Pure and Applied
UNDERGRADUATE TEXTS · 12

An Introduction to Complex Analysis and Geometry

John P. D'Angelo

American Mathematical Society
Providence, Rhode Island

Preface for the student

I hope that this book reveals the beauty and usefulness of complex numbers to you. I want you to enjoy both reading it and solving the problems in it. Perhaps you will spot something in your own area of interest and benefit from applying complex numbers to it. Students in my classes have found applications of ideas from this book to physics, music, engineering, and linguistics. Several students have become interested in historical and philosophical aspects of complex numbers. I have not yet seen anyone get excited about the hysterical aspects of complex numbers.

At the very least you should see many places where complex numbers shed a new light on things you have learned before. One of my favorite examples is trig identities. I found them rather boring in high school and later I delighted in proving them more easily using the complex exponential function. I hope you have the same experience. A second example concerns certain definite integrals. The techniques of complex analysis allow for stunningly easy evaluations of many calculus integrals and seem to lie within the realm of science fiction.

This book is meant to be readable, but at the same time it is precise and rigorous. Sometimes mathematicians include details that others feel are unnecessary or obvious, but do not be alarmed. If you do many of the exercises and work through the examples, then you should learn plenty and enjoy doing it. I cannot stress enough two things I have learned from years of teaching mathematics. First, students make too few sketches. You should strive to merge geometric and algebraic reasoning. Second, definitions are your friends. If a theorem says something about a concept, then you should develop both an intuitive sense of the concept and the discipline to learn the precise definition. When asked to verify something on an exam, start by writing down the definition of that something. Often the definition suggests exactly what you should do!

Some sections and paragraphs introduce more sophisticated terminology than is necessary at the time, in order to prepare for later parts of the book and even for subsequent courses. I have tried to indicate all such places and to revisit the crucial ideas. In case you are struggling with any material in this book, remain calm. The magician will reveal his secrets in due time.

Chapter 1

From the Real Numbers to the Complex Numbers

1. Introduction

Many problems throughout mathematics and physics illustrate an amazing principle: ideas expressed within the realm of real numbers find their most elegant expression through the unexpected intervention of complex numbers. Many of these delightful interventions arise in elementary, recreational mathematics. On the other hand most college students either never see complex numbers in action or they wait until the junior or senior year in college, at which time the sophisticated courses have little time for the elementary applications. Hence too few students witness the beauty and elegance of complex numbers. This book aims to present a variety of elegant applications of complex analysis and geometry in an accessible but precise fashion. We begin at the beginning, by recalling various number systems such as the integers $\mathbf{Z}$, the rational numbers $\mathbf{Q}$, and the real numbers $\mathbf{R}$, before even defining the complex numbers $\mathbf{C}$. We then provide three possible equivalent definitions. Throughout we strive for as much geometric reasoning as possible.

2. Number systems

The ancients were well aware of the so-called *natural numbers*, written $1, 2, 3, \ldots$. Mathematicians write $\mathbf{N}$ for the collection of natural numbers together with the usual operations of addition and multiplication. Partly because subtraction is not always possible, but also because negative numbers arise in many settings such as financial debts, it is natural to expand the natural number system to the larger system $\mathbf{Z}$ of integers. We assume that the reader has some understanding of the integers; the set $\mathbf{Z}$ is equipped with two distinguished members, written 1 and 0, and two operations, called addition $(+)$ and multiplication $(*)$, satisfying familiar laws. These operations make $\mathbf{Z}$ into what mathematicians call a *commutative ring* with unit 1. The integer 0 is special. We note that each n in $\mathbf{Z}$ has an additive

inverse $-n$ such that

$$n + (-n) = (-n) + n = 0. \tag{1}$$

Of course 0 is the only number whose additive inverse is itself.

Let a, b be given integers. As usual we write $a-b$ for the sum $a+(-b)$. Consider the equation $a + x = b$ for an unknown x. We learn to solve this equation at a young age; the idea is that subtraction is the inverse operation to addition. To solve $a + x = b$ for x, we first add $-a$ to both sides and use (1). We can then substitute b for $a + x$ to obtain the solution

$$x = 0 + x = (-a) + a + x = (-a) + b = b + (-a) = b - a.$$

This simple principle becomes a little more difficult when we work with multiplication. It is not always possible, for example, to divide a collection of n objects into two groups of equal size. In other words, the equation $2 * a = b$ does not have a solution in $\mathbf{Z}$ unless b is an even number. Within $\mathbf{Z}$, most integers (± 1 are the only exceptions) do not have multiplicative inverses.

To allow for division, we enlarge $\mathbf{Z}$ into the larger system $\mathbf{Q}$ of rational numbers. We think of elements of $\mathbf{Q}$ as fractions, but the definition of $\mathbf{Q}$ is a bit subtle. One reason for the subtlety is that we want $\frac{1}{2}$, $\frac{2}{4}$, and $\frac{50}{100}$ all to represent the same rational number, yet the expressions as fractions differ. Several approaches enable us to make this point precise. One way is to introduce the notion of equivalence class and then to define a rational number to be an equivalence class of pairs of integers. See [4] or [8] for this approach. A second way is to think of the rational number system as known to us; we then write elements of $\mathbf{Q}$ as letters, x, y, u, v, and so on, without worrying that each rational number can be written as a fraction in infinitely many ways. We will proceed in this second fashion. A third way appears in Exercise 1.2 below. Finally we emphasize that we cannot divide by 0. Surely the reader has seen alleged proofs that, for example, $1 = 2$, where the only error is a cleverly disguised division by 0.

▶ **Exercise 1.1.** Find an invalid argument that $1 = 2$ in which the only invalid step is a division by 0. Try to obscure the division by 0.

▶ **Exercise 1.2.** Show that there is a one-to-one correspondence between the set $\mathbf{Q}$ of rational numbers and the following set L of lines. The set L consists of all lines through the origin, except the vertical line $x = 0$, that pass through a nonzero point (a, b) where a and b are integers. (This problem sounds sophisticated, but one word gives the solution!)

The rational number system forms a *field*. A field consists of objects which can be added and multiplied; these operations satisfy the laws we expect. We begin our development by giving the precise definition of a field.

Definition 2.1. A field $\mathbf{F}$ is a mathematical system consisting of a collection of objects and two operations, addition and multiplication, subject to the following axioms.

1) For all x, y in $\mathbf{F}$, we have $x + y = y + x$ and $x * y = y * x$ (the commutative laws for addition and multiplication).

2) For all x, y, t in $\mathbf{F}$, we have $(x+y)+t = x+(y+t)$ and $(x*y)*t = x*(y*t)$ (the associative laws for addition and multiplication).

3) There are distinct distinguished elements 0 and 1 in $\mathbf{F}$ such that, for all x in $\mathbf{F}$, we have $0 + x = x + 0 = x$ and $1 * x = x * 1 = x$ (the existence of additive and multiplicative identities).

4) For each x in $\mathbf{F}$ and each y in $\mathbf{F}$ such that $y \neq 0$, there are $-x$ and $\frac{1}{y}$ in $\mathbf{F}$ such that $x + (-x) = 0$ and $y * \frac{1}{y} = 1$ (the existence of additive and multiplicative inverses).

5) For all x, y, t in $\mathbf{F}$ we have $t * (x + y) = (t * x) + (t * y) = t * x + t * y$ (the distributive law).

For clarity and emphasis we repeat some of the main points. The rational numbers provide a familiar example of a field. In any field we can add, subtract, multiply, and divide as we expect, although we cannot divide by 0. The ability to divide by a nonzero number distinguishes the rational numbers from the integers. In more general settings the ability to divide by a nonzero number distinguishes a field from a commutative ring. Thus every field is a commutative ring but a commutative ring need not be a field.

There are many elementary consequences of the field axioms. It is easy to prove that each element has a unique additive inverse and that each nonzero element has a unique multiplicative inverse, or reciprocal. The proof, left to the reader, mimics our early argument showing that subtraction is possible in $\mathbf{Z}$.

Henceforth we will stop writing $*$ for multiplication; the standard notation of xy for $x * y$ works adequately in most contexts. We also write x^2 instead of xx as usual. Let t be an element in a field. We say that x is a square root of t if $t = x^2$. In a field, taking square roots is not always possible. For example, we shall soon prove that there is no rational square root of 2 and that there is no real square root of -1.

At the risk of boring the reader we prove a few basic facts from the field axioms; the reader who wishes to get more quickly to geometric reasoning could omit the proofs, although writing them out gives one some satisfaction.

Proposition 2.1. *In a field the following laws hold:*

1) $0 + 0 = 0$.

2) *For all x, we have* $x0 = 0x = 0$.

3) $(-1)^2 = (-1)(-1) = 1$.

4) $(-1)x = -x$ *for all* x.

5) *If* $xy = 0$ *in* $\mathbf{F}$, *then either* $x = 0$ *or* $y = 0$.

Proof. Statement 1) follows from setting $x = 0$ in the axiom $0 + x = x$. Statement 2) uses statement 1) and the distributive law to write $0x = (0 + 0)x = 0x + 0x$. By property 4) of Definition 2.1, the object $0x$ has an additive inverse; we add this inverse to both sides of the equation. Using the meaning of additive inverse and then the associative law gives $0 = 0x$. Hence $x0 = 0x = 0$ and 2) holds. Statement 3) is a bit more interesting. We have $0 = 1 + (-1)$ by axiom 4) from Definition 2.1.

Multiplying both sides by -1 and using 2) yields

$$0 = (-1)0 = (-1)(1 + (-1)) = (-1)1 + (-1)^2 = -1 + (-1)^2.$$

Thus $(-1)^2$ is an additive inverse to -1; of course 1 also is. By the uniqueness of additive inverses, we see that $(-1)^2 = 1$. The proof of 4) is similar. Start with $0 = 1 + (-1)$ and multiply by x to get $0 = x + (-1)x$. Thus $(-1)x$ is an additive inverse of x and the result follows by uniqueness of additive inverses. Finally, to prove 5), we assume that $xy = 0$. If $x = 0$, the conclusion holds. If $x \neq 0$, we can multiply by $\frac{1}{x}$ to obtain

$$y = (\frac{1}{x}x)y = \frac{1}{x}(xy) = \frac{1}{x}0 = 0.$$

Thus, if $x \neq 0$, then $y = 0$, and the conclusion also holds. □

We note a point of language, where mathematics usage may differ with common usage. For us, the phrase "either $x = 0$ or $y = 0$" allows the possibility that **both** $x = 0$ and $y = 0$.

Example 2.1. A field with two elements. Let $\mathbf{F}_2$ consist of the two elements 0 and 1. We put $1 + 1 = 0$, but otherwise we add and multiply as usual. Then $\mathbf{F}_2$ is a field.

This example illustrates several interesting things. For example, the object 2 (namely $1 + 1$), can be 0 in a field. This possibility will prevent the quadratic formula from holding in a field for which $2 = 0$. In Theorem 2.1 we will derive the quadratic formula when it is possible to do so.

First we make a simple observation. We have shown that $(-1)^2 = 1$. Hence, when $-1 \neq 1$, it follows that 1 has two square roots, namely ± 1. Can an element of a field have more than two square roots? The answer is no.

Lemma 2.1. *In a field, an element t can have at most two square roots. If x is a square root of t, then so is $-x$, and there are no other possibilities.*

Proof. If $x^2 = t$, then $(-x)^2 = t$ by 3) and 4) of Proposition 2.1. To prove that there are no other possibilities, we assume that both x and y are square roots of t. We then have

$$0 = t - t = x^2 - y^2 = (x - y)(x + y). \tag{2}$$

By 5) of Proposition 2.1, we obtain either $x - y = 0$ or $x + y = 0$. Thus $y = \pm x$ and the result follows. □

The difference of two squares law stating that $x^2 - y^2 = (x - y)(x + y)$ is a gem of elementary mathematics. For example, suppose you are asked to multiply 88 times 92 in your head. You imagine $88 * 92 = (90 - 2)(90 + 2) = 8100 - 4 = 8096$ and impress some audiences. One can also view this algebraic identity for positive integers simply by removing a small square of dots from a large square of dots and rearranging the dots to form a rectangle. The author once used this kind of method when doing volunteer teaching of multiplication to third graders. See Figure 1.1 for a geometric interpretation of the identity in terms of area.

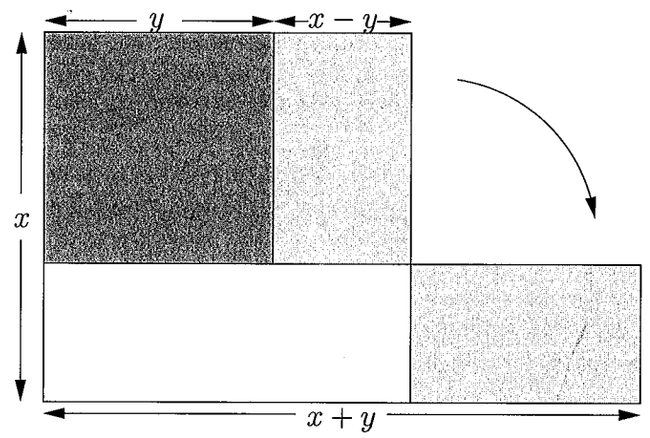

Figure 1.1. Difference of two squares.

We pause to make several remarks about square roots. The first remark concerns a notational convention; the discussion will help motivate the notion of ordered field defined below. The real number system will be defined formally below, and we will prove that positive real numbers have square roots. Suppose $t > 0$. We write $\sqrt{t}$ to denote the *positive* x for which $x^2 = t$. Thus both x and $-x$ are square roots of t, but the notation $\sqrt{t}$ means the positive square root. For the complex numbers, things will be more subtle. We will prove that each nonzero complex number z has two square roots, say $\pm w$, but there is no sensible way to prefer one to the other. We emphasize that the existence of square roots depends on more than the field axioms. Not all positive rational numbers have rational square roots, and hence it must be proved that each positive real number has a square root. The proof requires a limiting process. The quadratic formula, proved next, requires that the expression $b^2 - 4ac$ be a square. In an arbitrary field, the expression $\sqrt{t}$ usually means any x for which $x^2 = t$, but the ambiguity of signs can cause confusion. See Exercise 1.4.

Theorem 2.1. *Let $\mathbf{F}$ be a field. Assume that $2 \neq 0$ in $\mathbf{F}$. For $a \neq 0$ and arbitrary b, c we consider the quadratic equation*

$$ax^2 + bx + c = 0. \tag{3}$$

Then x solves (3) *if and only if*

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \tag{4}$$

If $b^2 - 4ac$ is not a square in $\mathbf{F}$, then (3) *has no solution.*

Proof. The idea of the proof is to *complete the square.* Since both a and 2 are nonzero elements of $\mathbf{F}$, they have multiplicative inverses. We therefore have

$$ax^2+bx+c = a(x^2+\frac{b}{a}x)+c = a(x^2+\frac{b}{a}x+\frac{b^2}{4a^2})+c-\frac{b^2}{4a}$$

$$= a(x+\frac{b}{2a})^2+\frac{4ac-b^2}{4a}. \tag{5}$$

We set (5) equal to 0 and we can easily solve for x. After dividing by a, we obtain

$$(x+\frac{b}{2a})^2 = \frac{b^2-4ac}{4a^2}. \tag{6}$$

The square roots of $4a^2$ are of course $\pm 2a$. Assuming that b^2-4ac has a square root in $\mathbf{F}$, we solve (6) for x by first taking the square root of both sides. We obtain

$$x+\frac{b}{2a} = \pm\frac{\sqrt{b^2-4ac}}{2a}. \tag{7}$$

After a subtraction and simplification we obtain (4) from (7). □

The reader surely has seen the quadratic formula before. Given a quadratic polynomial with real coefficients, the formula tells us that the polynomial will have no real roots when $b^2-4ac<0$. For many readers the first exposure to complex numbers arises when we introduce square roots of negative numbers in order to use the quadratic formula.

▶ **Exercise 1.3.** Show that additive and multiplicative inverses in a field are unique.

▶ **Exercise 1.4.** A subtlety. Given a field, is the formula

$$\sqrt{u}\sqrt{v} = \sqrt{uv}$$

always valid? In the proof of the quadratic formula, did we use this formula implicitly? If not, what did we use?

Example 2.2. One can completely analyze quadratic equations with coefficients in $\mathbf{F}_2$. The only such equations are $x^2=0$, $x^2+x=0$, $x^2+1=0$, and $x^2+x+1=0$. The first equation has only the solution 0. The second has the two solutions 0 and 1. The third has only the solution 1. The fourth has no solutions. We have given a complete analysis, even though Theorem 2.1 cannot be used in this setting.

Before introducing the notion of ordered field, we give a few other examples of fields. Several of these examples use modular (clock) arithmetic. The phrases *add modulo p* and *multiply modulo p* have the following meaning. Fix a positive integer p, called the *modulus.* Given integers m and n, we add (or multiply) them as usual and then take the remainder upon division by p. The remainder is called the sum (or product) modulo p. This natural notion is familiar to everyone; five hours after nine o'clock is two o'clock; we added modulo twelve. The subsequent examples can be skipped without loss of continuity.

Example 2.3. Fields with finitely many elements. Let p be a prime number, and let $\mathbf{F}_p$ consist of the numbers $0, 1, ..., p-1$. We define addition and multiplication modulo p. Then $\mathbf{F}_p$ is a field.

In Example 2.3, p needs to be a prime number. Property 5) of Proposition 2.1 fails when p is not a prime. We mention without proof that the number of elements in a finite field must be a power of a prime number. Furthermore, for each prime p and positive integer n, there exists a finite field with p^n elements.

▶ **Exercise 1.5.** True or false? Every quadratic equation in $\mathbf{F}_3$ has a solution.

Fields such as $\mathbf{F}_p$ are important in various parts of mathematics and computer science. For us, they will serve only as examples of fields. The most important examples of fields for us will be the real numbers and the complex numbers. To define these fields rigorously will take a bit more effort. We end this section by giving an example of a field built from the real numbers. We will not use this example in the logical development.

Example 2.4. Let K denote the collection of rational functions in one variable x with real coefficients. An element of K can be written $\frac{p(x)}{q(x)}$, where p and q are polynomials, and we assume that q is not the zero polynomial. (We allow $q(x)$ to equal 0 for some x, but not for all x.) We add and multiply such rational functions in the usual way. It is tedious but not difficult to verify the field axioms. Hence K is a field. Furthermore, K contains $\mathbf{R}$ in a natural way; we identify the real number c with the constant rational function $\frac{c}{1}$. As with the rational numbers, many different fractions represent the same element of K. To deal rigorously with such situations, one needs the notion of equivalence relation, discussed in Section 5.

3. Inequalities and ordered fields

Comparing the sizes of a pair of integers or of a pair of rational numbers is both natural and useful. It does not make sense however to compare the sizes of elements in an arbitrary ring or field. We therefore introduce a crucial property shared by the integers $\mathbf{Z}$ and the rational numbers $\mathbf{Q}$. For x, y in either of these sets, it makes sense to say that $x>y$. Furthermore, given the pair x, y, one and only one of three things must be true: $x>y$, $x<y$, or $x=y$. We need to formalize this idea in order to define the real numbers.

Definition 3.1. A field $\mathbf{F}$ is called *ordered* if there is a subset $P \subset \mathbf{F}$, called the *set of positive elements* of $\mathbf{F}$, satisfying the following properties:

1) For all x, y in P, we have $x+y \in P$ and $xy \in P$ (closure).

2) For each x in $\mathbf{F}$, one and only one of the following three statements is true: $x=0$, $x \in P$, $-x \in P$ (trichotomy).

▶ **Exercise 1.6.** Let $\mathbf{F}$ be an ordered field. Show that $1 \in P$.

▶ **Exercise 1.7.** Show that the trichotomy property can be rewritten as follows. For each x, y in $\mathbf{F}$, one and only one of the following three statements is true: $x=y$, $x-y \in P$, $y-x \in P$.

The rational number system is an ordered field; a fraction $\frac{p}{q}$ is positive if and only if p and q have the same sign. Note that q is never 0 and that a rational number is 0 whenever its numerator is 0. It is of course elementary to check in

this case that the set P of positive rational numbers is closed under addition and multiplication.

Once the set P of positive elements in a field has been specified, it is easier to work with inequalities than with P. We write $x > y$ if and only if $x - y \in P$. We also use the symbols $x \geq y$, $x \leq y$, $x < y$ as usual. The order axioms then can be written as follows:

1) If $x > 0$ and $y > 0$, then $x + y > 0$ and $xy > 0$.

2) Given $x \in \mathbf{F}$, one and only one of three things holds: $x = 0$, $x > 0$, $x < 0$.

Henceforth we will use inequalities throughout; we mention that these inequalities will compare real numbers. The complex numbers cannot be made into an ordered field. The following lemma about ordered fields does play an important role in our development of the complex number field $\mathbf{C}$.

Lemma 3.1. *Let $\mathbf{F}$ be an ordered field. For each $x \in \mathbf{F}$, we have $x^2 = x * x \geq 0$. If $x \neq 0$, then $x^2 > 0$. In particular, $1 > 0$.*

Proof. If $x = 0$, then $x^2 = 0$ by Proposition 2.1, and the conclusion holds. If $x > 0$, then $x^2 > 0$ by axiom 1) for an ordered field. If $x < 0$, then $-x > 0$, and hence $(-x)^2 > 0$. By 3) and 4) of Proposition 2.1 we get

$$x^2 = (-1)(-1)x^2 = (-x)(-x) = (-x)^2 > 0. \tag{8}$$

Thus, if $x \neq 0$, then $x^2 > 0$. □

By definition (see Section 3.1), the real number system $\mathbf{R}$ is an ordered field. The following simple corollary motivates the introduction of the complex number field $\mathbf{C}$.

Corollary 3.1. *There is no real number x such that $x^2 = -1$.*

3.1. The completeness axiom for the real numbers. In order to finally define the real number system $\mathbf{R}$, we require the notion of *completeness*. This notion is considerably more advanced than our discussion has been so far. The field axioms allow for algebraic laws, the order axioms allow for inequalities, and the completeness axiom allows for a good theory of *limits*. To introduce this axiom, we recall some basic notions from elementary real analysis. Let $\mathbf{F}$ be an ordered field. Let $S \subset \mathbf{F}$ be a subset. The set S is called *bounded* if there are elements m and M in $\mathbf{F}$ such that

$$m \leq x \leq M$$

for all x in S. The set S is called *bounded above* if there is an element M in $\mathbf{F}$ such that $x \leq M$ for all x in S, and it is called *bounded below* if there is an element m in $\mathbf{F}$ such that $m \leq x$ for all x in S. When these numbers exist, M is called an upper bound for S and m is called a lower bound for S. Thus S is bounded if and only if it is both bounded above and bounded below. The numbers m and M are not generally in S. For example, the set of negative rational numbers is bounded above, but the only upper bounds are 0 or positive numbers.

We can now introduce the completeness axiom for the real numbers. The fundamental notion is that of *least upper bound*. If M is an upper bound for S, then any number larger than M is of course also an upper bound. The term least

upper bound means the *smallest* possible upper bound; the concept juxtaposes small and big. The least upper bound α of S is the smallest number that is greater than or equal to any member of S. See Figure 1.2. One cannot prove that such a number exists based on the ordered field axioms; for example, if we work within the realm of rational numbers, the set of x such that $x^2 < 2$ is bounded above, but it has no least upper bound. Mathematicians often use the word *supremum* instead of least upper bound; thus $\sup(S)$ denotes the least upper bound of S. Postulating the existence of least upper bounds as in the next definition uniquely determines the real numbers.

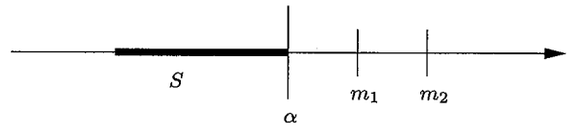

Figure 1.2. Upper bounds.

Definition 3.2. An ordered field $\mathbf{F}$ is *complete* if whenever S is a nonempty subset of $\mathbf{F}$ and S is bounded above, then S has a least upper bound in $\mathbf{F}$.

We could have instead decreed that each nonempty subset of $\mathbf{F}$ that is bounded below has a greatest lower bound (or *infimum*). The two statements are equivalent after replacing S with the set $-S$ of additive inverses of elements of S.

In a certain precise sense, called isomorphism, there is a unique complete ordered field. We will assume uniqueness and get the ball rolling by making the fundamental definition:

Definition 3.3. The real number system $\mathbf{R}$ is the unique complete ordered field.

3.2. What is a natural number? We pause to briefly consider how the natural numbers fit within the real numbers. In our approach, the real number system is taken as the starting point for discussion. From an intuitive point of view we can think of the natural numbers as the set $\{1, 1+1, 1+1+1, ...\}$. To be more precise, we proceed in the following manner.

Definition 3.4. A subset S of $\mathbf{R}$ is called *inductive* if whenever $x \in S$, then $x + 1 \in S$.

Definition 3.5. The set of natural numbers $\mathbf{N}$ is the intersection of all inductive subsets of $\mathbf{R}$ that contain 1.

Thus $\mathbf{N}$ is a subset of $\mathbf{R}$, and $1 \in \mathbf{N}$. Furthermore, if $n \in \mathbf{N}$, then n is an element of every inductive subset of $\mathbf{R}$. Hence $n + 1$ is also an element of every inductive subset of $\mathbf{R}$, and therefore $n+1$ is also in $\mathbf{N}$. Thus $\mathbf{N}$ is itself an inductive set; we could equally well have defined $\mathbf{N}$ to be the *smallest* inductive subset of $\mathbf{R}$ containing 1. As a consequence we obtain the *principle of mathematical induction*:

Proposition 3.1 (Mathematical induction). *Let S be an inductive subset of $\mathbf{N}$ such that $1 \in S$. Then $S = \mathbf{N}$.*

This proposition provides a method of proof, called induction, surely known to many readers. For each $n \in \mathbf{N}$, let P_n be a mathematical statement. To verify that P_n is a true statement for each n, it suffices to show two things: first, P_1 is true; second, for all k, whenever P_k is true, then P_{k+1} is true. The reason is that the set of n for which P_n is true is then an inductive set containing 1; by Proposition 3.1 this set is $\mathbf{N}$.

Figure 1.3. Induction.

▶ **Exercise 1.8.** Apply the principle of mathematical induction to establish the *well-ordering principle*: every nonempty subset of $\mathbf{N}$ contains a least element.

▶ **Exercise 1.9.** It is of course obvious that there is no natural number between 0 and 1. Prove it!

▶ **Exercise 1.10.** For a constant C put $f(x) = x + C$. Find a formula for the composition of f with itself n times. Prove the formula by induction.

▶ **Exercise 1.11.** For nonzero constants A and B, put $f(x) = A(x + B) - B$. Find a formula for the composition of f with itself n times. Prove the formula by induction. Find a short proof by expressing the behavior of f in simple steps.

▶ **Exercise 1.12.** For constants M, C with $M \neq 1$ put $f(x) = Mx + C$. Find a formula for the composition of f with itself n times. Suggestion: Write f in the notation of the previous exercise.

We close this section by proving a precise statement to the effect that many small things make a big thing. This seemingly evident but yet surprisingly subtle property of $\mathbf{R}$, as stated in Proposition 3.2, requires the completeness axiom for its proof. The proposition does not hold in all ordered fields. In other words, there exist ordered fields $\mathbf{F}$ with the following striking property: $\mathbf{F}$ contains the natural numbers, but it also contains *super numbers*, namely elements larger than any natural number. For the real numbers, however, things are as we believe. The natural numbers are an unbounded subset of the real numbers.

Proposition 3.2 (Archimedean property). *Given positive real numbers x and ϵ, there is a positive integer n such that $n\epsilon > x$. Equivalently, given $y > 0$, there is an $n \in \mathbf{N}$ such that $\frac{1}{n} < y$.*

Proof. If the first conclusion were false, then every natural number would be bounded above by $\frac{x}{\epsilon}$. If the second conclusion were false, then every natural number would be bounded above by $\frac{1}{y}$. Thus, in either case, $\mathbf{N}$ would be bounded above. We prove otherwise. If $\mathbf{N}$ were bounded above, then by the completeness axiom $\mathbf{N}$ would have a least upper bound K. But then $K - 1$ would not be an upper bound, and hence we could find an integer n with $K - 1 < n \leq K$. But then $K < n + 1$; since $n + 1 \in \mathbf{N}$, we contradict K being an upper bound. Thus $\mathbf{N}$ is unbounded above and the Archimedean property follows. □

▶ **Exercise 1.13.** Type "Non-Archimedean Ordered Field" into an internet search engine and see what you find. Then try to understand one of the examples.

3.3. Limits. Completeness in the sense of Definition 3.2 (for Archimedean ordered fields) is equivalent to a notion involving limits of Cauchy sequences. See Remark 3.1. We will carefully discuss these definitions from a calculus or beginning real analysis course. First we remind the reader of some elementary properties of the absolute value function. We gain intuition by thinking in terms of distance.

Definition 3.6. For $x \in \mathbf{R}$, we define $|x|$ by $|x| = x$ if $x \geq 0$ and $|x| = -x$ if $x < 0$. Thus $|x|$ represents the distance between x and 0. In general, we define the *distance* $\delta(x, y)$ between real numbers x and y by

$$\delta(x, y) = |x - y|.$$

▶ **Exercise 1.14.** Show that the absolute value function on $\mathbf{R}$ satisfies the following properties:

1) $|x| \geq 0$ for all $x \in \mathbf{R}$, and $|x| = 0$ if and only if $x = 0$.

2) $-|x| \leq x \leq |x|$ for all $x \in \mathbf{R}$.

3) $|x + y| \leq |x| + |y|$ for all $x, y \in \mathbf{R}$ (the triangle inequality).

4) $|a - c| \leq |a - b| + |b - c|$ for all $a, b, c \in \mathbf{R}$ (second form of the triangle inequality).

▶ **Exercise 1.15.** Why are properties 3) and 4) of the previous exercise called triangle inequalities?

We make several comments about Exercise 1.14. First of all, one can prove property 3) in two rather different ways. One way starts with property 2) for x and y and adds the results. Another way involves squaring. Property 4) is crucial because of its interpretation in terms of distances. Mathematicians have abstracted these properties of the absolute value function and introduced the concept of a *metric space*. See Section 6.

We recall that a sequence $\{x_n\}$ of real numbers is a function from $\mathbf{N}$ to $\mathbf{R}$. The real number x_n is called the n-th term of the sequence. The notation $x_1, x_2, ..., x_n, ...$, where we list the terms of the sequence in order, amounts to listing the values of the function. Thus $x : \mathbf{N} \to \mathbf{R}$ is a function, and we write x_n instead of $x(n)$. The intuition gained from this alteration of notation is especially valuable when discussing limits.

Definition 3.7. Let $\{x_n\}$ be a sequence of real numbers. Assume $L \in \mathbf{R}$.

- LIMIT. We say that "the limit of x_n is L" or that "x_n converges to L", and we write $\lim_{n\to\infty} x_n = L$ if the following statement holds: For all $\epsilon > 0$, there is an $N \in \mathbf{N}$ such that $n \geq N$ implies $|x_n - L| < \epsilon$.
- CAUCHY. We say that $\{x_n\}$ is a *Cauchy sequence* if the following statement holds: For all $\epsilon > 0$, there is an $N \in \mathbf{N}$ such that $m, n \geq N$ implies $|x_m - x_n| < \epsilon$.

When there is no real number L for which $\{x_n\}$ converges to L, we say that $\{x_n\}$ *diverges*.

The definition of the limit demands that the terms eventually get arbitrarily close to a given L. The definition of a Cauchy sequence states that the terms of the sequence eventually get arbitrarily close to each other. The most fundamental result in real analysis is that a sequence of real numbers converges if and only if it is a Cauchy sequence. The word *complete* has several similar uses in mathematics; it often refers to a metric space in which being Cauchy is a necessary and sufficient condition for convergence of a sequence. See Section 6. The following subtle remark indicates a slightly different way one can define the real numbers.

Remark 3.1. Consider an ordered field $\mathbf{F}$ satisfying the Archimedean property. In other words, given positive elements x and y, there is an integer n such that y added to itself n times exceeds x. Of course we write ny for this sum. It is possible to consider limits and Cauchy sequences in $\mathbf{F}$. Suppose that each Cauchy sequence in $\mathbf{F}$ has a limit in $\mathbf{F}$. One can then derive the least upper bound property, and $\mathbf{F}$ must be the real numbers $\mathbf{R}$. Hence we could give the definition of the real number system by decreeing that $\mathbf{R}$ is an ordered field satisfying the Archimedean property and that $\mathbf{R}$ is complete in the sense of Cauchy sequences.

We return to the real numbers. A sequence $\{x_n\}$ of real numbers is *bounded* if and only if its set of values is a bounded subset of $\mathbf{R}$. A convergent sequence must of course be bounded; with finitely many exceptions all the terms are within distance 1 from the limit. Similarly a Cauchy sequence must be bounded; with finitely many exceptions all the terms are within distance 1 of some particular x_N.

Proving that a convergent sequence must be Cauchy uses what is called an $\frac{\epsilon}{2}$ argument. Here is the idea: if the terms are eventually within distance $\frac{\epsilon}{2}$ of some limit L, then they are eventually within distance ϵ of each other. Proving the converse assertion is much more subtle; somehow one must find a candidate for the limit just knowing that the terms are close to each other. See for example [**8, 20**]. The proofs rely on the notion of subsequence, which we define now, but which we do not use meaningfully until Chapter 8. Let $\{x_n\}$ be a sequence of real numbers and let $k \to n_k$ be an increasing function. We write $\{x_{n_k}\}$ for the subsequence of $\{x_n\}$ whose k-th term is x_{n_k}. The proof that a Cauchy sequence converges amounts to first finding a convergent subsequence and then showing that the sequence itself converges to the same limit.

We next prove a basic fact that often allows us to determine convergence of a sequence without knowing the limit in advance. A sequence $\{x_n\}$ is called *nondecreasing* if, for each n, we have $x_{n+1} \geq x_n$. It is called *nonincreasing* if, for each n, we have $x_{n+1} \leq x_n$. It is called *monotone* if it is either nonincreasing or nondecreasing. The following fundamental result, illustrated by Figure 1.4, will get used occasionally in this book. It can be used also to establish that a Cauchy sequence of real numbers has a limit.

Proposition 3.3. *A bounded monotone sequence of real numbers has a limit.*

Proof. We claim that a nondecreasing sequence converges to its least upper bound (supremum) and that a nonincreasing sequence converges to its greatest lower bound (infimum). We prove the first, leaving the second to the reader. Suppose for all n we have

$$x_1 \leq ... \leq x_n \leq x_{n+1} \leq ... \leq M.$$

Let α be the least upper bound of the set $\{x_n\}$. Then, given $\epsilon > 0$, the number $\alpha - \epsilon$ is not an upper bound, and hence there is some x_N with $\alpha - \epsilon < x_N \leq \alpha$. By the nondecreasing property, if $n \geq N$, then

$$\alpha - \epsilon < x_N \leq x_n \leq \alpha < \alpha + \epsilon. \tag{9}$$

But (9) yields $|x_n - \alpha| < \epsilon$ and hence provides us with the needed N in the definition of the limit. Thus $\lim_{n\to\infty}(x_n) = \alpha$. □

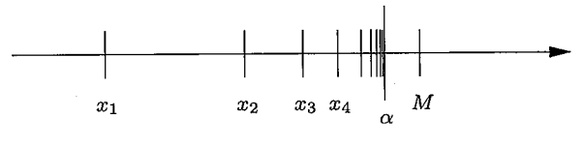

Figure 1.4. Monotone convergence.

Remark 3.2. Let $\{x_n\}$ be a monotone sequence of real numbers. Then $\{x_n\}$ converges if and only if it is bounded. Proposition 3.3 guarantees that it converges if it is bounded. Since a convergent sequence must be bounded, the converse holds as well. Monotonicity is required; for example, the sequence $(-1)^n$ is bounded but it does not converge.

The next few pages provide the basic real analysis needed as background material. In particular the material on square roots is vital to the development.

▶ **Exercise 1.16.** Finish the proof of Proposition 3.3; in other words, show that a nonincreasing bounded sequence converges to its greatest lower bound.

▶ **Exercise 1.17.** If c is a constant and $\{x_n\}$ converges, prove that $\{cx_n\}$ converges. Try to arrange your proof such that the special case $c = 0$ need not be considered separately. Prove that the sum and product of convergent sequences are convergent.

▶ **Exercise 1.18.** Assume $\{x_n\}$ converges to 0 and that $\{y_n\}$ is bounded. Prove that their product converges to 0.

An extension of the notion of limit of sequence is often useful in real analysis. We pause to introduce the idea and refer to [**20**] for applications and considerably more discussion. When S is a bounded and nonempty subset of $\mathbf{R}$, we write as usual $\inf(S)$ for the greatest lower bound of S and $\sup(S)$ for the least upper bound of S. Let now $\{x_n\}$ be a bounded sequence of real numbers. For each k, consider the set $X_k = \{x_n : n \geq k\}$. Then these sets are bounded as well. Furthermore the bounded sequence of real numbers defined by $\inf(X_k)$ is nondecreasing and the bounded sequence of real numbers $\sup(X_k)$ is nonincreasing. By the monotone convergence theorem these sequences necessarily have limits, called $\lim \inf(x_n)$ and $\lim \sup(x_n)$. These limits are equal **if and only if** $\lim(x_n)$ exists, in which case

all three values are the same. By contrast, let $x_n = (-1)^n$. Then $\liminf(x_n) = -1$ and $\limsup(x_n) = 1$. Occasionally in the subsequent discussion we can replace limit by lim sup and things still work.

Next we turn to the concept of continuity, which we also define in terms of sequences.

Definition 3.8. Let $f : \mathbf{R} \to \mathbf{R}$ be a function. Then f is *continuous at* a if whenever $\{x_n\}$ is a sequence and $\lim_{n\to\infty} x_n = a$, then $\lim_{n\to\infty} f(x_n) = f(a)$. Also, f is *continuous on a set* S if it is continuous at each point of the set. When S is $\mathbf{R}$ or when S is understood from the context to be the domain of f, we usually say "f is continuous" rather than the longer phrase "f is continuous on S".

▶ **Exercise 1.19.** Prove that the sum and product of continuous functions are continuous. If c is a constant and f is continuous, prove that cf is continuous.

▶ **Exercise 1.20.** Prove that f is continuous at a if and only if the following holds. For each $\epsilon > 0$, there is a $\delta > 0$ such that $|x - a| < \delta$ implies $|f(x) - f(a)| < \epsilon$.

We close this section by showing how the completeness axiom impacts the existence of square roots. First we recall the standard fact that there is no rational square root of 2, by giving a somewhat unusual proof. See Exercise 1.22 for a compelling generalization. These proofs are based on inequalities. For example, the order axioms yield the following: $0 < a < b$ implies $0 < a^2 < ab < b^2$; we use such inequalities without comment below.

Proposition 3.4. *There is no rational number whose square is* 2.

Proof. Seeking a contradiction, we suppose that there are integers m, n such that $(\frac{m}{n})^2 = 2$. We may assume that m and n are positive. Of all such representations we may assume that we have chosen the one for which n is the smallest possible positive integer. The equality $m^2 = 2n^2$ implies the inequality $2n > m > n$. Now we compute

$$\frac{m}{n} = \frac{m(m-n)}{n(m-n)} = \frac{m^2 - mn}{n(m-n)} = \frac{2n^2 - mn}{n(m-n)} = \frac{2n-m}{m-n}. \tag{10}$$

Thus $\frac{2n-m}{m-n}$ is also a square root of 2. Since $0 < m - n < n$, formula (10) provides a second way to write the fraction $\frac{m}{n}$; the second way has a positive denominator, smaller than n. We have therefore contradicted our choice of n. Hence there is no rational number whose square is 2. □

Although there is no *rational* square root of 2, we certainly believe that a positive *real* square root of 2 exists. For example, the length of the diagonal of the unit square should be $\sqrt{2}$. We next prove, necessarily relying on the completeness axiom, that each positive real number has a square root.

Theorem 3.1. *If $t \in \mathbf{R}$ and $t \geq 0$, then there is an $x \in \mathbf{R}$ with $x^2 = t$.*

Proof. This proof is somewhat sophisticated and can be omitted on first reading. If $t = 0$, then t has the square root 0. Hence we may assume that $t > 0$. Let S denote the set of real numbers x such that $x^2 < t$. This set is nonempty, because $0 \in S$. We claim that $M = \max(1, t)$ is an upper bound for S. To check the claim,

we note first that $x^2 < 1$ implies $x < 1$, because $x \geq 1$ implies $x^2 \geq 1$. Therefore if $t < 1$, then 1 is an upper bound for S. On the other hand, if $t \geq 1$, then $t \leq t^2$. Therefore $x^2 < t$ implies $x^2 \leq t^2$ and hence $x \leq t$. Therefore in this case t is an upper bound for S. In either case, S is bounded above by M and is nonempty. By the completeness axiom, S has a least upper bound α. We claim that $\alpha^2 = t$.

To prove the claim, we use the trichotomy property. We will rule out the cases $\alpha^2 < t$ and $\alpha^2 > t$. In each case we use the Archimedean property to find a positive integer n whose reciprocal is sufficiently small. Then we can add or subtract $\frac{1}{n}$ to α and obtain a contradiction. Here are the details. If $\alpha^2 > t$, then Proposition 3.2 guarantees that we can find an integer n such that

$$\frac{2\alpha}{n} < \alpha^2 - t.$$

We then have

$$(\alpha - \frac{1}{n})^2 = \alpha^2 - \frac{2\alpha}{n} + \frac{1}{n^2} > \alpha^2 - \frac{2\alpha}{n} > t.$$

Thus $\alpha - \frac{1}{n}$ is an upper bound for S, but it is smaller than α. We obtain a contradiction. Suppose next that $\alpha^2 < t$. We can find $n \in \mathbf{N}$ (Exercise 1.21) such that

$$\frac{2\alpha n + 1}{n^2} < t - \alpha^2. \tag{11}$$

This time we obtain

$$(\alpha + \frac{1}{n})^2 = \alpha^2 + \frac{2\alpha}{n} + \frac{1}{n^2} < t.$$

Since $\alpha + \frac{1}{n}$ is bigger than α and yet it is also an upper bound for S, again we obtain a contradiction. By trichotomy we must therefore have $\alpha^2 = t$. □

The kind of argument used in the proof of Theorem 3.1 epitomizes proofs in basic real analysis. In this setting one cannot prove an equality by algebraic reasoning; one requires the completeness axiom and analytic reasoning.

▶ **Exercise 1.21.** For $t - \alpha^2 > 0$, prove that there is an $n \in \mathbf{N}$ such that (11) holds.

▶ **Exercise 1.22.** Mimic the proof of Proposition 3.4 to prove the following statement. If k is a positive integer, then the square root of k must be either an integer or an irrational number. Suggestion: Multiply $\frac{m}{n}$ by $\frac{m-nq}{m-nq}$ for a suitable integer q.

4. The complex numbers

We are finally ready to introduce the complex numbers $\mathbf{C}$. The equation $x^2 + 1 = 0$ will have two solutions in $\mathbf{C}$. Once we allow a solution to this equation, we find via the quadratic formula and Lemma 4.1 below that we can solve all quadratic polynomial equations. With deeper work, we can solve any (nonconstant) polynomial equation over $\mathbf{C}$. We will prove this result, called the *fundamental theorem of algebra*, in Chapter 8.

Our first definition of $\mathbf{C}$ arises from algebraic reasoning. As usual, we write $\mathbf{R}^2$ for the set of ordered pairs (x, y) of real numbers. To think geometrically, we

identify the point (x, y) with the arrow from the origin $(0, 0)$ to the point (x, y). We know how to add vectors; hence we define

$$(x, y) + (a, b) = (x + a, y + b). \tag{12}$$

This formula amounts to adding vectors in the usual geometric manner. See Figure 1.5. More subtle is our definition of multiplication

$$(x, y) * (a, b) = (xa - yb, xb + ya). \tag{13}$$

Let us temporarily write $\mathbf{0}$ for $(0, 0)$ and $\mathbf{1}$ for $(1, 0)$. We claim that the operations in equations (12) and (13) turn $\mathbf{R}^2$ into a field.

We must first verify that both addition and multiplication are commutative and associative. The verifications are rather trivial, especially for addition:

$$(x, y) + (a, b) = (x + a, y + b) = (a + x, b + y) = (a, b) + (x, y),$$

$$\begin{aligned}((x, y) + (a, b)) + (s, t) &= (x + a, y + b) + (s, t) = (x + a + s, y + b + t) \\ &= (x, y) + (a + s, b + t) = (x, y) + ((a, b) + (s, t)).\end{aligned}$$

Here are the computations for multiplication:

$$(x, y) * (a, b) = (xa - yb, xb + ya) = (ax - by, ay + bx) = (a, b) * (x, y),$$

$$\begin{aligned}((x, y) * (a, b)) * (s, t) &= (xa - by, xb + ya) * (s, t) \\ &= (xas - bys - txb - tya, xbt + yat - (xas - bys)) = (x, y) * ((a, b) * (s, t)).\end{aligned}$$

We next verify that $\mathbf{0}$ and $\mathbf{1}$ have the desired properties.

$$(x, y) + (0, 0) = (x, y),$$

$$(x, y) * (1, 0) = (x1 - y0, x0 + y1) = (x, y).$$

The additive inverse of (x, y) is easily checked to be $(-x, -y)$. When $(x, y) \neq (0, 0)$, the multiplicative inverse of (x, y) is easily checked to be

$$\frac{\mathbf{1}}{(x, y)} = \left(\frac{x}{x^2 + y^2}, \frac{-y}{x^2 + y^2}\right). \tag{14}$$

Checking the distributive law is not hard, but it is tedious and left to the reader in Exercise 1.23.

These calculations provide the starting point for discussion.

Theorem 4.1. *Formulas* (12) *and* (13) *make* $\mathbf{R}^2$ *into a field.*

The verification of the field axioms given above is rather dull and uninspired. We do note, however, that $(-1, 0)$ is the additive inverse of $(1, 0) = \mathbf{1}$ and that $(0, 1) * (0, 1) = (-1, 0)$. Hence there is a square root of $-\mathbf{1}$ in this field.

The ordered pair notation for elements is a bit awkward. We wish to give two alternative definitions of $\mathbf{C}$ where things are more elegant.

What have we done so far? Our first definition of $\mathbf{C}$ as pairs of real numbers gave an unmotivated recipe for multiplication; it seems almost a fluke that we obtain a field using this definition. Furthermore computations seem clumsy. A more appealing approach begins by introducing a formal symbol $\mathbf{i}$ and defining $\mathbf{C}$

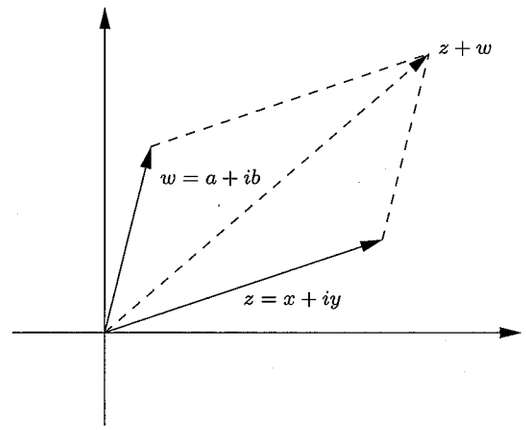

Figure 1.5. Addition of complex numbers.

to be the set of expressions of the form $a + \mathbf{i}b$ for real numbers a, b. We add and multiply as expected, using the distributive law; then we set $\mathbf{i}^2$ equal to -1. Thus

$$(x + \mathbf{i}y) + (a + \mathbf{i}b) = (x + a) + \mathbf{i}(y + b), \tag{15}$$

$$(x + \mathbf{i}y) * (a + \mathbf{i}b) = xa + \mathbf{i}(ya + xb) + \mathbf{i}^2(yb) = (xa - yb) + \mathbf{i}(ya + xb). \tag{16}$$

Equations (15) and (16) give the same results as (12) and (13). While this new approach is more elegant, it makes some readers feel uneasy. After all, we are assuming the existence of an object, namely $0 + \mathbf{i}1$, whose square is -1. In the first approach we never assume the existence of such a thing, but such a thing does exist: the square of $(0, 1)$ is $(-1, 0)$, which is the additive inverse of $(1, 0)$.

The reader will be on safe logical ground if he or she regards the above paragraph as an abbreviation for the previous discussion. In the next section we will give two additional equivalent ways of defining $\mathbf{C}$.

▶ **Exercise 1.23.** Prove the distributive law for addition and multiplication, as defined in (12) and (13). Do the same using (15) and (16). Compare.

The next lemma reveals a crucial difference between $\mathbf{R}$ and $\mathbf{C}$.

Lemma 4.1. *The complex numbers do not form an ordered field.*

Proof. Assume that a positive subset P exists. By Lemma 3.1, each nonzero square is in P. Since $1^2 = 1$ and $i^2 = -1$, both 1 and -1 are squares and hence must be positive, contradicting 2) of Definition 3.1. □

5. Alternative definitions of C

In this section we discuss alternative approaches to defining $\mathbf{C}$. We use some basic ideas from linear and abstract algebra that might be new to many students. The primary purpose of this section is to assuage readers who find the rules (12) and (13) unappealing but who find the rules (15) and (16) dubious, because we

introduced an object $\mathbf{i}$ whose square is -1. The first approach uses matrices of real numbers, and it conveys significant geometric information. The second approach fully justifies starting with (15) and (16) and it provides a quintessential example of what mathematicians call a *quotient space.*

A matrix approach to C. The matrix definition of $\mathbf{C}$ uses two-by-two matrices of real numbers and some of the ideas are crucial to subsequent developments. In this approach we think of $\mathbf{C}$ as the set of two-by-two matrices of the form (18), thereby presaging the Cauchy-Riemann equations which will appear throughout the book. In some sense we identify a complex number with the operation of multiplication by that complex number. This approach is especially useful in complex geometry.

We can regard a complex number as a special kind of linear transformation of $\mathbf{R}^2$. A general linear transformation $(x, y) \to (ax + cy, bx + dy)$ is given by a two-by-two matrix M of real numbers:

$$M = \begin{pmatrix} a & c \\ b & d \end{pmatrix}. \tag{17}$$

A complex number will be a special kind of two-by-two matrix. Given a pair of real numbers a, b and motivated by (13), we consider the mapping $L : \mathbf{R}^2 \to \mathbf{R}^2$ defined by

$$L(x, y) = (ax - by, bx + ay).$$

The matrix representation (in the standard basis) of this linear mapping L is the two-by-two matrix

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix}. \tag{18}$$

We say that a two-by-two matrix of real numbers satisfies the *Cauchy-Riemann equations* if it has the form (18). A real linear transformation from $\mathbf{R}^2$ to itself whose matrix representation satisfies (18) corresponds to a *complex linear* transformation from $\mathbf{C}$ to itself, namely multiplication by $a + ib$.

In this approach we *define* a complex number to be a two-by-two matrix (of real numbers) satisfying the Cauchy-Riemann equations. We add and multiply matrices in the usual manner. We then have an additive identity $\mathbf{0}$, a multiplicative identity $\mathbf{1}$, an analogue of $\mathbf{i}$, and inverses of nonzero elements, defined as follows:

$$\mathbf{0} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \tag{19}$$

$$\mathbf{1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{20}$$

$$\mathbf{i} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}. \tag{21}$$

If a and b are not both 0, then $a^2 + b^2 > 0$. Hence in this case the matrix

$$\begin{pmatrix} \frac{a}{a^2+b^2} & \frac{b}{a^2+b^2} \\ \frac{-b}{a^2+b^2} & \frac{a}{a^2+b^2} \end{pmatrix} \tag{22}$$

makes sense and satisfies the Cauchy-Riemann equations. Note that

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix} \begin{pmatrix} \frac{a}{a^2+b^2} & \frac{b}{a^2+b^2} \\ \frac{-b}{a^2+b^2} & \frac{a}{a^2+b^2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathbf{1}. \tag{23}$$

Thus (22) yields the formula $\frac{a-ib}{a^2+b^2}$ for the reciprocal of the nonzero complex number $a + ib$, expressed instead in matrix notation.

Thus $\mathbf{C}$ can be defined to be the set of two-by-two matrices satisfying the Cauchy-Riemann equations. Addition and multiplication are defined as usual for matrices. The additive identity $\mathbf{0}$ is given by (19) and the multiplicative identity $\mathbf{1}$ is given by (20). The resulting mathematical system is a field, and the element $\mathbf{i}$ defined by (21) satisfies $\mathbf{i}^2 + \mathbf{1} = \mathbf{0}$. This method of defining $\mathbf{C}$ should appease readers who on philosophical grounds question the existence of complex numbers.

▶ **Exercise 1.24.** Show that the square of the matrix in (21) is the negative of the matrix in (20); in other words, show that $\mathbf{i}^2 = -\mathbf{1}$.

▶ **Exercise 1.25.** Suppose $a^2 + b^2 = 1$ in (18). What is the geometric meaning of multiplication by L?

▶ **Exercise 1.26.** Suppose $b = 0$ in (18). What is the geometric meaning of multiplication by L?

▶ **Exercise 1.27.** Show that there are no real numbers x and y such that

$$\frac{1}{x} + \frac{1}{y} = \frac{1}{x+y}.$$

Show on the other hand that there are complex numbers z and w such that

$$\frac{1}{z} + \frac{1}{w} = \frac{1}{z+w}. \tag{24}$$

Describe all pairs (z, w) satisfying (24).

▶ **Exercise 1.28.** Describe all pairs A and B of two-by-two matrices of real numbers for which A^{-1} and B^{-1} exist and

$$A^{-1} + B^{-1} = (A + B)^{-1}.$$

Remark 5.1. Such pairs of n-by-n matrices exist if and only if n is even; the reason is intimately connected with complex analysis.

An algebraic definition of C. We next describe $\mathbf{C}$ as a *quotient space.* This approach allows us to regard a complex number as an expression $a + ib$, where $i^2 = -1$, as we wish to do. We will therefore define $\mathbf{C}$ in terms of the polynomial ring divided by an ideal. The reader may skip this section without loss of understanding.

First we recall the general notion of an *equivalence relation.* Let S be a set. We can think of an equivalence relation on S as being defined via a symbol $\cong$. We decree that certain pairs of elements $s, t \in S$ are *equivalent*; if so, we write $s \cong t$. The following three axioms must hold:

- For all $s \in S$, $s \cong s$ (reflexivity).
- For all $s, t \in S$, $s \cong t$ if and only if $t \cong s$ (symmetry).
- For all $s, t, u \in S$, $s \cong t$ and $t \cong u$ together imply $s \cong u$ (transitivity).

Given an equivalence relation $\cong$ on S, we partition S into *equivalence classes.* All the elements in a single equivalence class are equivalent, and no other member of S is equivalent to any of these elements. We have already seen two elementary examples. First, fractions $\frac{a}{b}$ and $\frac{c}{d}$ are equivalent if and only if they represent the same real number, that is, if and only if $ad = bc$. Thus a rational number may be regarded as an equivalence class of pairs of integers. Second, when doing arithmetic modulo p, we regard two integers as being in the same equivalence class if their difference is divisible by p.

▶ **Exercise 1.29.** The precise definition of modular arithmetic involves equivalence classes; we add and multiply equivalence classes (rather than numbers). Show that addition and multiplication modulo p are well-defined concepts. In other words, do the following. Assume m_1 and m_2 are in the same equivalence class modulo p and that n_1 and n_2 are also in the same equivalence class (not necessarily the same class m_1 and m_2 are in). Show that $m_1 + n_1$ and $m_2 + n_2$ are in the same equivalence class modulo p. Do the same for multiplication.

▶ **Exercise 1.30.** Let S be the set of students at a college. For $s, t \in S$, consider the relation $s \cong t$ if s and t take a class together. Is this relation an equivalence relation?

Let $\mathbf{R}[t]$ denote the collection of polynomials in one variable, with real coefficients. An element p of $\mathbf{R}[t]$ can be written

$$p = \sum_{j=0}^{d} a_j t^j,$$

where $a_j \in \mathbf{R}$. Notice that the sum is finite. Unless all the a_j are 0, there is a largest d for which $a_j \neq 0$. This number d is called the *degree* of the polynomial. When all the a_j equal 0, we call the resulting polynomial the *zero polynomial* and agree that it has no degree. (In some contexts, one assigns the symbol $-\infty$ to be the degree of the zero polynomial.) The sum and the product of polynomials are defined as in high school mathematics. In many ways $\mathbf{R}[t]$ resembles the integers $\mathbf{Z}$. Each is a commutative ring under the operations of sum and product. Unique factorization into irreducible elements holds in both settings, and the division algorithm works the same as well. See [4] or [8] for more details. Given polynomials p and g, we say that p is a *multiple* of g, or equivalently that g *divides* p, if there is a polynomial q with $p = gq$.

The polynomial $1 + t^2$ is irreducible, in the sense that it cannot be written as a product of two polynomials, each of lower degree, with real coefficients. The set I of polynomials divisible by $1 + t^2$ is called the *ideal generated by* $1 + t^2$. Given two polynomials p, q, we say that they are equivalent modulo I if $p - q \in I$, in other words, if $p - q$ is divisible by $1 + t^2$. We observe that the three properties of an equivalence relation hold:

- For all p, $p \cong p$.
- For all p, q, $p \cong q$ if and only if $q \cong p$.
- If $p \cong q$ and $q \cong r$, then $p \cong r$.

This equivalence relation partitions the set $\mathbf{R}[t]$ into equivalence classes; the situation is strikingly similar to modular arithmetic. Given a polynomial $p(t)$, we use the division algorithm to write $p(t) = q(t)(1 + t^2) + r(t)$, where the remainder r has degree at most one. Thus $r(t) = a + bt$ for some a, b, and this r is the unique first-degree polynomial equivalent to p. In the case of modular arithmetic we used the remainder upon division by the modulus; here we use the remainder upon division by $t^2 + 1$.

▶ **Exercise 1.31.** Verify the transitivity property of equivalence modulo I.

Standard notation in algebra writes $\mathbf{R}[t]/(1 + t^2)$ for the set of equivalence classes. We can add and multiply in $\mathbf{R}[t]/(1 + t^2)$. As usual, the sum (or product) of equivalence classes P and Q is defined to be the equivalence class of the sum $p + q$ (or the product pq) of members; the result is independent of the choice. An equivalence class then can be identified with a polynomial $a + bt$, and the sum and product of equivalence classes satisfies (15) and (16). In this setting we define $\mathbf{C}$ as the collection of equivalence classes with this natural sum and product:

$$\mathbf{C} = \mathbf{R}[t]/(1 + t^2). \tag{25}$$

Definition (25) allows us to set $t^2 = -1$ whenever we encounter a term of degree at least two. The irreducibility of $t^2 + 1$ matters. If we form $\mathbf{R}[t]/(p(t))$ for a reducible polynomial p, then the resulting object will not be a field. The reason is precisely parallel to the situation with modular arithmetic. If we consider $\mathbf{Z}/(n)$, then we get a field (written $\mathbf{F}_n$) if and only if n is prime.

▶ **Exercise 1.32.** Show that $\mathbf{R}[t]/(t^3 + 1)$ is not a field.

▶ **Exercise 1.33.** A polynomial $\sum_{k=0}^{d} c_k t^k$ in $\mathbf{R}[t]$ is equivalent to precisely one polynomial of the form $A + Bt$ in the quotient space. What is $A + Bt$ in terms of the coefficients c_k?

▶ **Exercise 1.34.** Prove the division algorithm in $\mathbf{R}[t]$. In other words, given polynomials p and g, with g not the zero polynomial, show that one can write $p = qg + r$ where either $r = 0$ or the degree of r is less than the degree of g. Show that q and r are uniquely determined by p and g.

▶ **Exercise 1.35.** For any polynomial p and any x_0, show that there is a polynomial q such that $p(x) = (x - x_0)q(x) + p(x_0)$.

6. A glimpse at metric spaces

Both the real number system and the complex number system provide intuition for the general notion of a metric space. This section can be omitted without impacting the logical development, but it should appeal to some readers.

Definition 6.1. Let X be a set. A *distance function* on X is a function $\delta : X \times X \to \mathbf{R}$ such that the following hold:

1) $\delta(x, y) \geq 0$ for all $x, y \in X$ (distances are nonnegative).

2) $\delta(x, y) = 0$ if and only if $x = y$ (distinct points have positive distance between them; a point has 0 distance to itself).

Chapter 2

Complex Numbers

The main point of our work in Chapter 1 was to provide a precise definition of the complex number field, based upon the existence of the real number field. While we will continue to work with the relationships between real numbers and complex numbers, our perspective will evolve toward thinking of complex numbers as the objects of interest. The reader will surely be delighted by how often this perspective leads to simpler computations, shorter proofs, and more elegant reasoning.

1. Complex conjugation

One of the most remarkable features of complex variable theory is the role played by complex conjugation. There are two square roots of -1, namely $\pm i$. When we make a choice of one of these, we create a kind of asymmetry. The mathematics must somehow keep track of the fundamental symmetry; these ideas lead to fascinating consequences.

Recall from Lemma 4.1 of Chapter 1 that $\mathbf{C}$ is not an ordered field. Therefore all inequalities used will compare real numbers. As we note below, real numbers are precisely those complex numbers unchanged by taking complex conjugates. Hence the fundamental issues involving inequalities also revolve around complex conjugation.

Definition 1.1. For x, y real, put $z = x + iy$. We write $x = \mathrm{Re}(z)$ and $y = \mathrm{Im}(z)$. The *complex conjugate* of z, written $\overline{z}$, is the complex number $x - iy$. The absolute value (or modulus) of z, written $|z|$, is the nonnegative real number $\sqrt{x^2 + y^2}$.

We make a few comments about the concepts in this definition. First we call x the *real part* and y the *imaginary part* of $x + iy$. Note that y is a real number; the imaginary part of z is not iy.

The absolute value function is fundamental in everything we do. Note first that $|z|^2 = z\overline{z}$. Next we naturally define the distance $\delta(z, w)$ between complex numbers z and w by $\delta(z, w) = |z - w|$. Then $\delta(z, w)$ equals the usual Euclidean distance

between these points in the plane. We can use the absolute value function to define *bounded set.* A subset S of $\mathbf{C}$ is bounded if there is a real number M such that $|z| \leq M$ for all z in $\mathbf{C}$. Thus S is bounded if and only if S is a subset of a ball about 0 of sufficiently large radius.

The function mapping z into its complex conjugate is called *complex conjugation.* Applying this function twice gets us back where we started; that is, $\overline{\overline{z}} = z$. This function satisfies many basic properties; see Lemma 1.1.

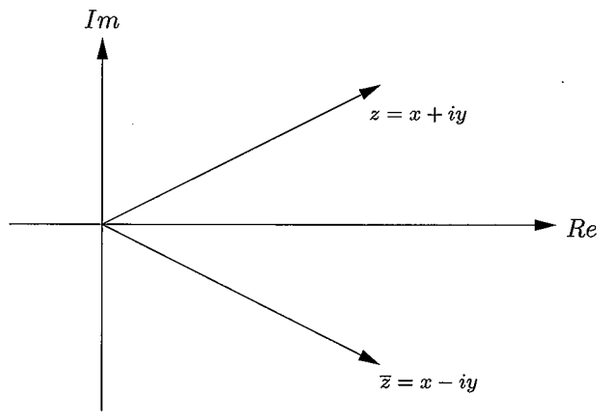

Figure 2.1. Complex conjugation.

Here is a way to stretch your imagination. Imagine that you have never heard of the real number system but that you know of a field called $\mathbf{C}$. Furthermore in this field there is a notion of convergent sequence making $\mathbf{C}$ complete in the sense of Cauchy sequences. Imagine also that there is a continuous function (called conjugation) $z \to \overline{z}$ satisfying properties 1), 2), and 3) from Lemma 1.1. Continuity guarantees that the conjugate of a limit of a sequence is the limit of the conjugates of the terms. We could then define the real numbers to be those complex numbers z for which $z = \overline{z}$.

For us the starting point was the real number system $\mathbf{R}$, and we constructed $\mathbf{C}$ from $\mathbf{R}$. We return to that setting.

Lemma 1.1. *The following formulas hold for all complex numbers z and w.*

1) $\overline{\overline{z}} = z$.
2) $\overline{z+w} = \overline{z} + \overline{w}$.
3) $\overline{zw} = \overline{z}\ \overline{w}$.
4) $|z|^2 = z\overline{z}$.
5) $\mathrm{Re}(z) = \frac{z+\overline{z}}{2}$.
6) $\mathrm{Im}(z) = \frac{z-\overline{z}}{2i}$.
7) *A complex number z is real if and only if $z = \overline{z}$.*
8) $|\overline{z}| = |z|$.

Proof. Left to the reader as an exercise. □

▶ **Exercise 2.1.** Prove Lemma 1.1. In each case, interpret the formula using Figure 2.1.

▶ **Exercise 2.2.** For a subset S of $\mathbf{C}$, define S^* by $z \in S^*$ if and only if $\overline{z} \in S$. Show that S^* is bounded if and only if S is bounded.

▶ **Exercise 2.3.** Show for all complex numbers z and w that

$$|z+w|^2 + |z-w|^2 = 2(|z|^2 + |w|^2).$$

Interpret geometrically.

▶ **Exercise 2.4.** Suppose that $|a| < 1$ and that $|z| \leq 1$. Prove that

$$\left|\frac{z-a}{1-\overline{a}z}\right| \leq 1.$$

Comment: This fact is important in non-Euclidean geometry.

▶ **Exercise 2.5.** Let c be real, and let $a \in \mathbf{C}$. Describe geometrically the set of z for which $az + \overline{az} = c$.

▶ **Exercise 2.6.** Let c be real, and let $a \in \mathbf{C}$. Suppose $|a|^2 \geq c$. Describe geometrically the set of z for which $|z|^2 + az + \overline{az} + c = 0$.

▶ **Exercise 2.7.** Let a and b be nonzero complex numbers. Call them *parallel* if one is a *real* multiple of the other. Find a simple algebraic condition for a and b to be parallel. (Use the imaginary part of something.)

▶ **Exercise 2.8.** Let a and b be nonzero complex numbers. Find an algebraic condition for a and b to be perpendicular. (Use a similar idea as in Exercise 2.7.)

done

▶ **Exercise 2.9.** What is the most general (defining) equation for a line in $\mathbf{C}$? (Hint: The imaginary part of something must be 0.) What is the most general (defining) equation of a circle in $\mathbf{C}$?

▶ **Exercise 2.10.** For $z, w \in \mathbf{C}$, prove that $|\mathrm{Re}(z)| \leq |z|$ and $|z+w| \leq |z| + |w|$. Then verify that the function $\delta(z,w) = |z-w|$ defines a distance function making $\mathbf{C}$ into a metric space. (See Definition 6.1 of Chapter 1.)

2. Existence of square roots

In this section we give an algebraic proof that we can find a square root of an arbitrary complex number. Some subtle points arise in the choice of signs. Later we give an easier geometric method.

Proposition 2.1. *For each $w \in \mathbf{C}$, there is a $z \in \mathbf{C}$ with $z^2 = w$.*

Proof. Given $w = a+bi$ with a, b real, we want to find $z = x+iy$ such that $z^2 = w$. If $a = b = 0$, then we put $z = 0$. Hence we may assume that $a^2 + b^2 \neq 0$. The equation $(x+iy)^2 = w$ yields the system of equations $x^2 - y^2 = a$ and $2xy = b$. We convert this system into a pair of linear equations for x^2 and y^2 by writing

$$(x^2+y^2)^2 = (x^2-y^2)^2 + 4x^2y^2 = a^2 + b^2. \tag{1}$$

The right-hand side of (1) is positive, and hence by Theorem 3.1 of Chapter 1 it has a positive real square root, and hence two real square roots. We choose the positive square root. We obtain the system

$$x^2 + y^2 = \sqrt{a^2 + b^2},$$
$$x^2 - y^2 = a.$$

We solve these two equations by adding and subtracting, obtaining

$$x^2 = \frac{a + \sqrt{a^2 + b^2}}{2}, \tag{2}$$

$$y^2 = \frac{-a + \sqrt{a^2 + b^2}}{2}. \tag{3}$$

First note that the right-hand sides of (2) and (3) are nonnegative, because $a^2 \leq a^2 + b^2$, and hence $\pm a \leq \sqrt{a^2 + b^2}$. Recall that we chose the positive square root of the expression $a^2 + b^2$. Now we would like to take the square roots of the right-hand sides of (2) and (3) to define x and y, but we are left with some ambiguity of signs. In general there are two possible signs for x and two possible signs for y, leading to four candidates for the solution. Yet we know from Lemma 2.1 of Chapter 1 that only two of these can work.

We resolve this ambiguity in the following manner, consistent with our convention that $\sqrt{t}$ denotes the positive square root of t when $t > 0$. First we deal with the case $b = 0$. When $b = 0$, we put $y = 0$ if $a > 0$; we obtain the two solutions $\pm\sqrt{|a|}$. When $b = 0$, we put $x = 0$ if $a < 0$; we obtain the two solutions $\pm i\sqrt{|a|}$. In both of these cases we use $|a|$ for the square root of a^2.

Next suppose $b > 0$. In taking the square roots of (2) and (3), we choose x and y to have the same sign. Squaring now shows that these two answers satisfy $(x+iy)^2 = a+ib$. Finally suppose $b < 0$. In taking the square roots of (2) and (3), we choose x and y to have opposite signs. Squaring again shows that both answers satisfy $(x + iy)^2 = a + ib$. □

In the proof of Proposition 2.1, we obtained four candidates $\pm x \pm iy$ for z. When x and y are both not 0, these four candidates are distinct. As we noted in the proof, at most two of them can be square roots of w. Thus two of them fail. Hence the delicate analysis involving the signs is required. Things seem too complicated! On the other hand, the existence of square roots follows easily from the polar representation of complex numbers in Section 6. At that time we will develop geometric intuition clarifying the subtleties in the proof of Proposition 2.1.

Example 2.1. We solve $z^2 = 11 + 60i$ by the method of Proposition 2.1. We have $a^2 + b^2 = 121 + 3600$ and hence $\sqrt{a^2 + b^2} = 61$. Therefore $x^2 = \frac{11+61}{2} = 36$ and $y^2 = \frac{-11+61}{2} = 25$. We then take $x = 6$ and $y = 5$ or $x = -6$ and $y = -5$ to obtain the answers $z = \pm(6+5i)$. The other combinations of signs fail. If instead we want the square root of $11 - 60i$, then we have $x^2 = 36$ and $y^2 = 25$ as before, but we need to choose x and y to have opposite signs.

▶ **Exercise 2.11.** Find the error in the following alleged proof that $-1 = 1$.

$$-1 = i^2 = \sqrt{-1}\sqrt{-1} = \sqrt{(-1)(-1)} = \sqrt{1} = 1.$$

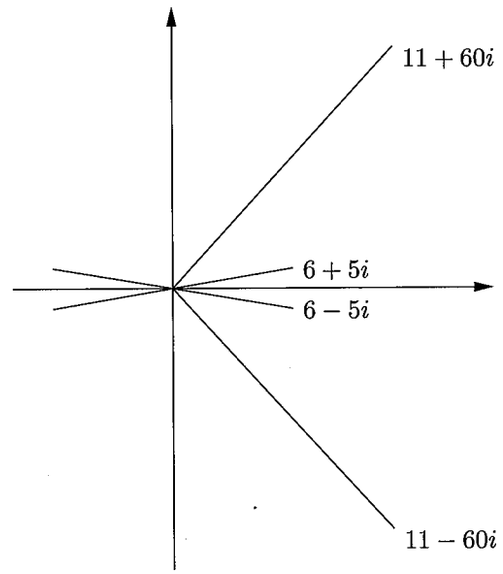

Figure 2.2. Finding square roots.

It is possible to solve cubic (third degree) and quartic (fourth degree) equations by using the quadratic formula cleverly. The history of this approach is quite interesting and relevant for the development of modern mathematics. See for example [**9**]. We limit ourselves here however to a few exercises about solving cubic polynomial equations. Cardano's solution of the cubic equation dates to 1545 and provided perhaps the first compelling argument in support of complex variables.

▶ **Exercise 2.12.** Let $z^3 + Az^2 + Bz + C$ be a cubic polynomial. What substitution reduces it to the form $w^3 + aw + b$?

▶ **Exercise 2.13.** Suppose we can solve the cubic $w^3 + aw + b = 0$, in the sense that we can find formulas for the roots in terms of a, b. By the previous exercise we can then solve the general cubic. To solve $w^3 + aw + b = 0$, we first make the substitution $w = \zeta + \frac{\alpha}{\zeta}$. If we choose α intelligently, then we get a sixth degree equation of the form

$$\zeta^6 + c_3\zeta^3 + c_0 = 0. \tag{4}$$

What is the intelligent choice for α? Why? Since (4) is a quadratic in ζ^3, one can solve it by the quadratic formula.

▶ **Exercise 2.14.** Solve $z^3 + 3z - 4 = 0$ by the method of the previous exercise. Also solve it by elementary means and compare what you get. Do the same for $z^3 + 6z - 20 = 0$.

Remark 2.1. The method of Exercises 2.12 and 2.13 gives a formula for the solution of the general cubic equation $z^3 + Az^2 + Bz + C = 0$ in terms of A, B, C. Unfortunately the solution will involve nested radicals. Trying to simplify these nested radicals often leads one back to the original equation. Hence the method is