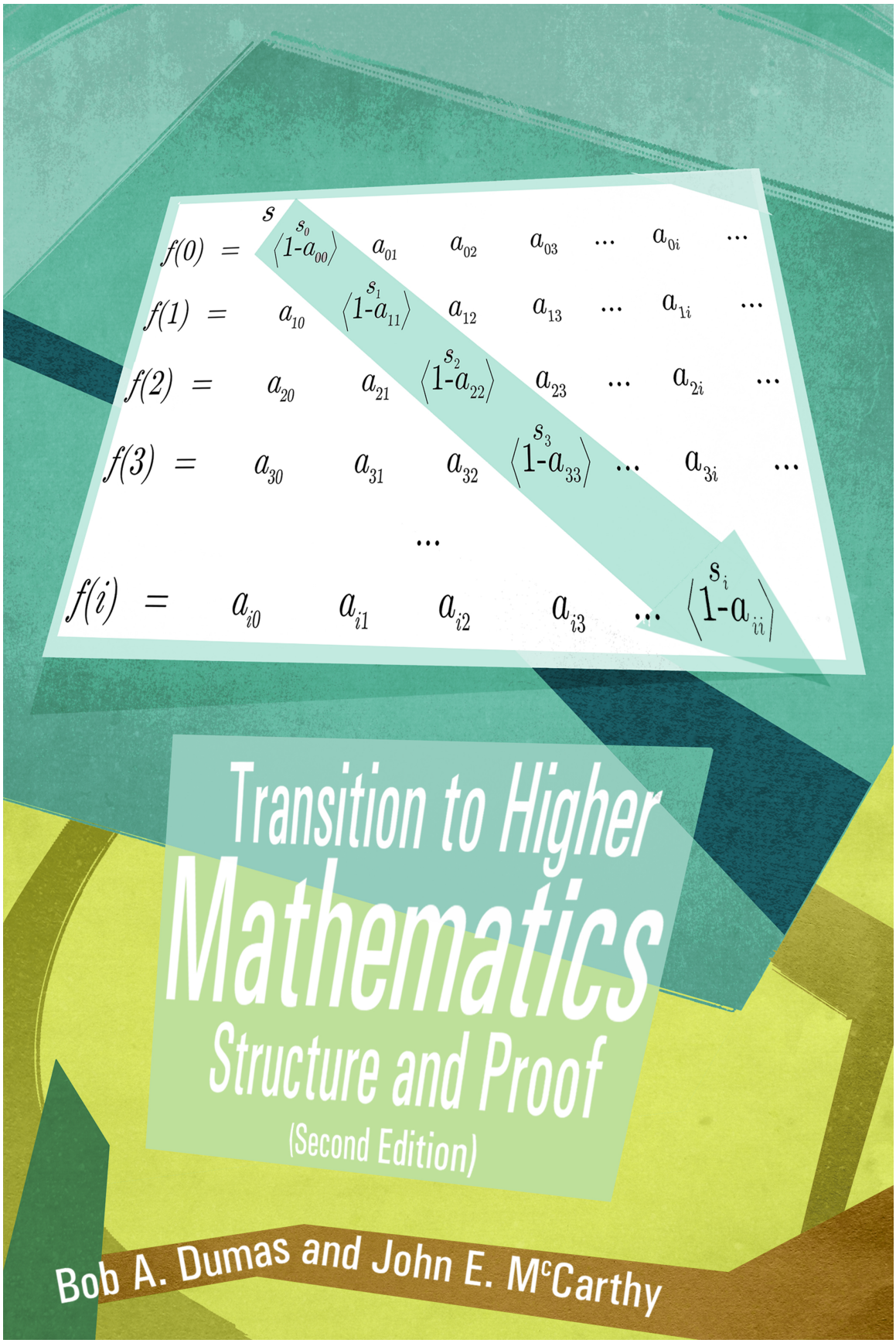

s
$f(0) = \langle 1\text{-}a_{00}^{s_0} \rangle \; a_{01} \; a_{02} \; a_{03} \; \ldots \; a_{0i} \; \ldots$
$f(1) = a_{10} \; \langle 1\text{-}a_{11}^{s_1} \rangle \; a_{12} \; a_{13} \; \ldots \; a_{1i} \; \ldots$
$f(2) = a_{20} \; a_{21} \; \langle 1\text{-}a_{22}^{s_2} \rangle \; a_{23} \; \ldots \; a_{2i} \; \ldots$
$f(3) = a_{30} \; a_{31} \; a_{32} \; \langle 1\text{-}a_{33}^{s_3} \rangle \; \ldots \; a_{3i} \; \ldots$
$\ldots$
$f(i) = a_{i0} \; a_{i1} \; a_{i2} \; a_{i3} \; \ldots \; \langle 1\text{-}a_{ii}^{s_i} \rangle$
Transition to Higher
Mathematics
Structure and Proof
(Second Edition)
Bob A. Dumas and John E. McCarthy

Transition to Higher Mathematics: Structure and Proof

Second Edition

Bob A. Dumas

John E. McCarthy

Cover design & illustration by Thomas Moore

ISBN: 978-1-941823-03-3

DOI: 10.7936/K7Z899HJ

To Gloria, Siena and William
B.D.

To Suzanne, Fiona and Myles
J.M^{c}C.

Contents

CHAPTER 0

Introduction

0.1. Why this book is

More students today than ever before take calculus in high school. This comes at a cost, however: fewer and fewer take a rigorous course in Euclidean geometry. Moreover, the calculus course taken by almost all students, whether in high school or college, avoids proofs, and often does not even give a formal definition of a limit. Indeed some students enter the university having never read or written a proof by induction, or encountered a mathematical proof of any kind.

As a consequence, teachers of upper level undergraduate mathematics courses in linear algebra, abstract algebra, analysis and topology have to work extremely hard inculcating the concept of proof while simultaneously trying to cover the syllabus. This problem has been addressed at many universities by introducing a bridge course, with a title like "Foundations for Higher Mathematics", taken by students who have completed the regular calculus sequence. Some of these students plan to become mathematics majors. Others just want to learn some more mathematics; but if what they are exposed to is interesting and satisfying, many will choose to major or double major in mathematics.

This book is written for students who have taken calculus and want to learn what "real mathematics" is. We hope you will find the material engaging and interesting, and that you will be encouraged to learn more advanced mathematics.

0.2. What this book is

The purpose of this book is to introduce you to the culture, language and thinking of mathematicians. We say "mathematicians", not

"mathematics", to emphasize that mathematics is, at heart, a human endeavor. If there is intelligent life in Erewhemos, then the Erewhemosians will surely agree that $2+2=4$. If they have thought carefully about the question, they will not believe that the square root of two can be exactly given by the ratio of two whole numbers, or that there are finitely many prime numbers. However we can only speculate about whether they would find these latter questions remotely interesting or what they might consider satisfying answers to questions of this kind.

Mathematicians have, after millennia of struggles and arguments, reached a widespread (if not quite universal) agreement as to what constitutes an acceptable mathematical argument. They call this a "proof", and it constitutes a carefully reasoned argument based on agreed premises. The methodology of mathematics has been spectacularly successful, and it has spawned many other fields. In the twentieth century, computer programming and applied statistics developed from offshoots of mathematics into disciplines of their own. In the nineteenth century, so did astronomy and physics. The increasing availability of data make the treatment of data in a sophisticated mathematical way one of the great scientific challenges of the twenty-first century.

In this book, we shall try to teach you what a proof is — what level of argument is considered convincing, what is considered overreaching, and what level of detail is considered too much. We shall try to teach you how mathematicians think — what structures they use to organize their thoughts. A structure is like a skeleton — if you strip away the inessential details you can focus on the real problem. A great example of this is the idea of number, the earliest human mathematical structure. If you learn how to count apples, and that two apples plus two apples make four apples, and if you think that this is about apples rather than counting, then you still don't know what two sheep plus two sheep make. But once you realize that there is an underlying structure of number, and that two plus two is four in the abstract, then adding wool or legs to the objects doesn't change the arithmetic.

0.3. What this book is not

There is an approach to teaching a transition course which many instructors favor. It is to have a problem-solving course, in which students learn to write proofs in a context where their intuition can help, such as in combinatorics or number theory. This helps to make the course interesting, and can keep students from getting totally lost.

We have not adopted this approach. Our reason is that in addition to teaching the skill of writing a logical proof, we also want to teach the skill of carefully analyzing definitions. Much of the instructor's labor in an upper-division algebra or analysis course consists of forcing the students to carefully read the definitions of new and unfamiliar objects, to decide which mathematical objects satisfy the definition and which do not, and to understand what follows "immediately" from the definitions. Indeed, the major reason that the epsilon-delta definition of limit has disappeared from most introductory calculus courses is the difficulty of explaining how the quantifiers $\forall\varepsilon\ \exists\delta$, in precisely this order, give the exact notion of limit for which we are striving. Thus, while students must work harder in this course to learn more abstract mathematics, they will be better prepared for advanced courses.

Nor is this a text in applied logic. The early chapters of the book introduce the student to the basic mathematical structures through formal definitions. Although we provide a rather formal treatment of first order logic and mathematical induction, our objective is to move to more advanced classical *mathematical* structures and arguments as soon as the student has an adequate understanding of the logic underlying mathematical proofs.

0.4. Advice to the Student

Welcome to higher mathematics! If your exposure to University mathematics is limited to calculus, this book will probably seem very different from your previous texts. Many students learn calculus by quickly scanning the text and proceeding directly to the problems.

When struggling with a problem, they seek similar problems in the text, and attempt to emulate the solution they find. Finally, they check the solution, usually found at the back of the text, to "validate" the methodology.

This book, like many texts addressing more advanced topics, is not written with computational problems in mind. Our objective is to introduce you to the various elements of higher undergraduate mathematics — the culture, language, methods, topics, standards and results. The problems in these courses are to prove true mathematical claims, or refute untrue claims. In the context of calculus, the mathematician must prove the results that you freely used. To most people, this activity seems very different from computation. For instance, you will probably find it necessary to think about a problem for some time before you begin writing. Unlike calculus, in which the general direction of the methods is usually obvious, trying to prove mathematical claims can feel directionless or accidental. However it is strategic rather than random. This is one of the great challenges of mathematics — at the higher levels, it is creative, not rote. With practice and disciplined thinking, you will learn to see your way to proving mathematical claims.

We shall begin our treatment of higher mathematics with a large number of definitions. This is usual in a mathematics course, and is necessary because mathematics requires precise expression. We shall try to motivate these definitions so that their usefulness will be obvious as early as possible. After presenting and discussing some definitions, we shall present arguments for some elementary claims concerning these definitions. This will give us some practice in reading, writing and discussing mathematics. In the early chapters of the book we include numerous discussions and remarks to help you grasp the basic direction of the arguments. In the later chapters of the book, you will read more difficult arguments for some deep classical results. We recommend that you read these arguments deliberately to ensure your thorough

understanding of the argument and to nurture your sense of the level of detail and rigor expected in an undergraduate mathematical proof.

There are exercises at the end of each chapter designed to direct your attention to the reading and compel you to think through the details of the proofs. Some of these exercises are straightforward, but many of them are very hard. We do not expect that every student will be able to solve every problem. However, spending an hour (or more) thinking about a difficult problem is time well-spent even if you do not solve the problem: it strengthens your mathematical muscles, and allows you to appreciate, and to understand more deeply, the solution if it is eventually shown to you. Ultimately, you will be able to solve some of the hard problems yourself after thinking deeply about them. Then you will be a real mathematician!

Mathematics is, from one point of view, a logical exercise. We define objects which do not physically exist, and use logic to draw the deepest conclusions we can concerning these objects. If this were the end of the story, mathematics would be no more than a game, and would be of little enduring interest. It happens, however, that interpreting physical objects, processes, behaviors, and other subjects of intellectual interest, as mathematical objects, and applying the conclusions and techniques from the study of these mathematical objects, allows us to draw reliable and powerful conclusions about practical problems. This method of using mathematics to understand the world is called mathematical modelling. The world in which you live, the way you understand this world, and how it differs from the world and understanding of your distant ancestors, is to a large extent the result of mathematical investigation. In this book, we try to explain how to draw mathematical conclusions with certainty. When you studied calculus, you used numerous deep theorems in order to draw conclusions that otherwise might have taken months rather than minutes. Now we shall develop an understanding of how results of this depth and power are derived.

0.5. Advice to the Instructor

Learning terminology — what do "contrapositive" and "converse" mean — comes easily to most students. Your challenge in the course is to teach them how to read definitions closely, and then how to manipulate them. This is much harder when there is no concrete image that students can keep in mind. Vectors in $\mathbb{R}^n$, for example, are more intimidating than in $\mathbb{R}^3$, not because of any great inherent increase in complexity, but because they are harder to think of geometrically, so students must trust the algebra alone. This trust takes time to build.

Chapter 1 is mainly to establish notation and discuss necessary concepts that some may have already seen (like injections and surjections). Unfortunately this may be the first exposure to some of these ideas for many students, so the treatment is rather lengthy. The speed at which the material is covered naturally will depend on the strength and background of the students. Take some time explaining why a sequence can be thought of as a function with domain $\mathbb{N}$ — variations on this idea will recur.

Chapter 2 introduces relations. These are hard to grasp, because of the abstract nature of the definition. Equivalences and linear orderings recur throughout the book, and students' comfort with these will increase.

Neither Chapter 1 nor Chapter 2 dwell on proofs. In fact mathematical proofs and elementary first order logic are not introduced until Chapter 3. Our objective is to get the student thinking about mathematical structures and definitions without the additional psychic weight of reading and writing proofs. We use examples to illustrate the definitions. The first Chapters provide basic conceptual foundations for later chapters, and we find that most students have their hands full just trying to read and understand the definitions and examples. In the exercises we ask the students to "show" the truth of some mathematical claims. Our intention is to get the student thinking about the task of proving mathematical claims. It is not expected that they will

write successful arguments before Chapter 3. We encourage the students to attempt the problems even though they will likely be uncertain about the requirements for a mathematical proof. If you feel strongly that mathematical proofs need to be discussed before launching into mathematical definitions, you can cover Chapter 3 first.

Chapter 3 is fairly formal, and should go quickly. Chapter 4 introduces students to the first major proof technique — induction. With practice, they can be expected to master this technique. We also introduce as an ongoing theme the study of polynomials, and prove for example that a polynomial has no more roots than its degree.

Chapters 5, 6 and 7 are completely independent of each other. Chapter 5 treats limits and continuity, up to proving that the uniform limit of a sequence of continuous functions is continuous. Chapter 6 is on infinite sets, proving Cantor's theorems and the Schröder-Bernstein theorem. By the end of the chapter, the students will have come to appreciate that it is generally much easier to construct two injections than one bijection!

Chapter 7 contains a little number theory — up to the proof of Fermat's little theorem. It then shows how much of the structure transfers to the algebra of real polynomials.

Chapter 8 constructs the real numbers, using Dedekind cuts, and proves that they have the least upper bound property. This is then used to prove the basic theorems of real analysis — the Intermediate Value theorem and the Extreme Value theorem. Sections 8.1 through 8.4 require only Chapters 1 - 4 and Section 6.1. Sections 8.5 - 8.8 require Sections 5.1 and 5.2. Section 8.9 requires Chapter 6.

In Chapter 9, we introduce the complex numbers. Sections 9.1 - 9.3 prove the Tartaglia-Cardano formula for finding the roots of a cubic, and point out how it is necessary to use complex numbers even

to find real roots of real cubics. These sections require only Chapters 1 - 4. In Section 9.4 we prove the Fundamental Theorem of Algebra. This requires Chapter 5 and the Bolzano-Weierstrass theorem from Section 8.6.

What is a reasonable course based on this book? Chapters 1 - 4 are essential for any course. In a one quarter course, one could also cover Chapter 6 and either Chapter 5 or 7. In a semester-long course, one could cover Chapters 1 - 6 and one of the remaining three chapters. Chapter 9 can be covered without Chapter 8 if one is willing to assert the Least Upper Bound property as an axiom of the real numbers, and then Section 8.6 can be covered before Section 9.4 without any other material from Chapter 8.

We suggest that you agree with your colleagues on a common curriculum for this course, so that topics that you cover thoroughly (*e.g.* cardinality) need not be repeated in successive courses.

This transition course is becoming one of the most important courses in the mathematics curriculum, and the first important course for the mathematics major. For the talented and intellectually discriminating first or second year student the standard early courses in the mathematics curriculum — calculus, differential equations, matrix algebra — provide little incentive for studying mathematics. Indeed, there is little mathematics in these courses, and less still with the evolution of lower undergraduate curricula towards the service of the sciences and engineering. This is particularly disturbing as it pertains to the talented student who has not yet decided on a major and may never have considered mathematics. We believe that the best students should be encouraged to take this course as early as possible — even concurrent with the second semester or third quarter of first year calculus. It is not just to help future math majors, but can also serve a valuable rôle in recruiting them, by letting smart students see that mathematics is challenging and, more to the point, interesting and deep. Mathematics

is its own best apologist. Expose the students early to authentic mathematical thinking and results and let them make an informed choice. It may come as a surprise to some, but good students still seek what mathematicians sought as students — the satisfaction of mastering a difficult, interesting and useful discipline.

0.6. Acknowledgements

We have received a lot of help in writing this book. In addition to the support of our families, we have received valuable advice and feedback from our students and colleagues, and from the reviewers of the manuscript. In particular we would like to thank Matthew Valeriote for many helpful discussions, and Alexander Mendez for drawing all the figures in the book.

CHAPTER 1

Preliminaries

To communicate mathematics you will need to understand and abide by the conventions of mathematicians. In this chapter we review some of these conventions.

1.1. "And" "Or"

Statements are declarative sentences; that is, a statement is a sentence which is true or false. Mathematicians make mathematical statements — sentences about mathematics which are true or false. For instance, the statement:

"All prime numbers, except the number 2, are odd."

is a true statement. The statement:

"$3 < 2$."

is false.

We use natural language connectives to combine mathematical statements. The connectives "and" and "or" have a particular usage in mathematical prose. Let P and Q be mathematical statements. The statement

P and Q.

is the statement that both P and Q are true.

Mathematicians use what is called the "inclusive or". In everyday usage the statement "P or Q" can sometimes mean that exactly one (but not both) of the statements P and Q is true. In mathematics, the statement

P or Q

is true when either or both statements are true, *i.e.* when any of the following hold:

P is true and Q is false.
P is false and Q is true.
P is true and Q is true.

1.2. Sets

Intuitively, a mathematical set is a collection of mathematical objects. Unfortunately this simple characterization of sets, carelessly handled, gives rise to contradictions. Some collections will turn out not to have the properties that we demand of mathematical sets. An example of how this can occur is presented in Section 1.7. We shall not develop formal set theory from scratch here. Instead, we shall assume that certain building block sets are known, and describe ways to build new sets out of these building blocks.

Our initial building blocks will be the sets of natural numbers, integers, rational numbers and real numbers. In Chapter 8, we shall show how to build all these from the natural numbers. One can't go much further than this, though: in order to do mathematics, one has to start with axioms that assert that the set of natural numbers exist.

DEFINITION. Element, $\in$ If X is a set and x is an object in X, we say that x is an element, or member, of X. This is written

$$x \in X.$$

We write $x \notin X$ if x is not a member of X.

There are numerous ways to define sets. If a set has few elements, it may be defined by listing. For instance,

$$X \;=\; \{2, 3, 5, 7\}$$

is the set of the first four prime numbers. In the absence of any other indication, a set defined by a list is assumed to have as elements only the objects in the list. For sets with too many elements to list, we

must provide the reader with a means to determine membership in the set. The author can inform the reader that not all elements of the set have been listed, but that enough information has been provided for the reader to identify a pattern for determining membership in the set. For example, let

$$X = \{2, 4, 6, 8, \ldots, 96, 98\}.$$

Then X is the set of positive even integers less than 100. However, using an ellipsis to define a set may not always work: it assumes that the reader will identify the pattern you wish to characterize. Although this usually works, it carries the risk that the reader is unable to correctly identify the pattern intended by the author.

Some sets are so important that they have standard names and notations that you will need to know.

NOTATION. *Natural numbers,* $\mathbb{N}$ *The natural numbers are the elements of the set*

$$\{0, 1, 2, 3, \ldots\}.$$

This set is denoted by $\mathbb{N}$.

Beware: Many authors call $\{1, 2, 3, \ldots\}$ the set of natural numbers. This is a matter of definition, and there is no universal convention; logicians tend to favor our convention, and algebraists the other. In this book, we shall use $\mathbb{N}^+$ to denote $\{1, 2, 3, \ldots\}$.

NOTATION. $\mathbb{N}^+$ $\mathbb{N}^+$ *is the set of positive integers,*

$$\{1, 2, 3, \ldots\}.$$

NOTATION. *Integers,* $\mathbb{Z}$ $\mathbb{Z}$ *is the set of integers,*

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

NOTATION. *Rational numbers,* $\mathbb{Q}$ $\mathbb{Q}$ *is the set of rational numbers,*

$$\left\{ \frac{p}{q} \text{ where } p, q \in \mathbb{Z} \text{ and } q \neq 0 \right\}.$$

NOTATION. *Real numbers,* $\mathbb{R}$ $\mathbb{R}$ *is the set of real numbers.*

A good understanding of the real numbers requires a bit of mathematical develoment. In fact, it was only in the nineteenth century that we really came to a modern understanding of $\mathbb{R}$. We shall have a good deal to say about the real numbers in Chapter 8.

DEFINITION. A number x is positive if $x > 0$. A number x is nonnegative if $x \geq 0$.

NOTATION. X^+ *If X is a set of real numbers, we use X^+ for the positive numbers in the set X.*

The notation we have presented for these sets is widely used. We introduce a final convention for set names which is not as widely recognized, but is useful for set theory.

NOTATION. $\ulcorner n \urcorner$ *is the set of all natural numbers less than n:*

$$\ulcorner n \urcorner \;=\; \{0, 1, 2, \ldots, n-1\}.$$

One purpose of this notation is to canonically associate any natural number n with a set having exactly n elements.

The reader should note that we have not *defined* the above sets. We are assuming that you are familiar with them, and some of their properties, by virtue of your previous experience in mathematics. We shall eventually define the sets systematically in Chapter 8.

A more precise method of defining a set is to use unambiguous conditions that characterize membership in the set.

NOTATION. $\{x \in X \mid P(x)\}$ *Let X be a (previously defined) set, and let $P(x)$ be a condition or property. Then the set*

$$Y \;=\; \{x \in X \mid P(x)\} \tag{1.1}$$

is the set of elements in X which satisfy condition P. The set X is called the domain of the variable.

In words, (1.1) is read: "Y equals the set of all (little) x in (capital) X such that P is true of x ". The symbol " | " in (1.1) is often written

instead with a colon, *viz.* $\{x \in X : P(x)\}$. In mathematics, $P(x)$ is a often a mathematical formula. For instance, suppose $P(x)$ is the formula "$x^2 = 4$". By $P(2)$ we mean the formula with 2 substituted for x, that is

$$\text{“}2^2 = 4\text{”}.$$

If the substitution results in a true statement, we say that $P(x)$ holds at 2, or $P(2)$ is true. If the statement that results from the substitution is false, for instance $P(1)$, we say that $P(x)$ does not hold at 1, or that $P(1)$ is false.

EXAMPLE 1.2. Consider the set

$$X = \{0, 1, 4, 9, \ldots\}.$$

A precise definition of the same set is the following:

$$X = \{x \in \mathbb{N} \mid \text{for some } y \in \mathbb{N},\ x = y^2\}.$$

EXAMPLE 1.3. Let Y be the set of positive even integers less than 100. Then Y can be written:

$$\{x \in \mathbb{N} \mid x < 100 \text{ and there is } n \in \mathbb{N}^+ \text{ such that } x = 2 \cdot n\}.$$

EXAMPLE 1.4. An interval I is a non-empty subset of $\mathbb{R}$ with the property that whenever $a, b \in I$ and $a < c < b$, then c is in I. A bounded interval must have one of the four forms

$$\begin{aligned}
(a, b) &= \{x \in \mathbb{R} \mid a < x < b\} \\
[a, b) &= \{x \in \mathbb{R} \mid a \leq x < b\} \\
(a, b] &= \{x \in \mathbb{R} \mid a < x \leq b\} \\
[a, b] &= \{x \in \mathbb{R} \mid a \leq x \leq b\},
\end{aligned}$$

where in the first three cases a and b are real numbers with $a < b$ and in the fourth case we just require $a \leq b$. Unbounded intervals have

five forms:

$$\begin{aligned}(-\infty, b) &= \{x \in \mathbb{R} \mid x < b\} \\ (-\infty, b] &= \{x \in \mathbb{R} \mid x \leq b\} \\ (b, \infty) &= \{x \in \mathbb{R} \mid x > b\} \\ [b, \infty) &= \{x \in \mathbb{R} \mid x \geq b\} \\ \mathbb{R}\end{aligned}$$

where b is some real number. An interval is called *closed* if it contains all its endpoints (both a and b in the first group of examples, just b in the first four examples of the second group), and *open* if it contains none of them. Notice that this makes $\mathbb{R}$ the only interval that is both closed and open.

For the sake of brevity, an author may not explicitly identify the domain of the variable. Be careful of this, as the author is relying on the reader to make the necessary assumptions. For instance, consider the set

$$X = \{x \mid (x^2 - 2)(x - 1)(x^2 + 1) = 0\}.$$

If the domain of the variable is assumed to be $\mathbb{N}$, then

$$X = \{1\}.$$

If the domain of the variable is assumed to be $\mathbb{R}$, then

$$X = \{1, \sqrt{2}, -\sqrt{2}\}.$$

If the domain of the variable is assumed to be the complex numbers, then,

$$X = \{1, \sqrt{2}, -\sqrt{2}, i, -i\},$$

where i is the complex number $\sqrt{-1}$. Remember, the burden of clear communication is on the author, not the reader.

Another alternative is to include the domain of the variable in the condition defining membership in the set. So, if X is the intended domain of the set and $P(x)$ is the condition for membership in the set,

$$\{x \in X \mid P(x)\} = \{x \mid x \in X \text{ and } P(x)\}.$$

As long as the definition is clear, the author has some flexibility with regard to notation.

1.2.1. Set Identity. When are two sets equal? You might be inclined to say that two sets are equal provided they are the *same* collection of objects. Of course this is true, but equality as a relation between objects is not very interesting. However, you have probably spent a lot of time investigating equations (which are just statements of equality), and we doubt that equality seemed trivial. This is because in general equality should be understood as a relationship between *descriptions* or *names* of objects, rather than between the objects themselves. The statement

$$a \;=\; b$$

is a claim that the object represented by a is the same object as that represented by b. For example, the statement

$$5-3 \;=\; 2$$

is the claim that the number represented by the arithmetic expression $5-3$ is the same number as that represented by the numeral 2.

In the case of sets, this notion of equality is called extensionality.

DEFINITION. Extensionality Let X and Y be sets. Then $X = Y$ provided that every element of X is also an element of Y and every element of Y is also an element of X.

There is flexibility in how a set is characterized as long as we are clear on which objects constitute the set. For instance, consider the set equation

$$\{\text{Mark Twain},\ \text{Samuel Clemens}\} \;=\; \{\text{Mark Twain}\}.$$

If by "Mark Twain" and "Samuel Clemens", we mean the deceased American author, these sets are equal, by extensionality, and the statement is true. The set on the left hand side of the equation has only one element since both names refer to the same person. If, however,

we consider "Mark Twain" and "Samuel Clemens" as names, the statement is false, since "Samuel Clemens" is a member of the set on the left hand side of the equation, but not the right hand side. You can see that set definitions can depend on the implicit domain of the variable even if the sets are defined by listing.

EXAMPLE 1.5. Consider the following six sets:

$$\begin{aligned} X_1 &= \{1, 2\} \\ X_2 &= \{2, 1\} \\ X_3 &= \{1, 2, 1\} \\ X_4 &= \{n \in \mathbb{N} \mid 0 < n < 3\} \\ X_5 &= \{n \in \mathbb{N} \mid \text{there exist } x, y, z \in \mathbb{N}^+ \text{ such that } x^n + y^n = z^n\} \\ X_6 &= \{0, 1, 2\}. \end{aligned}$$

The first five sets are all equal, and the sixth is different. However, while it is obvious that $X_1 = X_2 = X_3 = X_4$, the fact that $X_5 = X_1$ is the celebrated theorem of Andrew Wiles (his proof of Fermat's last theorem).

1.2.2. Relating Sets. In order to say anything interesting about sets, we need ways to relate them, and we shall want ways to create new sets from existing sets.

DEFINITION. Subset, $\subseteq$ Let X and Y be sets. X is a subset of Y if every element of X is also an element of Y. This is written

$$X \subseteq Y.$$

Superset, $\supseteq$ If $X \subseteq Y$, then Y is called a superset of X, written

$$Y \supseteq X.$$

In order to show two sets are equal (or that two descriptions of sets refer to the same set), you must show that they have precisely the same elements. It is often easier if the argument is broken into two simpler

arguments in which you show mutual containment of the sets. In other words, saying $X = Y$ is the same as saying

$$X \subseteq Y \text{ and } Y \subseteq X, \tag{1.6}$$

and verifying the two separate claims in (1.6) is often easier (or at least clearer) than showing that $X = Y$ all at once.

Let's add a few more elementary notions to our discussion of sets.

DEFINITION. Proper subset, $\subsetneq$, $\supsetneq$ Let X and Y be sets. X is a proper subset of Y if

$$X \subseteq Y \text{ and } X \neq Y.$$

We write this as

$$X \subsetneq Y$$

or

$$Y \supsetneq X.$$

DEFINITION. Empty set, $\emptyset$ The empty set is the set with no elements. It is denoted by $\emptyset$.

So for any set, X,

$$\emptyset \subseteq X.$$

(Think about why this is true). Just because $\emptyset$ is empty does not mean it is unimportant. Indeed, many mathematical questions reduce to asking whether a particular set is empty or not. Furthermore, as you will see in Chapter 8, we can build the entire real line from the empty set using set operations.

EXERCISE. (See Exercises 1.1). Show that

$$\{n \in \mathbb{N} \mid n \text{ is odd and } n = k(k+1) \text{ for some } k \in \mathbb{N}\}$$

is empty.

Let's discuss some ways to define new sets from existing sets.

DEFINITION. Union, $\cup$ Let X and Y be sets. The union of X and Y, written $X \cup Y$, is the set

$$X \cup Y = \{x \mid x \in X \text{ or } x \in Y\}.$$

(Recall our discussion in Section 1.1 about the mathematical meaning of the word "or".)

DEFINITION. Intersection, $\cap$ Let X and Y be sets. The intersection of X and Y, written $X \cap Y$, is the set

$$X \cap Y = \{x \mid x \in X \text{ and } x \in Y\}.$$

DEFINITION. Set difference, $\setminus$ Let X and Y be sets. The set difference of X and Y, written $X \setminus Y$, is the set

$$X \setminus Y = \{x \in X \mid x \notin Y\}.$$

DEFINITION. Disjoint Let X and Y be sets. X and Y are disjoint if

$$X \cap Y = \emptyset.$$

Oftentimes one deals with sets that are subsets of some fixed given set U. For example, when dealing with sets of natural numbers, the set U would be $\mathbb{N}$.

DEFINITION. Complement Let $X \subseteq U$. The complement of X in U is the set $U \setminus X$. When U is understood from the context, the complement of X is written X^c.

What about set operations involving more than two sets? Unlike arithmetic, in which there is a default order of operations (powers, products, sums), there is not a universal convention for the order in which set operations are performed. If intersections and unions appear in the same expression, then the order in which the operations are performed can matter. For instance, suppose X and Y are disjoint, nonempty sets, and consider the expression

$$X \cap X \cup Y.$$

If we mean for the intersection to be executed before the union, then

$$(X \cap X) \cup Y = X \cup Y.$$

If, however we intend the union to be computed before the intersection, then

$$X \cap (X \cup Y) = X.$$

Since Y is nonempty and disjoint from X,

$$(X \cap X) \cup Y \neq X \cap (X \cup Y).$$

Consequently, the order in which set operations are executed needs to be explicitly prescribed with parentheses.

EXAMPLE 1.7. Let $X = \mathbb{N}$ and $Y = \mathbb{Z} \setminus \mathbb{N}$. Then

$$(X \cap X) \cup Y = \mathbb{N} \cup Y = \mathbb{Z}.$$

However

$$X \cap (X \cup Y) = \mathbb{N} \cap \mathbb{Z} = \mathbb{N}.$$

DEFINITION. Cartesian product, Direct product, $X \times Y$ Let X and Y be sets. The Cartesian product of X and Y, written $X \times Y$, is the set of ordered pairs

$$\{(x, y) \mid x \in X \text{ and } y \in Y\}.$$

The Cartesian product is also called the direct product.

EXAMPLE 1.8. Let

$$X = \{1, 2, 3\}$$

and

$$Y = \{1, 2\}.$$

Then

$$X \times Y = \{(1,1), (1,2), (2,1), (2,2), (3,1), (3,2)\}.$$

Note that the order matters — that is

$$(1, 2) \neq (2, 1).$$

So $X \times Y$ is a set with six elements.

Since direct products are themselves sets, we can easily define the direct product of more than two factors. For example, let X, Y and Z be sets, then

$$(X \times Y) \times Z \;=\; \{((x, y), z) \mid x \in X,\, y \in Y,\, z \in Z\}. \tag{1.7}$$

Formally,

$$(X \times Y) \times Z \neq X \times (Y \times Z), \tag{1.8}$$

because $((x, y), z)$ and $(x, (y, z))$ are not the same. However in nearly every application, this distinction is not important, and mathematicians generally consider the direct product of more than two sets without regard to this detail. Therefore you will generally see the Cartesian product of three sets written without parentheses,

$$X \times Y \times Z.$$

In this event you may interpret the direct product as either side of statement 1.8.

With some thought, you can conclude that we have essentially described the Cartesian product of an arbitrary finite collection of sets. The elements of the Cartesian product $X \times Y$ are ordered pairs. Our characterization of the Cartesian product of three sets, X, Y and Z, indicates that its elements could be thought of as ordered pair of elements of $X \times Y$ and Z, respectively. From a practical point of view, it is simpler to think of elements of $X \times Y \times Z$ as ordered triples. We generalize this as follows.

DEFINITION. Cartesian product, Direct product, $\prod_{i=1}^{n} X_i$ Let $n \in \mathbb{N}^+$, and X_1, X_2, ...,X_n be sets. The Cartesian product of $X_1, \ldots, X_n$, written $X_1 \times X_2 \times \ldots \times X_n$, is the set

$$\{(x_1, x_2, \ldots, x_n) \mid x_i \in X_i,\; 1 \leq i \leq n\}.$$

This may also be written

$$\prod_{i=1}^{n} X_i.$$

When we take the Cartesian product of a set X with itself n times, we write it as X^n:

$$X^n := \overbrace{X \times X \times \cdots \times X}^{n \text{ times}}.$$

1.3. Functions

Like sets, functions are ubiquitous in mathematics.

DEFINITION. Function, $f : X \to Y$ Let X and Y be sets. A function f from X to Y, denoted by $f : X \to Y$, is an assignment of exactly one element of Y to each element of X.

For each element $x \in X$, the function f associates or selects a unique element $y \in Y$. The uniqueness condition does not allow x to be assigned to distinct elements of Y. It does allow different elements of X to be assigned to the same element of Y however. It is important to your understanding of functions that you consider this point carefully. The following examples may help illustrate this.

EXAMPLE 1.9. Let $f : \mathbb{Z} \to \mathbb{R}$ be given by

$$f(x) = x^2.$$

Then f is a function in which the element of $\mathbb{R}$ assigned to the element x of $\mathbb{Z}$ is specified by the expression x^2. For instance f assigns 9 to the integer 3. We express this by writing

$$f(3) = 9.$$

Observe that not every real number is assigned to a number from $\mathbb{Z}$. Furthermore, observe that 4 is assigned to both 2 and -2. Check that f does satisfy the definition of a function.

EXAMPLE 1.10. Let $g : \mathbb{R} \to \mathbb{R}$ be defined by $g(x) = \tan(x)$. Then g is not a function, because it is not defined when $x = \pi/2$ (or whenever $x - \pi/2$ is an integer multiple of π). This can be fixed by defining

$$X = \mathbb{R} \setminus \{\pi/2 + k\pi \mid k \in \mathbb{Z}\}.$$

Then $\tan : X \to \mathbb{R}$ is a function from X to $\mathbb{R}$.

EXAMPLE 1.11. Consider two rules, $f, g : \mathbb{R} \to \mathbb{R}$, defined by

$$f(x) = y \qquad \text{if } 3x = 2 - y$$
$$g(x) = y \qquad \text{if } x = y^4.$$

Then f is a function, and can be given explicitly as $f(x) = 2 - 3x$. But g does not define a function, because *e.g.* when $x = 16$, then $g(x)$ could be either 2 or -2.

DEFINITION. Image Let $f : X \to Y$. If $a \in X$, then the element of Y that f assigns to a is denoted by $f(a)$, and is called the image of a under f.

The notation $f : X \to Y$ is a statement that f is a function from X to Y. This statement has as a consequence that for every $a \in X$, $f(a)$ is a specific element of Y. We give an alternative characterization of functions based on Cartesian products.

DEFINITION. Graph of a function Let $f : X \to Y$. The graph of f, graph(f), is

$$\{(x, y) \mid x \in X \text{ and } f(x) = y\}.$$

EXAMPLE 1.12. Let $X \subseteq \mathbb{R}$ and $f : X \to \mathbb{R}$ be defined by $f(x) = -x$. Then the graph of f is

$$\{(x, -x) \mid x \in X\}.$$

EXAMPLE 1.13. The empty function f is the function with empty graph (that is the graph of f is the empty set). This means $f : \emptyset \to Y$ for some set Y.

If $f : X \to Y$, then,

$$\text{graph}(f) \subseteq X \times Y.$$

Let $Z \subseteq X \times Y$. Then Z is the graph of a function from X to Y if
(i) for any $x \in X$, there is some y in Y such that $(x, y) \in Z$
(ii) if (x, y) is in Z and (x, z) is in Z, then $y = z$.

Suppose X and Y are subsets of $\mathbb{R}$. Then Condition *(i)* is the condition that every vertical line through a point of X cuts the graph at least once. Condition *(ii)* is the condition that every vertical line through a point of X cuts the graph at most once.

DEFINITION. Domain, Codomain Let $f : X \to Y$. The set X is called the domain of f, and is written $\mathrm{Dom}(f)$. The set Y is called the codomain of f.

The domain of a function is a necessary component of the definition of a function. The codomain is a bit more subtle. If you think of functions as sets of ordered pairs, *i.e.* if you identified the function with its graph, then every function would have many possible codomains (take any superset of the original codomain). Set theorists think of functions this way, and if functions are considered as sets, extensionality requires that functions with the same graph are identical. However, this convention would make a discussion of surjections clumsy (see below), so we shall not adopt it.

When you write

$$f : X \to Y$$

you are explicitly naming the intended codomain, and this makes the codomain a crucial part of the definition of the function. You are indicating to the reader that your definition includes more than just the graph of the function. The definition of a function includes three pieces: the domain, the codomain, and the graph.

EXAMPLE 1.14. Let $f : \mathbb{N} \to \mathbb{N}$ be defined by

$$f(n) = n^2.$$

Let $g : \mathbb{N} \to \mathbb{R}$ be defined by

$$g(x) = x^2.$$

Then $\mathrm{graph}(f) = \mathrm{graph}(g)$. If $h : \mathbb{R} \to \mathbb{R}$ is defined by

$$h(x) = x^2$$

then $\text{graph}(f) \subsetneq \text{graph}(h)$, so $f \neq h$ and $g \neq h$. Although $\text{graph}(f) = \text{graph}(g)$, we consider f and g to be different functions because they have different codomains.

DEFINITION. Range Let $f : X \to Y$. The range of f, $\text{Ran}(f)$, is

$$\{y \in Y \mid \text{ for some } x \in X,\ f(x) = y\}.$$

So if $f : X \to Y$, then $\text{Ran}(f) \subseteq Y$, and is precisely the set of images under f of elements in X. That is

$$\text{Ran}(f) = \{f(x) \mid x \in X\}.$$

No proper subset of $\text{Ran}(f)$ can serve as a codomain for a function that has the same graph as f.

EXAMPLE 1.15. With the same notation as in Example 1.14, we have $\text{Ran}(f) = \text{Ran}(g) = \{n \in \mathbb{N} \mid n = k^2 \text{ for some } k \in \mathbb{N}\}$. The range of h is $[0, \infty)$.

DEFINITION. Real-valued function, Real function Let $f : X \to Y$. If $\text{Ran}(f) \subseteq \mathbb{R}$, we say that f is real-valued. If $X \subseteq \mathbb{R}$ and f is a real-valued function, then we call f a real function.

It is sometimes said that a function is a *rule* that assigns, to each element of a given set, some element from another set. If, by a rule, one means an instruction of some sort, you will see in Chapter 6 that there are "more" functions that cannot be characterized by rules than there are functions that can be. In practice, however, most of the functions we use are defined by rules.

If a function is given by a rule, it is common to write it in the form

$$\begin{aligned} f : X &\to Y \\ x &\mapsto f(x). \end{aligned}$$

The symbol $\mapsto$ is read "is mapped to". For example, the function g in the previous example could be defined by

$$\begin{aligned} g : \mathbb{N} &\to \mathbb{R} \\ n &\mapsto n^2. \end{aligned}$$

EXAMPLE 1.16. The function

$$\begin{array}{rcl} f : \mathbb{R} & \to & \mathbb{R} \\ x & \mapsto & \begin{cases} 0 & x < 0 \\ x+1 & x \geq 0 \end{cases} \end{array}$$

is defined by a rule, even though to apply the rule to a given x you must first check where in the domain x lies.

When a real function is defined by a rule and the domain is not explicitly stated, it is taken to be the largest set for which the rule is defined. This is the usual convention in calculus: real functions are defined by mathematical expressions and it is understood that the implicit domain of a function is the largest subset of $\mathbb{R}$ for which the expression makes sense. The codomain of a real function is assumed to be $\mathbb{R}$ unless explicitly stated otherwise.

EXAMPLE 1.17. Let $f(x) = \sqrt{x}$ be a real function. The domain of the function is assumed to be

$$\{x \in \mathbb{R} \mid x \geq 0\}.$$

DEFINITION. Operation Let X be a set, and $n \in \mathbb{N}^+$. An operation on X is a function from X^n to X.

Operations may be thought of as means of combining elements of a set to produce new elements of the set. The most common operations are binary operations (when $n = 2$).

EXAMPLE 1.18. $+$ and $\cdot$ are binary operations on $\mathbb{N}$.
$-$ and $\div$ are not operations on $\mathbb{N}$.

EXAMPLE 1.19. Let $X = \mathbb{R}^3$, thought of as the set of 3-vectors. The function $x \mapsto -x$ is a unary operation on X, the function $(x, y) \mapsto x+y$ is a binary operation, and the function $(x, y, z) \mapsto x \times y \times z$ is a ternary operation.

If $f : X \to Y$, $g : X \to Y$, and $\star$ is a binary operation on Y, then there is a natural way to define a new function on X using $\star$. Define $f \star g$ by

$$\begin{aligned} f \star g : X &\to Y \\ (f \star g)(x) &= f(x) \star g(x). \end{aligned}$$

EXAMPLE 1.20. Suppose f is the real function $f(x) = x^3$, and g is the real function $g(x) = 3x^2 - 1$. Then $f + g$ is the real function $x \mapsto x^3 + 3x^2 - 1$, and $f \cdot g$ is the real function $x \mapsto x^3(3x^2 - 1)$.

Another way to build new functions is by composition.

DEFINITION. Composition, $\circ$ Let $f : X \to Y$ and $g : Y \to Z$. Then the composition of g with f is the function,

$$\begin{aligned} g \circ f : X &\to Z \\ x &\mapsto g(f(x)). \end{aligned}$$

EXAMPLE 1.21. Let f be the real function

$$f(x) = x^2.$$

Let g be the real function

$$g(x) = \sqrt{x}.$$

Then

$$(g \circ f)(x) = |x|.$$

What is $f \circ g$? (Be careful about the domain).

EXAMPLE 1.22. Let

$$\begin{aligned} f : \mathbb{R} &\to \mathbb{R} \\ x &\mapsto 2x + 1 \end{aligned}$$

and let

$$\begin{aligned} g : \mathbb{R}^2 &\to \mathbb{R} \\ (x, y) &\mapsto x^2 + 3y^2. \end{aligned}$$

Then

$$\begin{aligned} f \circ g : \mathbb{R}^2 &\to \mathbb{R} \\ (x, y) &\mapsto 2x^2 + 6y^2 + 1. \end{aligned}$$

The function $g \circ f$ is not defined (why?).

1.4. Injections, Surjections, Bijections

Most basic among the characteristics a function may have are the properties of injectivity, surjectivity and bijectivity.

DEFINITION. Injection, One-to-one Let $f : X \to Y$. The function f is called an injection if, whenever x and y are distinct elements of X, we have $f(x) \neq f(y)$. Injections are also called one-to-one functions.

Another way of stating the definition (the contrapositive) is that if $f(x) = f(y)$ then $x = y$.

EXAMPLE 1.23. The real function $f(x) = x^3$ is an injection. To see this, let x and y be real numbers, and suppose that

$$f(x) = x^3 = y^3 = f(y).$$

Then

$$x = (x^3)^{1/3} = (y^3)^{1/3} = y.$$

So, for $x, y \in X$,

$$f(x) = f(y) \text{ only if } x = y.$$

EXAMPLE 1.24. The real function $f(x) = x^2$ is not an injection, since

$$f(2) = 4 = f(-2).$$

Observe that a single example suffices to show that f not an injection.

EXAMPLE 1.25. Suppose $f : X \to Y$ and $g : Y \to Z$. Prove that if f and g are injective, so is $g \circ f$.

PROOF. Suppose that $g \circ f(x) = g \circ f(y)$. Since g is injective, this means that $f(x) = f(y)$. Since f is injective, this in turn means that $x = y$. Therefore $g \circ f$ is injective, as desired. □

(See Exercise 1.20 below).

DEFINITION. Surjection, Onto Let $f : X \to Y$. We say f is a surjection from X to Y if $\text{Ran}(f) = Y$. We also describe this by saying that f is onto Y.

EXAMPLE 1.26. The function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$ is not a surjection. For instance, -1 is in the codomain of f, but $-1 \notin \text{Ran}(f)$. Therefore, $\text{Ran}(f) \subsetneq \mathbb{R}$.

EXAMPLE 1.27. Let $Y = \{x \in \mathbb{R} \mid x \geq 0\}$, and $f : \mathbb{R} \to Y$ be given by $f(x) = x^2$. Then f is a surjection. To prove this, we need to show that $Y = \text{Ran}(f)$. We know that $\text{Ran}(f) \subseteq Y$, so we must show $Y \subseteq \text{Ran}(f)$. Let $y \in Y$, so y is a non-negative real number. Then $\sqrt{y} \in \mathbb{R}$, and $f(\sqrt{y}) = y$. So $y \in \text{Ran}(f)$. Since y was an arbitrary element of Y, $Y \subseteq \text{Ran}(f)$. Hence $Y = \text{Ran}(f)$ and f is a surjection.

Whether a function is a surjection depends on the choice of the codomain. A function is always onto its range. You might wonder why one would not simply define the codomain as the range of the function (guaranteeing that the function is a surjection). One reason is that we may be more interested in relating two sets using functions than we are in any particular function between the sets. We study an important application of functions to relating sets in Chapter 6, where we use functions to compare the size of sets. This is of particular interest when comparing infinite sets, and has led to deep insights in the foundations of mathematics.

If we put the ideas of an injection and a surjection together, we arrive at the key idea of a bijection.

DEFINITION. Bijection, $\rightarrowtail$ Let $f : X \to Y$. If f is an injection and a surjection, then f is a bijection. This is written as $f : X \rightarrowtail Y$.

Why are bijections so important? From a theoretical point of view, functions may be used to relate the domain and the codomain of the function. If you are familiar with one set you may be able to develop

insights into a different set by finding a function between the sets which preserves some of the key characteristics of the sets. For instance, an injection can "interpret" one set into a different set. If the injection preserves the critical information from the domain, we can behave as if the domain of the function is virtually a subset of the codomain by using the function to "rename" the elements of the domain. If the function is a bijection, and it preserves key structural features of the domain, we can treat the domain and the codomain as virtually the same set. What the key structural features are depends on the area of mathematics you are studying. For example, if you are studying algebraic structures, you are probably most interested in preserving the operations of the structure. If you are studying geometry, you are interested in functions that preserve shape. The preservation of key structural features of the domain or codomain often allows us to translate knowledge of one set into equivalent knowledge of another set.

DEFINITION. Permutation Let X be a set. A permutation of X is a bijection $f : X \rightarrowtail X$.

EXAMPLE 1.28. Let $f : \mathbb{Z} \to \mathbb{Z}$ be defined by

$$f(x) \;=\; x + 1.$$

Then f is a permutation of $\mathbb{Z}$.

EXAMPLE 1.29. Let $X \;=\; \{0, 1, -1\}$. Then $f : X \to X$ given by $f(x) \;=\; -x$ is a permutation of X.

1.5. Images and Inverses

Functions can be used to define subsets of given sets.

DEFINITION. Image, $f[\ ]$ Let $f : X \to Y$ and $W \subseteq X$. The image of W under f, written $f[W]$, is the set

$$\{f(x) \mid x \in W\}.$$

So if $f : X \to Y$, then

$$\text{Ran}(f) \;=\; f[X].$$

EXAMPLE 1.30. Suppose f is the real function $f(x) = x^2 + 3$. Let $W = \{-2, 2, 3\}$, and $Z = (-1, 2)$. Then $f[W] = \{7, 12\}$, and $f[Z] = [3, 7)$.

In applications of mathematics, functions often describe numerical relationships between measurable observations. So if $f : X \to Y$ and $a \in X$, then $f(a)$ is the predicted or actual measurement associated with a. In this context, one is often interested in determining which elements of X are associated with a value, b, in the codomain of f.

DEFINITION 1.31. Inverse image, Pre-image, $f^{-1}(\,)$ Let $f : X \to Y$ and $b \in Y$. Then the inverse image of b under f, $f^{-1}(b)$, is the set

$$\{x \in X \mid f(x) = b\}.$$

This set is also called the pre-image of b under f.

Note that if $b \notin \text{Ran}(f)$, then $f^{-1}(b) = \emptyset$. If f is an injection, then for any $b \in \text{Ran}(f)$, $f^{-1}(b)$ has a single element.

DEFINITION. Inverse image, Pre-image, $f^{-1}[\,]$ Let $f : X \to Y$ and $Z \subseteq Y$. The inverse image of Z under f, or the pre-image of Z under f, is the set

$$f^{-1}[Z] \;=\; \{x \in X \mid f(x) \in Z\}.$$

We use $f^{-1}[\,]$ to mean the inverse image of a *subset* of the codomain, and $f^{-1}(\,)$ for the inverse image of an *element* of the codomain — both are subsets of the domain of f. If $Z \cap \text{Ran}(f) \;=\; \emptyset$, then

$$f^{-1}[Z] \;=\; \emptyset.$$

EXAMPLE 1.32. Let f be as in Figure 1.33 Then $f[\{b, c\}] = \{1, 3\}$, and $f^{-1}[\{1, 3\}] = \{a, b, c, d\}$.

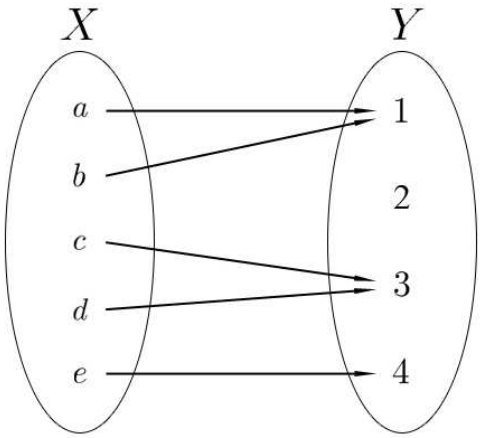


FIGURE 1.33. Picture of f

EXAMPLE 1.34. Let g be the real function $g(x) = x^2 + 3$. If $b \in \mathbb{R}$ and $b > 3$, then

$$g^{-1}(b) \;=\; \{\sqrt{b-3},\; -\sqrt{b-3}\}.$$

If $b = 3$, then $g^{-1}(3) = \{0\}$. If $b < 3$, then $g^{-1}(b)$ is empty.

EXAMPLE 1.35. Let h be the real function $h(x) = e^x$. If $b \in \mathbb{R}$ and $b > 0$, then

$$h^{-1}(b) \;=\; \{\log_e(b)\}.$$

For instance,

$$h^{-1}(1) \;=\; \{0\}.$$

Because h is strictly increasing, the inverse image of any element of the codomain ($\mathbb{R}$) is either a set with a single element or the empty set.

Let $I \;=\; (a, b)$, where $a, b \in \mathbb{R}$ and $0 < a < b$ (that is I is the open interval with end points a and b). Then

$$h^{-1}[I] \;=\; (\log_e(a), \log_e(b)).$$

We have discussed the construction of new functions from existing functions using algebraic operations and composition of functions.

Another tool for building new functions from known functions is the inverse function.

DEFINITION 1.36. Inverse function Let $f : X \rightarrowtail Y$ be a bijection. Then the inverse function of f, $f^{-1} : Y \to X$, is the function with graph

$$\{(b, a) \in Y \times X \mid (a, b) \in \text{graph}(f)\}.$$

The function f^{-1} is defined by "reversing the arrows". For this to make sense, $f : X \to Y$ must be bijective. Indeed, if f were not surjective, then there would be an element y of Y that is not in the range of f, so cannot be mapped back to anything in X. If f were not injective, there would be elements z of Y that were the image of distinct elements x_1 and x_2 in X. One could not define $f^{-1}(z)$ without specifying how to choose a particular pre-image. Both these problems can be fixed. If f is injective but not surjective, one can define $g : X \rightarrowtail \text{Ran}(f)$ by

$$g(x) = f(x)$$

for all $x \in X$. Then $g^{-1} : \text{Ran}(f) \rightarrowtail X$. If f is not injective, the problem is trickier; but if we can find some subset of X on which f is injective, we could restrict our attention to that set.

EXAMPLE 1.37. Let f be the real function $f(x) = x^2$. The function f is not an bijection, so it does not have an inverse function. However the function

$$\begin{aligned} g : [0, \infty) &\to [0, \infty) \\ x &\mapsto x^2 \end{aligned}$$

is a bijection. In this case,

$$g^{-1}(y) = \sqrt{y}.$$

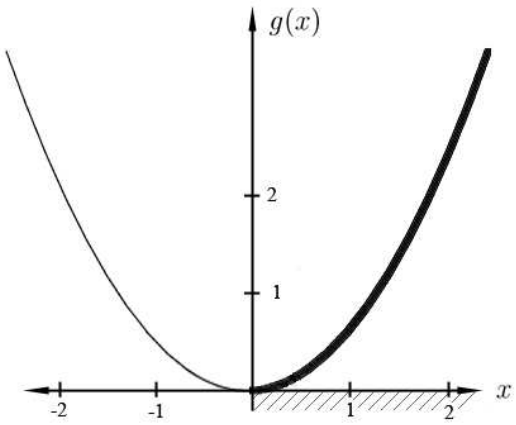


FIGURE 1.38. Picture of g

EXAMPLE 1.39. Let f be the real function, $f(x) = e^x$. You know from calculus that f is an injection, and that $\text{Ran}(f) = \mathbb{R}^+$. Hence f is not a surjection, since the implicit codomain of a real function is $\mathbb{R}$. The function

$$\begin{aligned} g : \mathbb{R} &\to \mathbb{R}^+ \\ x &\mapsto e^x \end{aligned}$$

is a bijection and

$$g^{-1}(x) = \log_e(x).$$

Warning: For $f : X \rightarrowtail Y$ a bijection we have assigned two different meanings to $f^{-1}(b)$. In Definition 1.31, it means the set of points in X that get mapped to b. In Definition 1.36, it means the inverse function, f^{-1}, of the bijection f applied to the point $b \in Y$. However, if f is a bijection, so that the second definition makes sense, then these definitions are closely related. Suppose $a \in \text{Dom}(f)$ and $f(a) = b$. According to Definition 1.31, $f^{-1}(b) = \{a\}$ and by Definition 1.36 $f^{-1}(b) = a$. In practice the context will make clear which definition is intended.

DEFINITION. Identity function, $\mathrm{id}|_X$ Let X be a set. The identity function on X, $\mathrm{id}|_X : X \rightarrowtail X$, is the function defined by

$$\mathrm{id}|_X(x) \;=\; x.$$

If $f : X \to Y$ is a bijection, then f^{-1} is the unique function such that

$$f^{-1} \circ f \;=\; \mathrm{id}|_X$$

and

$$f \circ f^{-1} \;=\; \mathrm{id}|_Y.$$

Because $f(x) \;=\; x^2$ is not an injection, it has no inverse, even after restricting the codomain to be the range. Therefore in order to "invert" f, we considered a different function $g(x)$, which was equal to f on a subset of the domain of f, and was an injection. In Example 1.37, we accomplished this by defining the function $g(x) \;=\; x^2$ with domain $\{x \in \mathbb{R} \mid x \geq 0\}$. Many of the functions that we need to invert for practical and theoretical reasons happen not to be injections, and hence do not have inverse functions. One way to address this obstacle is to consider the function on a smaller domain.

Given a function, $f : X \to Y$ we may wish to define an "inverse" of f on some subset of $W \subseteq X$ for which the *restriction* of f to W is an injection.

DEFINITION. Restricted domain, $f|_W$ Let $f : X \to Y$ and $W \subseteq X$. The restriction of f to W, written $f|_W$, is the function

$$\begin{aligned} f|_W : W &\to Y \\ x &\mapsto f(x). \end{aligned}$$

So if $f : X \to Y$ and $W \subseteq X$, then

$$\mathrm{graph}(f|_W) \;=\; [W \times Y] \cap [\mathrm{graph}(f)].$$

EXAMPLE 1.40. Let $f(x) = (x-2)^4$. Let $W = [2, \infty)$. Then

$$f|_W \;:\; W \;\to\; [0, \infty)$$

is a bijection.

EXAMPLE 1.41. Let f be the real function, $f(x) = \tan(x)$. Then

$$\mathrm{Dom}(f) = \{x \in \mathbb{R} \mid x \neq \pi/2 + k\pi,\ k \in \mathbb{Z}\},$$

and

$$\mathrm{Ran}(f) = \mathbb{R}.$$

The function f is periodic with period π, and is therefore not an injection. Nonetheless, it is important to answer the question,

"At what angle(s), x, does $\tan(x)$ equal a particular value, $a \in \mathbb{R}$?".

This is mathematically equivalent to asking,

"What is $\arctan(a)$?".

In calculus this need was met by restricting the domain to a largest interval, I such that

$$f|_I : I \rightarrowtail \mathbb{R}.$$

For any $k \in \mathbb{Z}$,

$$\left(\frac{(2k+1)\pi}{2}, \frac{(2k+3)\pi}{2}\right)$$

is such an interval. In order to define a specific function, the simplest of these intervals is selected, and we define

$$\mathrm{Tan} := \tan|_{(-\pi/2,\pi/2)}.$$

Observe that

$$\mathrm{Tan} : (-\pi/2, \pi/2) \rightarrowtail \mathbb{R}.$$

So the function is invertible, that is, Tan has an inverse function,

$$\mathrm{Arctan} = \mathrm{Tan}^{-1}.$$

1.6. Sequences

In calculus we think of a sequence as a (possibly infinite) list of objects. We shall expand on that idea somewhat, and express it in the language of functions.

DEFINITION. Finite sequence, $\langle a_n \mid n < N \rangle$ A finite sequence is a function f with domain $\ulcorner N \urcorner$, where $N \in \mathbb{N}$. We often identify the

sequence with the ordered finite set $\langle a_n \mid n < N\rangle$, where $a_n = f(n)$, for $0 \leq n < N$.

This interpretation of a sequence as a type of function is easily extended to infinite sequences.

DEFINITION. Infinite sequence, $\langle a_n \mid n \in \mathbb{N}\rangle$ An infinite sequence is a function f with domain $\mathbb{N}$. We often identify the sequence with the ordered infinite set $\langle a_n \mid n \in \mathbb{N}\rangle$, where $a_n = f(n)$, for $n \in \mathbb{N}$.

REMARK. Interval in $\mathbb{Z}$ Actually, the word sequence is normally used to mean any function whose domain is an interval in $\mathbb{Z}$, where an *interval in* $\mathbb{Z}$ is the intersection of some real interval with $\mathbb{Z}$. For convenience in this book, we usually assume that the first element of any sequence is indexed by 0 or 1.

EXAMPLE 1.42. The sequence $\langle 0, 1, 4, 9, \ldots\rangle$ is given by the function $f(n) = n^2$.

The sequence $\langle 1, -1, 2, -2, 3, -3, \ldots\rangle$ is given by the function

$$f(n) = \begin{cases} \frac{n}{2} + 1, & n \text{ even} \\ -\frac{n+1}{2}, & n \text{ odd.} \end{cases}$$

Sequences can take values in any set (the codomain of the function f that defines the sequence). We talk of a *real sequence* if the values are real numbers, an *integer sequence* if they are all integers, *etc.* It will turn out later that sequences with values in the two element set $\{0, 1\}$ occur quite frequently, so we have a special name for them: we call them binary sequences.

DEFINITION. Binary sequence A finite binary sequence is a function, $f : \ulcorner N \urcorner \to \ulcorner 2 \urcorner$, for some $N \in \mathbb{N}$. An infinite binary sequence is a function, $f : \mathbb{N} \to \ulcorner 2 \urcorner$.

We often use the expression $\langle a_n\rangle$ for the sequence $\langle a_n \mid n \in \mathbb{N}\rangle$.

Functions are also used to "index" sets in order to build more complicated sets with generalized set operations. We discussed the union

(or intersection) of more than two sets. You might ask whether it is possible to form unions or intersections of a large (infinite) collection of sets. There are two concerns that should be addressed in answering this question. We must be sure that the definition of the union of infinitely many sets is precise; that is, it uniquely characterizes an object in the mathematical universe. We also need notation for managing this idea — how do we specify the sets over which we are taking the union?

DEFINITION. Infinite union, Index set, $\bigcup_{n=1}^{\infty} X_n$ For $n \in \mathbb{N}^+$, let X_n be a set. Then

$$\bigcup_{n=1}^{\infty} X_n \;=\; \{x \mid \text{ for some } n \in \mathbb{N}^+,\; x \in X_n\}.$$

The set $\mathbb{N}^+$ is called the index set for the union.

This may be written in a few different ways.

NOTATION. $\bigcup_{n\in\mathbb{N}^+} X_n$ *The following three expressions are all equal:*

$$X_1 \cup X_2 \cup \ldots \cup X_n \cup \ldots$$
$$\bigcup_{n=1}^{\infty} X_n$$
$$\bigcup_{n\in\mathbb{N}^+} X_n.$$

We can use index sets other than $\mathbb{N}^+$.

DEFINITION. Family of sets, Indexed union, $\bigcup_{\alpha\in A} X_\alpha$ Let A be a set, and for $\alpha \in A$, let X_α be a set. The set

$$\mathcal{F} \;=\; \{X_\alpha \mid \alpha \in A\}$$

is called a family of sets indexed by A. Then

$$\bigcup_{\alpha\in A} X_\alpha \;=\; \{x \mid x \in X_\alpha \text{ for some } \alpha \in A\}.$$

The notation $\bigcup_{\alpha\in A} X_\alpha$ is read "the union over alpha in A of the sets X sub alpha".

So

$$x \in \bigcup_{\alpha \in A} X_\alpha \text{ if } x \in X_\alpha \text{ for some } \alpha \in A.$$

General intersections over a family of sets are defined analogously :

$$\bigcap_{\alpha \in A} X_\alpha \;=\; \{x \mid x \in X_\alpha \text{ for all } \alpha \in A\}.$$

EXAMPLE 1.43. Let $X_n = \{n+1, n+2, \ldots, 2n\}$ for each $n \in \mathbb{N}^+$. Then

$$\begin{aligned} \bigcup_{n=1}^{\infty} X_n &= \{k \in \mathbb{N} \mid k \geq 2\} \\ \bigcap_{n=1}^{\infty} X_n &= \emptyset. \end{aligned}$$

EXAMPLE 1.44. For each positive real number t, let $Y_t = [11/t, t]$. Then

$$\begin{aligned} \bigcup_{t \in (\sqrt{11}, \infty)} Y_t &= \mathbb{R}^+ \\ \bigcap_{t \in [\sqrt{11}, \infty)} Y_t &= \{\sqrt{11}\}. \end{aligned}$$

EXAMPLE 1.45. Let $f : X \to Y$, $A \subseteq X$ and $B \subseteq Y$. Then

$$\bigcup_{a \in A} \{f(a)\} \;=\; f[A].$$

and

$$\bigcup_{b \in B} f^{-1}(b) \;=\; f^{-1}[B].$$

1.7. Russell's Paradox

As the ideas for set theory were explored, there were attempts to define sets as broadly as possible. It was hoped that any collection of mathematical objects that could be defined by a formula would qualify as a set. This belief was known as the General Comprehension Principle (GCP) . Unfortunately, the GCP gave rise to conclusions which were unacceptable for mathematics.

Consider the collection defined by the following simple formula:

$$V \;=\; \{x \mid x \text{ is a set and } x \;=\; x\}.$$

If V is considered as a set, then since $V \;=\; V$,

$$V \in V.$$

If this is not an inconsistency, it is at least unsettling. Unfortunately, it gets worse. Consider the collection

$$X \;=\; \{x \mid x \notin x\}.$$

Then

$$X \in X \;\text{ if and only if }\; X \notin X.$$

This latter example is called Russell's paradox, and showed that the GCP is false. Clearly there would have to be some control over which definitions give rise to sets. Axiomatic set theory was developed to provide rules for rigorously defining sets. We give a brief discussion in Appendix B.

1.8. Exercises

EXERCISE 1.1. Show that

$$\{n \in \mathbb{N} \mid n \text{ is odd and } n = k(k+1) \text{ for some } k \in \mathbb{N}\}$$

is empty.

EXERCISE 1.2. Let X and Y be subsets of some set U. Prove de Morgan's laws:

$$\begin{aligned}(X \cup Y)^c &= X^c \cap Y^c \\ (X \cap Y)^c &= X^c \cup Y^c\end{aligned}$$

EXERCISE 1.3. Let X, Y and Z be sets. Prove

$$\begin{aligned}X \cap (Y \cup Z) &= (X \cap Y) \cup (X \cap Z) \\ X \cup (Y \cap Z) &= (X \cup Y) \cap (X \cup Z).\end{aligned}$$

EXERCISE 1.4. Let $X = \ulcorner 2 \urcorner, Y = \ulcorner 3 \urcorner$, and $Z = \ulcorner 1 \urcorner$. What are the following sets:
(i) $X \times Y$.
(ii) $X \times Y \times Z$.
(iii) $X \times Y \times Z \times \emptyset$.
(iv) $X \times X$.
(v) X^n.

EXERCISE 1.5. Suppose X is a set with m elements, and Y is a set with n elements. How many elements does $X \times Y$ have? Is the answer the same if one or both of the sets is empty?

EXERCISE 1.6. How many elements does $\emptyset \times \mathbb{N}$ have?

EXERCISE 1.7. Describe all possible intervals in $\mathbb{Z}$.

EXERCISE 1.8. Let X and Y be finite non-empty sets, with m and n elements, respectively. How many functions are there from X to Y? How many injections? How many surjections? How many bijections?

EXERCISE 1.9. What happens in Exercise 1.8 if m or n is zero?

EXERCISE 1.10. For each of the following sets, which of the operations addition, subtraction, multiplication, division and exponentiation are operations on the set:
(i) $\mathbb{N}$
(ii) $\mathbb{Z}$
(iii) $\mathbb{Q}$
(iv) $\mathbb{R}$
(v) $\mathbb{R}^+$.

EXERCISE 1.11. Let f and g be real functions, $f(x) = 3x + 8$, $g(x) = x^2 - 5x$. What are $f \circ g$ and $g \circ f$? Is $(f \circ g) \circ f = f \circ (g \circ f)$?

EXERCISE 1.12. Write down all permutations of $\{a, b, c\}$.

EXERCISE 1.13. What is the natural generalization of Exercise 1.2 to an arbitrary number of sets? Verify your generalized laws.

EXERCISE 1.14. What is the natural generalization of Exercise 1.3 to an arbitrary number of sets? Verify your generalized laws.

EXERCISE 1.15. Let X be the set of all triangles in the plane, Y the set of all right-angled triangles, and Z the set of all non-isosceles triangles. For any triangle T, let $f(T)$ be the longest side of T, and $g(T)$ be the maximum of the lengths of the sides of T. On which of the sets X, Y, Z is f a function? On which is g a function?

What is the complement of Z in X? What is $Y \cap Z^c$?

EXERCISE 1.16. For each positive real t, let $X_t = (-t, t)$ and $Y_t = [-t, t]$. Describe

(i) $\bigcup_{t>0} X_t$ and $\bigcup_{t>0} Y_t$.

(ii) $\bigcup_{0<t<10} X_t$ and $\bigcup_{0<t<10} Y_t$.

(iii) $\bigcup_{0<t\leq 10} X_t$ and $\bigcup_{0<t\leq 10} Y_t$.

(iv) $\bigcap_{t\geq 10} X_t$ and $\bigcap_{t\geq 10} Y_t$.

(v) $\bigcap_{t>10} X_t$ and $\bigcap_{t>10} Y_t$.

(vi) $\bigcap_{t>0} X_t$ and $\bigcap_{t>0} Y_t$.

EXERCISE 1.17. Let f be the real function cosine, and let g be the real function $g(x) = \dfrac{x^2+1}{x^2-1}$.

(i) What are $f \circ g, g \circ f, f \circ f, g \circ g$ and $g \circ g \circ f$?

(ii) What are the domains and ranges of the real functions $f, g, f \circ g$ and $g \circ f$?

EXERCISE 1.18. Let X be the set of vertices of a square in the plane. How many permutations of X are there? How many of these come from rotations? How many come from reflections in lines? How many come from the composition of a rotation and a reflection?

EXERCISE 1.19. Which of the following real functions are injective, and which are surjective:

(i) $f_1(x) = x^3 - x + 2$.

(ii) $f_2(x) = x^3 + x + 2$.

(iii) $f_3(x) = \dfrac{x^2+1}{x^2-1}$.

(iv) $f_4(x) = \begin{cases} -x^2 & x \leq 0 \\ 2x+3 & x > 0. \end{cases}$

EXERCISE 1.20. Suppose $f : X \to Y$ and $g : Y \to Z$. Prove that if $g \circ f$ is injective, then f is injective.

Give an example to show that g need not be injective.

EXERCISE 1.21. Suppose $f : X \to Y$ and $g : Y \to Z$.
(i) Show that if f and g are surjective, so is $g \circ f$.
(ii) Show that if $g \circ f$ is surjective, then one of the two functions f, g must be surjective (which one?). Give an example to show that the other function need not be surjective.

EXERCISE 1.22. For what $n \in \mathbb{N}$ is the function $f(x) = x^n$ an injection.

EXERCISE 1.23. Let $f : \mathbb{R} \to \mathbb{R}$ be a polynomial of degree $n \in \mathbb{N}$. For what values of n must f be a surjection, and for what values is it not a surjection?

EXERCISE 1.24. Write down a bijection from $(X \times Y)$ $timesZ$ to $X \times (Y$ $timesZ)$. Prove that it is one-to-one and onto.

EXERCISE 1.25. Let X be a set with n elements. How many permutations of X are there?

EXERCISE 1.26. Let $f : \mathbb{R} \to \mathbb{R}$ be a function built using only natural numbers and addition, multiplication and exponentiation (for instance f could be defined as $x \mapsto (x+3)^{x^2}$). What can you say about $f[\mathbb{N}]$? What can you say if we include subtraction or division?

EXERCISE 1.27. Let $f(x) = x^3 - x$. Find sets X and Y such that $f : X \to Y$ is a bijection. Is there a maximal choice of X? If there is, is it unique? Is there a maximal choice of y? If there is, is it unique?

EXERCISE 1.28. Let $f(x) = \tan(x)$. Use set notation to define the domain and range of f. What is $f^{-1}(1)$? What is $f^{-1}[\mathbb{R}^+]$?

EXERCISE 1.29. For each of the following real functions, find an interval X that contains more than one point and such that the function is a bijection from X to $f[X]$. Find a formula for the inverse function.
(i) $f_1(x) = x^2 + 5x + 6$.
(ii) $f_2(x) = x^3 - x + 2$.
(iii) $f_3(x) = \dfrac{x^2+1}{x^2-1}$.
(iv) $f_4(x) = \begin{cases} -x^2 & x \leq 0 \\ 2x+3 & x > 0 \end{cases}$

EXERCISE 1.30. Find formulas for the following sequences:
(i) $\langle 1, 2, 9, 28, 65, 126, \ldots \rangle$.
(ii) $\langle 1, -1, 1, -1, 1, -1, \ldots \rangle$.
(iii) $\langle 2, 1, 10, 27, 66, 125, 218, \ldots \rangle$.
(iv) $\langle 1, 1, 2, 3, 5, 8, 13, 21, \ldots \rangle$.

EXERCISE 1.31. Let the real function f be strictly increasing. Show that for any $b \in \mathbb{R}$, $f^{-1}(b)$ is either empty or consists of a single element, and that f is therefore an injection. If f is also a bijection, is the inverse function of f also strictly increasing?

EXERCISE 1.32. Let f be a real function that is a bijection. Show that the graph of f^{-1} is the reflection of the graph of f in the line $y = x$.

EXERCISE 1.33. Let $X_n = \{n+1, n+2, \ldots, 2n\}$ for each $n \in \mathbb{N}^+$ as in Example 1.43. What are
(i) $\cup_{n=1}^{5} X_n$.
(ii) $\cap_{n=4}^{6} X_n$.
(iii) $\cap_{k=1}^{5} \left[\cup_{n=1}^{k} X_n\right]$.
(iv) $\cap_{k=5}^{\infty} \left[\cup_{n=3}^{k} X_n\right]$.

EXERCISE 1.34. Verify the assertions of Example 1.44.

EXERCISE 1.35. Let $f : X \to Y$, and assume that $U_\alpha \subseteq X$ for every $\alpha \in A$, and $V_\beta \subseteq Y$ for every $\beta \in B$. Prove:

$$\begin{aligned}
(i) \quad & f\left(\bigcup_{\alpha\in A} U_\alpha\right) &&= \bigcup_{\alpha\in A} f(U_\alpha) \\
(ii) \quad & f\left(\bigcap_{\alpha\in A} U_\alpha\right) &&\subseteq \bigcap_{\alpha\in A} f(U_\alpha) \\
(iii) \quad & f^{-1}\left(\bigcup_{\beta\in B} V_\beta\right) &&= \bigcup_{\beta\in B} f^{-1}(V_\beta) \\
(iv) \quad & f^{-1}\left(\bigcap_{\beta\in B} V_\beta\right) &&= \bigcap_{\beta\in B} f^{-1}(V_\beta).
\end{aligned}$$

Note that (ii) has containment instead of equality. Give an example of proper containment in part (ii). Find a condition on f that would ensure equality in (ii).

1.9. Hints to get started on some exercises

Exercise 1.2. You could do this with a Venn diagram. However, once there are more than three sets (see Exercise 1.13), this approach will be difficult. An algebraic proof will generalize more easily, so try to find one here. Argue for the two inclusions

$$\begin{aligned}
(X \cup Y)^c &\subseteq X^c \cap Y^c \\
X^c \cap Y^c &\subseteq (X \cup Y)^c
\end{aligned}$$

separately. In the first one, for example, assume that $x \in (X \cup Y)^c$ and show that it must be in both X^c and Y^c.

Exercise 1.13. Part of the problem here is notation — what if you have more sets than letters? Start with a finite number of sets contained in U, and call them $X_1, \ldots, X_n$. What do you think the complement of their union is? Prove it as you did when $n = 2$ in Exercise 1.2. (See the advantage of having a proof in Exercise 1.2 that did not use Venn diagrams? One of the reasons mathematicians like to have multiple proofs of the same theorem is that each proof is likely to generalize in a different way).

Can you make the same argument work if your sets are indexed by some infinite index set?

Now do the same thing with the complement of the intersection.

Exercise 1.14. Again there is a notational problem, but while Y and Z play the same rôle in Exercise 1.3, X plays a different rôle. So rewrite the equations as

$$\begin{aligned} X \cap (Y_1 \cup Y_2) &= (X \cap Y_1) \cup (X \cap Y_2) \\ X \cup (Y_1 \cap Y_2) &= (X \cup Y_1) \cap (X \cup Y_2), \end{aligned}$$

and see if you can generalize these.

Exercise 1.35. (i) Again, this reduces to proving two containments. If y is in the left-hand side, then there must be some x_0 in some U_{α_0} such that $f(x) = y$. But then y is in $f(U_{\alpha_0})$, so y is in the right-hand side.

Conversely, if y is in the right-hand side, then it must be in $f(U_{\alpha_0})$ for some $\alpha_0 \in A$. But then y is in $f\left(\cup_{\alpha \in A} U_\alpha\right)$, and so is in the left-hand side.

CHAPTER 2

Relations

2.1. Definitions

DEFINITION. Relation Let X and Y be sets. A relation from X to Y is a subset of $X \times Y$.

Alternatively, any set of ordered pairs is a relation. If $Y = X$, we say that R is a relation on X.

NOTATION. *xRy Let X and Y be sets and R be a relation on $X \times Y$. If $x \in X$ and $y \in Y$, then we may express that x bears relation R to y (that is $(x, y) \in R$) by writing xRy.*

So for X and Y sets, $x \in X$, $y \in Y$, and R a relation on $X \times Y$,

$$xRy \quad \text{if and only if} \quad (x, y) \in R.$$

EXAMPLE 2.1. Let $\leq$ be the usual ordering on $\mathbb{Q}$. Then $\leq$ is a relation on $\mathbb{Q}$. We write

$$1/2 \leq 2$$

to express that $1/2$ bears the relation $\leq$ to 2.

EXAMPLE 2.2. Define a relation R from $\mathbb{Z}$ to $\mathbb{R}$ by xRy if $x > y+3$. Then we could write $7\ R\ \sqrt{2}$ or $(7, \sqrt{2}) \in R$ to say that $(7, \sqrt{2})$ is in the relation.

EXAMPLE 2.3. Let $X = \{2, 7, 17, 27, 35, 72\}$. Define a relation R by xRy if $x \neq y$ and x and y have a digit in common. Then

$$\begin{aligned} R = \{&(2,27),(2,72),(7,17),(7,27),(7,72),(17,7),(17,27),(17,72),\\ &(27,2),(27,7),(27,17),(27,72),(72,2),(72,7),(72,17),(72,27)\}. \end{aligned}$$

EXAMPLE 2.4. Let P be the set of all polygons in the plane. Define a relation E by saying $(x, y) \in E$ if x and y have the same number of sides.

How do mathematicians use relations? A relation on a set can be used to impose structure. In Example 2.1, the usual ordering relation $\leq$ on $\mathbb{Q}$ allows us to think of rational numbers as lying on a number line, which provides additional insight into rational numbers. In Example 2.4, we can use the relation to break polygons up into the sets of triangles, quadrilaterals, pentagons, etc.

A function $f : X \to Y$ can be thought of as a very special sort of relation from X to Y. Indeed, the graph of the function is a set of ordered pairs in $X \times Y$, but it has the additional property that every x in X occurs exactly once as a first element of a pair in the relation. As we discussed in Section 1.3, functions are a useful way to relate sets.

Let X be a set, and R a relation on X. Here are some important properties the relation may or may not have.

DEFINITION. Reflexive R is *reflexive* if for every $x \in X$,

$$xRx.$$

Symmetric R is *symmetric* if for any $x, y \in X$,

$$xRy \text{ implies } yRx.$$

Antisymmetric R is *antisymmetric* if for any $x, y \in X$,

$$[(x, y) \in R \text{ and } (y, x) \in R] \text{ implies } x = y.$$

Transitive R is *transitive* if for any $x, y, z \in X$,

$$[xRy \text{ and } yRz] \text{ implies } [xRz].$$

Which of these four properties apply to the relations given in Examples 2.1-2.4 (Exercise 2.1)?

2.2. Orderings

A relation on a set may be thought of as part of the structure imposed on the set. Among the most important relations on a set are order relations.

DEFINITION. Partial ordering Let X be a set and R a relation on X. We say that R is a partial ordering if:

(1) R is reflexive
(2) R is antisymmetric
(3) R is transitive.

EXAMPLE 2.5. Let X be a family of sets. The relation $\subseteq$ is a partial ordering on X. Every set is a subset of itself, so the relation is reflexive. If $Y \subseteq Z$ and $Z \subseteq Y$, then $Y = Z$, so the relation is antisymmetric. Finally, if $Y \subseteq Z$ and $Z \subseteq W$ then $Y \subseteq W$, so the relation is transitive.

EXAMPLE 2.6. Let R be the relation on $\mathbb{N}^+$ defined by xRy if and only if there is $z \in \mathbb{N}^+$ such that

$$xz = y.$$

Then R is a partial ordering of $\mathbb{N}^+$. (Prove this: Exercise 2.2).

DEFINITION. Linear ordering Let X be a set and R be a partial ordering of X. We say that R is a *linear ordering,* also called a *total ordering,* provided that, for any $x, y \in X$, either xRy or yRx.

Note that since a linear ordering is antisymmetric, for any distinct x and y, exactly one of xRy and yRx holds.

EXAMPLE 2.7. The ordering $\leq$ on $\mathbb{N}$ (or $\mathbb{R}$) is a linear ordering. So is the relation $\geq$. The relation $<$ is not (why?).

EXAMPLE 2.8. Let $X = \mathbb{R}^n$. We can define a reflexive relation on X as follows. Let $x = (a_1, \ldots, a_n)$ and $y = (b_1, \ldots, b_n)$ be distinct members of X. Let $k \in \mathbb{N}^+$ be the least number such that $a_k \neq b_k$. Then we define

$$xRy \quad \text{if and only if } a_k < b_k.$$

Then R is a linear ordering of X. It is called the *dictionary ordering*.

The notion of a linear ordering is probably natural for you, and you have used it intuitively since you began studying arithmetic. The relation $\leq$ helps you to visualize the set as a line in which the relative location of two elements of the set is determined by the linear ordering. If you are considering a set with operations, this in turn can help in visualizing how operations behave. For instance, think of using a number line to visualize addition, subtraction and multiplication of integers.

Partial orderings are generalizations of linear orderings, and $\leq$ is the most obvious example of a linear ordering. Because of this, the normal symbol for a partial ordering is $\preceq$ (it is also reminiscent of the symbol $\subseteq$, which is the example most mathematicians keep in mind when thinking about a partial ordering).

EXAMPLE 2.9. Let X be the set of all collections of apples and oranges. If x, y are in X, then say $x \preceq y$ if the number of apples in x is less than or equal to the number of apples in y, and the number of oranges in x is less than or equal to the number of oranges in y. This is a partial ordering. You may not be able to compare apples to oranges, but you can say that 2 apples and 5 oranges is inferior to 4 apples and 6 oranges!

One way to visualize a partial order $\preceq$ on a finite set X is to imagine arrows connecting distinct elements of X, x and y, if $x \preceq y$ and there is no third distinct point z satisfying $x \preceq z \preceq y$. Then two elements a and b in X will satisfy $a \preceq b$ if and only if you can get from a to b by following a path of arrows.

EXAMPLE 2.10. Consider the graph on the set $X = \{a, b, c, d, e, f\}$ give in Figure 2.11.

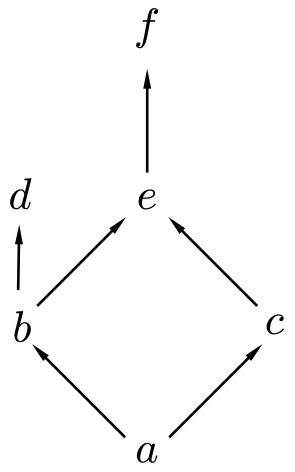


FIGURE 2.11. Picture of a partial order

It illustrates the partial order that could be described as the smallest reflexive, transitive relation $\preceq$ on X that satisfies $a \preceq b,\ a \preceq c,\ b \preceq d,\ b \preceq e,\ c \preceq e,\ e \preceq f$.

2.3. Equivalence Relations

DEFINITION. Equivalence relation Let X be a set and R a relation on X. We say R is an equivalence relation if

(1) R is reflexive
(2) R is symmetric
(3) R is transitive.

EXAMPLE 2.12. Define a relation R on $\mathbb{R}$ by xRy if and only if $x^2 = y^2$. Then R is an equivalence relation.

EXAMPLE 2.13. Let R be a relation defined on $\mathbb{Z} \times \mathbb{Z}$ as follows. If $a, b, c, d \in \mathbb{Z}$,

$$(a, b)\ R\ (c, d) \ \text{ if and only if } \ a + d \ = \ b + c. \tag{2.14}$$

Then R is an equivalence relation. Indeed, let us check the three properties.

Reflexive: By (2.14), we have $(a, b)\ R\ (a, b)$ if $a + b = a + b$, which clearly holds.

Symmetric: Suppose $(a, b)\ R\ (c, d)$, so $a + d = b + c$. To see if $(c, d)\ R\ (a, b)$, we must check whether $c + b = d + a$; but this holds by the commutativity of addition.

Transitive: Suppose $(a, b)\ R\ (c, d)$ and $(c, d)\ R\ (e, f)$. We must check that $(a, b)\ R\ (e, f)$, in other words that

$$a + f = b + e. \tag{2.15}$$

We have $a + d = b + c$ and $c + f = d + e$, and adding these two equations we get

$$a + d + c + f = b + c + d + e. \tag{2.16}$$

Cancelling $c + d$ from each side of (2.16), we get (2.15) as desired.

EXAMPLE 2.17. Let R be a relation on $X = \mathbb{Z} \times \mathbb{N}^+$ defined by

$$(a, b)\ R\ (c, d) \text{ if and only if } ad = bc.$$

Then R is an equivalence relation on X. (Prove this; Exercise 2.4).

EXAMPLE 2.18. Let $f : X \to Y$. Define a relation R_f on X by

$$x\, R_f\, y \text{ if and only if } f(x) = f(y).$$

Then R_f is an equivalence relation. We check the conditions for an equivalence relation:
R_f is clearly reflexive, since, for any $x \in X$,

$$f(x) = f(x).$$

R_f is symmetric since, for any $x \in X$ and $y \in X$,

$$f(x) = f(y) \text{ if and only if } f(y) = f(x).$$

To show R_f is transitive, let $x, y, z \in X$. If $f(x) = f(y)$ and $f(y) = f(z)$ then $f(x) = f(z)$.

Equivalence relations have three of the key properties of identity. They allow us to relate objects in a set that we wish to consider as "the same" in a given context. This allows us to focus on which differences between mathematical objects are relevant to the discussion

at hand, and which are not. For this reason, a common symbol for an equivalence relation is $\sim$.

DEFINITION. Equivalence class, $[x]_R$ Let R be an equivalence relation on a set X. If $x \in X$ then the equivalence class of x modulo R, denoted by $[x]_R$, is

$$[x]_R \;=\; \{y \in X \mid xRy\}.$$

If $y \in [x]_R$ we call y a representative element of $[x]_R$. The set of all equivalence classes $\{[x]_R \mid x \in X\}$ is written X/R. It is called the quotient space of X by R.

We may use $[x]$ for the equivalence class of x, provided that the equivalence relation is clear.

NOTATION. *Equivalence mod R, $\equiv_R$, $\sim$ Let R be an equivalence relation on a set X. We may express that xRy by writing*

$$x \equiv y \bmod R$$

or

$$x \equiv_R y$$

or

$$x \sim y.$$

PROPOSITION 2.19. *Suppose that $\sim$ is an equivalence relation on X. Let $x, y \in X$. If $x \sim y$, then*

$$[x] \;=\; [y]. \tag{2.20}$$

If x is not equivalent to y ($x \nsim y$), then

$$[x] \cap [y] \;=\; \emptyset.$$

PROOF. (i) Assume $x \sim y$. Let us show that $[x] \subseteq [y]$. Let $z \in [x]$. This means that $x \sim z$. Since $\sim$ is symmetric, and $x \sim y$, we have $y \sim x$. As $y \sim x$ and $x \sim z$, by transitivity of $\sim$ we get that $y \sim z$. Therefore $z \in [y]$. Since z is an arbitrary element of $[x]$, we have shown that $[x] \subseteq [y]$.

As $y \sim x$, the same argument with x and y swapped gives $[y] \subseteq [x]$, and therefore $[x] = [y]$.

(ii) Now assume that x and y are not equivalent. We must show that there is no z such that $z \in [x]$ and $z \in [y]$. We will argue by contradiction. Suppose there were such a z. Then we would have

$$x \sim z \qquad \text{and} \qquad y \sim z.$$

By symmetry, we have also that $z \sim y$, and by transitivity, we then have that $x \sim y$. This contradicts the assumption that x is not equivalent to y. So if x and y are not equivalent, no z can exist that is simultaneously in both $[x]$ and $[y]$. Therefore $[x]$ and $[y]$ are disjoint sets, as required. □

So what have we shown? We have not shown that any particular relation is an equivalence relation. Rather we have shown that any equivalence relation on a set partitions the set into disjoint equivalence classes.

As we shall see throughout this book, and you will see throughout your mathematical studies, this is a surprisingly powerful tool.

DEFINITION. Pairwise disjoint Let $\{X_\alpha \mid \alpha \in A\}$ be a family of sets. The family is pairwise disjoint if for any $\alpha, \beta \in A$, $\alpha \neq \beta$,

$$X_\alpha \cap X_\beta = \emptyset.$$

DEFINITION. Partition Let Y be a set and $\mathcal{F} = \{X_\alpha \mid \alpha \in A\}$ be a family of non-empty sets. The collection $\mathcal{F}$ is a partition of Y if $\mathcal{F}$ is pairwise disjoint and

$$Y = \bigcup_{\alpha \in A} X_\alpha.$$

Given an equivalence relation $\sim$ on a set X, the equivalence classes with respect to $\sim$ give a partition of X. Conversely, partitions give rise to equivalence relations.

THEOREM 2.21. *(i) Let X be a set, and $\sim$ an equivalence relation on X. Then $X/\sim$ is a partition of X.*

(ii) Conversely, let $\{X_\alpha \mid \alpha \in A\}$ be a partition of X. Let $\sim$ be the relation on X defined by $x \sim y$ whenever x and y are members of the same set in the partition. Then $\sim$ is an equivalence relation.

PROOF. Part (i) of the theorem is Proposition 2.19 restated, and we gave the proof above. To prove the converse, we must show that the relation $\sim$ defined as in part (ii) of the theorem is an equivalence.

Reflexivity: Let $x \in X$. Then x is in some X_{α_0}, as the union of all these sets is all of X. Therefore $x \sim x$.

Symmetry: Suppose $x \sim y$. Then there is some X_{α_0} such that $x \in X_{\alpha_0}$ and $y \in X_{\alpha_0}$. This implies that $y \sim x$.

Transitivity: Suppose $x \sim y$ and $y \sim z$. Then there are sets X_{α_0} and X_{α_1} such that both x and y are in X_{α_0}, and both y and z are in X_{α_1}. But since the sets X_α form a partition, and y is in both X_{α_0} and X_{α_1}, we must have that $X_{\alpha_0} = X_{\alpha_1}$. This implies that x and z are in the same member of the partition, and so $x \sim z$. □

2.4. Constructing Bijections

Let's consider a particularly interesting and important abstract application of equivalence classes. Let $f : X \to Y$. The function f need not be an injection or surjection. However, we have already discussed the desirability of finding an "inverse" for f , even when it fails to meet the necessary conditions for the existence of an inverse. In Section 1.3 we considered the function $f|_D$, where $D \subseteq X$ and $f|_D$ is an injection. Another approach is to use the function f to create a new function on a distinct domain that preserves much of the information of f.

We use f to induce an equivalence relation on X. Define a relation $\sim$ on X by

$$x \sim y \text{ if and only if } f(x) = f(y).$$

We showed in Example 2.18 that $\sim$ is an equivalence relation; it is the equivalence relation on X induced by f. The equivalence relation $\sim$ induces a partition of X, namely $X/\sim$ (which is the set $\{[x] \mid x \in X\}$ of all equivalence classes).

NOTATION. X/f *Let $f : X \to Y$ and $\sim$ be the equivalence relation on X induced by f. We write X/f for the set of equivalence classes induced by $\sim$ on X.*

An equivalence class in X/f is the inverse image of an element in $\text{Ran}(f)$. That is, if $x \in X$ and $f(x) = y$,

$$[x] = f^{-1}(y).$$

So

$$X/f = \{f^{-1}(y) \mid y \in \text{Ran}(f)\}.$$

The elements of X/f are called the level sets of f. The inspiration for this comes from thinking of a topographical map. The curves on a topographical map corresponding to fixed altitudes are called level curves. Consider the function from a point on a map to the altitude of the physical location represented by the point on the map. Level curves on the map are subsets of the level sets of this function.

NOTATION. Π_f *Let $f : X \to Y$. The function $\Pi_f : X \to X/f$ is defined by $\Pi_f(x) = [x]_f$, where $[x]_f$ is the equivalence class of x with respect to the equivalence relation induced by f.*

Let $Z \subseteq Y$ be the range of f. We define a new function, $\widehat{f} : X/f \to Z$ by

$$\widehat{f}([x]) = f(x).$$

The function $\widehat{f}$ is closely related to f; in fact, for every $x \in X$,

$$f(x) = \widehat{f} \circ \Pi_f(x).$$

This is sometimes illustrated with a diagram, as in Figure 2.22.

The function $\widehat{f}$ is a bijection. In this sense, every function can be canonically associated with a bijection. We consider the function that we looked at in Section 1.3.

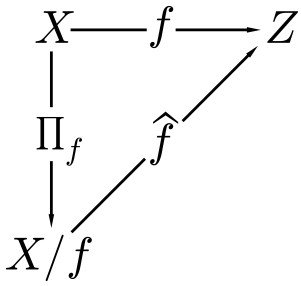


FIGURE 2.22. Making a function into a bijection

EXAMPLE 2.23. Let $f(x) = \tan(x)$. As we discussed earlier, we can "invert" this function by considering the function $\text{Tan} : (-\pi/2, \pi/2) \to \mathbb{R}$ by

$$\text{Tan} = \tan|_{(-\pi/2,\pi/2)}.$$

The function Tan is a bijection, and has an inverse,

$$\text{Arctan} : \mathbb{R} \to (-\pi/2, \pi/2).$$

For any $k \in \mathbb{Z}$ there is a corresponding restriction of tan,

$$\tan|_{\left(\frac{(2k+1)\pi}{2}, \frac{(2k+3)\pi}{2}\right)}$$

which is a bijection, and therefore has an inverse function.

Another bijection can be constructed on the equivalence classes induced by $f(x) = \tan(x)$. A level set of f is $[x]_f = \{x + k\pi \mid k \in \mathbb{Z}\}$. Let X be the domain of tan. Then

$$X/f = \{[x]_f \mid x \in X\}.$$

We can interpret an equivalence class $[x]_f$ with respect to angles in standard position in the Cartesian plane. The equivalence class of x is the set of angles in standard position that have terminal side collinear with the terminal side of the angle x — see Figure 2.24.

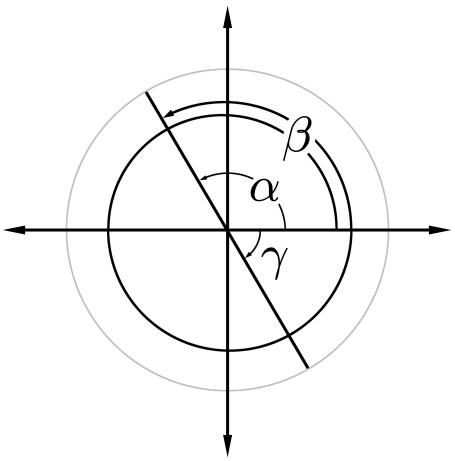


FIGURE 2.24. Collinear Angles

Following the construction outlined above, the function $\Pi_f : X \to X/f$ is the function

$$\Pi_f(x) \;=\; [x]_f \;=\; \{x + k\pi \mid \; k \in \mathbb{Z}\}.$$

The function $\widehat{f} : X/f \to \mathbb{R}$ given by

$$\widehat{f}([x]_f) \;=\; f(x)$$

is a bijection. Furthermore,

$$\tan \;=\; \widehat{f} \circ \Pi_f.$$

If $x \in X$, then $\Pi_f(x)$ is the set of all angles that have terminal side collinear with the terminal side of angle x in standard position. Thus Π_f tells us that tan can distinguish only the slope of the terminal side of the angle — not the quadrant of the angle or how many revolutions the angle subtended.

2.5. Modular Arithmetic

We define an equivalence relation that will help us derive insights in number theory.

DEFINITION. Divides, $a \mid b$ Let a and b be integers. Then a divides b, written $a \mid b$, if there is $c \in \mathbb{Z}$ such that

$$a \cdot c \;=\; b.$$

DEFINITION. Congruence, $x \equiv y \mod n$, $\equiv_n$ Let $x, y, n \in \mathbb{Z}$ and $n > 1$. Then

$$x \equiv y \mod n$$

(or $x \equiv_n y$) if

$$n|(x - y).$$

The relation $\equiv_n$ on $\mathbb{Z}$ is called congruence mod n.

THEOREM 2.25. *Congruence mod n is an equivalence relation on $\mathbb{Z}$.*

Exercise 2.5: Prove Theorem 2.25.

DEFINITION. Congruence class The equivalence classes of the relation $\equiv_n$ are called congruence classes, residue classes, or remainder classes mod n. The set of congruence classes mod n can be written $\mathbb{Z}_n$ or $\mathbb{Z}/n\mathbb{Z}$.

Of course $\mathbb{Z}_n$ is a partition of $\mathbb{Z}$. When $n = 2$, the residue classes are called the even and the odd numbers. Many of the facts you know about even and odd numbers generalize if you think of them as residue classes. What are the residue classes for $n = 3$?

We leave it as an exercise (Exercise 2.6) to prove that two integers are in the same remainder class mod n provided that they have the same remainder when divided by n.

NOTATION. [a] *Fix a natural number $n \geq 2$. Let a be in $\mathbb{Z}$. We represent the equivalence class of a with respect to the relation $\equiv_n$ by $[a]$.*

PROPOSITION 2.26. *If $a \equiv r \mod n$ and $b \equiv s \mod n$, then*

$$(i) \qquad a + b \equiv r + s \mod n$$

and

$$(ii) \qquad ab \equiv rs \mod n.$$

PROOF. (i) Assume that $a \equiv r \mod n$ and $b \equiv s \mod n$. Then $n|(a-r)$ and $n|(b-s)$. So

$$n|(a+b-(r+s)).$$

Therefore

$$a+b \equiv r+s \mod n,$$

proving (i).

To prove (ii), note that there are $i, j \in \mathbb{Z}$ such that

$$a = ni + r$$

and

$$b = nj + s.$$

Then

$$ab = n^2ji + rnj + sni + rs = n(nji + rj + si) + rs.$$

Therefore

$$n|(ab - rs)$$

and

$$ab \equiv rs \mod n.$$

□

Hence the algebraic operations that $\mathbb{Z}_n$ "inherits" from $\mathbb{Z}$ are well-defined. That is, we may define $+$ and $\cdot$ on $\mathbb{Z}_n$ by

$$[a] + [b] = [a+b] \tag{2.27}$$

and

$$[a] \cdot [b] = [a \cdot b]. \tag{2.28}$$

In mathematics, when you ask whether something is "well-defined", you mean that somewhere in the definition a choice was made, and you want to know whether a different choice would have resulted in the same final result. For example, let $X_1 = \{-2, 2\}$ and let $X_2 = \{-1, 2\}$. Define y_1 by: "Choose x in X_1 and let $y_1 = x^2$." Define y_2 by: "Choose

x in X_2 and let $y_2 = x^2$." Then y_1 is well-defined, and is the number 4; but y_2 is not well-defined, as different choices of x give rise to different numbers.

In (2.27) and (2.28), the right-hand sides depend *a priori* on a particular choice of elements from the equivalence classes $[a]$ and $[b]$. But Proposition 2.26 ensures that sum and product so defined are independent of the choice of representatives of the equivalence classes.

EXAMPLE 2.29. In Z_2 addition and multiplication are defined as follows:

(1) $[0] + [0] = [0]$
(2) $[0] + [1] = [1] + [0] = [1]$
(3) $[1] + [1] = [0]$
(4) $[0] \cdot [0] = [0] \cdot [1] = [1] \cdot [0] = [0]$
(5) $[1] \cdot [1] = [1]$.

Notice that if you read [0] as "even" and [1] as "odd", these are rules that you learned a long time ago.

When working with modular arithmetic we may pick the representatives of remainder classes which best suit our needs. For instance,

$$79 \cdot 23 \equiv 2 \cdot 2 \equiv 4 \mod 7.$$

In other words

$$[79 \cdot 23] = [79] \cdot [23] = [2] \cdot [2] = [4].$$

EXAMPLE 2.30. You may recall from your early exposure to multiplication tables that multiplication by nine resulted in a product whose digits summed to nine. This generalizes nicely with modular arithmetic. Specifically, if $a_n \in \ulcorner 10 \urcorner$ for $0 \leq n \leq N$ then

$$\sum_{n\,=\,0}^{N} a_n 10^n \equiv \sum_{n\,=\,0}^{N} a_n \mod 9. \tag{2.31}$$

The remainder of any integer divided by 9 equals the remainder of the sum of the digits of that integer when divided by 9.

PROOF. The key observation is that

$$10 \equiv 1 \mod 9.$$

Therefore

$$\begin{aligned} 10^2 &\equiv 1 \cdot 1 \equiv 1 \mod 9 \\ 10^3 &\equiv 1 \cdot 1 \cdot 1 \equiv 1 \mod 9, \end{aligned}$$

and so on for any power of 10:

$$10^n \equiv 1 \mod 9 \quad \text{for all } n \in \mathbb{N}.$$

(This induction to all powers of 10 is straightforward, but to prove it formally we shall need the notion of mathematical induction from Chapter 4). Therefore on the left-hand side of (2.31), working mod 9, we can replace all the powers of 10 by 1, and this gives us the right-hand side. □

EXAMPLE 2.32. The observation that a number's residue mod 9 is the same as that of the sum of the digits can be used in a technique called "casting out nines" to check arithmetic.

For example, consider the following (incorrect) sum. The number in the penultimate column is the sum of the digits, and the number in the last column is the repeated sum of the digits until reaching a number between 0 and 9.

1588	22	4
+1805	14	5
3493	19	1

If the addition had been correctly performed, the remainder mod 9 of the sum would equal the sum of the remainders; so we know a mistake was made.

EXAMPLE 2.33. What is the last digit of 7^7?

We want to know $7^7 \mod 10$. Note that, modulo 10, $7^0 \equiv 1,\ 7^1 \equiv 7,\ 7^2 \equiv 9,\ 7^3 \equiv 3,\ 7^4 \equiv 1$. So $7^7 = 7^4 7^3 \equiv 1 \cdot 3 \equiv 3$, and so 3 is the last digit of 7^7.

EXAMPLE 2.34. What is the last digit of 7^{7^7}?

By Example 2.33, we see that the residues of $7^n \mod 10$ repeat themselves every time n increases by 4. Therefore if $m \equiv n \mod 4$, then $7^m \equiv 7^n \mod 10$.

What is $7^7 \mod 4$? Well $7^1 \equiv 3 \mod 4$, $7^2 \equiv 1 \mod 4$, so $7^7 \equiv (7^2)^3 \cdot 7 \equiv 3 \mod 4$. Therefore

$$7^{7^7} \equiv 7^3 \equiv 3 \mod 10.$$

2.6. Exercises

EXERCISE 2.1. Which of the properties of reflexivity, symmetry, antisymmetry and transitivity apply to the relations given in Examples 2.1-2.4?

EXERCISE 2.2. Prove that the relation in Example 2.6 is a partial ordering.

EXERCISE 2.3. List every pair in the relation given in Example 2.10.

EXERCISE 2.4. Prove that the relation in Example 2.17 is an equivalence.

EXERCISE 2.5. Prove that congruence mod n is an equivalence relation on $\mathbb{Z}$.

EXERCISE 2.6. Prove that two integers are in the same congruence class mod n if and only if they have the same remainder when divided by n.

EXERCISE 2.7. Suppose R is a relation on X. What does it mean if R is both a partial order and an equivalence?

EXERCISE 2.8. Consider the relations on people "is a brother of", "is a sibling of", "is a parent of", "is married to", "is a descendant of". Which of the properties of reflexivity, symmetry, antisymmetry and transitivity do each of these relations have?

EXERCISE 2.9. Let $X = \{k \in \mathbb{N} : k \geq 2\}$. Consider the following relations on X:
(i) $j\ R_1\ k$ if and only if $\gcd(j,k) > 1$ (gcd stands for greatest common divisor).
(ii) $j\ R_2\ k$ if and only if j and k are coprime (*i.e.* $\gcd(j,k) = 1$).
(iii) $j\ R_3\ k$ if and only if $j|k$.
(iv) $j\ R_4\ k$ if and only if

$$\{p : p \text{ is prime and } p|j\} = \{q : q \text{ is prime and } q|k\}.$$

For each relation, say which of the properties of Reflexivity, Symmetry, Antisymmetry, Transitivity it has.

EXERCISE 2.10. For j, k in $\mathbb{N}^+$, define two relations R_1 and R_2 by jR_1k if j and k have a digit in common (but not necessarily in the same place) and $j\,R_2\,k$ if j and k have a common digit in the same place (so, for example, $108\ R_1\ 82$, but $(108, 82) \notin R_2$).

(i) If $j = \sum_{m=0}^{M} a_m 10^m$ and $k = \sum_{n=0}^{N} b_n 10^n$, with $a_M \neq 0$ and $b_N \neq 0$, how can one mathematically define R_1 and R_2 in terms of the coefficients a_m and b_n?

(ii) Which of the four properties of reflexivity, symmetry, antisymmetry and transitivity do R_1 and R_2 have?

EXERCISE 2.11. Let $X = \{a, b\}$. List all possible relations on X, and say which are reflexive, which are symmetric, which are antisymmetric, and which are transitive.

EXERCISE 2.12. How many relations are there on a set with 3 elements? How many of these are reflexive? How many are symmetric? How many are anti-symmetric?

EXERCISE 2.13. Repeat Exercise 2.12 for a set with N elements.

EXERCISE 2.14. The sum of two even integers is even, the sum of an even and an odd integer is odd, and the sum of two odd integers is even. What is the generalization of this statement to residue classes mod 3?

EXERCISE 2.15. What is the last digit of 3^{5^7}? Of 7^{5^3}? Of 11^{10^6}? Of 8^{5^4}?

EXERCISE 2.16. What is $2^{1000000} \mod 17$? What is $17^{77} \mod 14$?

EXERCISE 2.17. Show that a number's residue mod 3 is the same as the sum of its digits.

EXERCISE 2.18. Show that the assertion of Exercise 2.17 is not true mod n for any value of n except 3 and 9.

EXERCISE 2.19. Prove that there are an infinite number of natural numbers that cannot be written as the sum of three squares. (Hint: Look at the possible residues mod 8).

EXERCISE 2.20. Let $f : X \to Y$ and $g : Y \to Z$. What can you say about the relationship between X/f and $X/(g \circ f)$?

EXERCISE 2.21. Let R be the relation on $X = \mathbb{Z} \times \mathbb{N}^+$ defined in Example 2.17. Define an operation $\star$ on X/R as follows: for $x = (a, b)$ and $y = (c, d)$,

$$[x] \star [y] = [(ad + bc, cd)].$$

Is $\star$ well-defined?

EXERCISE 2.22. Let X be the set of functions from finite subsets of $\mathbb{N}$ to $\ulcorner 2 \urcorner$ (that is $f \in X$ iff there is a finite set $D \subseteq \mathbb{N}$ such that $f : D \to \ulcorner 2 \urcorner$). Define a relation R on X as follows: if $f, g \in X$, fRg iff $\mathrm{Dom}(g) \subseteq \mathrm{Dom}(f)$ and $g = f|_{\mathrm{Dom}(g)}$. Is R a partial ordering? Is R an equivalence relation?

EXERCISE 2.23. Let X be the set of all infinite binary sequences. Define a relation R on X as follows: For any $f, g \in X$, fRg iff $f^{-1}(1) \subseteq g^{-1}(1)$. Is R a partial ordering? Is R an equivalence relation?

EXERCISE 2.24. Let $X = \{\ulcorner n \urcorner \mid n \in \mathbb{N}\}$. Let R be a relation on X defined by $x, y \in R$ iff $x \subseteq y$. Prove that R is a linear ordering.

EXERCISE 2.25. Let $X = \{f : \mathbb{R} \to \mathbb{R} \mid f \text{ is a surjection}\}$. Define a relation R on X by fRg iff $f(0) = g(0)$. Prove that R is an equivalence relation. Let $F : X \to \mathbb{R}$ be defined by $F(f) = f(0)$. Show that the level sets of F are the equivalence classes of X/R. That is show that

$$X/R = X/F.$$

EXERCISE 2.26. Let $f : X \to Y$. Show that X/f is composed of singletons (sets with exactly one element) iff f is an injection.

CHAPTER 3

Proofs

3.1. Mathematics and Proofs

The primary activity of research mathematicians is proving mathematical claims. Depending on the depth of the claim, the relationship of the claim to other mathematical claims, and various other factors, a mathematical statement that has been proved is generally called a theorem, proposition, corollary or lemma. A mathematical statement that has not been proved, but that is expected to be true, is commonly called a conjecture. A statement that is accepted as a starting point for arguments without being proved is called an axiom.

Some mathematical results are so fundamental, deep, difficult, surprising or otherwise noteworthy that they are named. Part of your initiation as a member of the community of mathematicians is becoming familiar with some of these named statements — and we shall prove a few of them in this book.

It is likely that most of the mathematics you have studied has been the application of theorems to deriving solutions of relatively concrete problems. Here we begin learning how to prove theorems. Most students find the transition from computational mathematics to mathematical proofs very challenging.

What is a mathematical proof?

The nature of a mathematical proof depends on the context. There is a formal notion of a mathematical proof:

A finite sequence of formal mathematical statements such that each statement either

- is an axiom or assumption,

or

- follows by formal rules of logical deduction from previous statements in the sequence.

Most mathematicians do not think of mathematical proofs as formal mathematical proofs, and virtually no mathematician writes formal mathematical proofs. This is because a formal proof is a hopelessly cumbersome thing, and is generally outside the scope of human capability, even for the most elementary mathematical statements. Rather, mathematicians write proofs that are sequences of statements in a combination of natural language and formal mathematical symbols (interspersed with diagrams, questions, references and other devices that are intended to assist the reader in understanding the proof) that can be thought of as representing a purely formal argument. A good practical definition of a mathematical proof is:

An argument in favor of a mathematical statement that will convince the preponderance of knowledgeable mathematicians of the truth of the mathematical statement.

This definition is somewhat imprecise, and mathematicians can disagree on whether an argument is a proof, particularly for extremely difficult or deep arguments. However, for virtually all mathematical arguments, after some time for careful consideration, the mathematical community reaches a unanimous consensus on whether it is a proof.

The notion of a mathematical proof for the student is similar to the general idea of a mathematical proof. The differences are due to the type of statement that the student is proving, and the reasons for requesting that the student prove the statement. The statements that you will be proving are known to professional mathematicians or can be proved with relatively little effort by your instructors. Clearly the statements you will be proving require different conditions for a

satisfactory proof than those stated above for the professional mathematician. Let's define a successful argument by the student as follows: An argument for a mathematical statement that

- the instructor can understand
- the instructor cannot refute
- uses only assumptions that the instructor considers admissible.

Note that refuting an argument is not the same as refuting the original claim. The sentence "The square of every real number is non-negative because all real numbers are non-negative." is a false proof of a true statement. The sentence "The square of every real number is non-negative because all triangles have three sides." fails the first test: while both statements are true, your instructor will not see how the first follows from the second.

In this book, the solutions to the problems will be an exposition in natural language enhanced by mathematical expressions. The student is expected to learn the conventions of mathematical grammar and argument, and use them. Like most conventions, these are often determined by tradition or precedent. It can be quite difficult, initially, to determine whether your mathematical exposition meets the standards of your instructor. Practice, with feedback from a reader experienced in reading mathematics, is the best way to develop good proof-writing skills. Remember, readers of mathematics are quite impatient with trying to decipher what the author *means* to say — mathematics is sufficiently challenging when the author writes precisely what he or she intends. Most of the burden of communication is on the author of a mathematical proof, not the reader. A proof can be logically correct, but so difficult to follow that it is unacceptable to your instructor.

Why proofs?

Why are proofs the primary medium of mathematics? Mathematicians depend on proofs for certainty and explanation. Once a proof is accepted by the mathematical community, it is virtually unheard of that the result is subsequently refuted. This was not always the

case: in the 19^{th} century there were serious disputes as to whether results had really been proved or not (see Section 5.3 for an example, and the book [**4**] for a very extensive treatment of the development of rigor in mathematical reasoning). This led to our modern notion of a "rigorous" mathematical argument. While one might argue that it is possible that in the 21^{st} century a new standard of rigor will reject what we currently consider to be proofs, our current ideas have been stable for over a century, and most mathematicians (including the authors of this book) do not expect that there will be a philosophical shift.

For very complicated results, writing a detailed proof helps the author convince himself or herself of the truth of the claim. After a mathematician has hit upon the key idea behind an argument, there is a lot of hard work left developing the details of the argument. Many promising ideas fail as the author attempts to write a detailed argument based on the idea. Finally, proofs often provide a deeper insight into the result and the mathematical objects that are the subject of the proof. Indeed, even very clever proofs which fail to provide mathematical insights are held in lower regard, by some, than arguments that elucidate the topic.

Mathematical proofs are strongly related to formal proofs in a purely logical sense. It is supposed that the existence of an informal mathematical proof is overwhelming evidence for the existence of a formal mathematical proof. If it is not clear that the informal proof could conceivably be interpreted into a formal argument, it is doubtful that the informal argument will be accepted by the mathematical community. Consequently, mathematical arguments have a transparent underlying logical structure.

For this reason we shall begin our discussion of mathematical proofs with a brief discussion of propositional logic. Despite its abstractness, the topic is straightforward, and most of the claims of this section may be confirmed with some careful, patient thinking.

3.2. Propositional Logic

Propositional logic studies how the truth or falsehood of compound statements is determined by the truth or falsehood of the constituent statements. It gives us a way of reliably deriving true conclusions from true assumptions.

DEFINITION. Truth value If P is a statement which is true, then P has truth value 1. If P is a statement which is false, P has truth value 0. We write $T(P)$ for the truth value of P.

Truth values can be thought of as a function $T : S \to \ulcorner 2 \urcorner$, where S is the set of all statements. When investigating the abstract principles of propositional logic, we consider possible assignments of truth values to variables representing statements. We are interested in claims that are independent of any particular assignment of truth values to the propositional variables. We use the integers 0 and 1 to represent truth values because it allows us to use arithmetic operations in propositional logic. Other authors prefer F and T.

DEFINITION. Propositional connectives The symbols $\wedge$, $\vee$, $\neg$ and $\Rightarrow$ are propositional connectives. They are defined as follows for statements P and Q.

Connective	Name	Definition
$\neg$	negation	$T(\neg P) = 1 - T(P)$
$\wedge$	conjunction	$T(P \wedge Q) = T(P) \cdot T(Q)$
$\vee$	disjunction	$T(P \vee Q) = T(P) + T(Q) - T(P) \cdot T(Q)$
$\Rightarrow$	implication	$T(P \Rightarrow Q) = 1 - T(P) + T(P) \cdot T(Q)$

In the expression "$P \Rightarrow Q$", the statement P is called the *antecedent* or *hypothesis* and Q is called the *consequence* or *conclusion*.

Propositional connectives are formal equivalents of natural language connectives.

Connective	Natural Language Equivalent
$\neg$	not
$\wedge$	and
$\vee$	or
$\Rightarrow$	implies

Check that the formulas defining the propositional connectives give the meaning that you anticipate. For example, check that the definition of the truth value for $P \wedge Q$ means that $P \wedge Q$ is true if and only if both P and Q are true.

Propositional connectives approximate natural language connectives. Propositional connectives are formal and precise, while natural language connectives are imprecise and somewhat more expressive — consequently the approximation is imperfect. We saw an example of this when contrasting mathematicians' use of the connective "or" with its use in everyday language. For precision in mathematics we interpret the connectives formally — even when using natural language expressions.

We can build very complicated compound statements by using logical connectives. Naturally, there are rules for building correct statements with connectives.

DEFINITION. Atomic statement An atomic statement is a statement with no explicit propositional connectives.

An atomic statement is usually represented by a capital letter.

DEFINITION. Well-formed statement We define a well-formed propositional statement recursively as follows.
Atomic statements are well-formed.
If P and Q are well-formed statements, then the following are well-formed statements:

- $(\neg P)$
- $(P \wedge Q)$
- $(P \vee Q)$
- $(P \Rightarrow Q)$.

In practice the parentheses are dropped unless there is the potential for ambiguity. Additionally, "[" and "]" may be substituted for parentheses in the interests of readability. For any assignment of truth values to the atomic statements in a well-formed statement, the compound statement will have a well-defined truth value.

DEFINITION. Compound statement A compound statement is a well-formed statement composed of atomic statements and propositional connectives.

3.2.1. Propositional Equivalence. One purpose of propositional logic is to give tools for assessing the truth of a compound statement without necessarily having to understand the specific meaning of the atomic statements. That is, some statements are demonstrably true or false by virtue of their form. Central to this understanding is the idea of propositional equivalence.

DEFINITION. Propositional equivalence Let P and Q be well-formed statements built from atomic statements. We say that P and Q are propositionally equivalent provided that $T(P) = T(Q)$ for any assignment of truth values to the constituent atomic statements.

If P and Q are propositionally equivalent, we may write

$$P \equiv Q.$$

EXAMPLE 3.1.

$$[P \Rightarrow Q] \equiv [(\neg Q) \Rightarrow (\neg P)].$$

This is a very important example of a propositional equivalence. We will show this by considering all possible assignments of truth values to P and Q. Let's set this up in what is popularly called a truth table. We consider all possible assignments of truth values to P and Q, and compare the truth values of the compound statements under consideration:

$T(P)$	$T(Q)$	$T(P \Rightarrow Q)$	$T((\neg Q) \Rightarrow (\neg(P)))$
0	0	1	1
0	1	1	1
1	0	0	0
1	1	1	1

Each row of the truth table represents a particular assignment of truth values to the atomic statements P and Q. The four possible assignments are exhausted by the rows of the truth table. The truth values of the compound statements agree in each row of the truth table so the statements are equivalent.

Example 3.2.

$$[\neg(P \wedge Q)] \equiv [(\neg P) \vee (\neg Q)] \tag{3.3}$$

$$[\neg(P \vee Q)] \equiv [(\neg P) \wedge (\neg Q)] \tag{3.4}$$

Statements (3.3) and (3.4) are known as de Morgan's laws. (How are they related to Exercise 1.2?)

With two possible exceptions, once you carefully study what these connectives mean, you should understand them intuitively. One exception is that the logical and mathematical "or", $\vee$, is inclusive. We discussed this at the beginning of Chapter 2. The other exception is the logical connective "$\Rightarrow$".

3.2.2. Implication. Students often find it confusing that the implication $P \Rightarrow Q$ can be true when the consequence, Q, is false. This is understandable when we consider that implications are usually employed in argument in the following syllogism:

$$P$$
$$P \Rightarrow Q$$

therefore,

$$Q$$

(*i.e.* if P is true, and $P \Rightarrow Q$, then Q is true). This syllogism is the most important rule of logical deduction (called *Modus Ponens*). Logical implication is so often used to demonstrate the truth of the consequence that it is easy to understand why one might mistakenly think that the consequence must follow from the implication, rather than following from the *antecedent*. Consider the following statement:

If you are the king of France, then I am a monkey's uncle.

Is this statement true? Presumably you are not the king of France, and I don't believe that I am a monkey's uncle. So both the antecedent and the consequence are false. However the statement is true. In fact, this statement is logically equivalent to the statement:

If I am not a monkey's uncle, then you are not the king of France.

The definition of logical implication says that an implication in which the antecedent is false gives no information about the consequence. Hence, any logical implication with the antecedent "You are the king of France" will be true.

There is an additional concern with logical implication. In natural language (and intuitively in mathematics), the statement

$$P \Rightarrow Q$$

suggests a relationship between the statements P and Q — namely that the truth of P somehow forces the truth of Q. As a propositional connective, this relationship between P and Q is not required for logical implication. The truth of $P \Rightarrow Q$ is a function of the truth values of P and Q, not their *meanings*. In mathematical writing, it is understood that not only is the implication logically true, but that P and Q are related and that the truth of P indeed forces the truth of Q. For instance, consider the statement

$$\mathbb{N} \subset \mathbb{Q} \Rightarrow 3 > 2.$$

This statement is true by the formal definition of $\Rightarrow$. In fact, as a propositional statement, we could replace the antecedent with any other statement, true or false, and the conditional statement would

be true. However, such a statement is mathematically unacceptable, since the antecedent and the consequence have nothing to do with each other. We are not concerned with the accidental truth values of atomic statements, but the mathematical connections between these statements, which comply with, yet go beyond, the formal definition of logical connectives.

3.2.3. Converse and Contrapositive. Most mathematical claims have the form of an implication. Therefore you need to be familiar with the conventional nomenclature surrounding logical implication. Suppose we are interested in a particular logical implication,

$$P \Rightarrow Q.$$

There are two other logical implications which are naturally associated with $P \Rightarrow Q$. One is the contrapositive,

$$\neg Q \Rightarrow \neg P.$$

An implication and its contrapositive are propositionally equivalent.

Example 3.5. The statement,

"If this is an insect then it has six legs."

is propositionally equivalent to the statement

"If this does not have six legs, it is not an insect."

Example 3.6. The contrapositive of

"A whale is a fish"

is

"If it is not a fish then it is not a whale".

The latter example illustrates that a statement need not be true in order to have a contrapositive (which is, of course, still propositionally equivalent to the original conditional statement). It also illustrates that conditional statements in natural language need not include the word "if" or "then", nor be written in a particular form, in order to be a conditional statement.

The converse of a conditional statement,

$$P \Rightarrow Q$$

is the conditional statement,

$$Q \Rightarrow P.$$

A conditional statement and its converse are not propositionally equivalent. You can easily check that $P \Rightarrow Q$ and $Q \Rightarrow P$ have different truth values if $T(P) = 1$ and $T(Q) = 0$.

EXAMPLE 3.7. What is the converse to the statement

"All fish live in water"?

Since this is written in natural language, there is no unique answer. An obvious converse is

"If something lives in water, then it is a fish".

If we put together an implication and its converse, we get the biconditional connective.

DEFINITION. Biconditional, $\Longleftrightarrow$ Let P and Q be statements. The biconditional, written $\Longleftrightarrow$, is defined as follows.

Connective	Name	Definition
$\Longleftrightarrow$	biconditional	$T(P \iff Q) = T(P \Rightarrow Q) \cdot T(Q \Rightarrow P)$

The biconditional connective is the formal interpretation of "if and only if". This phrase is so commonly used in mathematics that it has its own abbreviation: iff.

Other natural language words that can be translated into propositional connectives are "necessary" and "sufficient". The statement

"In order for P to hold, it is necessary that Q holds"

is equivalent to $P \Rightarrow Q$. The statement

"In order for P to hold, it is sufficient that Q holds"

is equivalent to $Q \Rightarrow P$. Combining these two, we get that the statement

"In order for P to hold, it is necessary and sufficient that Q holds" is equivalent to $P \iff Q$.

3.3. Formulas

Loosely speaking, a formula is a mathematical expression with variables. Corresponding to each variable, x_i, appearing in a formula is a universe, U_i, from which that variable may be substituted.

DEFINITION. Open formula An open mathematical formula in variables x_1,...,x_n is a mathematical expression in which substitution of the x_i $(1 \leq i \leq n)$ by specific elements from U_i yields a mathematical statement.

EXAMPLE 3.8. Consider the formula,

$$x^2 + y^2 = z^2$$

in variables x, y and z, all with universe $\mathbb{N}$. Any substitution of the variables with natural numbers results in a statement. For instance,

$$3^2 + 4^2 = 5^2$$

or

$$1^2 + 1^2 = 2^2.$$

Of course, statements can be true or false, so some substitutions yield true statements, while others will yield false statements.

In discussing a general formula in n variables, we may use the notation $P(x_1, \ldots, x_n)$. For $1 \leq i \leq n$, let U_i be the universe of the variable x_i, and $a_i \in U_i$. The statement that results from the substitution of a_i for x_i, $1 \leq i \leq n$, is written $P(a_1, \ldots, a_n)$.

If $P(x_1, \ldots, x_n)$ is a formula in variables $x_1, \ldots, x_n$, and for $1 \leq i \leq n$, U_i is the universe of x_i, then we may think of $(x_1, \ldots, x_n)$ as a single variable with universe $U = \prod_{1 \leq i \leq n} U_i$.

Formulas can fulfill many purposes in mathematics:

(1) Characterize relationships between quantities
(2) Define computations

(3) Define sets
(4) Define functions.

EXAMPLE 3.9. Consider an open formula, $P(x, y)$, in two variables,

$$x^2 + y^2 \;=\; 1,$$

with universe $\mathbb{R}^2$. That is, the universe of x is $\mathbb{R}$ and the universe of y is $\mathbb{R}$. One way to think of $P(x, y)$ is as a means to partition $\mathbb{R}^2$ into two sets:

(1) the subset of the Cartesian Plane for which the equation is true, namely the unit circle;
(2) the subset of the Cartesian Plane for which the equation is false, the complement of the unit circle in $\mathbb{R}^2$.

DEFINITION. Characteristic set, χ_P Let $P(x)$ be a formula, and U the universe of the variable x. The subset of U for which the formula P holds is written χ_P. The set χ_P is called the characteristic set of $P(x)$.

So,

$$\chi_{\neg P} \;=\; U \setminus \chi_P.$$

3.3.1. Formulas and Propositional Connectives. Propositional logic is easily extended to formulas. Let $P(x)$ and $Q(x)$ be formulas in the variable x, with universe U. Let

$$R(x) = P(x) \wedge Q(x).$$

Then the characteristic set of $R(x)$ is given by

$$\chi_R = \{a \in U \mid T(\, P(a) \wedge Q(a)\,) \;=\; 1\,\}$$

Hence

$$\chi_R = \chi_P \cap \chi_Q.$$

The propositional connective $\wedge$ is strongly associated with the set operation $\cap$. Similarly $\vee$ may be associated with $\cup$, $\neg$ with complement (in U), and $\Rightarrow$ with $\subseteq$.

3.4. Quantifiers

Let $P(x)$ be a formula in one variable. If we substitute a constant, $a \in U$, for x we arrive at a statement $P(a)$. However, suppose that we are interested in $P(x)$ with regard to some set $X \subseteq U$, rather than a particular element of U. In particular, we ask if $P(a)$ is a true statement for all $a \in X$. Recall that one of the roles of a formula is to define sets. For any formula $P(x)$, universe U and $X \subseteq U$, $P(x)$ partitions X into two sets — those elements of X for which P is true, and those for which P is false. In this sense, asking whether P holds for all $x \in X$, or whether it holds for some $x \in X$ (which is complementary to asking whether $\neg P$ holds for all $x \in X$) is asking whether P defines a new or interesting subset of X.

Just as propositional connectives were introduced to formalize the linguistic behavior of certain widely employed natural language connectives (and, or, implies, not), we shall also formalize "quantification" over sets.

DEFINITION. Universal quantifier, $(\forall x \in X)\ P(x)$ Let $P(x)$ be a formula in one variable, with universe U. Let $X \subseteq U$. Let Q be the statement

$$(\forall x \in X)P(x).$$

Then Q is true if for every $a \in X$, $P(a)$ is true. Otherwise Q is false.

The notation

$$(\forall x \in X)\ P(x)$$

is a shorthand for

$$(\forall x)\ ([x \in X] \Rightarrow [P(x)]).$$

The statement "$(\forall x \in X)\ P(x)$" is read "for all x in X, $P(x)$".
We have

$$(\forall x \in X)\ P(x) \iff X \subseteq \chi_P.$$

DEFINITION. Existential quantifier, $(\exists x \in X)\ P(x)$ Let $P(x)$ be a formula in one variable with universe U. Let $X \subseteq U$, $X \neq \emptyset$. Let Q

be the statement

$$(\exists x \in X)\ P(x).$$

Then Q is true if there is some $a \in X$, for which $P(a)$ is true. Otherwise Q is false.

The expression

$$(\exists x \in X)\ P(x)$$

is a shorthand for

$$(\exists x)\ [(x \in X) \wedge P(x)].$$

The statement "$(\exists x \in X)\ P(x)$" is read "there exists x in X, such that $P(x)$". The quantifier "$\forall$" is the formal equivalent of the natural language expression "for all" or "every". The quantifier "$\exists$" is the formal equivalent of "for some" or "there exists ... such that ...".

Provided that the universe of a variable is clear, or not relevant to the discussion, it is common to suppress the universe in the expression of the statement. For instance, if $P(x)$ is a formula with universe U, we may write

$$(\forall x)\ P(x)$$

instead of

$$(\forall x \in U)\ P(x).$$

3.4.1. Multiple Quantifiers. Let $P(x_1, \ldots, x_n)$ be a formula in $n \geq 2$ variables. Then the formula

$$(\forall x_1)\ P(x_1, \ldots, x_n)$$

is a formula in the $n-1$ variables $x_2, \ldots, x_n$. Similarly, the formula

$$(\exists x_1)\ P(x_1, \ldots, x_n)$$

is a formula in $n-1$ variables.

Example 3.10. Consider the formula in five variables

$$P(x, x_0, L, \varepsilon, \delta)\ :=\ (0 <\mid x - x_0 \mid < \delta) \Rightarrow (\mid \sin(x) - L \mid < \varepsilon)$$

with all variables having universe $\mathbb{R}$.

Then $(\forall x_0)P(x, x_0, L, \varepsilon, \delta)$ is a formula in four variables, $(\forall x_0)(\exists L)P(x, x_0, L, \varepsilon, \delta)$ is a formula in three variables, and

$$(\forall x_0)(\exists L)(\forall \varepsilon)P(x, x_0, L, \varepsilon, \delta)$$

is a formula in two variables.

DEFINITION. Open variable, Bound variable In the formula $P(x)$, x is an open variable. In the formulas

$$(\forall x)\ P(x), \quad (\exists x)\ P(x), \quad (\forall x)\ Q(x, y), \quad (\exists x)\ Q(x, y)$$

x is a bound or quantified variable, and in the last two, y is an open variable.

3.4.2. Quantifier Order. In the discussion below, we need to discuss quantifiers generically, that is without regard to whether the quantifier under discussion is universal or existential. So we shall introduce some convenient notation just for this section.

NOTATION. $(\mathcal{Q}x)\ P(x)$ *We use the notation*

$$(\mathcal{Q}x)\ P(x)$$

to generically represent

$$(\forall x)\ P(x)$$

and

$$(\exists x)\ P(x).$$

Let $\mathcal{Q}_1, ..., \mathcal{Q}_n$ be logical quantifiers and $P(x_1, ..., x_n)$ be a formula with open variables $x_1, ..., x_n$. Then

$$(\mathcal{Q}_1 x_1)(\mathcal{Q}_2 x_2)(...)(\mathcal{Q}_n x_n)\ P(x_1, \ldots, x_n)$$

is a statement.

EXAMPLE 3.11. Consider a statement S in the form

$$S = (\forall x \in X)\ (\exists y \in Y)\ P(x, y).$$

S is true if for each $a \in X$,

$$(\exists y \in Y)\ P(a, y)$$

is true. This is satisfied provided that for each $a \in X$, there is an element of Y (let's call it b_a to remind us that this particular element of Y is associated with the previous choice, a) such that

$$P(a, b_a)$$

is true. So b_a is selected with a in mind. Statements in this form are especially important in mathematics because the definition of the limit in calculus is a statement in the form of this example.

Let's return to the statement

$$(\mathcal{Q}_1 x_1)(\ldots)(\mathcal{Q}_n x_n)\ P(x_1, \ldots, x_n).$$

The order of the quantifiers is significant. If $1 \leq i < j \leq n$, x_i behaves like a parameter from the point of view of x_j (that is, x_i is fixed from the point of view of x_j). Put another way, x_j is chosen with respect to the substitutions of $x_1, \ldots, x_{j-1}$, but without consideration for $x_{j+1}, \ldots, x_n$.

One always reads from the left. The statement

$$(\forall x_1)(\mathcal{Q}_2 x_2)(\ldots)(\mathcal{Q}_n x_n)\ P(x_1, \ldots, x_n)$$

is the same as

$$(\forall x_1)\ [\,(\mathcal{Q}_2 x_2)(\ldots)(\mathcal{Q}_n x_n)\ P(x_1, \ldots, x_n)\,],$$

or, in other words, for every choice of x_1, the statement

$$(\mathcal{Q}_2 x_2)(\ldots)(\mathcal{Q}_n x_n)\ P(x_1, \ldots, x_n)$$

is true. Similarly, the statement

$$(\exists x_1)(\mathcal{Q}_2 x_2)(\ldots)(\mathcal{Q}_n x_n)\ P(x_1, \ldots, x_n)$$

is the same as

$$(\exists x_1)\ [\,(\mathcal{Q}_2 x_2)(\ldots)(\mathcal{Q}_n x_n)\ P(x_1, \ldots, x_n)\,],$$

or in other words that there is some choice of x_1 for which the statement

$$(\mathcal{Q}_2 x_2)(\ldots)(\mathcal{Q}_n x_n)\ P(x_1, \ldots, x_n)$$

about the $n-1$ variables $x_2, \ldots, x_n$ is true.

EXAMPLE 3.12. Order of quantifiers is important, as you can see from the following:

$$(\forall x \in X)\ (\exists y \in Y)\ P(x, y)$$

is not equivalent to

$$(\exists y \in Y)\ (\forall x \in X)\ P(x, y).$$

For instance, the statement

$$(\forall x \in \mathbb{R})\ (\exists y \in \mathbb{R})\ (y\ =\ x^2)$$

is true. But

$$(\exists y \in \mathbb{R})\ (\forall x \in \mathbb{R})\ (y\ =\ x^2)$$

is false. The statement

$$[(\exists y \in Y)\ (\forall x \in X)\ P(x, y)] \Rightarrow [(\forall x \in X)\ (\exists y \in Y)\ P(x, y)]$$

is true. The converse clearly fails.

3.4.3. Negation of Quantifiers. In an important sense, $\wedge$ and $\vee$ are complementary. By de Morgan's identities (3.3) and (3.4), the negation of a simple conjunction is a disjunction of negations. Similarly, the negation of a simple disjunction is a conjunction of negations. Universal and existential quantifiers are also complementary. We observe that

$$[\neg(\forall x)\ P(x)] \equiv [(\exists x)\ \neg P(x)]$$

for any formula, $P(x)$. Similarly

$$[\neg(\exists x)\ P(x)] \equiv [(\forall x)\ \neg P(x)].$$

Of course, $P(x)$ itself may be a formula which has numerous quantifiers and bound variables. Let's suppose that

$$P(x)\ =\ (\exists y)\ Q(x, y). \tag{3.13}$$

Then the following statements are equivalent (for any choice of P and Q satisfying the identity (3.13)):

$$\neg(\forall x)\ P(x)$$

$$(\exists x)\ \neg P(x)$$

$$\neg(\forall x)\ (\exists y)\ Q(x,y)$$

$$(\exists x)\ \neg(\exists y)\ Q(x,y)$$

$$(\exists x)\ (\forall y)\ \neg Q(x,y).$$

This example suggests that it is permissible to permute a negation and a quantifier by changing the type of quantifier, and indeed this is so.

Let $\mathcal{Q}_i$ be a quantifier, for $1 \leq i \leq n$. For each $\mathcal{Q}_i$, let $\mathcal{Q}_i^*$ be the complementary quantifier. That is, if $\mathcal{Q}_i = \forall$, then let $\mathcal{Q}_i^* = \exists$; if $\mathcal{Q}_i = \exists$, then let $\mathcal{Q}_i^* = \forall$. Then,

$$\neg(\mathcal{Q}_1 x_1)(...)(\mathcal{Q}_n x_n)\ P(\bar{x}) \equiv (\mathcal{Q}_1^* x_1)(...)(\mathcal{Q}_n^* x_n)\ \neg P(\bar{x}).$$

3.5. Proof Strategies

There are two elementary logical forms that occur so commonly in mathematical claims that they warrant some general discussion.

3.5.1. Universal Statements. A logical form you are likely to encounter very often is

$$(\forall x)\ [H(x) \Rightarrow P(x)], \tag{3.14}$$

where $H(x)$ and $P(x)$ are formulas in one variable. Statements in this form are called universal statements. The formulas H and P are used to characterize properties of mathematical objects, so that the claims in this form may be thought of as stating:

If a mathematical object has property H, then it has property P as well.

This is particularly useful if we know a great deal about mathematical objects that have property P. Because the statement we are

endeavoring to prove is universal, examples do not suffice to prove such claims — the example you cite might accidentally have properties H and P. Rather, universal claims must be proved abstractly, arguing that satisfying a definition or set of properties implies the satisfaction of other properties. This generally requires carefully evaluating definitions. In practice, we often do this by assuming that we have an arbitrary element that satisfies a definition or explicit assumptions, and logically derive additional conclusions about this object. By arbitrary we mean that we are not allowed to make any claims about the element except those that follow immediately from definitions, explicit assumptions, or are logically derived from definitions and explicit assumptions. Since the object was arbitrary (except for the explicit assumptions you make at the outset of the argument), the conclusions you derive concerning the object will be true universally of all objects which satisfy the assumptions.

EXAMPLE 3.15. Suppose $F(x)$ is the formula:

"$x \in \mathbb{N}$ and x is a multiple of 4."

Let $E(x)$ be the formula:

"x is even."

Then

$$(\forall x)\ [F(x) \Rightarrow E(x)]. \tag{3.16}$$

It does not suffice to observe that 4, 8 and 12 are all even. In order to argue for the statement directly, you would argue abstractly that any object which satisfies $F(x)$ necessarily satisfies $E(x)$.

There are a couple of approaches that one commonly considers when proving conditional statements. Choosing an approach is choosing a strategy for the proof. Normally, more than one strategy can be made to work, but often one may be simpler than the others.

Claims of the form (3.14) are generally approached in one of the following ways:

(1) Direct Proof.

Let x be an object for which H holds. By decoding the property H, you might be able to show directly that P holds of x as well. Since x was an arbitrary object satisfying P, the universal claim will be proved.

EXAMPLE 3.17. Prove (3.16) directly.

Let $x \in \mathbb{N}$ (we treat x as a fixed but arbitrary element of the natural numbers). If $x = 4n$, then

$$x = 2 \cdot (2n),$$

and is therefore even.

EXAMPLE 3.18. Prove that any 3 points in the plane are either collinear or lie on a circle.

PROOF. Label the points A, B, C. Let L be the perpendicular bisector of AB. Every point on L is equidistant from A and B.

Let M be the perpendicular bisector of BC. Every point on M is equidistant from B and C.

If A, B and C are not collinear, the lines L and M are not parallel, so they intersect at some point D. The point D is equidistant from A, B and C, so these points lie on a circle centered at D. □

EXAMPLE 3.19. Pythagoras's theorem can be stated in the form (3.14). (What are H and P in this case?) Euclid's proof of Pythagoras's theorem is a direct proof (Euclid's Elements I.47).

(2) Contrapositive Proof.

It is sometimes easier to show that the failure of P implies the failure of H. Assume you have an object for which P fails (that is assume $\neg P$ holds of the object). Derive that H must fail for the object as well. In this case you will have demonstrated that

$$(\forall x)\ [\neg P(x) \Rightarrow \neg H(x)].$$

This is equivalent to the claim

$$(\forall x)\ [H(x) \Rightarrow P(x)].$$

EXAMPLE 3.20. Prove (3.16) by proving the contrapositive.

Let $x \in \mathbb{N}$, and assume $\neg E(x)$, so x is odd. As x is odd, then x divided by 4 has remainder 1 or 3. Then,

$$x \neq 4n.$$

So x is not a multiple of 4.

EXAMPLE 3.21. Prove that if x is an integer and x^2 is even, then x is even.

The contrapositive is the assertion that if x is an odd integer, then x^2 is odd. We shall prove this.

Suppose x is odd, so $x = 2n + 1$ for some integer n. Then $x^2 = 4n^2 + 4n + 1$, so $x^2 \equiv 1 \mod 2$, and x^2 is therefore odd.

(3) Contradiction.

This is a proof in which we show that $H \wedge \neg P$ is necessarily false. That is, assume that H holds for an arbitrary object and P fails for that object, and show that this gives rise to a contradiction. Since contradictions are logically impossible, it is logically necessary that

$$\neg(H \wedge \neg P)$$

which is propositionally equivalent to

$$\neg H \vee P$$

or, alternatively,

$$H \Rightarrow P.$$

Since we shall have shown that for any substitution of x, the statement $H \Rightarrow P$ holds, we shall have shown the universal claim.

EXAMPLE 3.22. Prove (3.16) by contradiction.

Assume that x is a multiple of 4 and that x is odd. Let r be the residue of x modulo 2. Since x is a multiple of $4 = 2 \cdot 2$, we have that $r \equiv 0 \mod 2$. Since r is odd, we have that $r \equiv 1 \mod 2$. This implies $0 \equiv 1 \mod 2$, a contradiction. Therefore the assumption that there was an x that was both a multiple of 4 and odd is false, and so ((3.16) must be true.

EXAMPLE 3.23. Prove that $\sqrt{2}$ is irrational.

PROOF. We restate this as an implication: If a number is rational, it's square cannot equal 2. We begin by considering the logical structure of the claim. Here the hypothesis $H(x)$ is that x is a rational number, and the conclusion $P(x)$ is that $x^2 \neq 2$. We wish to prove

$$(\forall x)\ H(x) \Rightarrow P(x).$$

We shall give a proof by contradiction. That is, we assume the statement is false and derive a contradiction. So we assume

$$\neg((\forall x)\ H(x) \Rightarrow P(x)).$$

This is logically equivalent to

$$(\exists x)H(x) \wedge \neg(P(x)).$$

Let's go back to mathematical prose now that we have fought through the logic. Assume that x is a rational number, and assume also that $x^2 = 2$; we wish to derive a logical contradiction. Write $x = m/n$, where m and n are non-zero integers that have no common factors. Then

$$x^2 \ = \ m^2/n^2 \ = \ 2,$$

so $m^2 = 2n^2$. Therefore m^2 is even, so by Example 3.21, m is even. Therefore $m = 2k$ for some integer k, and so

$$m^2 \ = \ 4k^2 \ = \ 2n^2.$$

Therefore $n^2 = 2k^2$ is even, so n is even. But then both m and n are even, and so have 2 as a common factor, which contradicts the assumption that m/n was the reduced form of the rational number x. □

Contrapositive proofs and proofs by contradiction are very similar. Indeed, any contrapositive proof, that $\neg P \Rightarrow \neg H$, automatically yields that $(H \wedge \neg P)$ is impossible. The distinction is more linguistic than logical. The reason for having names for different proof strategies is to provide guidance to the reader in order to make the proof easier to follow.

In Chapter 4 we shall see another powerful method for proving universal statements over $\mathbb{N}$, namely the Principle of Induction.

3.5.2. Existence Proofs. A second common form for a mathematical claim is an existential statement, that is, a statement in the form

$$(\exists x)\ P(x). \tag{3.24}$$

There are three common approaches to proving existential statements.

(1) Construction.

Obviously, the most direct way to show that something exists with certain properties is to introduce or construct an object with property P. For claims in this form, the example is the proof, although you will need to show that the object satisfies P, if it is not obvious.

EXAMPLE 3.25. Prove that there exists a real function whose first derivative is everywhere positive, and whose second derivative is everywhere negative.

PROOF. The easiest way to do this is to write down a function with these properties. One such function is $f(x) = 1 - e^{-x}$. The derivative is e^{-x}, which is everywhere positive, and the second derivative is $-e^{-x}$, which is everywhere negative.

(2) Counting.

Sometimes one can establish an object's existence by a counting argument.

EXAMPLE 3.26. Suppose there are 30 students in a class. Show that at least two of them share the same last initial.

PROOF. For each letter A,B,... group all the students with that letter as their last initial. As there are only 26 groups and $30 > 26$ students, at least one group must have more than on student in it. □

The argument we just gave is called the "pigeon-hole principle", based on the analogy of putting letters into pigeon-holes. If there are more letters than pigeon-holes, then some pigeon-hole must have more

than one letter. Notice that unlike a constructive proof, a counting proof does not tell you which group has more than one element in it.

For Cantor's spectacular generalization of the pigeon-hole principle to infinite sets, see Chapter 6.

(3) Contradiction.

It can be difficult to prove existential statements by construction. An alternative is to assume that the existential statement is false (that there is no object which satisfies $P(x)$). If it is impossible that no object has property P, then some object must. Again, this approach may not give us much insight into the objects that have property P. See *e.g.* Exercise 3.27.

EXAMPLE 3.27. Suppose all the points in the plane are colored either red or blue. Prove that there must be two points of the same color exactly one unit apart.

PROOF. Assume there are not. Draw an equilateral triangle of side 1. Label its vertices A, B and C. Then A and B must be different colors, B and C must be different colors, and C and A must be different colors. This is impossible with only two colors to choose from.

Notice that we have not said whether there is a red-red pair that is unit distance apart, or a blue-blue pair that is unit distance apart, just that one such pair must exist.

3.6. Exercises

EXERCISE 3.1. Prove de Morgan's laws, (3.3) and (3.4). (Hint: There are four possible assignments of truth values 0 and 1 to the two statements P and Q. For each such assignment, evaluate the truth values of the left-hand and right-hand sides of (3.3) and show they are always the same.)

EXERCISE 3.2. Prove that compound statements P and Q are propositionally equivalent iff $P \iff Q$.

EXERCISE 3.3. Give an example of a true conditional statement in which the consequence is false.

EXERCISE 3.4. If P, Q and R are statements, prove that the following are true:
a) $P \wedge \neg P \Rightarrow Q$
b) $[(P \Rightarrow Q) \wedge (Q \Rightarrow R)] \Rightarrow (P \Rightarrow R)$
c) $[P \Rightarrow (Q \wedge \neg Q)] \Rightarrow \neg P$
d) $[P \wedge (P \Rightarrow Q)] \Rightarrow Q$
e) $P \Rightarrow (Q \vee \neg Q)$.

EXERCISE 3.5. Let P and Q be statements. Prove that there are statements using only P, Q, $\neg$ and $\wedge$ that are propositionally equivalent to
a) $P \wedge Q$
b) $P \vee Q$
c) $P \Rightarrow Q$.

Prove that there are statements using only P, Q, $\neg$ and $\vee$ that are equivalent to the above.

EXERCISE 3.6. Prove the distributive laws for propositional logic: If P, Q and R are statements, then
a) $P \vee (Q \wedge R) \equiv (P \vee Q) \wedge (P \vee R)$
b) $P \wedge (Q \vee R) \equiv (P \wedge Q) \vee (P \wedge R)$.

EXERCISE 3.7. Prove the distributive law for sets: If X, Y and Z are sets, then
a) $X \cup (Y \cap Z) \;=\; (X \cup Y) \cap (X \cup Z)$
b) $X \cap (Y \cup Z) \;=\; (X \cap Y) \cup (X \cap Z)$.

EXERCISE 3.8. Let sets X, Y and Z be characteristic sets of formulas $P(x)$, $Q(x)$ and $R(x)$ respectively. For each possible region of the Venn diagram of X, Y and Z give a compound formula (with atomic formulas P, Q and R) that has that region as its characteristic set.

EXERCISE 3.9. Write a formula in one variable that defines the even integers.

EXERCISE 3.10. Write a formula that defines perfect squares.

EXERCISE 3.11. Write a formula in two variables that defines the points in $\mathbb{R}^2$ that have distance 1 from the point (π, e).

EXERCISE 3.12. Can you write a formula in one variable using only addition, multiplication, exponentiation, integers and equality, to define the set of all roots of a given polynomial with integer coefficients? How about the set of roots of all polynomials with integer coefficients?

EXERCISE 3.13. Which of the following statements are true?
a) $(\forall x \in \mathbb{R})\ x + 1 > x$
b) $(\forall x \in \mathbb{Z})\ x^2 > x$
c) $(\exists x \in \mathbb{Z})(\forall y \in \mathbb{Z})\ x \leq y$
d) $(\forall y \in \mathbb{Z})(\exists x \in \mathbb{Z})\ x \leq y$
e) $(\forall \varepsilon > 0)(\exists \delta > 0)(\forall x \in \mathbb{R})\ [0 <\mid x - 1 \mid< \delta] \Rightarrow [\mid x^2 - 1 \mid< \varepsilon]$.

EXERCISE 3.14. What is the negation of each statement in Exercise 3.13? Which of the negations are true?

EXERCISE 3.15. Let $a, L \in \mathbb{R}$ and f be a real function. Prove that the statements

$$(\forall \varepsilon > 0)(\exists \delta > 0)(\forall x \in \mathrm{Dom}(f))\ [0 <\mid x - a \mid< \delta] \Rightarrow [\mid f(x) - L \mid< \varepsilon]$$

and

$$(\exists \delta > 0)(\forall \varepsilon > 0)(\forall x \in \mathrm{Dom}(f))\ [0 <\mid x - a \mid< \delta] \Rightarrow [\mid f(x) - L \mid< \varepsilon]$$

are not equivalent. Which statement is a consequence of the other?

EXERCISE 3.16. Let $P(x, y)$ be a formula in two variables. Show that in general $(\forall x)(\exists y)\ P(x, y)$ need not be equivalent to $(\exists y)(\forall x)\ P(x, y)$. Show that $(\forall x)(\forall y)\ P(x, y)$ is equivalent to $(\forall y)(\forall x)\ P(x, y)$. What about $(\exists x)(\exists y)\ P(x, y)$ and $(\exists y)(\exists x)\ P(x, y)$?

EXERCISE 3.17. Consider the following statements. Write down the contrapositive and the converse to each one.

(i) All men are mortal.

(ii) I mean what I say.

(iii) Every continuous function on the interval $[0, 1]$ attains its maximum.

(iv) The sum of the angles of a triangle is 180^o.

EXERCISE 3.18. Prove that a number is divisible by 4 if and only if its last two digits are.

EXERCISE 3.19. Prove that a number is divisible by 8 iff its last three digits are.

EXERCISE 3.20. Prove that a number is divisible by 2^n iff its last n digits are.

EXERCISE 3.21. Suppose m is a number with the property that any natural number is divisible by m iff its last three digits are. What does this say about m? Prove your assertion.

EXERCISE 3.22. Prove that an integer is divisible by 11 iff the sum of the oddly placed digits minus the sum of the evenly placed digits is divisible by 11. (So $11 \,|\, 823493$ iff 11 divides $(2+4+3)-(8+3+9)$.)

EXERCISE 3.23. Show that every interval contains rational and irrational numbers.

EXERCISE 3.24. Prove that $\sqrt{3}$ is irrational.

EXERCISE 3.25. Prove that $\sqrt{10}$ is irrational.

EXERCISE 3.26. Prove that the square root of any natural number is either an integer or irrational.

EXERCISE 3.27. Prove that there exist irrational numbers x and y so that x^y is rational. (Hint: consider $\sqrt{2}^{\sqrt{2}}$ and $\left(\sqrt{2}^{\sqrt{2}}\right)^{\sqrt{2}}$.)

EXERCISE 3.28. Prove or disprove the following assertion: Any 4 points in the plane, no three of which are collinear, lie on a circle.

EXERCISE 3.29. Prove that there are an infinite number of primes.

EXERCISE 3.30. For $k = 0, 1, 2$, let P_k be the set of prime numbers that are congruent to k mod 3. By Exercise 3.29, $P_0 \cup P_1 \cup P_2$ is infinite. Can you say which of the sets P_0, P_1 and P_2 are infinite?

(Remark: For two of the three sets, this problem is not too difficult. For the third one, it is extremely difficult, and is a special case of a celebrated theorem of Dirichlet. See *e.g.* [**8**] for a treatment of Dirichlet's theorem.)

EXERCISE 3.31. Let the points in $\mathbb{R}^2$ be colored red, green and blue. Prove that either there are two points of the same color a distance 1 apart, or there is an equilateral triangle of side length $\sqrt{3}$ all of whose vertices are the same color.

EXERCISE 3.32. Prove that

$$e \;=\; \sum_{n=0}^{\infty} \frac{1}{n!}$$

is irrational. (Hint: Argue by contradiction. Assume $e = \frac{p}{q}$, and multiply both sides by $q!$. Rearrange the equation to get an integer equal to an infinite sum of rational numbers that converges to a number in $(0, 1)$.)

CHAPTER 4

Principle of Induction

4.1. Well-orderings

In this chapter we discuss the principle of mathematical induction. Be aware that the word induction has a different meaning in mathematics than in the rest of science. The principle of mathematical induction depends on the order structure of the natural numbers, and gives us a powerful technique for proving universal mathematical claims.

DEFINITION. Well-ordering Let X be a set, and $\preceq$ a linear ordering on X. We say that X is well-ordered with respect to $\preceq$ (or $\preceq$ is a well-ordering of X) if every non-empty subset of X has a least element with respect to $\preceq$. That is, for any non-empty subset Y of X

$$(\exists\, a \in Y)(\forall\, y \in Y)\ a \preceq y.$$

In general, linear orderings need not be well-orderings. Well-ordering is a universal property — a set X with an ordering $\preceq$ is well-ordered if *every* non-empty subset of X has a least element with respect to $\preceq$. If there is any non-empty subset which does not have a least element, then $\preceq$ does not well-order X.

EXAMPLE 4.1. $\mathbb{Z}$ is not well-ordered by $\leq$. The integers do not have a least element, which suffices to demonstrate that $\mathbb{Z}$ is not well-ordered by $\leq$.

EXAMPLE 4.2. Let $X = \{x \in \mathbb{R} \mid x \geq 2\}$. Let $\leq$ be the usual ordering on $\mathbb{R}$. X is linearly ordered by $\leq$, but X is not well-ordered by $\leq$. In this example, X has a least element, but any open interval contained in X will fail to have a least element.

The key order properties of $\mathbb{N}$ are that it is well-ordered and every element of $\mathbb{N}$, except 0, is the successor of a natural number:

WELL-ORDERING PRINCIPLE FOR THE NATURAL NUMBERS: *The set* $\mathbb{N}$ *is well-ordered by* $\leq$.

SUCCESSOR PROPERTY FOR THE NATURAL NUMBERS: *If* $n \in \mathbb{N}$ *and* $n \neq 0$, *then there is* $m \in \mathbb{N}$ *such that* $n = m + 1$.

If one accepts an intuitive understanding of the natural numbers, these principles are more or less obvious. Indeed, let Y be any non-empty subset of $\mathbb{N}$. Since it is non-empty, there is some m in Y. Now, consider each of the finitely many numbers $0, 1, 2, \ldots, m$ in turn. If $0 \in Y$, then 0 is the least element. If 0 is not in Y, proceed to 1. If this is in Y, it must be the least element; otherwise proceed to 2. Continue in this way, and you will find some number less than or equal to m that is the least element of Y.

This argument, though convincing, does rely on the fact that we have an idea of what $\mathbb{N}$ "is". If we wish to define $\mathbb{N}$ in terms of set operations, as we do in Chapter 8, we essentially have to include the well-ordering principle for the natural numbers as an axiom.

4.2. Principle of Induction

We begin by proving a theorem that is equivalent to the principle of induction.

THEOREM 4.3. *If*

(1) $X \subseteq \mathbb{N}$
(2) $0 \in X$
(3) $(\forall\, n \in \mathbb{N})\ n \in X \Rightarrow (n + 1) \in X$,

then

$$X = \mathbb{N}.$$

DISCUSSION. *We shall argue by contradiction. We assume that* $X \neq \mathbb{N}$. *Let* Y *be the complement of* X *in* $\mathbb{N}$. *Since* Y *is non-empty,*

it will have a least element. The third hypothesis of the theorem will not permit a least element in Y, other than 0, and this is impossible by the second hypothesis. Therefore Y is necessarily empty.

PROOF. Let X satisfy the hypotheses of the theorem. Let

$$Y = \mathbb{N} \setminus X.$$

We assume Y is non-empty. Since $Y \subseteq \mathbb{N}$, Y is well-ordered by $\leq$. Let $a \in Y$ be the least element of Y. We note that a is not 0, since $0 \in X$. Therefore $a \geq 1$ and is a successor, so $a-1$ is in $\mathbb{N}$ and not in Y. Hence $a-1$ is in X. But then by hypothesis (3) of the theorem, $a-1+1 \in X$. This is a contradiction, therefore Y is empty and $X = \mathbb{N}$. □

REMARK. We will occasionally include informal, labelled discussions in our proofs in order to guide you in your reading. This is not a usual practice. You should not include such discussions in your proofs unless your instructor requests it.

.

Theorem 4.3 is more easily applied in the following form.

COROLLARY 4.4. *Principle of induction* *Let $P(x)$ be a formula in one variable. If*

(1) $P(0)$
(2) $(\forall\, x \in \mathbb{N})\ P(x) \Rightarrow P(x+1)$,

then

$$(\forall\, x \in \mathbb{N})\ P(x).$$

PROOF. Let

$$\chi_P = \{x \in \mathbb{N} \mid P(x)\}.$$

We wish to show that $\chi_P = \mathbb{N}$. By assumption (1), $P(0)$, so $0 \in \chi_P$. Assume that $n \in \chi_P$. Then $P(n)$. By assumption (2)

$$P(n) \Rightarrow P(n+1).$$

Therefore $P(n+1)$ and $n+1 \in \chi_P$. Since n is arbitrary,

$$(\forall\, n \in \mathbb{N})\ n \in \chi_P \Rightarrow n+1 \in \chi_P.$$

By Theorem 4.3, $\chi_P = \mathbb{N}$ and

$$(\forall x \in \mathbb{N})\ P(x).$$

□

Suppose that you wish to show that a formula $P(x)$ holds for all natural numbers. When arguing by induction, the author must show that the hypotheses for the theorem are satisfied. Typically, the author first proves that $P(0)$. This is called the base case of the proof by induction. It is very often an easy, even trivial, conclusion. Nonetheless, it is necessary to prove a base case in order to argue by induction (can you demonstrate this?). Having proved the base case, the author will then prove the second hypothesis, namely, that the claim being true for an arbitrary natural number implies that it is true at the successor of that natural number. This is the induction step. The induction step requires proving a conditional statement, which is often proved directly. It is important to understand that the author is not claiming that P holds at an arbitrary natural number, otherwise the argument would be circular and invalid. Rather, the author will demonstrate that if the result *were* true at an arbitrary natural number, then it *would be* true for the subsequent natural number. The assumption that P holds at a fixed and arbitrary natural number is called the induction hypothesis. If the author successfully proves the base case and the induction step, then the assumptions of Corollary 4.4 are satisfied, and P holds at all natural numbers.

PROPOSITION 4.5. *Let $N \in \mathbb{N}$. Then*

$$\sum_{n\,=\,0}^{N} n \;=\; \frac{N(N+1)}{2}.$$

DISCUSSION. *This is a good first example of a proof by induction. The argument is a straightforward application of the technique and the result is of historical and practical interest.*

We argue by induction on the upper index of the sum. That is, the formula we are proving for all natural numbers is

$$P(x): \quad \sum_{n\,=\,0}^{x} n \;=\; \frac{x(x+1)}{2}.$$

It is important to identify the quantity over which you are applying the principle of induction, but some authors who are writing an argument for readers who are familiar with induction may not explicitly state the formula.

We prove a base case, $N = 0$, that corresponds to the sum with the single term 0. We then argue the induction step. This is our first argument using the principle of induction. Pay close attention to the structure of this proof. You should strive to follow the conventions for proofs by induction that we establish in this book.

PROOF. Base case: $N = 0$.

DISCUSSION. *Note that the base case is the statement $P(0)$.*

Since

$$\sum_{n\,=\,0}^{0} n \;=\; 0 \;=\; \frac{(0)(1)}{2},$$

$P(0)$ holds.

Induction step:

DISCUSSION. *We prove the universal statement*

$$(\forall\, x \in \mathbb{N})\ P(x) \Rightarrow P(x+1).$$

by showing that for an arbitrary natural number N

$$P(N) \Rightarrow P(N+1).$$

Thus we reduce proving a universal statement to proving an abstract conditional statement. We prove the resulting conditional statement

directly. That is, we assume $P(N)$ and derive $P(N+1)$. We remind the reader that we are not claiming the result holds at N — that is, we do not claim $P(N)$. Rather, we are proving the conditional statement by assuming the antecedent, the induction hypothesis, and deriving the consequence. If you do not use the induction hypothesis, you are not arguing by induction. Of course, in the body of the argument this is transparent, without reference to the underlying logical principles.

Let $N \in \mathbb{N}$ and assume that

$$\sum_{n\,=\,0}^{N} n \;=\; \frac{N(N+1)}{2}.$$

Then

$$\begin{aligned}\sum_{n\,=\,0}^{N+1} n \quad &= \quad \left(\sum_{n\,=\,0}^{N} n\right) + N + 1 \\ &=_{IH} \quad \frac{N(N+1)}{2} + N + 1\end{aligned}$$

by the induction hypothesis.

Discussion. *It is a good habit, and a consideration for your reader, to identify when you are invoking the induction hypothesis. We will use the subscript $_{IH}$ to indicate where we invoke the induction hypothesis.*

So

$$\begin{aligned}\sum_{n\,=\,0}^{N+1} n \;&=\; \frac{N(N+1)}{2} + N + 1 \\ &=\; \frac{N(N+1)}{2} + \frac{2N+2}{2} \\ &=\; \frac{N^2+3N+2}{2} \\ &=\; \frac{(N+1)((N+1)+1)}{2}.\end{aligned}$$

Therefore,

$$(\forall\, N \in \mathbb{N})\; P(N) \Rightarrow P(N+1).$$

By the principle of induction, the proposition follows. □

PROPOSITION 4.6. *Let $N \in \mathbb{N}$. Then*

$$\sum_{n\,=\,0}^{N} n^2 \;=\; \frac{N(N+1)(2N+1)}{6}. \tag{4.7}$$

PROOF. The assertion $P(N)$ is that the equation (4.7) holds. The base case, $N = 0$, is obvious:

$$\sum_{n\,=\,0}^{0} n^2 \;=\; \frac{0(0+1)(2\cdot 0+1)}{6}.$$

Induction step:

Assume that $N \in \mathbb{N}$ and

$$\sum_{n\,=\,0}^{N} n^2 \;=\; \frac{N(N+1)(2N+1)}{6}.$$

We prove that

$$\sum_{n\,=\,0}^{N+1} n^2 \;=\; \frac{(N+1)(N+2)(2N+3)}{6}.$$

Indeed

$$\begin{aligned}
\sum_{n\,=\,0}^{N+1} n^2 \quad &= \quad \left(\sum_{n\,=\,0}^{N} n^2\right) + (N+1)^2 \\
&=_{IH} \quad \frac{N(N+1)(2N+1)}{6} + (N+1)^2. \\
&= \quad \frac{N(N+1)(2N+1)}{6} + (N+1)^2 \\
&= \quad \frac{2N^3+9N^2+13N+6}{6} \\
&= \quad \frac{(N+1)(N+2)(2(N+1)+1)}{6}.
\end{aligned}$$

The proposition follows from the principle of induction. □

DISCUSSION. *The proof of Proposition 4.6 is very similar to the proof of Proposition 4.5. You may wish to confirm the algebraic identities in the latter portion of the proof, since they are not obvious. Just*

enough detail is included to guide you through the proof of the implication. The author of a proof by induction will assume that you are comfortable with the technique, and thereby may provide less detail than you like.

REMARK. There is more to Propositions 4.5 and 4.6 than just the proofs. There are also the formulas. Indeed, one use of induction is that if you *guess* a formula, you can use induction to prove your formula is correct. See Exercises 4.12 and 4.16.

Why is a base case necessary? Consider the following argument for the false claim $\sum_{n=0}^{N} n < \frac{N(N+1)}{2}$. Let $N \in \mathbb{N}$ and assume $P(N)$, where $P(N)$ is the statement

$$\sum_{n=0}^{N} n < \frac{N(N+1)}{2}.$$

Then

$$\begin{aligned}
\sum_{n=0}^{N+1} n &= \left(\sum_{n=0}^{N} n\right) + N + 1 \\
&<_{IH} \frac{N(N+1)}{2} + N + 1 \\
&= \frac{N^2 + 3N + 2}{2} \\
&= \frac{(N+1)((N+1)+1)}{2}.
\end{aligned}$$

Hence,

$$(\forall N \in \mathbb{N})\ P(N) \Rightarrow P(N+1).$$

Of course the inequality $P(N)$ is easily demonstrated to be false. What went wrong? Without a base case, proving

$$(\forall N \in \mathbb{N})\ P(N) \Rightarrow P(N+1)$$

is not sufficient to prove $(\forall N \in \mathbb{N})\, P(N)$. If $P(0)$ were true, then $P(1)$ would be true, and if $P(1)$ were true, then $P(2)$ would be, and so on. Indeed, if we are able to prove $P(N)$ for any $N \in \mathbb{N}$, then we know

$P(M)$ for any natural number $M > N$. But the sequence of statements $\langle P(0), P(1), P(2), \ldots \rangle$ never gets started. $P(N)$ fails for all N.

Another way to think of induction is in terms of guarantees. Suppose you decide to buy a car. First you go to Honest Bob's. Bob guarantees that any car he sells will go at least one mile. You buy a car, drive it off the lot, and after 3 miles it breaks down and can't be fixed. You walk back angrily, but Bob won't give you your money back because the car lived up to the guarantee.

Then you cross the road to Honest John's. John guarantees that if he sells you a car, once it starts it will never stop. This sounds pretty good, so you buy a car, put the keys in the ignition, and ... nothing. The car won't start. John won't give you your money back either, because the car did not fail to do what he claimed.

Feeling desperate, you end up at Honest Stewart's. Stewart's cars come with two guarantees:

(1) The car will start and go at least one mile.

(2) No matter how far the car has gone, it can always be driven an extra mile.

You think this over, and eventually decide that the car will go for ever. Best of all, the lease is only \$1 a month for the first two months. You sign the lease, and drive home rather pleased with yourself.[1]

There are many handy generalizations of the principle of induction. The first we discuss is called strong induction. It is so-named because the induction hypothesis is stronger than the induction hypothesis in standard induction, and hence the induction step is sometimes easier to prove in an argument by strong induction.

COROLLARY 4.8. *Strong induction* *Let $P(x)$ be a formula such that*

[1]You are correct that the Principle of Induction guarantees that your car will drive forever. However, as your mother points out when you show her the lease, after the first two months your payment each month is the sum of your payments in the previous two months. How much will you be paying after 5 years?

(1) $P(0)$
(2) *For each* $n \in \mathbb{N}$*,*

$$[(\forall\, x < n)\, P(x)] \Rightarrow [P(n)].$$

then

$$(\forall\, x \in \mathbb{N})\ P(x).$$

Intuitively this is not very different from basic induction. You start at a base case, and once started you can continue through the remainder of the natural numbers. The distinction is just in the number of assumptions you use when when proving something by strong induction. In practice, it gives the advantage that in the induction step you can reduce case N to any previous case, rather than the immediately preceding case, $N - 1$. In particular this simplifies arguments about divisibility and integers.

Discussion. *We reduce the principle of strong induction to the principle of induction. We accomplish this by introducing a formula,* $Q(x)$*, which says, "*$P(y)$ *is true for all* $y < x$*". Strong induction on* $P(x)$ *is equivalent to basic induction on* $Q(x)$*.*

Proof. Assume that $P(x)$ satisfies the hypotheses of the corollary. Let $Q(x)$ be the formula

$$(\forall\, y \leq x)\ P(y)$$

where the universe of y is $\mathbb{N}$. Then $Q(0) \equiv P(0)$, so is true. Let $N \in \mathbb{N}$, $N \geq 1$, and assume $Q(N)$. So

$$(\forall\, y \leq N)\ P(y)$$

and therefore $P(N + 1)$. Hence

$$(\forall\, n \leq N + 1)\ P(y)$$

and thus $Q(N + 1)$. Therefore

$$(\forall\, x \in \mathbb{N})\ Q(x) \Rightarrow Q(x + 1).$$

By the principle of induction,

$$(\forall\, x \in \mathbb{N})\ Q(x).$$

However, for any $N \in \mathbb{N}$, $Q(N) \Rightarrow P(N)$, so

$$(\forall\, x \in \mathbb{N})\ P(x).$$

□

Strong induction is particularly useful when proving claims about division. There are examples of the technique throughout Chapter 7. The results in Chapter 7 do not require Chapter 5 and Chapter 6, so you may easily skip ahead. See for example Section 7.1, where the Fundamental Theorem of Arithemetic is proved using strong induction.

Induction does not have to start at 0, or even at a natural number.

COROLLARY 4.9. *Let $k \in \mathbb{Z}$, and $P(x)$ be a formula in one variable such that*

(1) $P(k)$
(2) $(\forall\, x \geq k)\ P(x) \Rightarrow P(x+1)$.

Then

$$(\forall\, x \in \mathbb{Z})\ x \geq k \Rightarrow P(x).$$

DISCUSSION. *This can be proved by a defining a new formula that can be proved with standard induction. Can you define the formula?*

4.3. Polynomials

We now use the machinery developed in Section 4.2 to undertake a modest mathematical program. As we indicated in the first chapter of this book, most of you, until now, have used mathematical results to solve problems in computation. Here we are interested in proving a result with which you may be familiar.

This result concerns polynomials with real coefficients (i.e. coefficients that are real numbers). You have spent a good deal of your mathematical life investigating polynomials, and undoubtedly can make many interesting and truthful claims about them. But how confident

are you that these claims are true? It is possible that your belief in these claims is, by and large, mere confidence in the claims and beliefs of experts in the field. In practice, one can do worse than to acquiesce to the assertions of specialists, and practical limitations generally compel us to accept many claims on faith. Of course, this practice carries risks. For hundreds of years, the assertions of Aristotle were broadly accepted, often in spite of empirical evidence to the contrary. Naturally, we continue to accept many claims on faith. In the case of modern science, we generally do not have first hand access to primary evidence on which modern scientific theories are based. Mathematics is different from every other field of intellectual endeavor because you have the opportunity to verify virtually every mathematical claim you encounter. You are now at the point in your mathematical career at which you can directly confirm mathematical results.

The theorem we wish to prove is that the number of real roots of a real polynomial is at most the degree of the polynomial. You may be familiar with this claim, but uncertain of why it holds. This result is interesting, in part, because it guarantees that the graph of a polynomial will cross any horizontal line only finitely many times. Put another way, level sets of polynomials cannot have more elements than the degree of the polynomial.

NOTATION. $\mathbb{R}[x]$ $\mathbb{R}[x]$ *is the set of polynomials with real coefficients in the variable* x.

THEOREM 4.10. *Let* $N \in \mathbb{N}$ *and* $p \in \mathbb{R}[x]$ *have degree* $N \geq 1$. *Then* p *has at most* N *real roots.*

DISCUSSION. *This result is sufficiently difficult that we shall have to prove three preliminary results. These lemmas*[2] *are proved within the argument for the theorem. Throughout the argument we shall be investigating a general polynomial,* p, *of degree* N.

[2]A lemma is an auxiliary result that one uses in the proof of a theorem — sort of like a subroutine. In German, a theorem is called "Satz" and a lemma is called "Hilfsatz", a "helper theorem".

PROOF. We prove first that the distributive property generalizes to an arbitrary number of summands.

LEMMA 4.11. *Let $N \in \mathbb{N}^+$ and, for $0 \leq n \leq N$, $a_n \in \mathbb{R}$. If $c \in \mathbb{R}$, then*

$$\sum_{n\,=\,0}^{N} ca_n \;=\; c\left(\sum_{n\,=\,0}^{N} a_n\right).$$

DISCUSSION. *This result generalizes the distributive property to more than two summands. We are assuming the distributive property of real numbers: for $a, b, c \in \mathbb{R}$,*

$$c \cdot (a + b) \;=\; ca + cb.$$

We prove the lemma by induction. It is surprising that a claim that seems so obvious uses the powerful machinery of induction. But remember that we are proving this for all finite sums of arbitrarily many summands. Of course, you may feel that the lemma is altogether obvious. If so, you should try to produce your own proof, or read this one for practice in mathematical induction in a context where the mathematical content is easy.

We shall argue by induction on the number of terms in the sum. The base case is for sums with two summands — this is just the distributive property. In the induction step we prove the conditional result that if the lemma holds for all sums with N terms, then it holds for all sums with $N + 1$ terms. At each step of the argument (base and induction steps) we are arguing for infinitely many concrete claims by arguing for a single abstract claim.

PROOF. We argue by induction on N.
Base case: $N \;=\; 1$

Let $c, a_0, a_1 \in \mathbb{R}$. By the distributive property,

$$\begin{aligned}\sum_{n\,=\,0}^{1} ca_n &= ca_0 + ca_1 \\ &= c(a_0 + a_1) \\ &= c\left(\sum_{n\,=\,0}^{1} a_n\right).\end{aligned}$$

Induction step:
Let $c \in \mathbb{R}$ and $a_n \in \mathbb{R}$, for $0 \leq n \leq N+1$. We assume

$$\sum_{n\,=\,0}^{N} ca_n = c\left(\sum_{n\,=\,0}^{N} a_n\right).$$

We have

$$\begin{aligned}\sum_{n\,=\,0}^{N+1} ca_n &= \left(\sum_{n\,=\,0}^{N} ca_n\right) + ca_{N+1} \\ &=_{IH} c\left(\sum_{n\,=\,0}^{N} a_n\right) + ca_{N+1},\end{aligned}$$

By the distributive law (for two summands)

$$\begin{aligned}c\left(\sum_{n\,=\,0}^{N} a_n\right) + ca_{N+1} &= c\left(\sum_{n\,=\,0}^{N} a_n + a_{N+1}\right) \\ &= c\left(\sum_{n\,=\,0}^{N+1} a_n\right).\end{aligned}$$

Therefore,

$$\sum_{n\,=\,0}^{N+1} ca_n = c\left(\sum_{n\,=\,0}^{N+1} a_n\right).$$

By the induction principle the result holds for all $N \in \mathbb{N}$. □

LEMMA 4.12. *If $x, y \in \mathbb{R}$ and $n \in \mathbb{N}^+$, then*

$$\begin{aligned} x^n - y^n &= (x-y)(x^{n-1} + x^{n-2}y + \cdots + xy^{n-2} + y^{n-1}) \\ &= (x-y)\left(\sum_{\substack{i,j\in\mathbb{N} \\ i+j\,=\,n-1}} x^i y^j\right). \end{aligned}$$

DISCUSSION. *The notation in the last line of the lemma means that the sum is taken over all natural numbers i and j that have the property that $i + j = n - 1$.*

PROOF. By Lemma 4.11,

$$\begin{aligned} (x-y)\left(\sum_{\substack{i,j\in\mathbb{N} \\ i+j\,=\,n-1}} x^i y^j\right) &= x\left(\sum_{\substack{i,j\in\mathbb{N} \\ i+j\,=\,n-1}} x^i y^j\right) - y\left(\sum_{\substack{i,j\in\mathbb{N} \\ i+j\,=\,n-1}} x^i y^j\right) \\ &= \sum_{\substack{i,j\in\mathbb{N} \\ i+j\,=\,n-1}} x^{i+1} y^j - \sum_{\substack{i,j\in\mathbb{N} \\ i+j\,=\,n-1}} x^i y^{j+1} \\ &= x^n - y^n. \end{aligned}$$

□

The next lemma associates roots of polynomials and linear factors.

LEMMA 4.13. *Let p be a polynomial of degree N. A real number, c, is a root of p iff*

$$p(x) = (x-c)q(x),$$

where $q(x)$ is a polynomial of degree $N - 1$.

DISCUSSION. *This lemma is a biconditional statement. That is, the lemma is propositionally equivalent to the conjunction of two conditional statements. We prove the conditional statements independently. One of the conditional statements is obvious (can you determine which?). The more difficult conditional statement will use Lemma 4.12. When proving a biconditional, $P \iff Q$, by proving the conditional statements $P \Rightarrow Q$ and $Q \Rightarrow P$, we often use $(\Rightarrow)$ and $(\Leftarrow)$ to identify the conditional statement under consideration.*

PROOF. Let p be a polynomial of degree N. Then there are $a_0, a_1, \ldots, a_N \in \mathbb{R}$, $a_N \neq 0$, such that,

$$p(x) \;=\; \sum_{n\,=\,0}^{N} a_n x^n.$$

($\Leftarrow$) Assume that there is $c \in \mathbb{R}$ and a polynomial q of degree $N-1$ such that

$$p(x) \;=\; (x-c)q(x).$$

Then

$$p(c) \;=\; (c-c)q(c) \;=\; 0.$$

So c is a root of p.
($\Rightarrow$) Let $c \in \mathbb{R}$ be a root of p. Then

$$\begin{aligned} p(x) &= p(x) - p(c) \\ &= a_0 - a_0 + \sum_{n\,=\,1}^{N} a_n(x^n - c^n) \\ &= \sum_{n\,=\,1}^{N} a_n(x^n - c^n). \end{aligned}$$

By Lemma 4.12, for $n \geq 1$,

$$x^n - c^n \;=\; (x-c)q_n(x)$$

where

$$q_n(x) \;=\; x^{n-1} + cx^{n-2} + \cdots + c^{n-2}x + c^{n-1} \;=\; \sum_{\substack{i,j\in\mathbb{N} \\ i+j\,=\,n-1}} x^i c^j.$$

By Lemma 4.11,

$$p(x) \;=\; \sum_{n\,=\,1}^{N} a_n(x^n - c^n) \;=\; (x-c)\sum_{n\,=\,1}^{N} a_n q_n(x).$$

Let

$$q(x) \;=\; \sum_{n\,=\,1}^{N} a_n q_n(x).$$

For all n between 1 and N, $q_n(x)$ has degree $(n-1)$. So the degree of $q(x)$ is less than N. However the coefficient of x^{N-1} in $q(x)$ is a_N, and $a_N \neq 0$ by assumption. So the degree of $q(x)$ is $N-1$, and

$$p(x) \;=\; (x-c)q(x).$$

□

We complete the proof of the theorem. Let p be a polynomial of degree N. We argue by induction on the degree of p.
Base case: $N \;=\; 1$.
If p is a polynomial of degree 1, then it is of the form

$$p(x) \;=\; a_1 x + a_0,$$

and the only root is $-a_0/a_1$.
Induction step:
Assume that the theorem holds for $N \in \mathbb{N}^+$. Let p have degree $N+1$. If p has no roots, the theorem holds for p. So assume that p has a real root, $c \in \mathbb{R}$. By Lemma 4.13,

$$p(x) \;=\; (x-c)q(x), \tag{4.14}$$

where q is of degree N. By the induction hypothesis, q has at most N real roots. If x is a root of p, then by (4.14) either x is a root of q or $x \;=\; c$. Therefore p has at most $N+1$ roots, proving the induction step. □

As a function, a polynomial in a particular variable is the same as a polynomial with the same coefficients in a different variable. Let $p \in \mathbb{R}[x]$ be

$$p(x) \;=\; \sum_{n \,=\, 0}^{N} a_n x^n,$$

and $q \in \mathbb{R}[y]$ be

$$q(y) \;=\; \sum_{n \,=\, 0}^{N} a_n y^n.$$

Then as real functions p and q are the same function. That is,

$$\text{graph}(p) \;=\; \text{graph}(q).$$

As algebraic objects, however, one might occasionally wish to distinguish between polynomials in distinct variables.

We end this section by proving that polynomials are equal as functions if and only if they have the same coefficients.

COROLLARY 4.15. *Let $p, q \in \mathbb{R}[x]$. The coefficients of p and q are equal iff*

$$(\forall\, x \in \mathbb{R}) \;\; p(x) \;=\; q(x).$$

PROOF. ($\Rightarrow$) If the coefficients of p and q are all equal, then, letting a_n denote the n^{th} coefficient, we have

$$(\forall\, x \in \mathbb{R})\; p(x) \;=\; \sum_{n\,=\,0}^{N} a_n x^n \;=\; q(x).$$

($\Leftarrow$) Suppose $(\forall\, x \in \mathbb{R})\; p(x) \;=\; q(x)$. Then $p - q$ is a polynomial with infinitely many roots. If p and q disagree on any coefficient, then $p - q$ is a non-zero polynomial, has a degree, and by Theorem 4.10, finitely many roots. Therefore, p and q must agree on all coefficients. □

4.4. Arithmetic-Geometric Inequality

We have presented modest generalizations of basic mathematical induction (Corollary 4.8 and Corollary 4.9). The formality of our approach might suggest that induction is a rigid technique that must be applied inflexibly in a specific prescriptive way. To a mathematician induction is governed by two ideas:

(1) Induction uses the well ordering of the natural numbers, or more generally any well-ordered set, to prove universal statements quantified over the set.
(2) Every element in the set over which you quantify must be accounted for by the induction.

The formal characterizations of induction in Section 4.2 are sufficient but not necessary to achieve the objectives of a proof by induction. The theorem in this section will give you a sense about how the technique of induction can be extended.

DEFINITION. Arithmetic mean Let $a_1, \ldots, a_N$ be real numbers. The arithmetic mean of $a_1, \ldots, a_N$ is

$$\frac{1}{N}\left(\sum_{n\,=\,1}^{N} a_n\right).$$

DEFINITION. Geometric mean Let $a_1, \ldots, a_N$ be positive real numbers. The geometric mean of $a_1, \ldots, a_N$ is

$$\sqrt[N]{a_1 \cdots a_N}.$$

THEOREM 4.16. *Arithmetic-geometric mean inequality* *Let* $a_1, \ldots, a_N \in \mathbb{R}^+$. *Then*

$$\sqrt[N]{a_1 \cdots a_n} \leq \frac{1}{N}\left(\sum_{n\,=\,1}^{N} a_n\right). \tag{4.17}$$

DISCUSSION. *We prove this with an interesting argument due originally to Cauchy; our treatment is from the book* [**1**]. *We argue by induction on the size of the sample over which we are computing the means. After arguing the base case we show that if the inequality holds for the arithmetic and geometric means of* N *numbers, it necessarily holds for the means of* $2N$ *numbers. This implies that the theorem holds for the means of* 2^N *numbers for any* $N \in \mathbb{N}$ *(by a standard induction argument).*

We then show that the result holding for N *numbers implies that it holds for* $N-1$ *numbers. This implies that if the result holds at a natural number* N, *the inequality holds for all means of fewer than* N *numbers. Given any* $k \in \mathbb{N}$, $2^k > k$ *and since the theorem holds for means of* 2^k *numbers, it holds for means of* k *terms.*

PROOF. We argue by induction on the number of terms on each side of the inequality.

Base case: $(N = 2)$

Let $a_1, a_2 \in \mathbb{R}^+$. Then

$$(a_1 - a_2)^2 = a_1^2 - 2a_1a_2 + a_2^2 \geq 0.$$

Therefore

$$2a_1a_2 \leq a_1^2 + a_2^2,$$

and

$$\begin{aligned} 4a_1a_2 &\leq a_1^2 + 2a_1a_2 + a_2^2 \\ &= (a_1 + a_2)^2. \end{aligned}$$

Thus

$$2\sqrt{a_1a_2} \leq a_1 + a_2.$$

Therefore the inequality holds for two terms.

Induction step:

Let $P(N)$ be the statement that (4.17) holds for all $a_1, \ldots, a_N > 0$. We show that $P(N) \Rightarrow P(2N)$. Let

$$G_N = \prod_{n\,=\,1}^{N} a_n$$

and

$$A_N = \left(\frac{\sum_{n\,=\,1}^{N} a_n}{N} \right).$$

So

$$\begin{aligned} G_{2N} &= \prod_{n\,=\,1}^{2N} a_n \\ &= \left(\prod_{n\,=\,1}^{N} a_n \right) \left(\prod_{n\,=\,N+1}^{2N} a_n \right) \\ &\leq_{IH} \left(\sum_{n\,=\,1}^{N} \frac{a_n}{N} \right)^N \left(\sum_{n\,=\,N+1}^{2N} \frac{a_n}{N} \right)^N. \end{aligned}$$

Let

$$B = \sum_{n\,=\,N+1}^{2N} \frac{a_n}{N}.$$

By the base case

$$\begin{aligned} A_N B &\leq \left(\frac{A_N + B}{2} \right)^2 \\ &= (A_{2N})^2 \end{aligned}$$

So

$$\begin{aligned}(A_N)^N B^N &= (A_N B)^N \\ &\leq ((A_{2N})^2)^N \\ &= (A_{2N})^{2N}.\end{aligned}$$

Thus

$$G_{2N} \leq (A_{2N})^{2N}.$$

Therefore, for any $N \in \mathbb{N}^+$,

$$P(N) \Rightarrow P(2N).$$

DISCUSSION. *Let $Q(N)$ be the statement $P(2^N)$. Then the argument thus far is a standard proof by induction of $(\forall\, N \in \mathbb{N}^+)\, Q(N)$. Of course we wish to show $(\forall\, N \in \mathbb{N})\, P(N)$. We do this by proving*

$$(\forall\, N \in \mathbb{N}^+)\, P(N+1) \Rightarrow P(N).$$

Let $N > 2$. We prove that

$$P(N+1) \Rightarrow P(N).$$

Assume $P(N+1)$. Then

$$(G_N)(A_N) \leq \left(\frac{(\sum_{n\,=\,1}^{N} a_n) + A_N}{N+1} \right)^{N+1}. \tag{4.18}$$

DISCUSSION. *Recall that G_N is the product of $a_1, \ldots, a_N$. We are treating the sum A_N as the $N+1^{st}$ factor, a_{N+1}, and applying the inequality $P(N+1)$.*

As

$$\begin{aligned}\left(\frac{(\sum_{n\,=\,1}^{N} a_n) + A_N}{N+1} \right)^{N+1} &= \left(\frac{NA_N + A_N}{N+1} \right)^{N+1} \\ &= (A_N)^{N+1},\end{aligned}$$

Inequality 4.18 gives

$$G_N A_N \;\leq\; A_N^{N+1},$$

and so

$$G_N \leq (A_N)^N$$

which is the statement $P(N)$. So

$$(\forall\, N \in \mathbb{N}^+)\, P(N+1) \Rightarrow P(N).$$

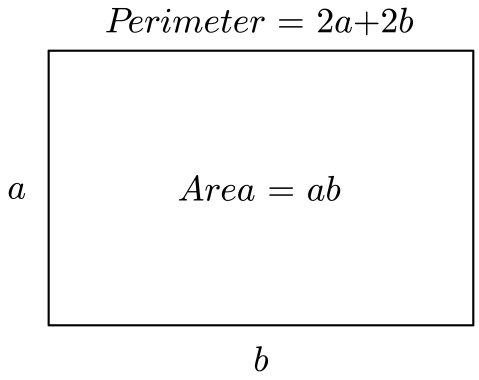
Hence for all $N \geq 2$, $P(N)$. □

The arithmetic mean and geometric mean are different ways of understanding averages. They are related by the arithmetic geometric mean inequality (called the AGM inequality). Can we apply the inequality? Let's consider an easy geometrical application of the case $N = 2$. Consider the rectangle with sides length a and b. The perimeter of the rectangle is

$$P = 2a + 2b$$

and the area is

$$A = ab.$$



In calculus you proved that the rectangle of fixed perimeter with the greatest area is the square. This can also be proved directly from

the AGM inequality:

$$\begin{aligned} P &= 2a + 2b \\ &= \frac{4a + 4b}{2} \\ &\geq \sqrt{16ab} \\ &= 4\sqrt{ab}. \end{aligned}$$

So

$$\frac{P^2}{16} \geq ab \;=\; A.$$

Recall that P is fixed, and therefore so is $\frac{P^2}{16}$, and we have shown that this is an upper bound for the area of the rectangle.

Is this upper bound achieved? The area A of the rectangle varies according to the dimensions of the rectangle and if $a \;=\; b$

$$\frac{P^2}{16} \;=\; \frac{(4a)^2}{16} \;=\; A.$$

Thus the maximum area of the rectangle is achieved when $a \;=\; b$. This result can be generalized to higher dimensions — without the need for multivariable calculus.

Proving theorems is not just a question of technique, though this must be mastered. It also requires creativity and insight. A beautiful collection of proofs is contained in the book [**1**] by Martin Aigner and Günter Ziegler.

4.5. Exercises

Exercise 4.1. Prove by induction that 3 divides $7^n - 4$ for every $n \in \mathbb{N}^+$.

Exercise 4.2. Prove by induction that

$$(\forall\, n \in \mathbb{N})\; 2^n > n.$$

Exercise 4.3. Prove that any subset of a well-ordered set is well-ordered.

EXERCISE 4.4. Prove that $(1+x)^n \geq 1+nx$ for every $n \in \mathbb{N}^+$ and every $x \in (-1, \infty)$.

EXERCISE 4.5. Prove by induction that every finite set of real numbers has a largest element.

EXERCISE 4.6. Let X and Y be sets with n elements each. How many bijections from X to Y are there? What does this tell you about the number of permutations of $\ulcorner n \urcorner$? Prove your claim.

EXERCISE 4.7. The binomial coefficients $\binom{n}{k}$ can be defined from Pascal's triangle by:

(i) $\forall\, n \in \mathbb{N},\ \binom{n}{0} = \binom{n}{n} = 1.$

(ii) $\forall\, 2 \leq n \in \mathbb{N},\ \forall\, 1 \leq k \leq n-1,\ \binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}.$

Prove by induction that

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

EXERCISE 4.8. Prove the binomial theorem: with $\binom{n}{k}$ defined by Exercise 4.7, for any $n \in \mathbb{N}$, the following identity holds

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

EXERCISE 4.9. Prove $\sum_{k=0}^{n} \binom{n}{k} = 2^n$.

EXERCISE 4.10. Prove, for all $n \in \mathbb{N}^+$,

$$\binom{2n}{n} \geq \frac{2^{2n-1}}{\sqrt{n}}.$$

EXERCISE 4.11. The Principle of Descent says that there is no strictly decreasing infinite sequence of natural numbers. Prove the Principle of Descent.

EXERCISE 4.12. The Fibonacci numbers are defined recursively by $F_1 = 1, F_2 = 1$, and for $n \geq 3$, $F_n = F_{n-1} + F_{n-2}$. Prove that the Fibonacci numbers are given by the equation

$$F_n = \frac{(1+\sqrt{5})^n - (1-\sqrt{5})^n}{2^n \sqrt{5}}. \tag{4.20}$$

This is an example of a formula that is hard to guess, but easy to verify. For an explanation of how the formula arises, see Exercise 5.29.

EXERCISE 4.13. Let X be a set well-ordered by a relation $\preceq$. We say that a sequence of elements in X, $\langle x_n \mid n \in \mathbb{N}\rangle$, is strictly decreasing (with respect to $\preceq$) if for all $m, n \in \mathbb{N}$

$$[m < n] \Rightarrow [x_n \preceq x_m \ \wedge \ x_n \neq x_m].$$

Prove that there is no strictly decreasing sequence of elements in X.

EXERCISE 4.14. Prove that the last digit of $7^{7^{\cdot^{\cdot^{\cdot^{7}}}}}$ is 3 for any tower of sevens of height more than 1.

EXERCISE 4.15. Give another example that illustrates the need for a base case in a valid proof by induction.

EXERCISE 4.16. Assume that there is a polynomial of degree 4 in N that gives $\sum_{n\,=\,0}^{N} n^3$. Find the polynomial and then prove that the formula is correct by induction.

Use Archimedes's method to prove that

EXERCISE 4.17. Let $\mathbb{N}[x]$ be the set of polynomials with natural number coefficients. Define a relation $\preceq$ on $\mathbb{N}[x]$ by:

Let $p(x) = \sum_{n=0}^{N} a_n x^n$, and $q(x) = \sum_{n=0}^{M} b_n x^n$. Say that $p \preceq q$ iff, if k is the coefficient of highest degree at which p and q differ, then $a_k \leq b_k$. Is $\preceq$ a linear ordering? Is it a well-ordering of $\mathbb{N}[x]$?

EXERCISE 4.18. Assume that there is a polynomial p of degree 5 such that

$$\sum_{n\,=\,0}^{N} n^4 \ = \ p(N).$$

Find p and prove that the formula you propose is correct.

EXERCISE 4.19. Determine the set of positive natural numbers n such that the sum of every n consecutive natural numbers is divisible by n.

EXERCISE 4.20. Let f be a real function such that, for $x, y \in \mathbb{R}$,

$$f(x+y) = f(x) + f(y).$$

Prove that
a) $f(0) = 0$
b) $f(n) = nf(1)$.

EXERCISE 4.21. Prove Corollary 4.9.

EXERCISE 4.22. Consider boxes with dimensions a, b and c in which the sum of the dimensions (*i.e.* $a+b+c$) is fixed. Prove that the box with largest possible volume has dimensions that satisfy $a = b = c$.

EXERCISE 4.23. Prove by induction that any well-formed propositional statement has a well-defined truth value.

EXERCISE 4.24. Prove by induction on the number of propositional connectives that every compound propositional statement is equivalent to a statement using only $\neg$ and $\vee$.

EXERCISE 4.25. Prove by induction on the number of propositional connectives that every compound propositional statement is equivalent to a statement using only $\neg$ and $\wedge$.

EXERCISE 4.26. Let Q_i be a quantifier, for $1 \leq i \leq n$. For each Q_i, let Q_i^* be the complementary quantifier. That is, if $Q_i = \forall$, then $Q_i^* = \exists$; if $Q_i = \exists$, then let $Q_i^* = \forall$. Prove by induction on the number of quantifiers that,

$$\neg(Q_1 x_1)(...)(Q_n x_n)P(x_1, \ldots, x_n) \equiv (Q_1^* x_1)(...)(Q_n^* x_n)\neg P(x_1, \ldots, x_n).$$

EXERCISE 4.27. Define the n^{th} Fermat number to be

$$F_n := 2^{2^n} + 1, \qquad n \in \mathbb{N}.$$

(i) Show that the Fermat numbers satisfy

$$\prod_{k=0}^{n} F_k = F_{n+1} - 2.$$

(ii) Conclude that any two distinct Fermat numbers are coprime.

EXERCISE 4.28. Let $\langle a_n : n \in \mathbb{N} \rangle$ be a sequence of positive numbers. Suppose that $a_0 \leq 1$, and that for all $N \in \mathbb{N}$,

$$a_{N+1} \leq \sum_{n=0}^{N} a_n. \tag{4.21}$$

Prove

$$(\forall\, N \in \mathbb{N})\ a_n \leq 2^N \tag{4.22}$$

EXERCISE 4.29. Let $\langle a_n : n \in \mathbb{N} \rangle$ be a sequence of positive numbers satisfying (4.21), and $a_0 \leq C$. What is the correct analogue of (4.22)? Prove your assertion.

EXERCISE 4.30. Let $\mathcal{F} = \{X_\alpha \mid \alpha \in A\}$ be an indexed family of pairwise disjoint sets. Suppose that each X_α is well-ordered by $\preceq_\alpha$ and that A is well-ordered by $\preceq$. Define a relation R on the union of all the sets in $\mathcal{F}$ by: for all $a, b \in \bigcup_{\alpha\in A} X_\alpha$, aRb iff
(a) $a \in X_{\alpha_1}$, $b \in X_{\alpha_2}$ and $\alpha_1 \preceq \alpha_2$,
or
(b) $(\exists\, \alpha \in A)$ $a, b \in X_\alpha$ and $a \preceq_\alpha b$.

Prove that R is a well ordering of $\bigcup_{\alpha\in A} X_\alpha$.

EXERCISE 4.31. Let X be a finite set and $f : X \to X$. Prove that f is an injection iff f is a surjection.

CHAPTER 5

Limits

The idea of a limit is the cornerstone of calculus. It is somewhat subtle, which is why, although it was implicit in the work of Archimedes[1], and essential to a proper understanding of Zeno's paradoxes, it took two thousand years to be understood fully. Calculus was developed in the 17^{th} century by Newton and Leibniz with a somewhat cavalier approach to limits; it was not until the 19^{th} century that a rigorous definition of limit was given, by Cauchy.

In Section 5.1 we define limits, and prove some elementary properties. In Section 5.2 we discuss continuous functions, and in Section 5.3 we look at limits of sequences of functions.

5.1. Limits

Given a real function $f : X \to \mathbb{R}$, the intuitive idea of the statement

$$\lim_{x \to a} f(x) \;=\; L \tag{5.1}$$

is that, as x gets closer and closer to a, the values of $f(x)$ get closer and closer to L. Making this notion precise is not easy — try to write down a mathematical definition now, before you read any further.

The idea behind the definition is to give a sequence of guarantees. Imagine yourself as an attorney, trying to defend the claim (5.1). For concreteness, let us fix $g(x) = \frac{\sin(x)}{x}$, and try to defend the claim that

$$\lim_{x \to 0} g(x) \;=\; 1 \tag{5.2}$$

[1]Archimedes (287-212 BC) calculated the area under a parabola (what we would now call $\int_0^1 x^2 dx$) by calculating the area of the rectangles of width $1/N$ under the parabola and letting N tend to infinity. This is identical to the modern approach of finding an integral by taking a limit of Riemann sums.

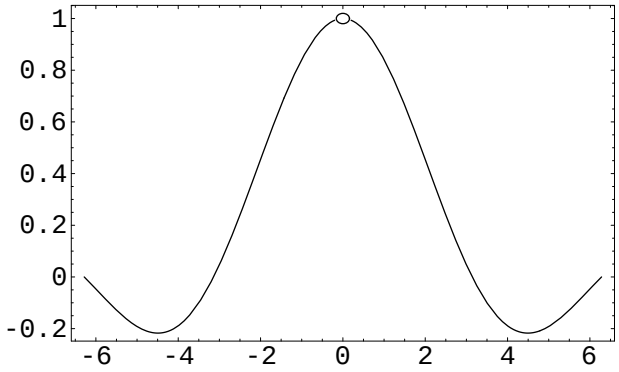


FIGURE 5.3. Plot of $\sin(x)/x$

The skeptical judge asks "Can you guarantee that $g(x)$ is within .1 of 1?"

"Yes, your honor, provided that $|x| < .7$."

"Hmm, well can you guarantee that $g(x)$ is within .01 of 1?"

"Yes, your honor, provided that $|x| < .2$."

And so it goes. If, for every possible tolerance the judge poses, you can find a precision (*i.e.* an allowable deviation of x from a) that guarantees that the difference between the function value and the limit is within the allowable tolerance, then you can successfully defend the claim.

EXERCISE. Now try to give a mathematical definition of a limit, without reading any further.

We shall start with the case that the function is defined on an open interval.

DEFINITION. Limit, $\lim_{x\to a} f(x)$ Let I be an open interval and a some point in I. Let f be a real-valued function defined on $I \setminus \{a\}$. (It doesn't matter whether f is defined at a or not). Then we say

$$\lim_{x\to a} f(x) \;=\; L$$

(in words, "the limit as x tends to a of $f(x)$ is L") if, for every $\varepsilon > 0$, there exists $\delta > 0$, so that

$$0 < |x - a| < \delta \quad \Longrightarrow \quad |f(x) - L| < \varepsilon. \tag{5.4}$$

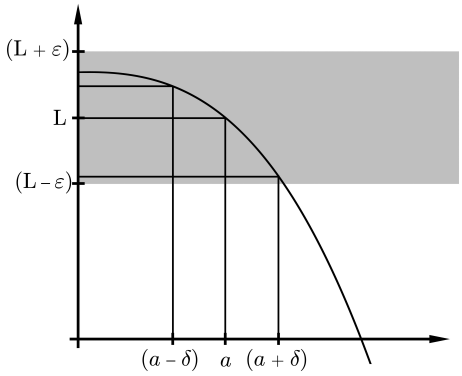


FIGURE 5.5. One choice of δ for a given ε

The condition $0 < |x - a| < \delta$ means we exclude $x = a$ from consideration. *Limits are about the behavior of a function near the point, not at the point.* For a function like $g(x) = \sin(x)/x$, the value at 0 is undefined; nevertheless $\lim_{x\to 0} g(x)$ exists, and is the same as $\lim_{x\to 0}$ of the function

$$h(x) \;=\; \begin{cases} \sin(x)/x & x \neq 0 \\ 5 & x = 0. \end{cases}$$

REMARK. The use of ε for the allowable error and δ for the corresponding precision required is hallowed by long usage. Mathematicians need all the convenient symbols they can find. The Greek alphabet has long been used as a supplement to the Roman alphabet in Western mathematics, and you need to be familiar with it (see Appendix A for the Greek alphabet).

The main point to note in the definition is the order of the quantifiers: $\forall\,\varepsilon,\ \exists\,\delta$. What would it mean to say

$$(\exists\,\delta>0)\ (\forall\,\varepsilon>0)\quad [0<|x-a|<\delta\quad\Longrightarrow\quad |f(x)-L|<\varepsilon]\ ?$$

To talk comfortably about limits, it helps to have some words that describe inequalities (5.4). Let us say that the ε-neighborhood of L is the set of points within ε of L, *i.e.* the interval $(L-\varepsilon, L+\varepsilon)$. The punctured δ-neighborhood of a is the set of points within δ of a, excluding a itself, *i.e.* $(a-\delta, a)\cup(a, a+\delta)$. When we speak of ε-neighborhoods and punctured δ-neighborhoods, we always assume that ε and δ are positive, so that the neighborhoods are non-empty.

Then the definition of limit can be worded as "every ε–neighborhood of L has an inverse image under f that contains some punctured δ-neighborhood of a".

Remark. We can revisit our court-room analogy, and say that to prove that f has limit L at a, we need a strategy that produces a workable δ for any ε. So a proof is essentially a function F that takes any positive ε and spits out a positive $\delta = F(\varepsilon)$ for which (5.4) works.

Example 5.6. Let $f(x)=5x+2$. Prove $\lim_{x\to 3} f(x)=17$.

Let $\varepsilon>0$. We want to find a $\delta>0$ so that the punctured δ-neighborhood of 3 is mapped into the ε-neighborhood of 17.

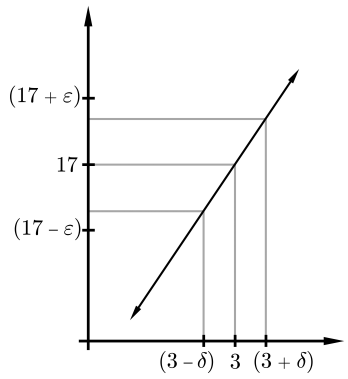


Figure 5.7. Relationship between δ and ε

Taking $\delta = \varepsilon/5$ will work, as will any smaller choice of δ. Indeed, if $0 < |x-3| < \delta$, then $|f(x) - 17| < 5\delta = \varepsilon$.

EXAMPLE 5.8. This time, let $g(x) = 55x+2$. To prove $\lim_{x\to 3} g(x) = 167$, we must take $\delta \leq \varepsilon/55$.

If two functions f and g both have limits at the point a, then so do all the algebraic combinations $f+g, f-g, f\cdot g$ and cf for c a constant. The quotient f/g also has a limit at a, provided $\lim_{x\to a} g(x) \neq 0$. Moreover, these limits are what you would expect.

THEOREM 5.9. *Suppose f and g are functions on an open interval I, and at the point a in I both $\lim_{x\to a} f(x)$ and $\lim_{x\to a} g(x)$ exist. Let c be any real number. Then*

$$\begin{aligned}
&(i) && \lim_{x\to a}[f(x)+g(x)] &&= \lim_{x\to a} f(x) + \lim_{x\to a} g(x)\\
&(ii) && \lim_{x\to a}[f(x)-g(x)] &&= \lim_{x\to a} f(x) - \lim_{x\to a} g(x)\\
&(iii) && \lim_{x\to a} cf(x) &&= c\left[\lim_{x\to a} f(x)\right]\\
&(iv) && \lim_{x\to a}[f(x)g(x)] &&= \left[\lim_{x\to a} f(x)\right]\cdot\left[\lim_{x\to a} g(x)\right]\\
&(v) && \lim_{x\to a}\tfrac{f(x)}{g(x)} &&= \frac{\lim_{x\to a} f(x)}{\lim_{x\to a} g(x)}, \qquad \text{provided } \lim_{x\to a} g(x) \neq 0.
\end{aligned}$$

DISCUSSION. *How do we go about proving a theorem like this? Well, to start with, don't be intimidated by its length. Let's start on part* (i). *We only have the definition of limit to work with, so we only have one strategic option: prove directly that the definition is satisfied.*

PROOF OF *(i)*. Let L_1 and L_2 be the limits of f and g respectively at a. Let ε be an arbitrary positive number. We must find a $\delta > 0$ so that

$$0 < |x-a| < \delta \quad \Longrightarrow \quad |f(x)+g(x)-(L_1+L_2)| < \varepsilon.$$

The key idea, common to many limit arguments, is to use the observation that

$$|f(x)+g(x)-(L_1+L_2)| \;\leq\; |f(x)-L_1| + |g(x)-L_2|. \tag{5.10}$$

This is an application of the so-called triangle inequality, which you are asked later to prove (Lemma 5.14). It is the assertion that for any real numbers c and d, we have

$$|c+d| \ \leq \ |c|+|d|.$$

(What values of c and d yield (5.10)?) So if we can make *both* $|f(x)-L_1|$ and $|g(x)-L_2|$ small, then Inequality (5.10) forces

$$|f(x)+g(x)-(L_1+L_2)|$$

to be small too, which is what we want.

Since f and g have limits L_1 and L_2 at a, we know that there exist positive numbers δ_1 and δ_2 such that

$$\begin{aligned} 0<|x-a|<\delta_1 \qquad &\Longrightarrow \qquad |f(x)-L_1|<\varepsilon \\ 0<|x-a|<\delta_2 \qquad &\Longrightarrow \qquad |g(x)-L_2|<\varepsilon. \end{aligned}$$

If $|x-a|$ is less than both δ_1 and δ_2, then both inequalities are satisfied, and we get

$$|f(x)+g(x)-(L_1+L_2)| \ \leq \ |f(x)-L_1|+|g(x)-L_2| \ \leq \ \varepsilon+\varepsilon. \quad (5.11)$$

This isn't quite good enough; we want the left-hand side of (5.11) to be bounded by ε, not 2ε. We are saved, however, by the requirement that for *any* positive number η, we can guarantee that f and g are in an η-neighborhood of L_1 and L_2, respectively. In particular, let η be $\varepsilon/2$. Since f and g have limits at a, there are positive numbers δ_3 and δ_4 so that

$$\begin{aligned} 0<|x-a|<\delta_3 \qquad &\Longrightarrow \qquad |f(x)-L_1| \ < \ \frac{\varepsilon}{2} \\ 0<|x-a|<\delta_4 \qquad &\Longrightarrow \qquad |g(x)-L_2| \ < \ \frac{\varepsilon}{2}. \end{aligned}$$

So we set δ equal to the smaller of δ_3 and δ_4, and we get

$$\begin{aligned} &0<|x-a|<\delta \qquad \Longrightarrow \\ &\qquad |f(x)+g(x)-(L_1+L_2)| \ \leq \ |f(x)-L_1|+|g(x)-L_2| \ < \ \varepsilon, \end{aligned}$$

as required. ◁

EXERCISE. Explain in words how the preceding proof worked. In short-hand, one could say that if F_1 and F_2 are strategies for proving $\lim_{x\to a} f(x) = L_1$ and $\lim_{x\to a} g(x) = L_2$ respectively, then

$$F \;=\; \min\left\{F_1\left(\frac{\varepsilon}{2}\right),\, F_2\left(\frac{\varepsilon}{2}\right)\right\}$$

is a strategy for proving $\lim_{x\to a} f(x) + g(x) = L_1 + L_2$.

DISCUSSION. *What next? We could prove* (ii) *in a similar way, but mathematicians like shortcuts. Notice that if we prove* (iii) *and let* $c = -1$, *then we can apply* (i) *to* $f + (-g)$ *and get* (ii) *that way. Moreover,* (iii) *is just a special case of* (iv), *if we know that the constant function* $g(x) = c$ *has the limit* c *at every point. So let us prove* (iv) *next.*

PROOF OF *(iv).* Again, let ε be an arbitrary positive number. We must find a $\delta > 0$ so that

$$0 < |x - a| < \delta \quad \Longrightarrow \quad |f(x)g(x) - (L_1L_2)| < \varepsilon.$$

It is not quite clear how close f and g have to be to L_1 and L_2 to get that their product is close enough to L_1L_2, so let's play it safe by not choosing yet. For every $\varepsilon_1, \varepsilon_2 > 0$, we know there exist $\delta_1, \delta_2 > 0$ such that

$$\begin{aligned} 0 < |x - a| < \delta_1 \quad &\Longrightarrow \quad |f(x) - L_1| < \varepsilon_1 \\ 0 < |x - a| < \delta_2 \quad &\Longrightarrow \quad |g(x) - L_2| < \varepsilon_2. \end{aligned}$$

Now we use the second common trick in proving the existence of limits: add and subtract the same quantity so that one can factor.

$$\begin{aligned} |f(x)g(x) - L_1L_2| \;&=\; |f(x)g(x) - L_1g(x) + L_1g(x) - L_1L_2| \\ &\leq\; |f(x)g(x) - L_1g(x)| + |L_1g(x) - L_1L_2| \\ &\leq\; |f(x) - L_1||g(x)| + |g(x) - L_2||L_1|. \quad (5.12) \end{aligned}$$

Now if both summands on the last line can be made less than $\varepsilon/2$, we win. The second term is easy: we choose

$$\varepsilon_2 = \frac{\varepsilon}{2|L_1| + 1}.$$

Then there is a δ_2 so that

$$\begin{aligned} 0 < |x-a| < \delta_2 \quad &\Longrightarrow \quad |g(x) - L_2| \quad < \varepsilon_2 \\ &\Longrightarrow \quad |g(x) - L_2||L_1| \ < \frac{\varepsilon|L_1|}{2|L_1|+1} \ < \frac{\varepsilon}{2}. \end{aligned}$$

(If $L_1 \neq 0$, we could have chosen $\varepsilon_2 = \frac{\varepsilon}{2|L_1|}$; we added 1 to the denominator just so we did not have to consider the two cases separately.)

What about the first summand in (5.12), the term $|f(x)-L_1||g(x)|$? First let us get some bound on how big $|g|$ can be. We know that if $0 < |x-a| < \delta_2$, then $|g(x) - L_2| < \varepsilon/(2|L_1|+1)$, so

$$|g(x)| \ < \ |L_2| + \frac{\varepsilon}{2|L_1|+1} \ =: M.$$

If we let $\varepsilon_1 = \varepsilon/(2M)$, we know that there exists $\delta_1 > 0$ so that

$$0 < |x-a| < \delta_1 \quad \Longrightarrow \quad |f(x) - L_1||g(x)| \ < \ \varepsilon_1|g(x)|. \tag{5.13}$$

Finally, we let $\delta = \min(\delta_1, \delta_2)$. For $0 < |x-a| < \delta$, both summands in (5.12) are less than $\varepsilon/2$: the second summand because $\delta \leq \delta_2$, and the first because when $0 < |x-a| < \delta$, Inequality 5.13 is strengthened to

$$|f(x) - L_1||g(x)| \ < \ \varepsilon_1|g(x)| \ < \ \varepsilon_1 \, M \ = \ \varepsilon/2.$$

Therefore, for $0 < |x-a| < \delta$, we have $|f(x)g(x) - L_1L_2| < \varepsilon$, as desired. ◁

Proof of *(iii)*. This is a special case of *(iv)*, once we know that constant functions have limits. Let us state this as a lemma. Given Lemma 5.15, *(iii)* is proved, and hence so is *(ii)*.

Proof of *(v)*. Exercise. □

Lemma 5.14. *Triangle inequality Let c, d be real numbers. Then $|c+d| \leq |c| + |d|$.*

Proof. Exercise. □

Lemma 5.15. *Let $g(x) \equiv c$ be the constant function c. Then,*

$$(\forall\, a \in \mathbb{R}) \ \lim_{x \to a} g(x) = c.$$

PROOF. Exercise. □

EXAMPLE 5.16. The Heaviside function $H(t)$ is defined by

$$H(t) \;=\; \begin{cases} 0 & t < 0 \\ 1 & t \geq 0. \end{cases}$$

Show that H does not have a limit at 0.

DISCUSSION. *To prove that a limit does not exist, we must prove the opposite of* $\forall\,\varepsilon\,\exists\,\delta$, i.e. *that* $\exists\,\varepsilon\,\nexists\,\delta$. *As the gap between the function on* $[0,\infty)$ *and* $(-\infty,0)$ *is* 1, *it is clear that any band of width* < 1 *cannot be wide enough to contain values of* $H(t)$ *for* t *on both sides of* 0. *So we will choose some* $\varepsilon < .5$, *and argue by contradiction.*

Suppose the limit exists and equals L. Let $\varepsilon = \frac{1}{4}$. By hypothesis, there exists $\delta > 0$ such that

$$0 < |t| < \delta \quad\Longrightarrow\quad |H(t) - L| < \frac{1}{4}.$$

But for t negative, this means $|L| < \frac{1}{4}$; and for t positive, this means $|L-1| < \frac{1}{4}$. Thus we get a contradiction. □

If the function is defined on the closed interval $[c,d]$, we may still want to ask if it has a limiting value at c; if so, however, we only want to consider points near c that are in the domain of definition. More generally, we are led to the following definition of a restricted limit.

DEFINITION. Restricted limit, $\lim\limits_{X\ni x\to a} f(x)$ Suppose f is a real function and $X \subseteq \mathrm{Dom}(f)$. Let $a \in \mathbb{R}$. We say that $\lim\limits_{X\ni x\to a} f(x) = L$ if

$$(\forall\,\varepsilon > 0)\ (\exists\,\delta > 0)\ (\forall\,x \in X) \quad [0 < |x-a| < \delta] \quad\Longrightarrow\quad |f(x) - L| < \varepsilon.$$

We read "$\lim\limits_{X\ni x\to a} f(x) = L$" as "the limit as x tends to a inside X of $f(x)$ is L." An important special case of restricted limits are the following:

DEFINITION. Right-hand limit, $\lim_{x\to a^+} f(x)$ Let $a, b, L \in \mathbb{R}$, $a < b$ and f be a real function defined on (a,b). We say that

$$\lim_{x\to a^+} f(x) \;=\; L$$

if

$$(\forall \varepsilon > 0)(\exists \delta > 0)\ [x \in (a, a+\delta)] \Rightarrow [\mid f(x) - L \mid < \varepsilon].$$

The number L is the right-hand limit of $f(x)$ at a. The left-hand limit is defined analogously. If $a, c, L \in \mathbb{R}$, $c < a$ and f is a real function defined on (c, a), we say that $\lim_{x \to a^-} f(x) = L$ if

$$(\forall \varepsilon > 0)(\exists \delta > 0)[x \in (a-\delta, a)] \Rightarrow [\mid f(x) - L \mid < \varepsilon].$$

Right-hand limits and left-hand limits are called one-sided limits. One sided limits are examples of restricted limits.

EXAMPLE 5.17. Let $H(t)$ be the Heaviside function. Then

$$\begin{aligned} \lim_{t \to 0^+} H(t) &= 1 \\ \lim_{t \to 0^-} H(t) &= 0. \end{aligned}$$

5.2. Continuity

Most functions you have encountered have the property that at (almost) every point the function has a limit that agrees with its value there. This is a very useful feature of a function, and it is called *continuity*.

DEFINITION. Continuous Let f be a real function with domain $X \subseteq \mathbb{R}$. Let $a \in X$. Then we say f is continuous at a if $\lim_{X \ni x \to a} f(x) = f(a)$. We say f is continuous on X if it is continuous at every point of X.

Intuitively, the idea of a continuous function on an interval is that it has no jumps. We shall make this precise in Chapter 8 when we prove the Intermediate Value Theorem 8.10, which asserts that if a continuous function on an interval takes on two distinct values c and d, it must also take on every value between c and d.

EXAMPLE 5.18. Prove that the function $f(x) = x^2$ is continuous on $\mathbb{R}$.