, if f is a differentiable function of a complex variable on an open set $U \subseteq \mathbb{C}$, then we will see that f is actually expandable in a Taylor series around every point in U. In particular, a function $f$of a complex variable is guaranteed to have infinitely many derivatives on U if it merely has the first one on U. This is in marked contrast with functions of a real variable. See part (3) of Theorem 4.17.
The main points of this chapter are:

(1) **The Cauchy-Riemann Equations** (Theorem 7.1),
(2) **Cauchy's Theorem** (Theorem 7.3),
(3) **Cauchy Integral Formula** (Theorem 7.4),
(4) **A complex-valued function that is differentiable on an open set is expandable in a Taylor series around each point of the set** (Theorem 7.5),
(5) **The Identity Theorem** (Theorem 7.6),
(6) **The Fundamental Theorem of Algebra** (Theorem 7.7),
(7) **Liouville's Theorem** (Theorem 7.8),
(8) **The Maximum Modulus Principle** (corollary to Theorem 7.9),
(9) **The Open Mapping Theorem** (Theorem 7.10),
(10) **The uniform limit of analytic functions is analytic** (Theorem 7.12), and
(11) **The Residue Theorem** (Theorem 7).17.

CAUCHY'S THEOREM

We begin with a simple observation connecting differentiability of a function of a complex variable to a relation among of partial derivatives of the real and imaginary parts of the function. Actually, we have already visited this point in Exercise 4.8.

THEOREM 7.1. (Cauchy-Riemann equations) Let $f = u + iv$ be a complex-valued function of a complex variable $z = x + iy \equiv (x, y)$, and suppose f is differentiable, as a function of a complex variable, at the point $c = (a, b)$. Then the following two partial differential equations, known as the *Cauchy-Riemann Equations*, hold:

$$\frac{\partial u}{\partial x}(a, b) = \frac{\partial v}{\partial y}(a, b),$$

and

$$\frac{\partial u}{\partial y}(a, b) = -\frac{\partial v}{\partial x}(a, b).$$

PROOF. We know that

$$f'(c) = \lim h \to 0 \frac{f(c + h) - f(c)}{h},$$

and this limit is taken as the complex number h approaches 0. We simply examine this limit for real h's approaching 0 and then for purely imaginary h's approaching 0. For real h's, we have

$$\begin{aligned}
f'(c) &= f'(a+ib) \\
&= \lim_{h\to 0} \frac{f(a+h+ib)-f(a+ib)}{h} \\
&= \lim h \to 0 \frac{u(a+h,b)+iv(a+h,b)-u(a,b)-iv(a,b)}{h} \\
&= \lim_{h\to 0} \frac{u(a+h,b)-u(a,b)}{h} + i \lim_{h\to 0} \frac{v(a+h,b)-v(a,b)}{h} \\
&= \frac{\partial u}{\partial x}(a,b) + i\frac{\partial v}{\partial x}(a,b).
\end{aligned}$$

For purely imaginary h's, which we write as $h = ik$, we have

$$\begin{aligned}
f'(c) &= f'(a+ib) \\
&= \lim_{k\to 0} \frac{f(a+i(b+k))-f(a+ib)}{ik} \\
&= \lim_{k\to 0} \frac{u(a,b+k)+iv(a,b+k)-u(a,b)-iv(a,b)}{ik} \\
&= -i \lim_{k\to 0} \frac{u(a,b+k)-u(a,b)}{k} + \frac{v(a,b+k)-v(a,b)}{k} \\
&= -i\frac{\partial u}{\partial y}(a,b) + \frac{\partial v}{\partial y}(a,b).
\end{aligned}$$

Equating the real and imaginary parts of these two equivalent expressions for $f'(c)$ gives the Cauchy-Riemann equations.

As an immediate corollary of this theorem, together with Green's Theorem (Theorem 6.14), we get the following result, which is a special case of what is known as Cauchy's Theorem.

COROLLARY. *Let S be a piecewise smooth geometric set whose boundary C_S has finite length. Suppose f is a complex-valued function that is continuous on S and differentiable at each point of the interior S^0 of S. Then the contour integral $\int_{C_S} f(\zeta)\,d\zeta = 0$.*

Exercise 7.1. (a) Prove the preceding corollary. See Theorem 6.12.
(b) Suppose $f = u+iv$ is a differentiable, complex-valued function on an open disk $B_r(c)$ in $\mathbb{C}$, and assume that the real part u is a constant function. Prove that f is a constant function. Derive the same result assuming that v is a constant function.
(c) Suppose f and g are two differentiable, complex-valued functions on an open disk $B_r(c)$ in $\mathbb{C}$. Show that, if the real part of f is equal to the real part of g, then there exists a constant k such that $f(z) = g(z) + k$, for all $z \in B_r(c)$.

For future computational purposes, we give the following implications of the Cauchy-Riemann equations. As with Theorem 7.1, this next theorem mixes the notions of differentiability of a function of a complex variable and the partial derivatives of its real and imaginary parts.

THEOREM 7.2. *Let $f = u + iv$ be a complex-valued function of a complex variable, and suppose that f is differentiable at the point $c = (a, b)$. Let A be the 2×2 matrix*

$$A = \begin{pmatrix} u_x(a,b) & v_x(a,b) \\ u_y(a,b) & v_y(a,b) \end{pmatrix}.$$

Then:

(1) $|f'(c)|^2 = \det(A)$.

(2) *The two vectors*

$$\vec{V_1} = (u_x(a,b), u_y(a,b)) \text{ and } \vec{V_2} = (v_x(a,b), v_y(a,b))$$

are linearly independent vectors in $\mathbb{R}^2$ if and only if $f'(c) \neq 0$.

(3) *The vectors*

$$\vec{V_3} = (u_x(a,b), v_x(a,b)) \text{ and } \vec{V_4} = (u_y(a,b), v_y(a,b))$$

are linearly independent vectors in $\mathbb{R}^2$ if and only if $f'(c) \neq 0$.

PROOF. Using the Cauchy-Riemann equations, we see that the determinant of the matrix A is given by

$$\begin{aligned}\det A &= u_x(a,b)v_y(a,b) - u_y(a,b)v_x(a,b) \\ &= (u_x(a,b))^2 + (v_x(a,b))^2 \\ &= (u_x(a,b) + iv_x(a,b))(u_x(a,b) - iv_x(a,b)) \\ &= f'(c)\overline{f'(c)} \\ &= |f'(c)|^2,\end{aligned}$$

proving part (1).

The vectors $\vec{V_1}$ and $\vec{V_2}$ are the columns of the matrix A, and so, from elementary linear algebra, we see that they are linearly independent if and only if the determinant of A is nonzero. Hence, part (2) follows from part (1). Similarly, part (3) is a consequence of part (1).

It may come as no surprise that the contour integral of a function f around the boundary of a geometric set S is not necessarily 0 if the function f is not differentiable at each point in the interior of S. However, it is exactly these kinds of contour integrals that will occupy our attention in the rest of this chapter, and we shouldn't jump to any conclusions.

Exercise 7.2. Let c be a point in $\mathbb{C}$, and let S be the geometric set that is a closed disk $\overline{B}_r(c)$. Let ϕ be the parameterization of the boundary C_r of S given by $\phi(t) = c + re^{it}$ for $t \in [0, 2\pi]$. For each integer $n \in \mathbb{Z}$, define $f_n(z) = (z - c)^n$.

(a) Show that $\int_{C_r} f_n(\zeta \, d\zeta = 0$ for all $n \neq -1$.

(b) Show that

$$\int_{C_r} f_{-1}(\zeta)\, d\zeta = \int_{C_r} \frac{1}{\zeta - c}\, d\zeta = 2\pi i.$$

There is a remarkable result about contour integrals of certain functions that aren't differentiable everywhere within a geometric set, and it is what has been called the Fundamental Theorem of Analysis, or Cauchy's Theorem. This theorem has many general statements, but we present one here that is quite broad and certainly adequate for our purposes.

THEOREM 7.3. (Cauchy's Theorem, Fundamental Theorem of Analysis) Let S be a piecewise smooth geometric set whose boundary C_S has finite length, and let $\widehat{S} \subseteq S^0$ be a piecewise smooth geometric set, whose boundary $C_{\widehat{S}}$ also is of finite length. Suppose f is continuous on $S \cap \widetilde{\widehat{S}^0}$, i.e., at every point z that is in S but not in $\widehat{S}^0$, and assume that f is differentiable on $S^0 \cap \widetilde{\widehat{S}}$, i.e., at every point z in S^0 but not in $\widehat{S}$. (We think of these sets as being the points "between" the boundary curves of these geometric sets.) Then the two contour integrals $\int_{C_S} f(\zeta)\, d\zeta$ and $\int_{C_{\widehat{S}}} f(\zeta)\, d\zeta$ are equal.

PROOF. Let the geometric set S be determined by the interval $[a, b]$ and the two bounding functions u and l, and let the geometric set $\widehat{S}$ be determined by the subinterval $[\widehat{a}, \widehat{b}]$ of $[a, b]$ and the two bounding functions $\widehat{u}$ and $\widehat{l}$. Because $\widehat{S} \subseteq S^0$, we know that $\widehat{u}(t) < u(t)$ and $l(t) < \widehat{l}(t)$ for all $t \in [\widehat{a}, \widehat{b}]$. We define four geometric sets $S_1, \dots , S_4$ as follows:

(1) S_1 is determined by the interval $[a, \widehat{a}]$ and the two bounding functions u and l restricted to that interval.
(2) S_2 is determined by the interval $[\widehat{a}, \widehat{b}]$ and the two bounding functions u and $\widehat{u}$ restricted to that interval.
(3) S_3 is determined by the interval $[\widehat{a}, \widehat{b}]$ and the two bounding functions $\widehat{l}$ and l restricted to that interval.
(4) S_4 is determined by the interval $[\widehat{b}, b]$ and the two bounding functions u and l restricted to that interval.

Observe that the five sets $\widehat{S}, S_1, \dots , S_4$ constitute a partition of the geometric set S. The corollary to Theorem 7.1 applies to each of the four geometric sets $S_1, \dots , S_4$. Hence, the contour integral of f around each of the four boundaries of these geometric sets is 0. So, by Exercise 6.20,

$$\begin{aligned} \int_{C_S} f(\zeta)\, d\zeta &= \int_{C_{\widehat{S}}} f(\zeta)\, d\zeta + \sum_{k=1}^{4} \int_{C_{S_k}} f(\zeta)\, d\zeta \\ &= \int_{C_{\widehat{S}}} f(\zeta)\, d\zeta, \end{aligned}$$

as desired.

Exercise 7.3. (a) Draw a picture of the five geometric sets in the proof above and justify the claim that the sum of the four contour integrals around the geometric sets $S_1, \dots , S_4$ is the integral around C_S minus the integral around $C_{\widehat{S}}$.
(b) Let $S_1, \dots , S_n$ be pairwise disjoint, piecewise smooth geometric sets, each having a boundary of finite length, and each contained in a piecewise smooth geometric set S whose boundary also has finite length. Prove that the S_k's are some of the elements of a partition $\{\widetilde{S}_l\}$ of S, each of which is piecewise smooth and has a boundary of finite length. Show that, by reindexing, $S_1, \dots , S_n$ can be chosen to be the first n elements of the partition $\{\widehat{S}_l\}$.
HINT: Just carefully adjust the proof of Theorem 5.25.
(c) Suppose S is a piecewise smooth geometric set whose boundary has finite length, and let $S_1, \dots , S_n$ be a partition of S for which each S_k is piecewise smooth and has a boundary C_{S_k} of finite length. Suppose f is continuous on each of the boundaries

C_{S_k} of the S_k's as well as the boundary C_S of S, and assume that f is continuous on each of the S_k's, for $1 \leq k \leq m$, and differentiable at each point of their interiors. Prove that

$$\int_{C_S} f(\zeta)\,d\zeta = \sum_{k=m+1}^{n} \int_{C_{S_k}} f(\zeta)\,d\zeta.$$

(d) Prove the following generalization of the Cauchy Theorem: Let $S_1, \ldots, S_n$ be pairwise disjoint, piecewise smooth geometric sets whose boundaries have finite length, all contained in the interior of a piecewise smooth geometric set S whose boundary also has finite length. Suppose f is continuous at each point of S that is not in the interior of any of the S_k's, and that f is differentiable at each point of S^0 that is not an element of any of the S_k's. Prove that

$$\int_{C_S} f(\zeta)\,d\zeta = \sum_{k=1}^{n} \int_{C_{S_k}} f(\zeta)\,d\zeta.$$

Perhaps the main application of Theorem 7.3 is what's called the Cauchy Integral Formula. It may not appear to be useful at first glance, but we will be able to use it over and over throughout this chapter. In addition to its theoretical uses, it is the basis for a technique for actually evaluating contour integrals, line integrals, as well as ordinary integrals.

THEOREM 7.4. (Cauchy Integral Formula) Let S be a piecewise smooth geometric set whose boundary C_S has finite length, and let f be a continuous function on S that is differentiable on the interior S^0 of S. Then, for any point $z \in S^0$, we have

$$f(z) = \frac{1}{2\pi i} \int_{C_S} \frac{f(\zeta)}{\zeta - z}\,d\zeta.$$

REMARK. This theorem is an initial glimpse at how differentiable functions of a complex variable are remarkably different from differentiable functions of a real variable. Indeed, Cauchy's Integral Formula shows that the values of a differentiable function f at all points in the interior of a geometric set S are completely determined by the values of that function on the boundary of the set. The analogous thing for a function of a real variable would be to say that all the values of a differentiable function f at points in the open interval (a, b) are completely determined by its values at the endpoints a and b. This is patently absurd for functions of a real variable, so there surely is something marvelous going on for differentiable functions of a complex variable.

PROOF. Let r be any positive number such that $\overline{B}_r(z)$ is contained in the interior S^0 of S, and note that the close disk $\overline{B}_r(z)$ is a piecewise smooth geometric set $\widehat{S}$ contained in S^0. We will write C_r instead of $C_{\widehat{S}}$ for the boundary of this disk, and we will use as a parameterization of the curve C_r the function $\phi : [0, 2\pi] \to C_r$ given by $\phi(t) = z + re^{it}$. Now the function $g(\zeta) = f(\zeta)/(\zeta - z)$ is continuous on $S \cap \widetilde{\widehat{S}^0}$ and differentiable on $S^0 \cap \widetilde{\widehat{S}}$, so that Theorem 7.3 applies to the function g.

Hence

$$\begin{aligned}\frac{1}{2\pi i}\int_{C_S}\frac{f(\zeta)}{\zeta-z}\,d\zeta &= \frac{1}{2\pi i}\int_{C_S} g(\zeta)\,d\zeta\\ &= \frac{1}{2\pi i}\int_{C_R} g(\zeta),d\zeta\\ &= \frac{1}{2\pi i}\int_{C_r}\frac{f(\zeta)}{\zeta-z}\,d\zeta\\ &= \frac{1}{2\pi i}\int_0^{2\pi}\frac{f(z+re^{it})}{z+re^{it}-z}ire^{it}\,dt\\ &= \frac{1}{2\pi}\int_0^{2\pi} f(z+re^{it})\,dt.\end{aligned}$$

Since the equality established above is valid, independent of r, we may take the limit as r goes to 0, and the equality will persist. We can evaluate such a limit by replacing the r by $1/n$, in which case we would be evaluating

$$\lim_{n\to\infty}\frac{1}{2\pi}\int_0^{2\pi} f(z+\frac{1}{n}e^{it})\,dt = \lim_{n\to\infty}\frac{1}{2\pi}\int_0^{2\pi} f_n(t)\,dt,$$

where $f_n(t) = f(z + frac1ne^{it})$. Finally, because the function f is continuous at the point z, it follows that the sequence $\{f_n\}$ converges uniformly to the constant function $f(z)$ on the interval $[0, 2\pi]$. So, by Theorem 5.6, we have that

$$\lim_{n\to\infty}\frac{1}{2\pi}\int_0^{2\pi} f_n(t)\,dt = \frac{1}{2\pi}\int_0^{2\pi} f(z)\,dt = f(z).$$

Therefore,

$$\frac{1}{2\pi i}\int_{C_S}\frac{f(\zeta}{\zeta-z}\,d\zeta = \lim_{r\to 0}\frac{1}{2\pi}\int_0^{2\pi} f(z+re^{it})\,dt = f(z),$$

and the theorem is proved.

The next exercise gives two simple but strong consequences of the Cauchy Integral Formula, and it would be wise to spend a few minutes deriving other similar results.

Exercise 7.4. (a) Let S and f be as in the preceding theorem, and assume that $f(z) = 0$ for every point on the boundary C_S of S. Prove that $f(z) = 0$ for every $z \in S$.
(b) Let S be as in part (a), and suppose that f and g are two continuous functions on S, both differentiable on S^0, and such that $f(\zeta) = g(\zeta)$ for every point on the boundary of S. Prove that $f(z) = g(z)$ for all $z \in S$.

The preceding exercise shows that two differentiable functions of a complex variable are equal everywhere on a piecewise smooth geometric set S if they agree on the boundary of the set. More is true. We will see below in the Identity Theorem that they are equal everywhere on a piecewise smooth geometric set S if they agree just along a single convergent sequence in the interior of S.
Combining part (b) of Exercise 7.3, Exercise 6.20, and Theorem 7.3, we obtain the following corollary:

COROLLARY. *Let $S_1, \dots, S_n$ be pairwise disjoint, piecewise smooth geometric sets whose boundaries have finite length, all contained in the interior of a piecewise smooth geometric set S whose boundary has finite length. Suppose f is continuous at each point of S that is not in the interior of any of the S_k's, and that f is differentiable at each point of S^0 that is not an element of any of the S_k's. Then, for any $z \in S^0$ that is not an element of any of the S_k's, we have*

$$f(z) = \frac{1}{2\pi i}\left(\int_{C_S} \frac{f(\zeta)}{\zeta - z}\,d\zeta - \sum_{k=1}^{n} \int_{C_{S_k}} \frac{f(\zeta)}{\zeta - z}\,d\zeta\right).$$

PROOF. Let $r > 0$ be such that $\overline{B}_r(z)$ is disjoint from all the S_k's. By part (b) of Exercise 7.3, let $T_1, \dots, T_m$ be a partition of S such that $T_k = S_k$ for $1 \leq k \leq n$, and $T_{n+1} = \overline{B}_r(z)$. By Exercise 6.20, we know that

$$\int_{C_S} \frac{f(\zeta)}{\zeta - z}\,d\zeta = \sum_{k=1}^{m} \int_{C_{T_k}} \frac{f(\zeta)}{\zeta - z}\,d\zeta.$$

From the Cauchy Integral Formula, we know that

$$\int_{C_{T_{n+1}}} \frac{f(\zeta)}{\zeta - z}\,d\zeta = 2\pi i f(z).$$

Also, since $f(\zeta)/(\zeta - z)$ is differentiable at each point of the interior of the sets T_k for $k > n+1$, we have from Theorem 7.2 that for all $k > n+1$

$$\int_{C_{t_k}} \frac{f(\zeta)}{\zeta - z}\,d\zeta = 0.$$

Therefore,

$$\int_{C_S} \frac{f(\zeta)}{\zeta - z}\,d\zeta = \sum_{k=1}^{n} \int_{C_{S_k}} \frac{f(\zeta)}{\zeta - z}\,d\zeta + 2\pi i f(z),$$

which completes the proof.

Exercise 7.5. Suppose S is a piecewise smooth geometric set whose boundary has finite length, and let $c_1, \dots, c_n$ be points in S^0. Suppose f is a complex-valued function that is continuous at every point of S except the C_k's and differentiable at every point of S^0 except the c_k's. Let $r_1, \dots, r_n$ be positive numbers such that the disks $\{\overline{B}_{R_k}(c_k)\}$ are pairwise disjoint and all contained in S^0.
(a) Prove that

$$\int_{C_S} f(\zeta)\,d\zeta, = \sum_{k=1}^{n} \int_{C_k} f(\zeta)\,d\zeta$$

where C_k denotes the boundary of the disk $\overline{B}_{r_k}(c_k)$.
(b) For any $z \in S^0$ that is not in any of the closed disks $\overline{B}_{r_k}(c_k)$, show that

$$\int_{C_S} \frac{f(\zeta)}{\zeta - z}\,d\zeta = 2\pi i f(z) + \sum_{k=1}^{n} \int_{C_k} \frac{f(\zeta)}{\zeta - z}\,d\zeta.$$

(c) Specialize part (b) to the case where $S = \overline{B}_r(c)$, and f is analytic at each point of $B_r(c)$ except at the central point c. For each $z \neq c$ in $B_r(c)$, and any $0 < \delta < |z - c|$, derive the formula

$$f(z) = \frac{1}{2\pi i}\int_{C_r} \frac{f(\zeta)}{\zeta - z}\,d\zeta - \frac{1}{2\pi i}\int_{C_\delta} \frac{f(\zeta)}{\zeta - z}\,d\zeta.$$

BASIC APPLICATIONS OF THE CAUCHY INTEGRAL FORMULA

As a major application of the Cauchy Integral Formula, let us show the much alluded to remarkable fact that a function that is a differentiable function of a complex variable on an open set U is actually expandable in a Taylor series around every point in U, i.e., is an analytic function on U.

THEOREM 7.5. *Suppose f is a differentiable function of a complex variable on an open set $U \subseteq \mathbb{C}$, and let c be an element of U. Then f is expandable in a Taylor series around c. In fact, for any $r > 0$ for which $\overline{B}_r(c) \subseteq U$, we have*

$$f(z) = \sum_{n=0}^{\infty} a_n (z - c)^n$$

for all $z \in B_r(c)$.

PROOF. Choose an $r > 0$ such that the closed disk $\overline{B}_r(c) \subseteq U$, and write C_r for the boundary of this disk. Note that, for all points ζ on the curve C_r, and any fixed point z in the open disk $B_r(c)$, we have that $|z - c| < r = |\zeta - c|$, whence $|z - c|/|\zeta - c| = |z - c|/r < 1$. Therefore the geometric series

$$\sum_{n=0}^{\infty} \left(\frac{z - c}{\zeta - c}\right)^n \text{ converges to } \frac{1}{1 - \frac{z-c}{\zeta-c}}.$$

Moreover, by the Weierstrass M-Test, as functions of the variable ζ, this infinite series converges uniformly on the curve C_r. We will use this in the calculation below. Now, according to Theorem 7.4, we have that

$$\begin{aligned}
f(z) &= \frac{1}{2\pi i}\int_{C_r} \frac{f(\zeta)}{\zeta - z}\,d\zeta \\
&= \frac{1}{2\pi i}\int_{C_r} \frac{f(\zeta)}{\zeta - c + c - z}\,d\zeta \\
&= \frac{1}{2\pi i}\int_{C_r} \frac{f(\zeta)}{(\zeta - c)(1 - \frac{z-c}{\zeta-c})}\,d\zeta \\
&= \frac{1}{2\pi i}\int_{C_r} \frac{f(\zeta)}{\zeta - c}\sum_{n=0}^{\infty}\left(\frac{z - c}{\zeta - c}\right)^n\,d\zeta \\
&= \frac{1}{2\pi i}\int_{C_r} \sum_{n=0}^{\infty} \frac{f(\zeta)}{(\zeta - c)^{n+1}}(z - c)^n\,d\zeta \\
&= \frac{1}{2\pi i}\sum_{n=0}^{\infty}\int_{C_r} \frac{f(\zeta)}{(\zeta - c)^{n+1}}(z - c)^n\,d\zeta \\
&= \sum_{n=0}^{\infty} a_n (z - c)^n,
\end{aligned}$$

where we are able to bring the summation sign outside the integral by part (3) of Theorem 6.10, and where

$$a_n = \frac{1}{2\pi i}\int_{C_r} \frac{f(\zeta)}{(\zeta - c)^{n+1}}\, d\zeta.$$

This proves that f is expandable in a Taylor series around the point c, as desired.

Using what we know about the relationship between the coefficients of a Taylor series and the derivatives of the function, together with the Cauchy Integral Theorem, we obtain the following formulas for the derivatives of a differentiable function f of a complex variable. These are sometimes also called the Cauchy Integral Formulas.

COROLLARY. *Suppose f is a differentiable function of a complex variable on an open set U, and let c be an element of U. Then f is infinitely differentiable at c, and*

$$f^{(n)}(c) = \frac{n!}{2\pi i}\int_{C_s} \frac{f(\zeta)}{(\zeta - c)^{n+1}}\, d\zeta,$$

for any piecewise smooth geometric set $S \subseteq U$ whose boundary C_S has finite length, and for which c belongs to the interior S^0 of S.

Exercise 7.6. (a) Prove the preceding corollary.
(b) Let f, U, and c be as in Theorem 7.5. Show that the radius of convergence r of the Taylor series expansion of f around c is at least as large as the supremum of all s for which $B_s(c) \subseteq U$.
(c) Conclude that the radius of convergence of the Taylor series expansion of a differentiable function of a complex variable is as large as possible. That is, if f is differentiable on a disk $B_r(c)$, then the Taylor series expansion of f around c converges on all of $B_r(c)$.
(d) Consider the real-valued function of a real variable given by $f(x) = 1/(1+x^2)$. Show that f is differentiable at each real number x. Show that f is expandable in a Taylor series around 0, but show that the radius of convergence of this Taylor series is equal to 1. Does this contradict part (c)?
(e) Let f be the complex-valued function of a complex variable given by $f(z) = 1/(1+z^2)$. We have just replaced the real variable x of part (d) by a complex variable z. Explain the apparent contradiction that parts (c) and (d) present in connection with this function.

Exercise 7.7. (a) Let S be a piecewise smooth geometric set whose boundary C_S has finite length, and let f be a continuous function on the curve C_S. Define a function F on S^0 by

$$F(z) = \int_{C_S} \frac{f(\zeta)}{\zeta - z}\, d\zeta.$$

Prove that F is expandable in a Taylor series around each point $c \in S^0$. Show in fact that $F(z) = \sum a_n(z-c)^n$ for all z in a disk $B_r(c) \subseteq S^0$, where

$$a_n = \frac{n!}{2\pi i}\int_{C_S} \frac{f(\zeta)}{(\zeta - c)^{n+1}}\, d\zeta.$$

HINT: Mimic the proof of Theorem 7.5.

(b) Let f and F be as in part (a). Is F defined on the boundary C_S of S? If z belongs to the boundary C_S, and $z = \lim z_n$, where each $z_n \in S^0$, Does the sequence $\{F(z_n)\}$ converge, and, if so, does it converge to $f(z)$?
(c) Let S be the closed unit disk $\overline{B}_1(0)$, and let f be defined on the boundary C_1 of this disk by $f(z) = \bar{z}$, i.e., $f(x+iy) = x - iy$. Work out the function F of part (a), and then re-think about part (b).
(d) Let f and F be as in part (a). If, in addition, f is continuous on all of S and differentiable on S^0, show that $F(z) = 2\pi i f(z)$ for all $z \in S^0$. Think about this "magic" constant $2\pi i$. Review the proof of the Cauchy Integral Formula to understand where this constant comes from.

Theorem 3.14 and Exercise 3.26 constitute what we called the "identity theorem" for functions that are expandable in a Taylor series around a point c. An even stronger result than that is actually true for functions of a complex variable.

THEOREM 7.6. (Identity Theorem) Let f be a continuous complex-valued function on a piecewise smooth geometric set S, and assume that f is differentiable on the interior S^0 of S. Suppose $\{z_k\}$ is a sequence of distinct points in S^0 that converges to a point c in S^0. If $f(z_k) = 0$ for every K, then $f(z) = 0$ for every $z \in S$.

PROOF. It follows from Exercise 3.26 that there exists an $r > 0$ such that $f(z) = 0$ for all $z \in B_r(c)$. Now let w be another point in S^0, and let us show that $f(w)$ must equal 0. Using part (f) of Exercise 6.2, let $\phi : [\widehat{a}, \widehat{b}] \to C$ be a piecewise smooth curve, joining c to w, that lies entirely in S^0. Let A be the set of all $t \in [\widehat{a}, \widehat{b}]$ such that $f(\phi(s)) = 0$ for all $s \in [\widehat{a}, t)$. We claim first that A is nonempty. Indeed, because ϕ is continuous, there exists an $\epsilon > 0$ such that $|\phi(s) - c| = |\phi(s) - \phi(\widehat{a})| < r$ if $|s - \widehat{a}| < \epsilon$. Therefore $f(\phi(s)) = 0$ for all $s \in [\widehat{a}, \widehat{a} + \epsilon)$, whence, $\widehat{a} + \epsilon \in A$. Obviously, A is bounded above by $\widehat{b}$, and we write t_0 for the supremum of A. We wish to show that $t_0 = \widehat{b}$, whence, since ϕ is continuous at $\widehat{B}$, $f(w) = f(\phi(\widehat{b})) = f(\phi(t_0)) = 0$. Suppose, by way of contradiction, that $t_0 < \widehat{b}$, and write $z_0 = \phi(t_0)$. Now $z_0 \in S^0$, and $z_0 = \lim \phi(t_0 - 1/k)$ because ϕ is continuous at t_0. But $f(\phi(t_0 - 1/k)) = 0$ for all k. So, again using Exercise 3.26, we know that there exists an $r' > 0$ such that $f(z) = 0$ for all $z \in B_{r'}(z_0)$. As before, because ϕ is continuous at t_0, there exists a $\delta > 0$ such that $t_0 + \delta < \widehat{b}$ and $|\phi(s) - \phi(t_0)| < r'$ if $|s - t_0| < \delta$. Hence, $f(\phi(s)) = 0$ for all $s \in (t_0 - \delta, t_0 + \delta)$, which implies that $t_0 + \delta$ belongs to A. But then t_0 could not be the supremum of A, and therefore we have arrived at a contradiction. Consequently, $t_0 = \widehat{b}$, and therefore $f(w) = 0$ for all $w \in S^0$. Of course, since every point in S is a limit of points from S^0, and since f is continuous on S, we see that $f(z) = 0$ for all $z \in S$, and the theorem is proved.

The next exercise gives some consequences of the Identity Theorem. Part (b) may appear to be a contrived example, but it will be useful later on.

Exercise 7.8. (a) Suppose f and g are two functions, both continuous on a piecewise smooth geometric set S and both differentiable on its interior. Suppose $\{z_k\}$ is a sequence of elements of S^0 that converges to a point $c \in S^0$, and assume that $f(z_k) = g(z_k)$ for all k. Prove that $f(z) = g(z)$ for all $z \in S$.
(b) Suppose f is a nonconstant differentiable function defined on the interior of a piecewise smooth geometric set S. If $c \in S^0$ and $B_\epsilon(c) \subseteq S^0$, show that there must exist an $0 < r < \epsilon$ for which $f(c) \neq f(z)$ for all z on the boundary of the disk $B_r(c)$.

THE FUNDAMENTAL THEOREM OF ALGEBRA

We can now prove the Fundamental Theorem of Algebra, the last of our primary goals. One final trumpet fanfare, please!

THEOREM 7.7. (Fundamental Theorem of Algebra) Let $p(z)$ be a nonconstant polynomial of a complex variable. Then there exists a complex number z_0 such that $p(z_0) = 0$. That is, every nonconstant polynomial of a complex variable has a root in the complex numbers.

PROOF. We prove this theorem by contradiction. Thus, suppose that p is a nonconstant polynomial of degree $n \geq 1$, and that $p(z)$ is never 0. Set $f(z) = 1/p(z)$, and observe that f is defined and differentiable at every point $z \in \mathbb{C}$. We will show that f is a constant function, implying that $p = 1/f$ is a constant, and that will give the contradiction. We prove that f is constant by showing that its derivative is identically 0, and we compute its derivative by using the Cauchy Integral Formula for the derivative.
From part (4) of Theorem 3.1, we recall that there exists a $B > 0$ such that $\frac{|c_n|}{2}|z|^n \leq |p(z)|$, for all z for which $|z| \geq B$, and where c_n is the (nonzero) leading coefficient of the polynomial p. Hence, $|f(z)| \leq \frac{M}{|z|^n}$ for all $|z| \geq B$, where we write M for $2/|c_n|$. Now, fix a point $c \in \mathbb{C}$. Because f is differentiable on the open set $U = \mathbb{C}$, we can use the corollary to Theorem 7.4 to compute the derivative of f at c by using any of the curves C_r that bound the disks $B_r(c)$, and we choose an r large enough so that $|c + re^{it}| \geq B$ for all $0 \leq t \leq 2\pi$. Then,

$$\begin{aligned}
|f'(c)| &= |\frac{1}{2\pi i}\int_{C_r} \frac{f(\zeta)}{(\zeta - c)^2}\,d\zeta| \\
&= \frac{1}{2\pi}|\int_0^{2\pi} \frac{f(c+re^{it})}{(c+re^{it}-c)^2} ire^{it}\,dt| \\
&\leq \frac{1}{2\pi r}\int_0^{2\pi} |f(c+re^{it})|\,dt \\
&\leq \frac{1}{2\pi r}\int_0^{2\pi} \frac{M}{|c+re^{it}|^n}\,dt \\
&\leq \frac{M}{rB^n}.
\end{aligned}$$

Hence, by letting r tend to infinity, we get that

$$|f'(c)| \leq \lim_{r\to\infty} \frac{M}{rB^n} = 0,$$

and the proof is complete.

REMARK. The Fundamental Theorem of Algebra settles a question first raised back in Chapter I. There, we introduced a number I that was a root of the polynomial $x^2 + 1$. We did this in order to build a number system in which negative numbers would have square roots. We adjoined the "number" i to the set of real numbers to form the set of complex numbers, and we then saw that in fact every complex number z has a square root. However, a fear was that, in order to build a system in which every number has an nth root for every n, we would continually need to be adjoining new elements to our number system. However, the Fundamental Theorem of Algebra shows that this is not necessary. The set of complex numbers is already rich enough to contain all nth roots and even more.

Practically the same argument as in the preceding proof establishes another striking result.

THEOREM 7.8. (Liouville) Suppose f is a bounded, everywhere differentiable function of a complex variable. Then f must be a constant function.

Exercise 7.9. Prove Liouville's Theorem.

THE MAXIMUM MODULUS PRINCIPLE

Our next goal is to examine so-called "max/min" problems for coplex-valued functions of complex variables. Since order makes no sense for complex numbers, we will investigate max/min problems for the absolute value of a complex-valued function. For the corresponding question for real-valued functions of real variables, we have as our basic result the First Derivative Test (Theorem 4.8). Indeed, when searching for the poinhts where a differentiable real-valued function f on an interval $[a, b]$ attains its extreme values, we consider first the poinhts where it attains a local max or min, to which purpose end Theorem 4.8 is useful. Of course, to find the absolute minimum and maximum, we must also check the values of the function at the endpoints.

An analog of Theorem 4.8 holds in the complex case, but in fact a much different result is really valid. Indeed, it is nearly impossible for the absolute value of a differentiable function of a complex variable to attain a local maximum or minimum.

THEOREM 7.9. *Let f be a continuous function on a piecewise smooth geometric set S, and assume that f is differentiable on the interior S^0 of S. Suppose c is a point in S^0 at which the real-valued function $|f|$ attains a local maximum. That is, there exists an $\epsilon > 0$ such that $|f(c)| \geq |f(z)|$ for all z satisfying $|z - c| < \epsilon$. Then f is a constant function on S; i.e., $f(z) = f(c)$ for all $z \in S$. In other words, the only differentiable functions of a complex variable, whose absolute value attains a local maximum on the interior of a geometric set, are constant functions on that set.*

PROOF. If $f(c) = 0$, then $f(z) = 0$ for all $z \in B_\epsilon(c)$. Hence, by the Identity Theorem (Theorem 7.6), $f(z)$ would equal 0 for all $z \in S$. so, we may as well assume that $f(c) \neq 0$. Let r be any positive number for which the closed disk $\overline{B}_r(c)$ is contained in $B_\epsilon(c)$. We claim first that there exists a point z on the boundary C_r of the disk $\overline{B}_r(c)$ for which $|f(z)| = |f(c)|$. Of course, $|f(z| \leq |f(c)|$ for all z on this boundary by assumption. By way of contradiction, suppose that $|f(\zeta)| < |f(c)|$ for all ζ on the boundary C_r of the disk. Write M for the maximum value of the function $|f|$ on the compact set C_r. Then, by our assumption, $M < |f(c)|$. Now, we use the Cauchy Integral Formula:

$$\begin{aligned}|f(c)| &= |\frac{1}{2\pi i}\int_{C_r}\frac{f(\zeta)}{\zeta - c}\,d\zeta| \\ &= \frac{1}{2\pi}|\int_0^{2\pi}\frac{f(c+re^{it})}{re^{it}}ire^{it}\,dt| \\ &\le \frac{1}{2\pi}\int_0^{2\pi}|f(c+re^{it})|\,dt \\ &\le \frac{1}{2\pi}\int_0^{2\pi}M\,dt \\ &= M \\ &< |f(c)|,\end{aligned}$$

and this is a contradiction.
Now for each natural number n for which $1/n < \epsilon$, let z_n be a point for which $|z_n - c| = 1/n$ and $|f(z_n)| = |f(c)|$. We claim that the derivative $f'(z_n)$ of f at $z_n = 0$ for all n. What we know is that the real-valued function $F(x,y) = |f(x+iy)|^2 = (u(x,y)^2 + (v(x,y))^2$ attains a local maximum value at $z_n = (x_n, y_n)$. Hence, by Exercise 4.34, both partial derivatives of F must be 0 at (x_n, y_n). That is

$$2u(x_n,y_n)\frac{\partial u}{\partial x}(x_n,y_n) + 2v(x_n,y_n)\frac{\partial v}{\partial x}(x_n,y_n) = 0$$

and

$$2u(x_n,y_n)\frac{\partial u}{\partial y}(x_n,y_n) + 2v(x_n,y_n)\frac{\partial v}{\partial y}(x_n,y_n) = 0.$$

Hence the two vectors

$$\vec{V_1} = (\frac{\partial u}{\partial x}(x_n,y_n), \frac{\partial v}{\partial x}(x_n,y_n))$$

and

$$\vec{V_2} = (\frac{\partial u}{\partial y}(x_n,y_n), \frac{\partial v}{\partial y}(x_n,y_n))$$

are both perpendicular to the vector $\vec{V_3} = (u(x_n,y_n), v(x_n,y_n))$. But $\vec{V_3} \neq 0$, because $\|\vec{V_3}\| = |f(z_n)| = |f(c)| > 0$, and hence $\vec{V_1}$ and $\vec{V_2}$ are linearly dependent. But this implies that $f'(z_n) = 0$, according to Theorem 7.2.
Since $c = \lim z_n$, and f' is analytic on S^0, it follows from the Identity Theorem that there exists an $r > 0$ such that $f'(z) = 0$ for all $z \in B_r(c)$. But this implies that f is a constant $f(z) = f(c)$ for all $z \in B_r(c)$. And thenm, again using the Identity Theorem, this implies that $f(z) = f(c)$ for all $z \in S$, which completes the proof.

REMARK. Of course, the preceding proof contains in it the verification that if $|f|$ attains a maximum at a point c where it is differentiable, then $f'(c) = 0$. This is the analog for functions of a complex variable of Theorem 4.8. But, Theorem 7.9 certainly asserts a lot more than that. In fact, it says that it is impossible for the absolute value of a nonconstant differentiable function of a complex variable to attain a local maximum. Here is the coup d'grâs:

COROLLARY. (Maximum Modulus Principle) Let f be a continuous, nonconstant, complex-valued function on a piecewise smooth geometric set S, and suppose that f is differentiable on the interior S^0 of S. Let M be the maximum value of the continuous, real-valued function $|f|$ on S, and let z be a point in S for which $|f(z)| = M$. Then, z does not belong to the interior S^0 of S; it belongs to the boundary of S. In other words, $|f|$ attains its maximum value only on the boundary of S.

Exercise 7.10. (a) Prove the preceding corollary.
(b) Let f be an analytic function on an open set U, and let $c \in U$ be a point at which $|f|$ achieves a local minimum; i.e., there exists an $\epsilon > 0$ such that $|f(c)| \leq |f(z)|$ for all $z \in B_\epsilon(c)$. Show that, if $f(c) \neq 0$, then f is constant on $B_\epsilon(c)$. Show by example that, if $f(c) = 0$, then f need not be a constant on $B_\epsilon(c)$.
(c) Prove the "Minimum Modulus Principle:" Let f be a nonzero, continuous, nonconstant, function on a piecewise smooth geometric set S, and let m be the minimum value of the function $|f|$ on S. If z is a point of S at which this minimum value is atgtained, then z belongs to the boundary C_S of S.

THE OPEN MAPPING THEOREM AND THE INVERSE FUNCTION THEOREM

We turn next to a question about functions of a complex variable that is related to Theorem 4.10, the Inverse Function Theorem. That result asserts, subject to a couple of hypotheses, that the inverse of a one-to-one differentiable function of a real variable is also differentiable. Since a function is only differentiable at points in the interior of its domain, it is necessary to verify that the point $f(c)$ is in the interior of the domain $f(S)$ of the inverse function f^{-1} before the question of differentiability at that point can be addressed. And, the peculiar thing is that it is this point about $f(c)$ being in the interior of $f(S)$ that is the subtle part. The fact that the inverse function is differentiable there, and has the prescribed form, is then only a careful $\epsilon - \delta$ argument. For continuous real-valued functions of real variables, the fact that $f(c)$ belongs to the interior of $f(S)$ boils down to the fact that intervals get mapped onto intervals by continuous functions, which is basically a consequence of the Intermediate Value Theorem. However, for complex-valued functions of complex variables, the situation is much deeper. For instance, the continuous image of a disk is just not always another disk, and it may not even be an open set. Well, all is not lost; we just have to work a little harder.

THEOREM 7.10. (Open Mapping Theorem) Let S be a piecewise smooth geometric set, and write U for the (open) interior S^0 of S. Suppose f is a nonconstant differentiable, complex-valued function on the set U. Then the range $f(U)$of f is an open subset of $\mathbb{C}$.

PROOF. Let c be in U. Because f is not a constant function, there must exist an $r > 0$ such that $f(c) \neq f(z)$ for all z on the boundary C_r of the disk $B_r(c)$. See part (b) of Exercise 7.8. Let z_0 be a point in the compact set C_r at which the continuous real-valued function $|f(z) - f(c)|$ attains its minimum value s. Since $f(z) \neq f(c)$ for any $z \in C_r$, we must have that $s > 0$. We claim that the disk $B_{s/2}(f(c))$ belongs to the range $f(U)$ of f. This will show that the point $f(c)$ belongs to the ihnterior of the set $f(U)$, and that will finish the proof.
By way of contradiction, suppose $B_{s/2}(f(c)$ is not contained in $f(U)$,, and let $w \in B_{s/2}(f(c))$ be a complex number that is not in $f(U)$. We have that $|w - f(c)| < s/2$,

which implies that $|w - f(z)| > s/2$ for all $z \in C_r$. Consider the function g defined on the closed disk $\overline{B}_r(c)$ by $g(z) = 1/(w - f(z))$. Then g is continuous on the closed disk $\overline{B}_r(c)$ and differentiable on $B_r(c)$. Moreover, g is not a constant function, for if it were, f would also be a constant function on $B_r(c)$ and therefore, by the Identity Theorem, constant on all of U, whichg is not the case by hypothesis. Hence, by the Maximum Modulus Principle, the maximum value of $|g|$ only occurs on the boundary C_r of this disk. That is, there exists a point $z' \in C_r$ such that $|g(z)| < |g(z')|$ for all $z \in B_r(c)$. But then

$$\frac{2}{s} = \frac{1}{s/2} < \frac{1}{|w - f(c)|} < \frac{1}{|w - f(z')|} \leq \frac{1}{s},$$

which gives the desired contradiction. Therefore, the entire disk $B_{s/2}(f(c))$ belongs to $f(U)$, and hence the point $f(c)$ belongs to the interior of the set $f(U)$. Since this holds for any point $c \in U$, it follows that $f(U)$ is open, as desired.

Now we can give the version of the Inverse Function Theorem for complex variables.

THEOREM 7.11. *Let S be a piecewise smooth geometric set, and suppose $f : S \to \mathbb{C}$ is continuously differentiable at a point $c = a+bi$, and assume that $f'(c) \neq 0$. Then:*

1. *There exists an $r > 0$, such that $\overline{B}_r(c) \subseteq S$, for which f is one-to-one on $\overline{B}_r(c)$.*
2. *$f(c)$ belongs to the interior of $f(S)$.*
3. *If g denotes the restriction of the function f to $B_r(c)$, then g is one-to-one, g^{-1} is differentiable at the point $f(c)$, and ${g^{-1}}'(f(c) = 1/f'(c)$.*

PROOF. Arguing by contradiction, suppose that f is not one-to-one on any disk $\overline{B}_r(c)$. Then, for each natural number n, there must exist two points $z_n = x_n + iy_n$ and $z'_n = x'_n + iy'_n$ such that $|z_n - c| < 1/n$, $|z'_n - c| < 1/n$, and $f(z_n) = f(z'_n)$. If we write $f = u + iv$, then we would have that $u(x_n, y_n) - u(x'_n, y'_n) = 0$ for all n. So, by part (c) of Exercise 4.35, there must exist for each n a point $(\widehat{x}_n, \widehat{y}_n)$, such that $(\widehat{x}_n, \widehat{y}_n)$ is on the line segment joining z_n and z'_n, and for which

$$0 = u(x_n, y_n) - u(x'_n, y'_n) = \frac{\partial u}{\partial x}(\widehat{x}_n, \widehat{y}_n)(x_n - x'_n) + \frac{\partial u}{\partial y}(\widehat{x}_n, \widehat{y}_n)(y_n - y'_n).$$

Similarly, applying the same kind of reasoning to v, there must exist points $(\widetilde{x}_n, \widetilde{y}_n)$ on the segment joining z_n to z'_n such that

$$0 = \frac{\partial v}{\partial x}(\widetilde{x}_n, \widetilde{y}_n)(x_n - x'_n) + \frac{\partial v}{\partial y}(\widetilde{x}_n, \widetilde{y}_n)(y_n - y'_n).$$

If we define vectors $\vec{U}_n$ and $\vec{V}_n$ by

$$\vec{U}_n = (\frac{\partial u}{\partial x}(\widehat{x}_n, \widehat{y}_n), \frac{\partial u}{\partial y}(\widehat{x}_n, \widehat{y}_n))$$

and

$$\vec{V}_n = (\frac{\partial v}{\partial x}(\widetilde{x}_n, \widetilde{y}_n), \frac{\partial v}{\partial y}(\widetilde{x}_n, \widetilde{y}_n)),$$

then we have that both $\vec{U}_n$ and $\vec{V}_n$ are perpendicular to the nonzero vector $((x_n - x'_n), (y_n - y'_n))$. Therefore, $\vec{U}_n$ and $\vec{V}_n$ are linearly dependent, whence

$$\det(\begin{pmatrix} \frac{\partial u}{\partial x}(\widehat{x}_n, \widehat{y}_n) & \frac{\partial u}{\partial y}(\widehat{x}_n, \widehat{y}_n) \\ \frac{\partial v}{\partial x}(\widetilde{x}_n, \widetilde{y}_n) & \frac{\partial v}{\partial y}(\widetilde{x}_n, \widetilde{y}_n) \end{pmatrix}) = 0.$$

Now, since both $\{\widehat{x}_n + i\widehat{y}_n\}$ and $\{\widetilde{x}_n + i\widetilde{y}_n\}$ converge to the point $c = a + ib$, and the partial derivatives of u and v are continuous at c, we deduce that

$$\det(\begin{pmatrix} \frac{\partial u}{\partial x}(a, b) & \frac{\partial u}{\partial y}(a, b) \\ \frac{\partial v}{\partial x}(a, b) & \frac{\partial v}{\partial y}(a, b) \end{pmatrix}) = 0.$$

Now, from Theorem 7.2, this would imply that $f'(c) = 0$, and this is a contradiction. Hence, there must exist an $r > 0$ for which f is one-to-one on $\overline{B}_r(c)$, and this proves part (1).

Because f is one-to-one on $B_r(c)$, f is obviously not a constant function. So, by the Open Mapping Theorem, the point $f(c)$ belongs to the interior of the range of f, and this proves part (2).

Now write g for the restriction of f to the disk $B_r(c)$. Then g is one-to-one. According to part (2) of Theorem 4.2, we can prove that g^{-1} is differentiable at $f(c)$ by showing that

$$\lim_{z \to f(c)} \frac{g^{-1}(z) - g^{-1}(f(c))}{z - f(c)} = \frac{1}{f'(c)}.$$

That is, we need to show that, given an $\epsilon > 0$, there exists a $\delta > 0$ such that if $0 < |z - f(c)| < \delta$ then

$$|\frac{g^{-1}(z) - g^{-1}(f(c))}{z - f(c)} - \frac{1}{f'(c)}| < \epsilon.$$

First of all, because the function $1/w$ is continuous at the point $f'(c)$, there exists an $\epsilon' > 0$ such that if $|w - f'(c)| < \epsilon'$, then

$$|\frac{1}{w} - \frac{1}{f'(c)}| < \epsilon.$$

Next, because f is differentiable at c, there exists a $\delta' > 0$ such that if $0 < |y - c| < \delta'$ then

$$|\frac{f(y) - f(c)}{y - c} - f'(c)| < \epsilon'.$$

Now, by Theorem 3.10, g^{-1} is continuous at the point $f(c)$, and therefore there exists a $\delta > 0$ such that if $|z - f(c)| < \delta$ then

$$|g^{-1}(z) - g^{-1}(f(c)| < \delta'.$$

So, if $|z - f(c)| < \delta$, then

$$|g^{-1}(z) - c| = |g^{-1}(z) - g^{-1}(f(c))| < \delta'.$$

But then,

$$|\frac{f(g^{-1}(z)) - f(c)}{g^{-1}(z) - c} - f'(c)| < \epsilon',$$

from which it follows that

$$|\frac{g^{-1}(z) - g^{-1}(f(c))}{z - f(c)} - \frac{1}{f'(c)}| < \epsilon,$$

as desired.

UNIFORM CONVERGENCE OF ANALYTIC FUNCTIONS

Part (c) of Exercise 4.26 gives an example showing that the uniform limit of a sequence of differentiable functions of a real variable need not be differentiable. Indeed, when thinking about uniform convergence of functions, the fundamental result to remember is that the uniform limit of continuous functions is continuous (Theorem 3.17). The functions in Exercise 4.26 were differentiable functions of a real variable. The fact is that, for functions of a complex variable, things are as usual much more simple. The following theorem is yet another masterpiece of Weierstrass.

THEOREM 7.12. *Suppose U is an open subset of $\mathbb{C}$, and that $\{f_n\}$ is a sequence of analytic functions on U that converges uniformly to a function f. Then f is analytic on U. That is, the uniform limit of differentiable functions on an open set U in the complex plane is also differentiable on U.*

PROOF. Though this theorem sounds impressive and perhaps unexpected, it is really just a combination of Theorem 6.10 and the Cauchy Integral Formula. Indeed, let c be a point in U, and let $r > 0$ be such that $\overline{B}_r(c) \subseteq U$. Then the sequence $\{f_n\}$ converges uniformly to f on the boundary C_r of this closed disk. Moreover, for any $z \in B_r(c)$, the sequence $\{f_n(\zeta)/(\zeta - z)\}$ converges uniformly to $f(\zeta)/(\zeta - z)$ on C_r. Hence, by Theorem 6.10, we have

$$\begin{aligned} f(z) &= \lim f_n(z) \\ &= \lim_n \frac{1}{2\pi i} \int_{C_r} \frac{f_n(\zeta)}{\zeta - z}\, d\zeta \\ &= \frac{1}{2\pi i} \int_{C_r} \frac{f(\zeta)}{\zeta - z}\, d\zeta. \end{aligned}$$

Hence, by part (a) of Exercise 7.7, f is expandable in a Taylor series around c, i.e., f is analytic on U.

ISOLATED SINGULARITIES, AND THE RESIDUE THEOREM

The first result we present in this section is a natural extension of Theorem 7.3. However, as we shall see, its consequences for computing contour integrals can hardly be overstated.

THEOREM 7.13. *Let S be a piecewise smooth geometric set whose boundary C_S has finite length. Suppose $c_1, \dots, c_n$ are distinct points in the interior S^0 of S, and that $r_1, \dots, r_n$ are positive numbers such that the closed disks $\{\overline{B}_{r_k}(c_k)\}$ are contained in S^0 and pairwise disjoint. Suppose f is continuous on $S \setminus \cup B_{r_k}(c_k)$, i.e., at each point of S that is not in any of the open disks $B_{r_k}(c_k)$, and that f is differentiable on $S^0 \setminus \cup\overline{B}_{r_k}(c_k)$, i.e., at each point of S^0 that is not in any of the closed disks $\overline{B}_{r_k}(c_k)$. Write C_k for the circle that is the boundary of the closed disk $\overline{B}_{r_k}(c_k)$. Then*

$$\int_{C_S} f(\zeta)\,d\zeta = \sum_{k=1}^{n} \int_{C_k} f(\zeta)\,d\zeta.$$

PROOF. This is just a special case of part (d) of Exercise 7.3.

Let f be continuous on the punctured disk $\overline{B'}_r(c)$, analytic at each point z in $B'_r(c)$, and suppose f is undefined at the central point c. Such points c are called *isolated singularities* of f, and we wish now to classify these kinds of points. Here is the first kind:

DEFINITION. A complex number c is called a *removable singularity* of an analytic function f if there exists an $r > 0$ such that f is continuous on the punctured disk $\overline{B'}_r(c)$, analytic at each point in $B'_r(c)$, and $\lim_{z\to c} f(z)$ exists.

Exercise 7.11. (a) Define $f(z) = \sin z/z$ for all $z \neq 0$. Show that 0 is a removable singularity of f.
(b) For $z \neq c$, define $f(z) = (1 - \cos(z - c))/(z - c)$. Show that c is a removable singularity of f.
(c) For $z \neq c$, define $f(z) = (1-\cos(z-c))/(z-c)^2$. Show that c is still a removable singularity of f.
(d) Let g be an analytic function on $B_r(c)$, and set $f(z) = (g(z) - g(c))/(z - c)$ for all $z \in B'_r(c)$. Show that c is a removable singularity of f.

The following theorem provides a good explanation for the term "removable singularity." The idea is that this is not a "true" singularity; it's just that for some reason the natural definition of f at c has not yet been made.

THEOREM 7.14. *Let f be continuous on the punctured disk $\overline{B}'_r(c)$ and differentiable at each point of the open punctured disk $B'_r(c)$, and assume that c is a removable singularity of f. Define $\widetilde{f}$ by $\widetilde{f}(z) = f(z)$ for all $z \in B'_r(c)$, and $\widetilde{f}(c) = \lim_{z\to c} f(z)$. Then*

(1) *$\widetilde{f}$ is analytic on the entire open disk $B_r(c)$, whence*

$$f(z) = \sum_{k=0}^{\infty} c_k (z - c)^k$$

for all $z \in B'_r(c)$.

(2) *For any piecewise smooth geometric set $S \subseteq B_r(c)$, whose boundary C_S has finite length, and for which $c \in S^0$,*

$$\int_{C_S} f(\zeta)\,d\zeta = 0.$$

PROOF. As in part (a) of Exercise 7.7, define F on $B_r(c)$ by

$$F(z) = \frac{1}{2\pi i}\int_{C_r} \frac{f(\zeta)}{\zeta - z}\, d\zeta.$$

Then, by that exercise, F is analytic on $B_r(c)$. We show next that $F(z) = \widetilde{f}(z)$ on $B_r(c)$, and this will complete the proof of part (1).
Let z be a point in $B_r(c)$ that is not equal to c, and let $\epsilon > 0$ be given. Choose $\delta > 0$ such that $\delta < |z - c|/2$ and such that $|\widetilde{f}(\zeta) - \widetilde{f}(c)| < \epsilon$ if $|\zeta - c| < \delta$. Then, using part (c) of Exercise 7.5, we have that

$$\begin{aligned}
\widetilde{f}(z) &= f(z)\\
&= \frac{1}{2\pi i}\int_{C_r} \frac{f(\zeta)}{\zeta - z}\, d\zeta - \frac{1}{2\pi i}\int_{C_\delta} \frac{f(\zeta)}{\zeta - z}\, d\zeta\\
&= F(z) - \frac{1}{2\pi i}\int_{C_\delta} \frac{f(\zeta) - \widetilde{f}(c)}{\zeta - z}\, d\zeta - \frac{1}{2\pi i}\int_{C_\delta} \frac{\widetilde{f}(c)}{\zeta - z}\, d\zeta\\
&= F(z) - \frac{1}{2\pi i}\int_{C_\delta} \frac{\widetilde{f}(\zeta) - \widetilde{f}(c)}{\zeta - z}\, d\zeta,
\end{aligned}$$

where the last equality holds because the function $\widetilde{f}(c)/(\zeta - z)$ is an analytic function of ζ on the disk $B_\delta(c)$, and hence the integral is 0 by Theorem 7.3. So,

$$\begin{aligned}
|\widetilde{f}(z) - F(z)| &= |\frac{1}{2\pi i}\int_{C_\delta} \frac{\widetilde{f}(\zeta) - \widetilde{f}(c)}{\zeta - z}\, d\zeta|\\
&\leq \frac{1}{2\pi}\int_{C_\delta} \frac{|\widetilde{f}(\zeta) - \widetilde{f}(c)|}{|\zeta - z|}\, ds\\
&\leq \frac{1}{2\pi}\int_{C_\delta} \frac{\epsilon}{\delta/2}\, ds\\
&= \frac{2\epsilon}{\delta} \times \delta\\
&= 2\epsilon.
\end{aligned}$$

Since this holds for arbitrary $\epsilon > 0$, we see that $\widetilde{f}(z) = F(z)$ for all $z \neq c$ in $B_r(c)$. Finally, since

$$\widetilde{f}(c) = \lim_{z \to c} \widetilde{f}(z) = \lim z \to cF(z) = F(c),$$

the equality of F and $\widetilde{f}$ on all of $B_r(c)$ is proved. This finishes the proof of part (1).

Exercise 7.12. Prove part (2) of the preceding theorem.

Now, for the second kind of isolated singularity:

DEFINITION. A complex number c is called a *pole* of a function f if there exists an $r > 0$ such that f is continuous on the punctured disk $\overline{B'}_r(c)$, analytic at each point of $B'_r(c)$, the point c is not a removable singularity of f, and there exists

a positive integer k such that the analytic function $(z-c)^k f(z)$ has a removable singularity at c.
A pole c of f is said to be *of order n*, if n is the smallest positive integer for which the function $\widetilde{f}(z) \equiv (z-c)^n f(z)$ has a removable singularity at c.

Exercise 7.13. (a) Let c be a pole of order n of a function f, and write $\widetilde{f}(z) = (z-c)^n f(z)$. Show that $\widetilde{f}$ is analytic on some disk $B_r(c)$.
(b) Define $f(z) = \sin z/z^3$ for all $z \neq 0$. Show that 0 is a pole of order 2 of f.

THEOREM 7.15. *Let f be continuous on a punctured disk $\overline{B'}_r(c)$, analytic at each point of $B'_r(c)$, and suppose that c is a pole of order n of f. Then*

(1) *For all $z \in B'_r(c)$,*

$$f(z) = \sum_{k=-n}^{\infty} a_k (z-c)^k.$$

(2) *The infinite series of part (1) converges uniformly on each compact subset K of $B'_r(c)$.*

(3) *For any piecewise smooth geometric set $S \subseteq B_r(c)$, whose boundary C_S has finite length, and satisfying $c \in S^0$,*

$$\int_{C_S} f(\zeta)\, d\zeta = 2\pi i a_{-1},$$

where A_{-1} is the coefficient of $(z-c)^{-1}$ in the series of part (1).

PROOF. For each $z \in B'_r(c)$, write $\widetilde{f}(z) = (z-c)^n f(z)$. Then, by Theorem 7.14, $\widetilde{f}$ is analytic on $B_r(c)$, whence

$$\begin{aligned} f(z) &= \frac{\widetilde{f}(z)}{(z-c)^n} \\ &= \frac{1}{(z-c)^n} \sum_{k=0}^{\infty} c_k (z-c)^k \\ &= \sum_{k=-n}^{\infty} a_k (z-c)^k, \end{aligned}$$

where $a_k = c_{n+k}$. This proves part (1).
We leave the proof of the uniform convergence of the series on each compact subset of $B'_r(c)$, i.e., the proof of part (2), to the exercises.
Part (3) follows from Cauchy's Theorem (Theorem 7.3) and the computations in Exercise 7.2. Thus:

$$\begin{aligned} \int_{C_S} f(\zeta)\, d\zeta &= \int_{C_r} f(\zeta)\, d\zeta \\ &= \int_{C_r} \sum_{k=-n}^{\infty} a_k (z-c)^k \, d\zeta \\ &= \sum_{k=-n}^{\infty} a_k \int_{C_r} (\zeta - c)^k \, d\zeta \\ &= a_{-1} 2\pi i, \end{aligned}$$

as desired. The summation sign comes out of the integral because of the uniform convergence of the series on the compact circle C_r.

Exercise 7.14. (a) Complete the proof to part (2) of the preceding theorem. That is, show that the infinite series $\sum_{k=-n}^{\infty} a_k(z-c)^k$ converges uniformly on each compact subset K of $B'_r(c)$.
HINT: Use the fact that the Taylor series $\sum_{n=0}^{\infty} c_n(z-c)^n$ for $\widetilde{f}$ converges uniformly on the entire disk $\overline{B}_r(c)$, and that if c is not in a compact subset K of $B_r(c)$, then there exists a $\delta > 0$ such that $|z-c| > \delta$ for all $z \in K$.
(b) Let f, c, and $\widetilde{f}$ be as in the preceding proof. Show that

$$a_{-1} = \frac{\widetilde{f}^{(n-1)}(c)}{(n-1)!}.$$

(c) Suppose g is a function defined on a punctured disk $B'_r(c)$ that is given by the formula

$$g(z) = \sum_{k=-n}^{\infty} a_k(z-c)^k$$

for some positive integer n and for all $z \in B'_r(c)$. Suppose in addition that the coefficient $a_{-n} \neq 0$. Show that c is a pole of order n of g.

Having defined two kinds of isolated singularities of a function f, the removable ones and the polls of finite order, there remain all the others, which we collect into a third type.

DEFINITION. Let f be continuous on a punctured disk $\overline{B'}_r(c)$, and analytic at each point of $B'_r(c)$. The point c is called an *essential singularity* of f if it is neither a removable singularity nor a poll of any finite order. Singularities that are either poles or essential singularities are called *nonremovable singularities.*

Exercise 7.15. For $z \neq 0$, define $f(z) = e^{1/z}$. Show that 0 is an essential singularity of f.

THEOREM 7.16. *Let f be continuous on a punctured disk $\overline{B'}_r(c)$, analytic at each point of $B'_r(c)$, and suppose that c is an essential singularity of f. Then*

(1) *For all $z \in B'_r(c)$,*

$$f(z) = \sum_{k=-\infty}^{\infty} a_k(z-c)^k,$$

where the sequence $\{a_k\}_{-\infty}^{\infty}$ has the property that for any negative integer N there is a $k < N$ such that $a_k \neq 0$.

(2) *The infinite series in part (1) converges uniformly on each compact subset K of $B'_r(c)$. That is, if F_n is defined by $F_n(z) = \sum_{k=-n}^{n} a_k(z-c)^k$, then the sequence $\{F_n\}$ converges uniformly to f on the compact set K.*

(3) *For any piecewise smooth geometric set $S \subseteq B_r(c)$, whose boundary C_S has finite length, and satisfying $c \in S^0$, we have*

$$\int_{C_S} f(\zeta)\, d\zeta = 2\pi i a_{-1},$$

where a_{-1} is the coefficient of $(z-c)^{-1}$ in the series of part (1).

PROOF. Define numbers $\{a_k\}_{-\infty}^{\infty}$ as follows.

$$a_k = \frac{1}{2\pi i}\int_{C_r} \frac{f(\zeta)}{(\zeta - c)^{k+1}}\, d\zeta.$$

Note that for any $0 < \delta < r$ we have from Cauchy's Theorem that

$$a_k = \frac{1}{2\pi i}\int_{C_\delta} \frac{f(\zeta)}{(\zeta - c)^{k+1}}\, d\zeta,$$

where C_δ denotes the boundary of the disk $\overline{B}_\delta(c)$.
Let $z \neq c$ be in $B_r(c)$, and choose $\delta > 0$ such that $\delta < |z - c|$. Then, using part (c) of Exercise 7.5, and then mimicking the proof of Theorem 7.5, we have

$$\begin{aligned}
f(z) &= \frac{1}{2\pi i}\int_{C_r} \frac{f(\zeta)}{\zeta - z}\, d\zeta - \frac{1}{2\pi i}\int_{C_\delta} \frac{f(\zeta)}{\zeta - z}\, d\zeta \\
&= \frac{1}{2\pi i}\int_{C_r} \frac{f(\zeta)}{(\zeta - c) - (z - c)}\, d\zeta + \frac{1}{2\pi i}\int_{C_\delta} \frac{f(\zeta)}{(z - c) - (\zeta - c)}\, d\zeta \\
&= \frac{1}{2\pi i}\int_{C_r} \frac{f(\zeta)}{\zeta - c}\frac{1}{1 - \frac{z-c}{\zeta-c}}\, d\zeta + \frac{1}{2\pi i}\int_{C_\delta} \frac{f(\zeta)}{z - c}\frac{1}{1 - \frac{\zeta-c}{z-c}}\, d\zeta \\
&= \frac{1}{2\pi i}\int_{C_r} \frac{f(\zeta)}{\zeta - c}\sum_{k=0}^{\infty}(\frac{z - c}{\zeta - c})^k\, d\zeta + \frac{1}{2\pi i}\int_{C_\delta} \frac{f(\zeta)}{z - c}\sum_{j=0}^{\infty}(\frac{\zeta - c}{z - c})^j\, d\zeta \\
&= \sum_{k=0}^{\infty}\frac{1}{2\pi i}\int_{C_r} \frac{f(\zeta)}{(\zeta - c)^{k+1}}\, d\zeta (z - c)^k + \sum_{j=0}^{\infty}\frac{1}{2\pi i}\int_{C_\delta} f(\zeta)(\zeta - c)^j\, d\zeta (z - c)^{-j-1} \\
&= \sum_{k=0}^{\infty} a_k (z - c)^k + \sum_{k=-\infty}^{-1}\frac{1}{2\pi i}\int_{C_\delta} \frac{f(\zeta)}{(\zeta - c)^{k+1}}\, d\zeta (z - c)^k \\
&= \sum_{k=0}^{\infty} a_k (z - c)^k + \sum_{k=-\infty}^{-1}\frac{1}{2\pi i}\int_{C_r} \frac{f(\zeta)}{(\zeta - c)^{k+1}}\, d\zeta (z - c)^k \\
&= \sum_{k=-\infty}^{\infty} a_k (z - c)^k,
\end{aligned}$$

which proves part (1).
We leave the proofs of parts (2) and (3) to the exercises.

Exercise 7.16. (a) Justify bringing the summation signs out of the integrals in the calculation in the preceding proof.
(b) Prove parts (2) and (3) of the preceding theorem. Compare this with Exercise 7.14.

REMARK. The representation of $f(z)$ in the punctured disk $B'_r(c)$ given in part (1) of Theorems 7.15 and 7.16 is called the *Laurent expansion* of f around the singularity c. Of course it differs from a Taylor series representation of f, as this one contains negative powers of $z - c$. In fact, which negative powers it contains indicates what kind of singularity the point c is.

Non removable isolated singularities of a function f share the property that the integral of f around a disk centered at the singularity equals $2\pi i a_{-1}$, where the number a_{-1} is the coefficient of $(z-c)^{-1}$ in the Laurent expansion of f around c. This number $2\pi i a_{-1}$ is obviously significant, and we call it the *residue of f at c*, and denote it by $R_f(c)$.
Combining Theorems 7.13, 7.15, and 7.16, we obtain:

THEOREM 7.17. (Residue Theorem) Let S be a piecewise smooth geometric set whose boundary has finite length, let $c_1, \ldots, c_n$ be points in S^0, and suppose f is a complex-valued function that is continuous at every point z in S except the c_k's, and differentiable at every point $z \in S^0$ except at the c_k's. Assume finally that each c_k is a nonremovable isolated singularity of f. Then

$$\int_{C_S} f(\zeta)\, d\zeta = \sum_{k=1}^{n} R_f(c_k).$$

That is, the contour integral around C_S is just the sum of the residues inside S.

Exercise 7.17. Prove Theorem 7.17.

Exercise 7.18. Use the Residue Theorem to compute $\int_{C_S} f(\zeta)\, d\zeta$ for the functions f and geometric sets S given below. That is, determine the poles of f inside S, their orders, the corresponding residues, and then evaluate the integrals.
(a) $f(z) = \sin(3z)/z^2$, and $S = \overline{B}_1(0)$.
(b) $f(z) = e^{1/z}$, and $S = \overline{B}_1(0)$.
(c) $f(z) = e^{1/z^2}$, and $S = \overline{B}_1(0)$.
(d) $f(z) = (1/z(z-1))$, and $S = \overline{B}_2(0)$.
(e) $f(z) = ((1-z^2)/z(1+z^2)(2z+1)^2)$, and $S = \overline{B}_2(0)$.
(f) $f(z) = 1/(1+z^4) = (1/(z^2-i)(z^2+i))$, and $S = \overline{B}_r(0)$ for any $r > 1$.

The Residue Theorem, a result about contour integrals of functions of a complex variable, can often provide a tool for evaluating integrals of functions of a real variable.

EXAMPLE 1. Consider the integral

$$\int_{-\infty}^{\infty} \frac{1}{1+x^4}\, dx.$$

Let us use the Residue Theorem to compute this integral.
Of course what we need to compute is

$$\lim_{B\to\infty} \int_{-B}^{B} \frac{1}{1+x^4}\, dx.$$

The first thing we do is to replace the real variable x by a complex variable Z, and observe that the function $f(z) = 1/(1+z^4)$ is analytic everywhere except at the four points $\pm e^{i\pi/4}$ and $\pm e^{3i\pi/4}$. See part (f) of the preceding exercise. These are the four points whose fourth power is -1, and hence are the poles of the function f.
Next, given a positive number B, we consider the geometric set (rectangle) S_B that is determined by the interval $[-B, B]$ and the two bounding functions $l(x) = 0$ and

$u(x) = B$. Then, as long as $B > 1$, we know that f is analytic everywhere in S^0 except at the two points $c_1 = e^{i\pi/4}$ and $c_2 = e^{3i\pi/4}$, so that the contour integral of f around the boundary of S_B is given by

$$\int_{C_{S_B}} \frac{1}{1+\zeta^4}\, d\zeta = R_f(c_1) + R_f(c_2).$$

Now, this contour integral consists of four parts, the line integrals along the bottom, the two sides, and the top. The magic here is that the integrals along the sides, and the integral along the top, all tend to 0 as B tends to infinity, so that the integral along the bottom, which after all is what we originally were interested in, is in the limit just the sum of the residues inside the geometric set.

Exercise 7.19. Verify the details of the preceding example.
(a) Show that

$$\lim_{B\to\infty} \int_0^B \frac{1}{1+(B+it)^4}\, dt = 0.$$

(b) Verify that

$$\lim_{B\to\infty} \int_{-B}^B \frac{1}{1+(t+iB)^4}\, dt = 0.$$

(c) Show that

$$\int_{-\infty}^{\infty} \frac{1}{1+x^4}\, dx = \pi\sqrt{2}.$$

Methods similar to that employed in the previous example and exercise often suffice to compute integrals of real-valued functions. However, the method may have to be varied. For instance, sometimes the appropriate geometric set is a rectangle below the x-axis instead of above it, sometimes it should be a semicircle instead of a rectangle, etc. Indeed, the choice of contour (geometric set) can be quite subtle. The following exercise may shed some light.

Exercise 7.20. (a) Compute

$$\int_{-\infty}^{\infty} \frac{e^{ix}}{1+x^4}\, dx$$

and

$$\int_{-\infty}^{\infty} \frac{e^{-ix}}{1+x^4}\, dx.$$

(b) Compute

$$\int_{-\infty}^{\infty} \frac{\sin(-x)}{1+x^3}\, dx$$

and

$$\int_{-\infty}^{\infty} \frac{\sin x}{1+x^3}\, dx.$$

EXAMPLE 2. An historically famous integral in analysis is $\int_{-\infty}^{\infty} \sin x/x\,dx$. The techniques described above don't immediately apply to this function, for, even replacing the x by a z, this function has no poles, so that the Residue Theorem wouldn't seem to be much help. Though the point 0 is a singularity, it is a removable one, so that this function $\sin z/z$ is essentially analytic everywhere in the complex plane. However, even in a case like this we can obtain information about integrals of real-valued functions from theorems about integrals of complex-valued functions. Notice first that $\int_{-\infty}^{\infty} \sin x/x\,dx$ is the imaginary part of $\int_{-\infty}^{\infty} e^{ix}/x\,dx$, so that we may as well evaluate the integral of this function. Let f be the function defined by $f(z) = e^{iz}/z$, and note that 0 is a pole of order 1 of f, and that the residue $R_f(0) = 2\pi i$. Now, for each $B > 0$ and $\delta > 0$ define a geometric set $S_{B,\delta}$, determined by the interval $[-B, B]$, as follows: The upper bounding function $u_{B,\delta}$ is given by $u_{B,\delta}(x) = B$, and the lower bounding function $l_{B,\delta}$ is given by $l_{B,\delta}(x) = 0$ for $-B \leq x \leq -\delta$ and $\delta \leq x \leq B$, and $l_{B,\delta}(x) = \delta e^{i\pi x/\delta}$ for $-\delta < x < \delta$. That is, $S_{B,\delta}$ is just like the rectangle S_B in Example 1 above, except that the lower boundary is not a straight line. Rather, the lower boundary is a straight line from $-B$ to $-\delta$, a semicircle below the x-axis of radius δ from $-\delta$ to δ, and a straight line again from δ to B.

By the Residue Theorem, the contour integral

$$\int_{C_{S_{B,\delta}}} f(\zeta)\,d\zeta = R_f(0) = 2\pi i.$$

As in the previous example, the contour integrals along the two sides and across the top of $S_{B,\delta}$ tend to 0 as B tends to infinity. Finally, according to part (e) of Exercise 6.15, the contour integral of f along the semicircle in the lower boundary is πi independent of the value of δ. So,

$$\lim_{B\to\infty} \lim_{\delta\to 0} \int_{\mathrm{graph}(l_{B,\delta})} \frac{e^{i\zeta}}{\zeta}\,d\zeta = \pi i,$$

implying then that

$$\int_{-\infty}^{\infty} \frac{\sin x}{x}\,dx = \pi.$$

Exercise 7.21. (a) Justify the steps in the preceding example. In particular, verify that

$$\lim_{B\to\infty} \int_0^B \frac{e^{i(B+it)}}{B+it}\,dt = 0,$$

$$\lim_{B\to\infty} \int_{-B}^B \frac{e^{i(t+iB)}}{t+iB}\,dt = 0,$$

and

$$\int_{C_\delta} \frac{e^{i\zeta}}{\zeta}\,d\zeta = \pi i,$$

where C_δ is the semicircle of radius δ, centered at the origin and lying below the x-axis.

(b) Evaluate

$$\int_{-\infty}^{\infty} \frac{\sin^2 x}{x^2}\,dx.$$

APPENDIX
EXISTENCE AND UNIQUENESS OF A COMPLETE ORDERED FIELD

This appendix is devoted to the proofs of Theorems 1.1 and 1.2, which together assert that there exists a unique complete ordered field. Our construction of this field will follow the ideas of Dedekind, which he presented in the late 1800's.

DEFINITION. By a *Dedekind cut,* or simply a *cut,* we will mean a pair (A, B) of nonempty (not necessarily disjoint) subsets of the set $\mathbb{Q}$ of rational numbers for which the following two conditions hold.

(1) $A \cup B = \mathbb{Q}$. That is, every rational number is in one or the other of these two sets.

(2) For every element $a \in A$ and every element $b \in B$, $A \leq b$. That is, every element of A is less than or equal to every element of B.

Recall that when we define the rational numbers as quotients (ordered pairs) of integers, we faced the problem that two different quotients determine the same rational number, e.g., $2/3 \equiv 6/9$. There is a similar equivalence among Dedekind cuts.

DEFINITION. Two Dedekind cuts (A_1, b_1) and (A_2, B_2) are called *equivalent* if $a_1 \leq b_2$ for all $a_1 \in A_1$ and all $b_2 \in B_2$, and $a_2 \leq b_1$ for all $a_2 \in A_2$ and all $b_1 \in B_1$. In such a case, we write $(A_1, B_1) \equiv (A_2, B_2)$.

bf Exercise A.1. (a) Show that every rational number r determines three distinct Dedekind cuts that are mutually equivalent.
(b) Let B be the set of all positive rational numbers r whose square is greater than 2, and let A comprise all the rationals not in B. Prove that the pair (A, B) is a Dedekind cut. Do you think this cut is not equivalent to any cut determined by a rational number r as in part (a)? Can you prove this?
(c) Prove that the definition of equivalence given above satisfies the three conditions of an equivalence relation. Namely, show that
(i) (Reflexivity) (A, B) is equivalent to itself.
(ii) (Symmetry) If $(A_1, B_1) \equiv (A_2, B_2)$, then $(A_2, B_2) \equiv (A_1, B_1)$.
(iii) (Transitivity) If $(A_1, B_1) \equiv (A_2, B_2)$ and $(A_2, B_2) \equiv (A_3, B_3)$, then $(A_1, B_1) \equiv (A_3, B_3)$.

There are three relatively simple-sounding and believable properties of cuts, and we present them in the next theorem. It may be surprising that the proof seems to be more difficult than might have been expected.

THEOREM A.1. *Let (A, B) be a Dedekind cut. Then*

(1) *If $a \in A$ and $a' < a$, then $a' \in A$.*

(2) *If $b \in B$ and $b' > b$, then $b' \in B$.*

(3) *Let ϵ be a positive rational number. Then there exists an $a \in A$ and a $b \in B$ such that $b - a < \epsilon$.*

PROOF. Suppose a is an element of A, and let $a' < a$ be given. By way of contradiction suppose that a' does not belong to A. Then, by Condition (1) of the definition of a cut, it must be that $a' \in B$. But then, by Condition (2) of the definition of a cut, we must have that $a \leq a'$, and this is a contradiction, because $a' < a$. This proves part (1). Part (2) is proved in a similar manner.

To prove part (3), let the rational number $\epsilon > 0$ be given, and set $r = \epsilon/2$. Choose an element $a_0 \in A$ and an element $b_0 \in B$. Such elements exist, because A and B are nonempty sets. Choose a natural number N such that $a_0 + Nr > b_0$. Such a natural number N must exist. For instance, just choose N to be larger than the rational number $(b_0 - a_0)/r$. Now define a sequence $\{a_k\}$ of rational numbers by $a_k = a_0 + kr$, and let K be the first natural number for which $a_K \in B$. Obviously, such a number exists, and in fact K must be less than or equal to N. Now, a_{K-1} is not in B, so it must be in A. Set $a = A_{K-1}$ and $b = A_K$. Clearly, $a \in A$, $b \in B$, and

$$b - a = a_K - a_{K-1} = a_0 + Kr - a_0 - (K-1)r = r = \frac{\epsilon}{2} < \epsilon,$$

and this proves part (3).

We will make a complete ordered field F whose elements are the set of equivalence classes of Dedekind cuts. We will call this field the *Dedekind field*. To make this construction, we must define addition and multiplication of equivalence classes of cuts, and verify the six required field axioms. Then, we must define the set P that is to be the positive elements of the Dedekind field F, and then verify the required properties of an ordered field. Finally, we must prove that this field is a complete ordered field; i.e., that every nonempty set that is bounded above has a least upper bound. First things first.

DEFINITION. If (A_1, B_1) and (A_2, B_2) are Dedekind cuts, define the *sum* of (A_1, B_1) and (A_2, B_2) to be the cut (A_3, B_3) described as follows: B_3 is the set of all rational numbers b_3 that can be written as $b_1 + b_2$ for some $b_1 \in B_1$ and $b_2 \in B_2$, and A_3 is the set of all rational numbers r such that $r < b_3$ for all $b_3 \in B_3$.

Several things need to be checked. First of all, the pair (A_3, B_3) is again a Dedekind cut. Indeed, it is clear from the definition that every element of A_3 is less than or equal to every element of B_3, so that Condition (2) is satisfied. To see that Condition (1) holds, let r be a rational number, and suppose that it is not in A_3. We must show that r belongs to B_3. Now, since $r \notin A_3$, there must exist an element $b_3 = b_1 + b_2 \in B_3$ for which $r > b_3$. Otherwise, r would be in A_3. But this means that $r - b_2 > b_1$, and so by part (2) of Theorem A.1, we have that $r - b_2$ is an element b_1' of B_1. Therefore, $r = b_1' + b_2$, implying that $r \in B_3$, as desired.
We define the *0 cut* to be the pair $A_0 = \{r : r \leq 0\}$ and $B_0 = \{r : r > 0\}$. This cut is one of the three determined by the rational number 0.

bf Exercise A.2. (a) Prove that addition of Dedekind cuts is commutative and associative.
(b) Prove that if $(A_1, B_1) \equiv (C_1, D_1)$ and $(A_2, B_2) \equiv (C_2, D_2)$, then $(A_1, B_1) + (A_2, B_2) \equiv (C_1, D_1) + (C_2, D_2)$.
(c) Find an example of a cut (A, B) such that $(A, B) + 0 \neq (A, B)$.
(d) Prove that $(A, B) + 0 \equiv (A, B)$ for every cut (A, B).

We define addition in the set F of all equivalence classes of Dedekind cuts as follows:

DEFINITION. If x is the equivalence class of a cut (A, b) and y is the equivalence class of a cut (C, D), then $x + y$ is the equivalence class of the cut $(A, B) + (C, D)$.

It follows from the previous exercise, that addition in F is well-defined, commutative, and associative. We are on our way.

We define the element 0 of F to be the equivalence class of the 0 cut. The next theorem establishes one of the important field axioms for F, namely, the existence of an additive inverse for each element of F.

THEOREM A.2. *If (A, B) is a Dedekind cut, then there exists a cut (A', B') such that $(A, B) + (A', B')$ is equivalent to the 0 cut. Therefore, if x is an element of F, then there exists an element y of F such that $x + y = 0$.*

PROOF. Let $A' = -B$, i.e., the set of all the negatives of the elements of B, and let $B' = -A$, i.e., the set of all the negatives of the elements of A. It is immediate that the pair (A', B') is a Dedekind cut. Let us show that $(A, B) + (A', B')$ is equivalent to the zero cut. Let $(C, D) = (A, B) + (A', B')$. Then, by the definition of the sum of two cuts, we know that D consists of all the elements of the form $d = b + b' = b - a$, where $b \in B$ and $a \in A$. Since $a \leq b$ for all $a \in A$ and $b \in B$, we see then that the elements of D are all greater than or equal to 0. To see that (C, D) is equivalent to the 0 cut, it will suffice to show that D contains all the positive rational numbers. (Why?) Hence, let $\epsilon > 0$ be given, and choose an $a \in A$ and a $b \in B$ such that $b - a < \epsilon$. This can be done by Condition (3) of Theorem A.1. Then, the number $b - a \in D$, and hence, by part (2) of Theorem A.1, $\epsilon \in D$. It follows then that the cut (C, D) is equivalent to the zero cut (A_0, B_0), as desired.

We will write $-(A, B)$ for the cut (A', B') of the preceding proof.

bf Exercise A.3. (a) Suppose (A, B) is a cut, and let (C, D) be a cut for which $(A, B) + (C, D)$ is equivalent to the 0 cut. Show that $(C, D) \equiv (A', B') = -(A, B)$.
(b) Prove that the additive inverse of an element x of the Dedekind field F is unique.

The definition of multiplication of cuts, as well as multiplication in F, is a bit more tricky. In fact, we will first introduce the notion of positivity among Dedekind cuts.

DEFINITION. A Dedekind cut $x = (A, B)$ is called *positive* if A contains at least one positive rational number.

bf Exercise A.4. (a) Suppose (A, B) and (C, D) are equivalent cuts, and assume that (A, B) is positive. Prove that (C, D) also is positive. Make the obvious definition of positivity in the set F.
(b) Show that the sum of two positive cuts is positive. Conclude that the sum of two positive elements of F, i.e., the sum of two equivalence classes of positive cuts, is positive.
(c) Let (A, B) be a Dedekind cut. Show that one and only one of the following three properties holds for (A, B). (i) (A, B) is a positive cut, (ii) $-(A, B)$ is a positive cut, or (iii) (A, B) is equivalent to the 0 cut.
(d) Establish the law of tricotomy for F : That is, show that one and only one of the following three properties holds for an element $x \in F$. (i) x is positive, (ii) $-x$ is positive, or (iii) $x = 0$.

We first define multiplication of cuts when one of them is positive.

DEFINITION. Let (A_1, B_1) and (A_2, B_2) be two Dedekind cuts, and suppose that one of these cuts is a positive cut. We define the *product* (A_3, B_3) of (A_1, B_1) and (A_2, B_2) as follows: Set B_3 equal to the set of all b_3 that can be written as $b_1 b_2$ for some $b_1 \in B_1$ and $b_2 \in B_2$. Then set A_3 to be all the rational numbers r for which $r < b_3$ for all $b_3 \in B_3$.

Again, things need to be checked.

bf Exercise A.5. (a) Show that the pair (A_3, B_3) of the preceding definition for the product of positive cuts is in fact a Dedekind cut.
(b) Prove that multiplication of Dedekind cuts, when one of them is positive, is commutative.
(c) Suppose (A_1, B_1) is a positive cut. Prove that

$$(A_1, B_1)((A_2, B_2) + (A_3, B_3)) = (A_1, B_1)(A_2, B_2) + (A_1, B_1)(A_3, B_3)$$

for any cuts (A_2, B_2) and (A_3, B_3).
(d) Show that, if $(A_1, B_1) \equiv (A_2, B_2)$ and $(C_1, D_1) \equiv (C_2, D_2)$ and (a_1, B_1) and (A_2, B_2) are positive cuts, then $(A_1, B_1)(C_1, D_1) \equiv (A_2, B_2)(C_2, D_2)$.
(e) Show that the product of two positive cuts is again a positive cut.

We are ready to define multiplication in F.

DEFINITION. Let x and y be elements of F.
If either x or y is positive, define the product $x \times y$ to be the equivalence class of the cut $(A, B)(C, D)$, where x is the equivalence class of (A, B) and y is the equivalence class of (C, D).
If either x or y is 0, define $x \times y$ to be 0.
If both x and y are negative, i.e., both $-x$ and $-y$ are positive, define $x \times y = (-x) \times (-y)$.

The next exercise is tedious. It amounts to checking a bunch of cases.

bf Exercise A.6. (a) Prove that multiplication in F is commutative.
(b) Prove that multiplication in F is associative.
(c) Prove that multiplication in F is distributive over addition.
(d) Prove that the product of two positive elements of F is again positive.

We define the element 1 of F to be the equivalence class of the cut (A^1, B^1), where $A^1 = \{r : r \leq 1\}$ and $B^1 = \{r : r > 1\}$.

bf Exercise A.7. (a) Prove that the elements 0 and 1 of F are not equal.
(b) Prove that $x \times 1 = x$ for every element $x \in F$.
(c) Use the associative law and part (b) to prove that if $xy = 1$ and $xz = 1$, then $y = z$.

THEOREM A.3. *With respect to the operations of addition and multiplication defined above, together with the definition of positive elements, F is an ordered field.*

PROOF. The first five axioms for a field, given in Chapter I, have been established for F in the preceding exercises, so that we need only verify axiom 6 to complete the proof that F is a field. Thus, let $x \in F$ be a nonzero element. We must show the existence of an element y of F for which $x \times y = 1$. Suppose first that x is a positive element of F. Then x is the equivalence class of a positive cut (A, B), and therefore A contains some positive rational numbers. Let a_0 be a positive number that is contained in A. It follows then that every element of B is greater than or equal to a_0 and hence is positive. Define $\widehat{B}$ to be the set of all rational numbers r for which $r \geq 1/b$ for every $b \in B$. Then define $\widehat{A}$ to be the set of all rationals r for which $r \leq \widehat{b}$ for every $\widehat{b} \in \widehat{B}$. It follows directly that the pair $(\widehat{A}, \widehat{B})$ is a Dedekind cut.

Let $(C,D) = (A,B) \times (\widehat{A},\widehat{B})$, and note that every element $d \in D$ is of the form $d = b\widehat{b}$, and hence is greater than or equal to 1. We claim that (C,D) is equivalent to the cut (A^1,B^1) that determines the element 1 of F. To see this we must verify that D contains every rational number r that is greater than 1. Thus, let $r > 1$ be given, and set $\epsilon = a_0(r-1)$. From Condition (3) of Theorem A.1, choose an $a' \in A$ and a $b' \in B$ such that $b' - a' < \epsilon$. Without loss of generality, we may assume that $a' \geq a_0$. Finally, set $\widehat{b} = 1/a'$. Clearly $\widehat{b} \geq 1/b$ for all $b \in B$, so that $\widehat{b} \in \widehat{B}$. Also $d = b'\widehat{b} \in D$, and

$$d = b'\widehat{b} = \frac{b'}{a'} = \frac{a' + b' - a'}{a'} < 1 + \frac{\epsilon}{a'} \leq 1 + \frac{\epsilon}{a_0} = r,$$

implying that $r \in D$. Therefore, (C,D) is equivalent to the cut (A^1,B^1), implying that $(A,B) \times (\widehat{A},\widehat{B})$ is equivalent to the cut (A^1,B^1). Therefore, if y is the element of F that is the equivalence class of the cut $(\widehat{A},\widehat{B})$, then $x \times y = 1$, as desired.
If x is negative, then $-x$ is positive. If we write z for the multiplicative inverse of the positive element $-x$, then $-z$ is the multiplicative inverse of the element x. Indeed, by the definition of the product of two negative elements of F, $x \times (-z) = (-x) \times z = 1$.
The properties that guarantee that F is an ordered field also have been established in the preceding exercises, so that the proof of this theorem is complete.

So, the Dedekind field is an ordered field, but we have left to prove that it is complete. This means we must examine upper bounds of sets, and that requires us to understand when one cut is less than another one. We say that a cut (A,B) is *less than or equal* to a cut $C,D)$ if $a \leq d$ for every $a \in A$ and $d \in D$. We say that an element x in the ordered field F is *less than or equal* to an element y if $y - x$ is either positive or 0.

THEOREM A.4. *Let x and y be elements of F, and suppose x is the equivalence class of the cut $(A,B()$ and y is the equivalence class of the cut (C,D). Then $x \leq y$ if and only if $(A,B) \leq (C,D)$.*

PROOF. We have that $x \leq y$ if and only if the element $y - x = y + -x$ is positive or 0. Writing, as before, (A',B') for the cut $-(A,B)$, we have that $y - x$ is the equivalence class of the cut $(C,d) - (A,B) = (C,D) + (A',B')$, so we need to determine when the cut $(G,H) = (C,D) + (A',B')$ is a positive cut or the 0 cut; which is the case when the set H only contains nonnegative numbers. By definition of addition, the set H contains all numbers of the form $h = d + b'$ for some $d \in D$ and some $b' \in B'$. Since $B' = -A$, this means that H consists of all elements of the form $h = d - a$ for some $d \in D$ and $a \in A$. Now these numbers h are all greater than or equal to 0 if and only if each $a \in A$ is less than or equal to each $d \in D$, i.e., if and only if $(A,B) \leq (C,D)$. This proves the theorem

We are now ready to present the first of the two main theorems of this appendix, that is Theorem 1.1 in Chapter I.

THEOREM A.5. *There exists a complete ordered field. Indeed, the Dedekind field F is a complete ordered field.*

PROOF. Let S be a nonempty subset of F, and suppose that there exists an upper bound for S; i.e., an element M of F such that $x \leq M$ for all $x \in S$. Write (A,B)

for a cut such that M is the equivalence class of (A, B). We must show that there exists a least upper bound for S.

For each $x \in S$, let (A_x, B_x) be a Dedekind cut for which x is the equivalence class of (A_x, B_x), and note that $a_x \leq b$ for all $a_x \in A_x$ and all $b \in B$. Let A_0 be the union of all the sets A_x for $x \in S$. Let B_0 be the set of all rational numbers r for which $r \geq a_0$ for every $a_0 \in A_0$. we claim first that the pair (A_0, B_0) is a Dedekind cut. Both sets are nonempty; A_0 because it is the union of nonempty sets, and B_0 because it contains all the elements of the nonempty set B. Clearly Condition (2) for a cut holds from the very definition of this pair. To see Condition (1), let r be a rational number that is not in B_0. We must show that it is in A_0. Now, since r is not in B_0, there must exist some $a_0 \in A_0$ for which $r < a_0$. But $a_0 \in \cup_{x \in S} A_x$, so that there must exist an $x \in S$ such that $a_0 \in A_x$, and hence r is also in A_x. But then $r \in A_0$, and this proves that (A_0, B_0) is a Dedekind cut.

Let M_0 be the equivalence class determined by the cut (A_0, B_0). Since each $A_x \subseteq A_0$, we see that $a_x \leq b_0$ for every $a_x \in A_x$ and every $b_0 \in B_0$. Hence, $(A_x, B_x) \leq (A_0, B_0)$ for every $x \in S$, and therefore, by Theorem A.4, $x \leq M_0$ for all $x \in S$. This shows that M_0 is an upper bound for S.

Finally, suppose M' is another upper bound for S, and let (A', B') be a cut for which M' is the equivalence class of (A', B'). Then $a_x \leq b'$ for every $a_x \in A_x$ and every $b' \in B'$, implying that $a_0 \leq b'$ for every $a_0 \in A_0$ and every $b' \in B'$. Therefore, $(A_0, B_0) \leq (A', B')$, implying that $M_0 \leq M'$. This shows that M_0 is the least upper bound for S, and the theorem is proved.

We come now to the second major theorem of this appendix, i.e., Theorem 1.2 of Chapter I. This one asserts the uniqueness, up to isomorphism, of complete ordered fields.

THEOREM A.6. *Let $\widehat{F}$ be a complete ordered field. Then there exists an isomorphism of $\widehat{F}$ onto the Dedekind field F. That is, there exists a one-to-one function $J : \widehat{F} \to F$ that is onto all of F, and that satisfies*

(1) $J(x + y) = J(x) + J(y)$.
(2) $J(xy) = J(x)J(y)$.
(3) *If* $x > 0$, *then* $J(x) > 0$.

PROOF. We know from Chapter I that, inside any ordered field, there is a subset that is isomorphic to the field $\mathbb{Q}$ of rational numbers. We will therefore identify this special subset of $\widehat{F}$ with $\mathbb{Q}$.

If x is an element of $\widehat{F}$, let $A_x = \{r \in \mathbb{Q} : r \leq x\}$ and let $B_x = \{r \in \mathbb{Q} : r > x\}$. We claim first that the pair (A_x, B_x) is a Dedekind cut. Indeed, from the definition of A_x and B_x, we see that Condition (2), i.e., that each $a_x \in A_x$ is less than or equal to each $b_x \in B_x$, holds. To see that Condition (1) also holds, let r be a rational number in $\widehat{F}$. Then, because $\widehat{F}$ is an ordered field, either $r \leq x$ or $r > x$, i.e., $r \in A_x$ or $r \in B_x$. Hence, (A_x, B_x) is a Dedekind cut.

We define a function J from $\widehat{F}$ into F by setting $J(x)$ equal to the equivalence class determined by the cut (A_x, B_x). We must check several things.

First of all, J is one-to-one. Indeed, let x and y be elements of $\widehat{F}$ that are not equal. Assume without loss of generality that $x < y$. Then, according to Theorem 1.8, which is a theorem about complete ordered fields and hence applicable to $\widehat{F}$,, there exist two rational numbers r_1 and r_2 such that $x < r_1 < r_2 < y$, which implies that $r_1 \in B_x$ and $r_2 \in A_y$. Since $r_2 > r_1$, the cut (A_y, B_y) is not equivalent to the cut (A_x, B_x), and therefore $J(x) \neq J(y)$.

Next, we claim that the function J is onto all of the Dedekind field F. Indeed, let z be an element of F, and let (A, B) be a Dedekind cut for which z is the equivalence class determined by (A, B). Think of A as a subset of the complete ordered field $\widehat{F}$. Then A is nonempty and is bounded above. In fact, every element of B is an upper bound of A. Let $x = \sup A$. (Here is another place where we are using the completeness of the field $\widehat{F}$.) We claim that the cut (A, B) is equivalent to the cut (A_x, B_x), which will imply that $J(x) = z$. Thus, if $a_x \in A_x$, then $a_x \leq x$, and $x \leq b$ for every $b \in B$, because x is the least upper bound of A. Similarly, if $a \in A$, then $a \leq x$, and $x < b_x$ for every $b_x \in B_x$. This proves that the cuts (A, B) and (A_x, B_x) are equivalent, as desired.

If x and y are elements of $\widehat{F}$, and $b_x \in B_x$ and $b_y \in B_y$, then $b_x > x$ and $b_y > y$, so that $b_x + b_y > x + y$, and therefore $b_x + b_y \in B_{x+y}$ for every $b_x \in B_x$ and $b_y \in B_y$. On the other hand, if $r \in B_{x+y}$, then $r > x + y$. Therefore, $r - x > y$, implying, again by Theorem 1.8, that there exists an element $b_y \in B_y$ such that $y < b_y < r - x$. But then $r - b_y > x$, which means that $r - b_y = b_x$ for some $b_x \in B_x$. So, $r = b_x + b_y$, and this shows that $B_{x+y} = b_x + B_y$. It follows from this that the cuts (A_{x+y}, B_{x+y}) and $(A_x, B_x) + (A_y, B_y)$ are equal, and therefore $J(x+y) = J(x) + J(y)$. A consequence of this is that $J(-x) = -J(x)$ for all $x \in \widehat{F}$.

If x and y are two positive elements of $\widehat{F}$, then an argument just like the one in the preceding paragraph shows that $J(xy) = J(x)J(y)$. Then, since $J(-x) = -J(x)$, the fact that $J(xy) = J(x)J(y)$ for all $x, y \in \widehat{F}$ follows.

Finally, if x is a positive element of $\widehat{F}$, then the set A_x must contain some positive rationals, and hence the cut (A_x, B_x) is a positive cut, implying that $J(x) > 0$.

We have verified all the requirements for an isomorphism between the two fields $\widehat{F}$ and F, and the theorem is proved.