Intuitive Set Theory

From Membership to the Transfinite

ADUM Project

Ariel Daley Undergraduate Mathematics Project

Contact: ariel@ly4i.com

Contents

Preface vii

I Learning the Language of Sets 1

1 Entering Mathematical Language 3
1.1 Why Set Theory Begins in Ordinary Language . . . 3
1.2 Propositions, Connectives, and Quantifiers . . . 5
1.3 Direct Proof, Contrapositive, Contradiction, and Cases . . . 8
1.4 Definitions, Theorems, Examples, and Numbering Conventions . . . 11
1.5 The Informal Viewpoint and the Promise of Later Axioms . . . 14

2 Sets, Subsets, and Elementary Operations 17
2.1 Membership and Extensionality . . . 17
2.2 Empty Set, Singletons, Pair Sets, and Subsets . . . 20
2.3 Union, Intersection, Difference, and Complement . . . 22
2.4 Power Sets . . . 25
2.5 The Algebra of Sets . . . 27

3 Ordered Pairs, Cartesian Products, and Functions 33
3.1 Ordered Pairs and Cartesian Products . . . 33
3.2 Functions as Assignments and as Graphs . . . 36
3.3 Domain, Codomain, Image, and Preimage . . . 38
3.4 Composition, Identity, Inverse, Injection, Surjection, Bijection . . . 41
3.5 Equality of Functions and Extensional Reasoning . . . 45

4 Families of Sets: General Unions, Intersections, and Products 47
4.1 Indexed Families . . . 47
4.2 General Unions and General Intersections . . . 49
4.3 Disjoint Unions and Partitions of a Set . . . 53
4.4 General Cartesian Products . . . 56
4.5 Choice Functions as a First Preview . . . 59

5 Relations, Equivalence, and Order 63
5.1 Relations and Their Basic Properties . . . 64
5.2 Equivalence Relations and Partitions . . . 67
5.3 Partial Orders . . . 71
5.4 Total Orders, Lexicographic Orders, and Hasse Diagrams . . . 73

5.5 Well-Ordered Sets as a Preview of the Transfinite . . . 76

II Numbers Built from Sets and the First Infinite Worlds 79

6 The Natural Numbers as Sets 81
6.1 Why Define Numbers Inside Set Theory? . . . 82
6.2 Successor and the von Neumann Naturals . . . 83
6.3 Induction . . . 85
6.4 Recursion on the Natural Numbers . . . 88
6.5 Addition, Multiplication, and Exponentiation on $\mathbb{N}_0$ . . . 91

7 Finite Sets and Counting 99
7.1 Bijections and the Meaning of "Same Number of Elements" . . . 99
7.2 Finite Sets . . . 101
7.3 The Pigeonhole Principle . . . 106
7.4 Counting Functions, Subsets, and Permutations of Finite Sets . . . 108
7.5 Where Finite Intuition Begins to Fail . . . 112

8 Infinite, Countable, and Countably Infinite Sets 115
8.1 Infinite Sets and Dedekind's Idea . . . 116
8.2 Countably Infinite and Countable Sets . . . 117
8.3 Standard Countable Sets . . . 120
8.4 Countable Unions and Products . . . 123
8.5 Lists, Enumerations, and the Shape of Countability . . . 124

9 Uncountability and the Power Set 127
9.1 Cantor's Diagonal Argument . . . 127
9.2 The Real Numbers Are Uncountable . . . 129
9.3 Cantor's Theorem for Power Sets . . . 131
9.4 Binary Sequences, Intervals, and the Continuum . . . 132
9.5 The New Landscape of Infinite Size . . . 136

III Choice and the Transfinite Viewpoint 139

10 The Axiom of Choice 141
10.1 Why Choice Appears Naturally . . . 142
10.2 The Axiom of Choice and Choice Functions . . . 143
10.3 Equivalent Forms: Well-Ordering and Zorn's Lemma . . . 145
10.4 Representatives, Products, and Maximal Principles . . . 151
10.5 What Choice Changes—and What It Does Not . . . 154

11 Well-Ordered Sets and Ordinal Numbers 157
11.1 Order Isomorphism and Order Type . . . 157
11.2 Well-Ordered Sets and Initial Segments . . . 160
11.3 Ordinals as Transitive Well-Ordered Sets . . . 163
11.4 Finite Ordinals, ω, Successor Ordinals, and Limit Ordinals . . . 167
11.5 Suprema of Sets of Ordinals . . . 170

12 Transfinite Induction, Recursion, and Ordinal Arithmetic **175**
12.1 Transfinite Induction . . . 176
12.2 Transfinite Recursion . . . 177
12.3 Ordinal Addition and Multiplication . . . 180
12.4 Ordinal Exponentiation . . . 186
12.5 Countable Ordinals and the First Uncountable Ordinal . . . 188

IV Cardinality Beyond Countability **193**

13 Cardinal Numbers **195**
13.1 Equinumerosity, Injections, Surjections, and Comparison . . . 195
13.2 The Cantor–Bernstein Theorem . . . 199
13.3 Cardinals as Initial Ordinals . . . 201
13.4 Aleph Numbers and Familiar Infinite Cardinals . . . 204
13.5 Hartogs' Theorem as a Glimpse of Deeper Structure . . . 205

14 Cardinal Arithmetic **209**
14.1 Sum, Product, and Exponentiation of Cardinals . . . 209
14.2 Finite versus Infinite Arithmetic . . . 214
14.3 Infinite Cardinal Arithmetic with Choice . . . 216
14.4 The Continuum and $2^{\aleph_0}$ . . . 220
14.5 A Glimpse of the Continuum Hypothesis and Independence . . . 222

V Foundations, Axioms, and Outlook **225**

15 From Intuition to Axioms: ZF and ZFC **227**
15.1 Why Naive Set Theory Needs Restraint . . . 228
15.2 Russell's Paradox and Related Warnings . . . 229
15.3 The Core Axioms: Extensionality, Empty Set, Pairing, Union, Power Set, Infinity . . . 231
15.4 Separation, Replacement, and Foundation . . . 234
15.5 Choice, Classes, and the Picture of ZF versus ZFC . . . 237

16 Looking Back and Further Directions **241**
16.1 From Counting to the Transfinite . . . 241
16.2 Set Theory as a Language for the Rest of Mathematics . . . 244
16.3 Independence, Large Cardinals, and Further Foundational Questions . . . 246
16.4 Further Directions in Set Theory . . . 248
16.5 Suggested Further Reading . . . 251

Bibliography **253**

Index **255**

Preface

What this book is about

Set theory is one of the simplest subjects in mathematics to state and one of the deepest subjects to pursue. At first sight it seems to ask for very little. We speak of collections of objects, of whether an object belongs to a collection, of whether two collections are equal, and of how one collection may be contained in another. Soon after that, however, the subject expands. We begin to speak about functions between sets, relations on sets, ways of arranging sets in order, methods for defining the natural numbers inside set theory itself, and ways of comparing infinite collections whose elements can never be counted one by one in practice. By the time we reach the later parts of the book, we are discussing the axiom of choice, well-ordered sets, ordinal numbers, cardinal numbers, and the first glimpse of the axiomatic foundations that keep the whole subject from collapsing into paradox.

This book is an introduction to that arc. It begins from the *intuitive* viewpoint that most readers naturally bring to the subject: a set is a collection, a function is a rule or assignment, an order tells us what comes before what, and counting is a way of comparing sizes. Those first intuitions are not the whole story, but neither are they a mistake. They are the natural point of entry. The aim of the book is to take those first ideas seriously, to refine them slowly, and to lead the reader from familiar finite examples to the first genuinely transfinite ones.

Two summits guide the narrative. The first is Cantor's theory of size: finite and infinite sets, countable and uncountable sets, cardinal numbers, and cardinal arithmetic. The second is the transfinite viewpoint: well-ordered sets, ordinal numbers, transfinite induction, and transfinite recursion. These two themes are different but deeply intertwined. Cardinals tell us how large a set is, ordinals tell us how a well-ordered process is arranged, and both are central to the modern language of mathematics.

The book is therefore about more than just notation or elementary set manipulation. It is about learning to think structurally. When we ask whether two sets have the same size, we shall learn that the correct comparison is often not literal equality but the existence of a bijection. When we ask what it means to multiply infinitely many sets or take the union of an indexed family, we shall learn that functions quietly stand behind the definitions. When we ask why unrestricted phrases such as "the set of all objects with a given property" can be dangerous, we shall discover that intuition needs axiomatic guidance.

> *We begin with intuition not because rigor is unimportant, but because intuition is how a subject first becomes meaningful. We return to axioms later not because the early intuition was worthless, but because mathematics eventually asks us to say exactly what our intuition is allowed to mean.*

For whom this book is written

This book is written for a reader who may be meeting proof-based university mathematics for the first time. We assume only a modest technical background: high-school calculus, routine algebraic manipulation, and elementary matrix operations such as matrix addition and multiplication. We do *not* assume prior knowledge of set theory, formal logic, linear algebra, abstract algebra, topology, real analysis, or the style of reading and writing proofs that is standard in later undergraduate mathematics.

That assumption affects the exposition at every stage. We do not rush past the language of statements, quantifiers, and implication. We do not treat the first proof methods as obvious. We do not assume that a reader already knows how to interpret the difference between a definition and a theorem, or how a theorem number is meant to function inside a textbook. When a new concept first appears, we try to say not only what it is, but also why one might care about it and what familiar examples should be kept in mind.

At the same time, this is not a watered-down book. The intended reader is a beginner, but the mathematics is real mathematics. The subject will gradually become deeper: countability gives way to uncountability, finite induction gives way to transfinite induction, and the ordinary idea of "size" is refined into several different notions that must be handled with care. A first-year student can learn such material, but only if the road toward it is laid carefully. That is the road this book is trying to build.

A more advanced reader may also use the book profitably, especially if that reader wants a slower and more conceptually staged entry into set theory than many standard texts provide. The early chapters deliberately spend time on examples, language, and proof patterns that more advanced books often compress into a few pages.

How the book is organized

The book is divided into five parts.

Part	*Chapters*	*Main themes*
Learning the Language of Sets	1–5	Mathematical language, proof methods, basic set operations, ordered pairs, Cartesian products, functions, indexed families, relations, and order.
Numbers Built from Sets and the First Infinite Worlds	6–9	The natural numbers as sets, finite sets and counting, infinite and countable sets, and the first major uncountability results.
Choice and the Transfinite Viewpoint	10–12	Choice functions, the axiom of choice, well-orders, ordinals, transfinite induction, recursion, and ordinal arithmetic.
Cardinality Beyond Countability	13–14	Cardinal numbers, comparison of sizes, Cantor–Bernstein, Hartogs' theorem, and cardinal arithmetic.
Foundations, Axioms, and Outlook	15–16	Russell's paradox, the basic axioms of ZF and ZFC, the cumulative hierarchy, and directions for further study.

Part I is deliberately slow. Before we ask difficult questions about infinite sets, we must learn to read mathematical statements carefully and to manipulate basic objects with confidence. Chapters 1 through 5 therefore establish the language of the subject: propositions and quantifiers, sets and subsets, functions, indexed families, and relations. These chapters are not mere preliminaries to be rushed through. They teach the habits on which everything later depends.

Part II begins the first substantial ascent. In Chapter 6 we define the natural numbers inside set theory, not because this is the only way to understand numbers, but because it reveals how much arithmetic can be built from a small amount of set-theoretic structure. Chapters 7, 8, and 9 then develop the first theory of infinity: finite sets, countably infinite sets, countable sets, and uncountable sets. The contrast between what can be listed and what cannot be listed is one of the turning points of the subject.

Part III introduces the axiom of choice and the transfinite viewpoint. This is where many students feel that set theory becomes both more powerful and more surprising. We learn that well-ordered sets can be studied through their order types, that transfinite induction really is a natural extension of ordinary induction, and that some constructions with infinitely many sets depend on a choice principle that is far from trivial.

Part IV returns to the question of size at a deeper level. It is one thing to know that some sets are uncountable and others are not. It is another to compare large infinities systematically and perform arithmetic with them. Cardinals make that possible.

Part V steps back from the intuitive development and asks what justifies it. There we explain why naive comprehension must be restricted, why modern set theory is usually developed axiomatically, and how the subject opens onto further directions. This final part is not a retreat from intuition; it is a clarification of intuition.

How to read this book

A beginning reader often feels that a mathematics textbook is asking for two kinds of learning at once. One must learn the mathematics, but one must also learn how a mathematics textbook is organized. That is normal. If this is your first proof-based book, some of the difficulty comes not from the concepts themselves, but from the unfamiliar format in which they are presented.

A *definition* tells us what a word or phrase means. We do not prove a definition; we learn how to use it. A *theorem* is a statement that has to be justified. A *proof* is the justification. An *example* shows the definition or theorem at work in a concrete case. A *remark* usually explains how to interpret what has just happened, warns about a common misunderstanding, or places the result in a broader setting. These four ingredients play different roles, and it is worth learning to recognize those roles early.

The numbering of items is purely a navigational aid. If you see a label such as *Theorem 1.1.3*, the string "1.1.3" does *not* mean that the number is prime, mysterious, or worthy of memorization in itself. It simply means that you are looking at the third numbered item in Chapter 1, Section 1. Another book may use a different system. Some books number results only by chapter, some by section, and some do not number examples at all. The important thing is not to memorize the numbers, but to know how to use them to find your way back.

For many readers, the most useful strategy is to read in layers. On a first pass through a section, try to identify the main ideas: what is being defined, what is being proved, and what examples are supposed to stay in mind. On a second pass, fill in the details of the proofs. When a proof feels difficult, it often helps to rewrite the statement in your own words, list the

hypotheses separately from the conclusion, and test the statement on a small example before reading further.

Examples are not decorative. In a subject such as set theory, an example is often where the definition first becomes intelligible. If a definition concerns unions of indexed families, try it first when the index set has two or three elements. If a theorem concerns countable sets, test it on the integers or the rational numbers before worrying about the general form. We repeatedly move from small explicit cases to general statements because that is how the subject is most naturally learned.

This book does not contain formal exercise sets at the ends of chapters. Instead, it contains many examples, remarks, and brief "invitations to compute" inside the exposition. A reader who wants to learn actively should stop at those moments and carry out the small verifications. Set theory becomes much clearer when one actually checks, for instance, what a power set looks like for a three-element set, how a relation on a finite set decomposes into equivalence classes, or why a proposed enumeration of a family of sets fails.

A final practical remark: do not be discouraged if the early pages feel slower than you expected. They are meant to be slow. Chapter 1 is not postponing the mathematics; it is teaching the language in which the rest of the mathematics will be written.

Why we begin informally

A completely formal treatment of set theory is usually placed inside a larger logical framework. One specifies a formal language, a deductive system, and an axiomatic theory, and then one proves theorems within that system. This is an important subject. It belongs to the foundations of mathematics and deserves careful study in its own right. But it is not the easiest way to meet sets for the first time.

A beginner usually understands the phrase "the set of even integers" long before understanding what it means to formalize predicates, separation schemes, or satisfaction in a first-order structure. It would therefore be pedagogically backwards to insist that every ounce of formal background be mastered before any real set-theoretic ideas are allowed to appear. In this book we choose the opposite order. We start with the intuitive language that working mathematicians actually use in their everyday reasoning, while remaining honest that this language has limits.

Those limits matter. If one allows absolutely every property to define a set, contradictions arise. The later discussion of Russell's paradox shows this vividly. So the informal viewpoint is not the final viewpoint. It is a carefully managed beginning. We shall speak, early on, as if sets are simply collections and as if any familiar collection can be discussed without anxiety. In ordinary mathematical practice, that level of informality is usually harmless and often illuminating. But we keep in mind from the start that something more disciplined will be needed.

The role of Chapter 15 is precisely to make good on that promise. By the time we reach the axioms of ZF and ZFC, the reader will have seen many constructions already: unions, power sets, functions, products, relations, natural numbers, ordinals, and cardinals. The axioms then arrive not as abstract bureaucratic rules, but as a way of explaining which earlier moves are legitimate and why. This, we hope, makes the axiomatic viewpoint feel earned rather than imposed.

> *The book begins by saying, in effect, "let us learn how sets behave." It ends by asking, "what assumptions allow us to say all of that safely?" The distance between those two questions is*

the educational journey of the book.

Notation and conventions

A few conventions should be stated at the outset.

In this book,

$$\mathbb{N} = \{1, 2, 3, \dots\} \qquad \text{and} \qquad \mathbb{N}_0 = \{0, 1, 2, 3, \dots\}.$$

This distinction is deliberate. Many authors include 0 in the natural numbers and many authors do not. Both conventions are common. To avoid unnecessary ambiguity, we use $\mathbb{N}$ for the positive integers and $\mathbb{N}_0$ when we explicitly want the version that begins at 0. This becomes especially useful when we define the von Neumann natural numbers inside set theory.

When we introduce a function formally, we write it in the style $f : A \to B$. The set A is the domain, the set B is the codomain, and the image of a subset will later be distinguished carefully from the codomain itself. Set-builder notation will usually appear in the form

$$\{x \in A \mid P(x)\},$$

which should be read as "the elements x of A such that $P(x)$ holds."

The word "or" is inclusive unless we explicitly say otherwise. Thus, if we say that an integer is "even or divisible by three," we allow the possibility that both properties hold. Logical phrases such as "for every" and "there exists" will appear constantly. One of the aims of the opening chapter is to make such phrases feel natural rather than intimidating.

Finally, equality in mathematics must always be read with care. Two sets are equal when they have the same elements. Two ordered pairs are equal when both coordinates agree in the correct order. Two functions are equal when they have the same domain, the same codomain when the context requires it, and the same value at every input. Different objects are compared in different ways, and set theory becomes clearer once those comparisons are kept distinct.

A final word before Chapter 1

Set theory sometimes creates a curious double impression. On one hand, its basic statements can look almost childlike: a point is either in a set or not in it; a subset is contained in a larger set; a function assigns one output to each input. On the other hand, the generality of these ideas quickly leads to questions of real depth. How many real numbers are there? Can every set be well-ordered? Is there a set of all ordinals? Can one speak of "the set of all sets" without running into contradiction? These are not elementary questions merely because their wording is simple.

The chapters that follow therefore ask for patience and curiosity in equal measure. We begin with the language of mathematics itself, because learning that language well is the surest way to make the later ideas accessible. If at some point a definition feels abstract, return to the nearby examples. If a proof feels long, identify first what it is trying to prove before worrying about every detail. If a theorem about infinity feels surprising, remember that surprise is often a sign that the subject is beginning to show its true character.

We now turn to Chapter 1, where the first task is not yet to manipulate large infinite sets, but to learn how mathematical statements, definitions, and proofs are read. That is the right beginning. Once we can read the language with confidence, the sets themselves can begin to speak.

Part I

Learning the Language of Sets

Chapter 1

Entering Mathematical Language

Set theory is often introduced as the study of *sets*, that is, of collections of objects. But before we can speak comfortably about collections, membership, functions, or infinite processes, we must learn to read the language in which mathematics is written. A beginning student sometimes feels that a university mathematics book is not merely teaching new ideas; it is also written in a slightly unfamiliar dialect. Definitions appear in boxes, theorems are numbered, proofs unfold line by line, and small words such as "if," "for every," and "there exists" suddenly carry a great deal of weight.

This chapter is meant to slow that experience down. We shall not assume that the reader already knows how proof-based mathematics is supposed to look. Instead, we begin with the sentences one already meets in school mathematics: "every square has four sides," "there is a prime number greater than one hundred," "if a number is divisible by four, then it is even." Such sentences already contain logical structure. They have subjects, predicates, conditions, quantifiers, and hidden choices of what objects we are talking about. Set theory will later make many of these matters explicit, but the first step is simply to notice that ordinary mathematical language is more structured than it first appears.

There is another reason for beginning here. This book develops set theory first at an *intuitive* level. We shall say things such as "the set of divisors of 12" or "the set of real numbers between 0 and 1" long before we give an axiomatic account of which collections are allowed to count as sets. That is deliberate. A student meeting the subject for the first time usually understands examples and patterns before formal foundations. Still, we must be honest: unrestricted talk about "the set of all objects with a given property" eventually leads to trouble. So this chapter also begins a promise that will be fulfilled much later: we shall work informally at first, but we shall not forget that a later axiomatic treatment is needed.

The five sections of this chapter therefore play a double role. They teach the reader how to read the book, and they also introduce the logical habits on which the rest of the subject depends. We begin with propositions and predicates, move to connectives and quantifiers, study several common proof patterns, explain how textbook structure should be read, and close by clarifying why informal set-theoretic talk is useful at the beginning but insufficient at the end.

1.1 Why Set Theory Begins in Ordinary Language

A first encounter with set theory should not begin by pretending that mathematics is made only of symbols. Long before one learns formal logic, one already understands many mathematical statements in plain language. The job of this section is to make that familiar language more explicit. In particular, we want to distinguish complete statements from expressions that still

depend on a variable, and we want to notice that every mathematical sentence is spoken within some background collection of objects.

Statements and open sentences

The simplest building blocks of mathematical discourse are sentences that say something definite.

Definition 1.1.1 (Proposition)**.** A *proposition* is a declarative sentence that is either true or false.

The point of the definition is modest but important. A proposition is not required to be easy to decide, and we do not need to know whether it is true or false at the moment we write it down. It is enough that it *has* a truth value.

Example 1.1.2. Each of the following is a proposition.

(i) "7 is a prime number."

(ii) "$12 < 5$."

(iii) "There exists an integer whose square is 49."

The first and third are true, while the second is false.

By contrast, the following are *not* propositions.

(iv) "Is 7 prime?"

(v) "Please compute $3 + 4$."

(vi) "$x + 3 = 5$."

The first is a question, the second is a command, and the third still depends on the unspecified symbol x.

The last item is especially important. In mathematics we constantly write expressions such as "x is even," "$x^2 = 1$," or "$x < y$." These are not yet full propositions. They become propositions only after the variables are specified or quantified.

Definition 1.1.3 (Predicate, or open sentence)**.** A *predicate* (or *open sentence*) is an expression involving one or more variables that becomes a proposition when specific values are substituted for those variables.

Example 1.1.4. Let $P(n)$ be the sentence "n is even," where n is an integer. Then $P(6)$ is the proposition "6 is even," which is true, while $P(7)$ is the proposition "7 is even," which is false.

Similarly, let $Q(x)$ be the sentence "$x^2 = 1$." Then $Q(1)$ and $Q(-1)$ are true propositions, whereas $Q(2)$ is a false proposition.

Predicates will reappear throughout the book. In Chapter 2 we shall write sets in forms such as

$$\{x \in A \mid P(x)\},$$

which means "the elements x of A for which the predicate $P(x)$ is true." For the moment, it is enough to see that the language of properties already stands in the background.

The ambient domain matters

Mathematical sentences are not spoken in a vacuum. One must know what sorts of objects the variables are supposed to denote.

Definition 1.1.5 (Domain of discourse)**.** The *domain of discourse* of a statement is the collection of objects from which its variables are understood to range.

Example 1.1.6. Consider the sentence

$$\text{"There exists } x \text{ such that } x^2 = 2\text{."}$$

Its truth depends on the domain of discourse.

(i) If the domain is $\mathbb{Q}$, the sentence is false.

(ii) If the domain is $\mathbb{R}$, the sentence is true.

So the same symbolic expression can represent different propositions when the background domain changes.

This observation may seem elementary, but it is one of the main reasons set theory matters. Later we shall insist on naming the set or class of objects under discussion, because many statements become clearer once their ambient domain is made explicit.

Remark 1.1.7. Ordinary mathematical English already contains the seeds of logic. When we say "every integer has a successor" or "some polynomial has no real root," we are already using quantifiers, variables, and an implicit domain of discourse. Formal symbolism does not replace this language; it refines it.

1.2 Propositions, Connectives, and Quantifiers

To read mathematical prose with confidence, one must learn to hear the logical shape of a sentence. Small words such as "not," "and," "or," "if," and "for every" are not decorative. They tell us how simpler statements are being assembled into more complicated ones. This section introduces the basic logical connectives and the two most important quantifiers. We keep the discussion intuitive, but we do so carefully enough that these ideas can be used immediately in proofs.

Connectives

Definition 1.2.1 (Logical connectives)**.** Let P and Q be propositions.

(i) The *negation* of P, written $\neg P$, is the proposition "not P."

(ii) The *conjunction* of P and Q, written $P \wedge Q$, is the proposition "P and Q."

(iii) The *disjunction* of P and Q, written $P \vee Q$, is the proposition "P or Q," where "or" is inclusive unless stated otherwise.

(iv) The *implication* from P to Q, written $P \to Q$, is the proposition "if P, then Q."

(v) The *biconditional* $P \iff Q$ means "P if and only if Q."

The first three items are usually read without difficulty, but the last two deserve more attention. An implication $P \to Q$ makes a promise: whenever the condition P happens, the conclusion Q must follow. The biconditional $P \iff Q$ is stronger; it asserts that each of the two statements implies the other.

For quick reference, the most common connectives can be summarized as follows.

Symbol	*How to read it*	*What it asserts*
$\neg P$	not P	P is false
$P \wedge Q$	P and Q	both are true
$P \vee Q$	P or Q	at least one is true
$P \to Q$	if P, then Q	Q follows from P
$P \iff Q$	P iff Q	each implies the other

Example 1.2.2 (Reading an implication)**.** Consider the statement

"If an integer is divisible by 4, then it is even."

This is an implication of the form $P \to Q$, where

$$P: \text{ "the integer is divisible by 4,"}$$
$$Q: \text{ "the integer is even."}$$

To show that this statement is true, one must explain why every integer satisfying P also satisfies Q. To show that it is false, one would need an integer for which P is true but Q is false.

Definition 1.2.3 (Converse)**.** The *converse* of an implication $P \to Q$ is the implication $Q \to P$.

Remark 1.2.4. A statement and its converse need not have the same truth value. "If an integer is divisible by 4, then it is even" is true, but its converse "if an integer is even, then it is divisible by 4" is false. The integer 6 is a counterexample to the converse.

Remark 1.2.5 (On inclusive "or"). In mathematics, "or" is usually inclusive. Thus "n is even or n is a multiple of 3" means that at least one of the two properties holds, and possibly both. If an exclusive meaning is intended, it must be stated explicitly.

Quantified statements

The next step is to bind variables by saying how broadly a claim is supposed to range.

Definition 1.2.6 (Quantifiers)**.** Let $P(x)$ be a predicate on a domain D.

(i) The statement
$$\forall x \in D,\ P(x)$$
is read "for every x in D, $P(x)$ holds." This is the *universal quantifier*.

(ii) The statement
$$\exists x \in D \text{ such that } P(x)$$
is read "there exists an x in D such that $P(x)$ holds." This is the *existential quantifier*.

More briefly, we often refer to $\forall$ and $\exists$ as *quantifiers*.

Example 1.2.7 (Quantifier order matters)**.** The two statements
$$\forall n \in \mathbb{N}\, \exists m \in \mathbb{N} \text{ such that } m > n$$
and
$$\exists m \in \mathbb{N}\, \forall n \in \mathbb{N} \text{ such that } m > n$$
look similar, but they say different things.

The first is true: given any natural number n, we may choose $m = n + 1$. The second is false: it would assert that there is one natural number larger than every natural number.

The lesson is simple and important: changing the order of quantifiers can change the meaning of a sentence completely.

A reader should become comfortable translating ordinary mathematical sentences into quantified form and then back again. That is often the fastest way to understand what a theorem is really saying.

Proposition 1.2.8 (Negating quantified statements)**.** *Let $P(x)$ be a predicate on a domain D. Then:*

(i) $\neg(\forall x \in D,\ P(x))$ *is equivalent to* $\exists x \in D$ *such that* $\neg P(x)$.

(ii) $\neg(\exists x \in D$ *such that* $P(x))$ *is equivalent to* $\forall x \in D,\ \neg P(x)$.

Proof. The first statement says that it is *not* the case that every object in D satisfies P. But that means exactly that at least one object in D fails to satisfy P; in other words, there exists $x \in D$ such that $\neg P(x)$ holds.

The second statement says that there does *not* exist any object in D satisfying P. This is the same as saying that every object in D fails to satisfy P, that is, $\forall x \in D, \neg P(x)$. □

Example 1.2.9. The negation of

"Every real number has a real square root"

is

"There exists a real number that does not have a real square root."

The number -1 shows that the original statement is false.

Proposition 1.2.10 (Witnesses and counterexamples)**.** *Let $P(x)$ be a predicate on a domain D.*

(i) To prove $\exists x \in D$ such that $P(x)$, it is enough to produce a specific element $a \in D$ for which $P(a)$ is true.

(ii) To disprove $\forall x \in D,\ P(x)$, it is enough to produce a specific element $a \in D$ for which $P(a)$ is false.

Proof. For (i), a single example with property P is exactly what an existential statement asks for. Such an example is often called a *witness*.

For (ii), if we can produce an element a with $\neg P(a)$, then by Proposition 1.2.8 the universal statement $\forall x \in D, P(x)$ cannot be true. Such an element is called a *counterexample*. □

Example 1.2.11. The number 11 is a witness for the true statement

$$\exists n \in \mathbb{N} \text{ such that } n > 10 \text{ and } n \text{ is prime.}$$

The number 2 is a counterexample to the false statement

$$\forall n \in \mathbb{N}, \text{ if } n \text{ is prime then } n \text{ is odd.}$$

Remark 1.2.12 (Necessary and sufficient conditions). When we say "P is sufficient for Q," we mean that $P \to Q$. When we say "P is necessary for Q," we mean that $Q \to P$. Thus a biconditional $P \iff Q$ says that each of P and Q is both necessary and sufficient for the other.

Remark 1.2.13 (A mild warning about implication). Students sometimes expect that an implication $P \to Q$ says something only when P actually occurs. In logic, however, an implication is false only in one situation: P is true and Q is false. This is why a statement such as

$$\forall n \in \mathbb{N}, \text{ if } n < 0 \text{ then } n^2 = 17$$

is considered true: there is no natural number n with $n < 0$. This phenomenon is called *vacuous truth*. The terminology may sound odd at first, but it becomes natural with use.

1.3 Direct Proof, Contrapositive, Contradiction, and Cases

Once we understand the logical form of a statement, the next question is how to prove it. A proof is not a mysterious performance reserved for experts. It is a chain of reasons. Some proofs begin directly from the hypotheses, some prove an equivalent contrapositive, some assume the negation and derive an impossibility, and some split the problem into a small number of natural cases. The examples in this section use only elementary arithmetic, because the goal is to make proof structure visible before the mathematics becomes more elaborate.

What a proof is

Definition 1.3.1 (Proof)**.** A *proof* of a proposition is a logically coherent argument that shows the proposition must be true, using accepted facts, definitions, and previously established results.

A proof is not merely a sequence of true sentences. Its steps must be connected. Each sentence must either come from the hypotheses, follow from an earlier line, or appeal to a known fact. One also learns what *does not* count as a proof: checking a few examples, drawing a suggestive picture, or saying that something is "obvious" when the reason has not been explained.

Definition 1.3.2 (Common proof patterns)**.** Among the most common proof patterns are the following.

(i) A *direct proof* begins from the assumptions and derives the desired conclusion.

(ii) A *proof by contrapositive* proves $P \to Q$ by showing the logically equivalent statement $\neg Q \to \neg P$.

(iii) A *proof by contradiction* assumes that the desired conclusion is false and then derives an impossibility.

(iv) A *proof by cases* splits the problem into several possibilities and proves the statement in each case.

Direct proof

Proposition 1.3.3. *If m and n are even integers, then $m + n$ is even.*

Proof. Because m is even, there exists an integer a such that $m = 2a$. Because n is even, there exists an integer b such that $n = 2b$. Therefore

$$m + n = 2a + 2b = 2(a + b).$$

Since $a + b$ is an integer, the number $m + n$ has the form $2k$ for some integer k. Hence $m + n$ is even. □

Example 1.3.4. Take $m = 8$ and $n = 14$. Then $m + n = 22$, which is even. Of course, this single calculation does not prove Proposition 1.3.3; the proof works for all even integers because it explains the general structure.

The proof above illustrates a standard habit of mathematical writing: we begin by *unpacking the definitions*. To use the fact that m is even, we rewrite it in the defining form $m = 2a$. Many elementary proofs are nothing more than the patient and orderly use of definitions.

Contrapositive

Proposition 1.3.5 (Contrapositive equivalence)**.** *For any propositions P and Q, the implication $P \to Q$ is logically equivalent to the implication $\neg Q \to \neg P$.*

Proof. Suppose first that $P \to Q$ is true. To show $\neg Q \to \neg P$, assume $\neg Q$. If P were true, then by $P \to Q$ the statement Q would also be true, contradicting $\neg Q$. Hence P cannot be true, so $\neg P$ holds.

Conversely, suppose that $\neg Q \to \neg P$ is true. To prove $P \to Q$, assume P. If Q were false, then $\neg Q$ would hold, and our hypothesis would imply $\neg P$, contradicting the assumption P. Therefore Q must be true. So $P \to Q$ holds. □

Theorem 1.3.6. *Let n be an integer. If n^2 is even, then n is even.*

Proof. Instead of proving the statement directly, we prove its contrapositive: if n is odd, then n^2 is odd.

Assume that n is odd. Then $n = 2k + 1$ for some integer k. Therefore

$$n^2 = (2k + 1)^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1.$$

This has the form $2\ell + 1$ for an integer ℓ, so n^2 is odd. By Proposition 1.3.5, the original implication follows. □

Corollary 1.3.7. *For every integer n, the following are equivalent:*

(i) n is even;

(ii) n^2 is even.

Proof. We must prove both directions.

If n is even, then $n = 2k$ for some integer k, and hence

$$n^2 = (2k)^2 = 4k^2 = 2(2k^2),$$

so n^2 is even.

Conversely, if n^2 is even, then by Theorem 1.3.6 the integer n is even. Therefore (i) and (ii) are equivalent. □

Example 1.3.8. The corollary says, for example, that 15^2 is odd because 15 is odd, and that if someone tells us an integer square is even, then we already know the original integer must have been even.

Contradiction

Theorem 1.3.9. *There are infinitely many prime numbers.*

Proof. We argue by contradiction. Assume that there are only finitely many prime numbers. Then we can list them as

$$p_1, p_2, \ldots, p_r.$$

Now form the integer

$$N = p_1 p_2 \cdots p_r + 1.$$

Because $N > 1$, it has a prime divisor. Since our list was supposed to contain every prime number, this prime divisor must be one of the p_i.

But if p_i divides both $p_1 p_2 \cdots p_r$ and N, then p_i also divides their difference:

$$N - p_1 p_2 \cdots p_r = 1.$$

No prime divides 1, so this is impossible. The assumption that there are only finitely many primes must therefore be false. Hence there are infinitely many prime numbers. □

Example 1.3.10. If we start with the first four primes $2, 3, 5, 7$, then

$$2 \cdot 3 \cdot 5 \cdot 7 + 1 = 211.$$

In this case 211 is itself prime. The theorem does *not* say that the number obtained in Euclid's construction is always prime; it says only that it has a prime divisor not already on the list.

Remark 1.3.11 (A classical proof). The proof of Theorem 1.3.9 goes back to Euclid. It is often one of the first examples that shows students how powerful a contradiction argument can be: a very simple assumption, pushed carefully, destroys itself.

Proof by cases

Proposition 1.3.12. *For every integer n, the product $n(n+1)$ is even.*

Proof. Every integer is either even or odd.

If n is even, then $n(n+1)$ is even because it has an even factor.

If n is odd, then $n+1$ is even, so again $n(n+1)$ is even because it has an even factor.

In both cases the product is even, and therefore the proposition holds for every integer n. □

Example 1.3.13. When $n = 6$, the product is $6 \cdot 7 = 42$, which is even. When $n = 7$, the product is $7 \cdot 8 = 56$, which is also even. The proof explains why this must happen regardless of which integer n we choose.

Remark 1.3.14 (How to prove "if and only if"). A biconditional $P \iff Q$ should almost always be read as two separate tasks: prove $P \to Q$, and then prove $Q \to P$. Corollary 1.3.7 is a model example. One direction was a direct proof, and the other used a contrapositive.

Remark 1.3.15 (Further practice). Readers who want a more systematic introduction to proof patterns, especially with many short exercises, may find Velleman's text [1] very helpful. The present chapter is not a substitute for practice; it is a guided beginning.

1.4 Definitions, Theorems, Examples, and Numbering Conventions

A proof-based textbook is not just a list of facts. It is a carefully organized conversation. Definitions introduce language, theorems make claims in that language, proofs justify those claims, examples show how the ideas behave in concrete cases, and remarks supply context, warnings, or perspective. A beginning student often tries to read only the statements and skip the examples or remarks. That is usually a mistake. The examples are where the language starts to feel real.

What the common environments are for

The following table summarizes the most common pieces of structure that will appear throughout this book.

Name	*Typical role in the exposition*
Definition	introduces a term or concept precisely
Theorem	records an important result
Proposition	states a useful but usually smaller result
Corollary	gives an immediate consequence of a theorem
Example	shows the idea in a concrete situation
Remark	adds context, warning, or interpretation
Proof	explains why a statement is true

These names are not absolute laws of nature. One author's "proposition" may be another author's "theorem." What matters is the role the item plays in the exposition.

Definition 1.4.1 (Counterexample)**.** A *counterexample* to a universal statement is a specific example for which that statement fails.

Example 1.4.2. The false statement

"Every prime number is odd"

has the counterexample 2.

The false statement

"Every real number is positive"

has many counterexamples, for instance -1 and 0.

Proposition 1.4.3. *If there exists an element a in the domain of discourse such that $P(a)$ is false, then the universal statement $\forall x, P(x)$ is false.*

Proof. This is exactly the content of the negation rule in Proposition 1.2.8. If $P(a)$ is false for some a, then $\exists x\, \neg P(x)$ holds. Therefore $\forall x, P(x)$ cannot be true. □

Remark 1.4.4. One counterexample is enough to destroy a universal statement, but no finite number of favorable examples proves such a statement. Seeing that $n^2 + n + 41$ is prime for many small values of n may be suggestive, but suggestion is not proof.

How numbering works

A student opening a mathematics book for the first time is often surprised by labels such as "Theorem 1.3.4" or "Definition 2.1.7." There is nothing mystical about them. They are simply addresses.

Remark 1.4.5 (How to read a number like 1.1.3). A label such as "Theorem 1.1.3" means: Chapter 1, Section 1, third major numbered item in that section. Likewise, "Definition 2.4.5" means: Chapter 2, Section 4, fifth major numbered item there.

These numbers are navigational tools, not mathematical facts to be memorized. Different books use different numbering systems, so one should never spend study time trying to memorize theorem numbers as if they were part of the mathematics itself.

The same philosophy applies to equation numbers and cross-references. A label exists so that later passages can point back efficiently. The goal is not to burden the reader with bookkeeping, but to make long arguments readable.

How to read a proof-based chapter

Reading mathematics is slower than reading a novel, and that is not a sign of failure. Here are several habits that help.

(i) Before reading a proof, make sure you understand the statement. Identify the hypotheses and the conclusion.

(ii) Rewrite a new definition in your own words, then check a few examples and nonexamples.

(iii) When possible, test the statement on small concrete cases before reading the proof.

(iv) While reading a proof, ask at each step: where did this line come from?

(v) After finishing a proof, close the book and try to summarize the main idea in one or two sentences.

These habits are part of learning mathematics itself. One does not first master a separate skill called "how to read proofs" and only then begin mathematics. The two processes happen together.

Remark 1.4.6 (Notation used from the start). We adopt the following conventions throughout the book:

$\mathbb{N}$	the natural numbers $\{1, 2, 3, \dots\}$
$\mathbb{N}_0$	the nonnegative integers $\{0, 1, 2, 3, \dots\}$
$\mathbb{Z}$	the integers
$\mathbb{Q}$	the rational numbers
$\mathbb{R}$	the real numbers
$\mathbb{C}$	the complex numbers

When we introduce a function formally, we write it in the form $f : A \to B$. These are merely conventions, but using them consistently makes the exposition easier to follow.

Remark 1.4.7 (Definitions and examples belong together). When a new term is introduced, do not try to remember only the exact wording of the definition. Immediately ask what the simplest examples are, what the simplest nonexamples are, and why anyone wanted this concept in the first place. In a good mathematics text, the examples are not ornaments; they are part of the meaning.

1.5 The Informal Viewpoint and the Promise of Later Axioms

We are now ready to say clearly how this book will treat set theory at the beginning. We shall speak about sets informally, the way one does in most early mathematical conversations: the set of divisors of 12, the set of prime numbers less than 100, the set of points on a line segment, the set of students in a classroom. This language is natural, efficient, and pedagogically valuable. But it is not the whole story. If taken without restriction, it leads to contradictions. The purpose of this final section is therefore not to abandon the informal viewpoint, but to place it under an honest warning label.

Why the informal viewpoint is attractive

Definition 1.5.1 (Informal set-theoretic viewpoint)**.** The *informal set-theoretic viewpoint* (or *naive set-theoretic viewpoint*) is the practice of speaking of sets as collections of objects and describing them by ordinary mathematical properties, without first reducing everything to a formal axiomatic framework.

Example 1.5.2. The following descriptions are perfectly natural from the informal point of view.

(i) the set of divisors of 12;

(ii) the set of vowels in the English alphabet;

(iii) the set of real numbers between 0 and 1;

(iv) the set of prime numbers less than 100.

At an intuitive level, no confusion arises when we speak this way.

This is why beginning informally is reasonable. The first goal is not to burden the reader with foundational machinery, but to help the reader become fluent with the basic objects and operations of the subject.

Why later axioms are necessary

Definition 1.5.3 (Axiom and axiomatic system)**.** An *axiom* is a basic principle accepted within a mathematical framework. An *axiomatic system* is a collection of axioms from which further statements are derived.

The reason axioms eventually become necessary is that not every verbal description should be allowed to determine a set. If one assumes that *every* property gives rise to a set, one runs into a paradox.

Proposition 1.5.4 (Russell's paradox in naive form)**.** *Suppose that every property $P(x)$ determines a set $\{x \mid P(x)\}$. Then a contradiction follows.*

Proof. Assume that every property determines a set. In particular, consider the property

$$P(x) : \Longleftrightarrow \ x \notin x.$$

Then there should exist a set

$$R = \{x \mid x \notin x\}.$$

Now ask whether $R \in R$.

If $R \in R$, then by the defining property of R we must have $R \notin R$. If $R \notin R$, then again by the defining property of R we must have $R \in R$. In both directions we obtain a contradiction. Therefore the original assumption was impossible. □

Remark 1.5.5 (What the paradox teaches). Proposition 1.5.4 does *not* show that sets are impossible. It shows that unrestricted set formation is impossible. In other words, the phrase "the set of all objects with property P" cannot be accepted automatically for every property P.

Remark 1.5.6 (Historical note). Russell's paradox played a central role in the early twentieth-century reconsideration of set-theoretic foundations; see Russell's classic book [21]. One response was Zermelo's axiomatic approach [23], in which set formation is carefully restricted. We shall return to that point of view in Chapter 15.

Why we still begin informally

> *Intuition is not the enemy of rigor. For a beginner, intuition is the path by which rigor becomes meaningful.*

That principle guides the early chapters of this book. We shall use informal language because it allows the reader to gain fluency with membership, subsets, operations on sets, functions, relations, and infinite processes without carrying the full technical burden of formal logic and axiomatic set theory from the first page. But we shall also remain honest about what is happening. When we write a collection such as

$$\{x \in A \mid P(x)\},$$

we are already using a disciplined version of set formation: the objects x are being selected from a previously given ambient set A. Much later, the axioms will tell us precisely why that kind of construction is legitimate.

Remark 1.5.7 (Why this book takes this route). Many friendly introductions to set theory begin in this spirit; see, for example, Halmos [2] and Devlin [5]. A more explicitly axiomatic development appears in texts such as Enderton [3]. Our plan is to benefit from both perspectives: intuitive access first, axiomatic clarification later.

Remark 1.5.8 (A promise for later chapters). The reader should not worry that the present informality makes the whole subject vague. On the contrary, the early chapters are laying conceptual groundwork. Once the basic ideas of membership, subset, function, order, countability, ordinal, and cardinal have become familiar, the axioms in Chapter 15 will answer a more meaningful question: not merely "what are the axioms?", but "why were axioms needed in the first place?"

Looking ahead

In the next chapter we finally begin to work directly with sets. There we introduce membership notation, subsets, the empty set, power sets, and the basic operations of union, intersection, and set difference. The logical vocabulary of the present chapter will be used immediately: a subset is defined by a quantified statement, set operations are understood through predicates, and

proofs about sets will rely on the proof patterns we have just studied. In that sense, this first chapter is not a preliminary obstacle to the mathematics; it is the beginning of the mathematics.

Chapter 2

Sets, Subsets, and Elementary Operations

In Chapter 1 we learned how to read the language of mathematical statements. We now begin the subject itself. At first sight, the basic object of set theory seems almost too simple to deserve a whole chapter: a set is just a collection of things. But one of the recurring lessons of mathematics is that simple objects can support rich structure once we begin to combine them, compare them, and reason about them systematically.

The first surprise is that a set remembers less than many beginners expect. It remembers which objects belong to it, but it does not remember the order in which those objects were written down, and it does not record repeated appearances of the same object. Thus the set $\{1, 2, 3\}$ is the same as the set $\{3, 2, 1\}$, and writing an object twice does not create a new element. This apparently modest observation is the source of a central principle, called *extensionality*, according to which a set is determined entirely by its elements.

The second surprise is that sets have an algebra of their own. Starting from two sets A and B, we can form their union, intersection, difference, and power set. These operations satisfy laws that resemble familiar algebraic identities. There are commutative laws, associative laws, distributive laws, and De Morgan laws. The proofs, however, do not use calculation with numbers. They use the logical methods of Chapter 1: quantifiers, implications, and the practice of fixing an arbitrary object and asking whether it belongs to a given set.

Throughout this chapter we continue the informal but disciplined viewpoint explained in Section 1.5. We speak of sets intuitively, and we work with them confidently, but we also keep our constructions tethered to clearly described ambient collections. That habit is not a burden. It is part of learning to think clearly about sets from the very beginning.

2.1 Membership and Extensionality

We begin with the most basic relation in the subject: the relation of an object to a set that contains it. Every later idea in the book will build on this relation in one way or another. Before speaking about operations on sets, we therefore pause to ask a foundational question: what does a set actually remember, and what does it forget?

What a set remembers

Definition 2.1.1 (Set, element, and membership)**.** Intuitively, a *set* is a collection of distinct objects regarded as a single whole. The objects belonging to a set are called its *elements* (or *members*). If x is an element of a set A, we write $x \in A$; if x is not an element of A, we write

$x \notin A$.

The word "distinct" is important. A set does not remember how many times an object was listed when the set was first written down. It records only whether the object is present or absent.

Example 2.1.2. The following membership statements are true:

(i) $3 \in \{1, 3, 5, 7\}$.

(ii) $\pi \in \mathbb{R}$.

(iii) $-2 \in \{n \in \mathbb{Z} \mid n^2 < 5\}$.

The following membership statements are false:

(iv) $4 \in \{1, 3, 5, 7\}$.

(v) $\sqrt{2} \in \mathbb{Q}$.

(vi) $5 \in \{n \in \mathbb{Z} \mid n^2 < 5\}$.

A second point, which feels strange at first, is that sets may contain other sets as elements.

Example 2.1.3 (Sets can themselves be elements)**.** Let

$$A = \{\varnothing, \{\varnothing\}\}.$$

Then both $\varnothing \in A$ and $\{\varnothing\} \in A$ are true. However,

$$\{\{\varnothing\}\} \notin A.$$

The set A has exactly two elements, namely the empty set and the singleton whose only element is the empty set.

Examples such as this one show why we must look carefully at braces. A set and a singleton built from that set are usually different objects.

Definition 2.1.4 (Roster notation and set-builder notation)**.** Two common ways of describing a set are the following.

(i) In *roster notation*, we list the elements explicitly inside braces, as in $\{1, 3, 5, 7\}$.

(ii) In *set-builder notation*, we specify a property that the elements must satisfy, as in $\{x \in A \mid P(x)\}$.

Example 2.1.5. The set of positive divisors of 12 can be written in roster form as

$$\{1, 2, 3, 4, 6, 12\}.$$

The same set can be written in set-builder form as

$$\{n \in \mathbb{N} \mid n \text{ divides } 12\}.$$

Similarly, the set of even integers whose absolute value is at most 4 can be written as

$$\{-4, -2, 0, 2, 4\}$$

or as

$$\{n \in \mathbb{Z} \mid n \text{ is even and } |n| \leq 4\}.$$

Remark 2.1.6 (Why the ambient set is helpful). Whenever possible, it is wise to use set-builder notation in the form $\{x \in A \mid P(x)\}$, where the ambient set A is named explicitly. This habit connects naturally with the discussion in Section 1.5: later axiomatic set theory is much happier with "select the elements of an already given set that satisfy property P" than with unrestricted expressions of the form $\{x \mid P(x)\}$. Friendly introductions such as Halmos [2] and Pinter [6] already encourage this disciplined style.

Equality by extensionality

When are two sets equal? Not when they are written in the same order, and not when they were produced by the same recipe. They are equal when they contain exactly the same elements.

Definition 2.1.7 (Equality of sets)**.** Two sets A and B are said to be *equal* if they have exactly the same elements. Equivalently,

$$A = B \quad \text{means} \quad \text{for every object } x,\ x \in A \iff x \in B.$$

The idea that a set is determined entirely by its elements is called *extensionality*.

Proposition 2.1.8 (Order and repetition do not matter)**.** *For any objects a and b,*

$$\{a, b\} = \{b, a\}.$$

For any object a,

$$\{a, a\} = \{a\}.$$

Proof. We prove each equality by Definition 2.1.7.

For the first claim, let x be any object. Then $x \in \{a, b\}$ means that $x = a$ or $x = b$. But this is equivalent to saying that $x = b$ or $x = a$, which is exactly the statement $x \in \{b, a\}$. Hence $x \in \{a, b\} \iff x \in \{b, a\}$ for every x, and so $\{a, b\} = \{b, a\}$.

For the second claim, let x be any object. Then $x \in \{a, a\}$ means that $x = a$ or $x = a$, which is equivalent simply to $x = a$. But that is the condition for $x \in \{a\}$. Hence $x \in \{a, a\} \iff x \in \{a\}$ for every x, so $\{a, a\} = \{a\}$. □

Example 2.1.9 (Sets versus lists and tuples)**.** The expressions

$$\{1, 2, 3\}, \qquad \{3, 2, 1\}, \qquad \{1, 1, 2, 3\}$$

all describe the same set. By contrast, the lists

$$(1, 2, 3) \qquad \text{and} \qquad (3, 2, 1)$$

are different because lists remember order. Likewise, a multiset can record repeated appearances, whereas a set cannot. In Chapter 3 we shall build mathematical objects that remember order explicitly.

Remark 2.1.10 (Extensionality later becomes an axiom). At present extensionality is being used as the most natural informal rule for deciding equality of sets. In axiomatic set theory, however, it is elevated to a formal statement, the *axiom of extensionality*. See Enderton [3] or Hrbacek and Jech [4] for that later viewpoint. The crucial idea is already visible here: a set has no hidden interior structure beyond its membership relation.

2.2 Empty Set, Singletons, Pair Sets, and Subsets

Once membership is available, the next step is to understand the smallest and most basic kinds of sets. These are not merely toy examples. The empty set explains what it means for a set to have no elements at all, singletons isolate one object, pair sets hold two objects together, and the notion of subset gives us our first genuine way of comparing sets. Many later definitions will be phrased in the language of subsets.

The smallest sets

Definition 2.2.1 (The empty set, singleton, and pair set)**.** (i) The *empty set*, written $\varnothing$, is the set with no elements.

(ii) If a is an object, then $\{a\}$ is the *singleton* whose only element is a.

(iii) If a and b are objects, then $\{a, b\}$ is the *pair set* containing a and b.

Example 2.2.2. The following are all correct examples of the objects just defined:

(i) $\varnothing$ has no elements.

(ii) $\{5\}$ is a singleton.

(iii) $\{1, 2\}$ is a pair set with two distinct elements.

(iv) $\{\varnothing\}$ is a singleton whose unique element is the empty set.

(v) $\{3, 3\} = \{3\}$, so a pair set may collapse to a singleton when the two listed objects are the same.

Proposition 2.2.3 (The empty set is unique)**.** *There is exactly one empty set.*

Proof. Suppose E and F are both empty sets. Because E has no members, there is no object x for which $x \in E$. The same is true for F. Hence, for every object x, the statements $x \in E$ and $x \in F$ are both false. Therefore $x \in E \iff x \in F$ for every x. By Definition 2.1.7, we conclude that $E = F$. □

Remark 2.2.4. The notation $\{a, b\}$ does not guarantee that the set has two members. If $a = b$, then $\{a, b\} = \{a\}$. In other words, "pair set" names the way the set is formed, not the number of distinct elements it must have.

Inclusion

To say that one set is contained inside another is to say that every element of the first is already an element of the second. This simple idea is so important that it receives its own notation.

Definition 2.2.5 (Subset and proper subset)**.** Let A and B be sets.

(i) We say that A is a *subset* of B, and write $A \subseteq B$, if every element of A is an element of B.

(ii) We say that A is a *proper subset* of B, and write $A \subsetneq B$, if $A \subseteq B$ and $A \neq B$.

Example 2.2.6. (i) $\{1, 3\} \subseteq \{1, 2, 3, 4\}$.

(ii) $\{1, 2, 3\} \subseteq \{1, 2, 3\}$. A set is allowed to be a subset of itself.

(iii) $\{1, 2, 3\} \subsetneq \{1, 2, 3, 4\}$.

(iv) $\{2, 4, 6\} \not\subseteq \{1, 2, 3, 4, 5\}$, because 6 is an element of the first set but not of the second.

Remark 2.2.7 (Do not confuse $\in$ with $\subseteq$). This is one of the most common early confusions.
If $A = \{1, 2, 3\}$, then

$$1 \in A \qquad \text{and} \qquad \{1\} \subseteq A,$$

but

$$\{1\} \notin A.$$

The symbol $\in$ relates an object to a set. The symbol $\subseteq$ relates one set to another set. They answer different questions.

Proposition 2.2.8 (Basic facts about subsets)**.** *For all sets A, B, and C, the following hold:*

(i) $A \subseteq A$.

(ii) $\varnothing \subseteq A$.

(iii) If $A \subseteq B$ and $B \subseteq C$, then $A \subseteq C$.

Proof. For (i), let $x \in A$. Then of course $x \in A$. So every member of A is a member of A, which means $A \subseteq A$.

For (ii), we must show that every element of $\varnothing$ lies in A. But $\varnothing$ has no elements at all, so there is nothing to check. This is an instance of the idea, discussed in Chapter 1, that a universally quantified statement over an empty collection is automatically true.

For (iii), suppose $A \subseteq B$ and $B \subseteq C$. Let $x \in A$. Since $A \subseteq B$, we have $x \in B$. Since $B \subseteq C$, we then have $x \in C$. Thus every element of A is an element of C, so $A \subseteq C$. □

Theorem 2.2.9 (Double inclusion criterion)**.** *For any sets A and B,*

$$A = B \quad \textit{if and only if} \quad A \subseteq B \textit{ and } B \subseteq A.$$

Proof. First suppose that $A = B$. If $x \in A$, then, because the two sets are equal, we also have $x \in B$. Hence $A \subseteq B$. By the same reasoning, $B \subseteq A$.

Conversely, assume that $A \subseteq B$ and $B \subseteq A$. Let x be any object. If $x \in A$, then $x \in B$ because $A \subseteq B$. If $x \in B$, then $x \in A$ because $B \subseteq A$. Therefore $x \in A \iff x \in B$ for every object x. By Definition 2.1.7, it follows that $A = B$. □

Example 2.2.10 (A first identity proved by inclusion)**.** Let

$$S = \{n \in \mathbb{Z} \mid n^2 < 2\}.$$

Then

$$S = \{-1, 0, 1\}.$$

Proof. We prove the equality by double inclusion.

First let $n \in S$. Then n is an integer and $n^2 < 2$. If $|n| \geq 2$, then $n^2 \geq 4$, which is impossible. Therefore $|n| < 2$. Since n is an integer, this leaves only the possibilities $n = -1$, $n = 0$, and $n = 1$. Hence $n \in \{-1, 0, 1\}$, and so $S \subseteq \{-1, 0, 1\}$.

Conversely, each of the integers -1, 0, and 1 has square strictly less than 2. Thus every element of $\{-1, 0, 1\}$ lies in S, so $\{-1, 0, 1\} \subseteq S$.

By Theorem 2.2.9, the two sets are therefore equal. □

Remark 2.2.11 (Subset as comparison). The relation $\subseteq$ already behaves like a form of comparison. It is reflexive and transitive by Proposition 2.2.8, and it becomes antisymmetric by Theorem 2.2.9. In Chapter 5 we shall recognize this pattern as an order relation, and sets of subsets such as $\mathcal{P}(A)$ will become important examples of ordered structures; see also Davey and Priestley [8] for the broader order-theoretic viewpoint.

2.3 Union, Intersection, Difference, and Complement

A single set is already a mathematical object, but much of the subject comes alive when we learn to build new sets from old ones. The most important elementary constructions are union, intersection, and set-theoretic difference. Together with complements relative to an ambient set, they form the basic calculus of subsets.

Venn diagrams provide useful geometric intuition for these operations, but intuition alone is not enough. In proofs we must return to the definitions and ask, element by element, what it means to belong to each set under discussion.

Building new sets from old ones

Definition 2.3.1 (Union and intersection)**.** Let A and B be sets.

(i) The *union* of A and B is the set

$$A \cup B = \{x \mid x \in A \text{ or } x \in B\}.$$

(ii) The *intersection* of A and B is the set

$$A \cap B = \{x \mid x \in A \text{ and } x \in B\}.$$

Example 2.3.2. If

$$A = \{1, 2, 3, 4\} \qquad \text{and} \qquad B = \{3, 4, 5\},$$

then

$$A \cup B = \{1, 2, 3, 4, 5\} \qquad \text{and} \qquad A \cap B = \{3, 4\}.$$

The union gathers all elements that occur in at least one of the two sets; the intersection keeps only the elements that occur in both.

Definition 2.3.3 (Difference and complement)**.** Let A and B be sets.

(i) The *difference* (or *relative complement*) of B in A is the set

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

(ii) If U is an ambient set and $A \subseteq U$, then the *complement* of A in U is the set

$$A^c = U \setminus A.$$

Example 2.3.4. Let

$$A = \{1, 2, 3, 4\}, \qquad B = \{3, 4, 5\}, \qquad U = \{1, 2, 3, 4, 5, 6\}.$$

Then

$$A \setminus B = \{1, 2\}, \qquad B \setminus A = \{5\}, \qquad A^c = \{5, 6\}.$$

The first two sets are different, which shows that difference is not a symmetric operation.

Definition 2.3.5 (Disjoint sets)**.** Two sets A and B are said to be *disjoint* if

$$A \cap B = \varnothing.$$

Example 2.3.6. The sets $\{1, 3, 5\}$ and $\{2, 4, 6\}$ are disjoint. The sets $\{1, 2\}$ and $\{2, 3\}$ are not disjoint, because their intersection is $\{2\}$.

Remark 2.3.7 (Complements are always relative). The notation A^c is convenient only when the ambient set U has been fixed in advance. Without such an ambient set, the phrase "the complement of A" is ambiguous. In this book we therefore use complement notation only when the surrounding universe of discourse is clear from context.

What the definitions really say

The definitions above become useful only when we can translate them quickly into membership statements.

Proposition 2.3.8 (Membership criteria for the basic operations)**.** *For every object x, the following*

equivalences hold:

$$\begin{aligned} x \in A \cup B &\iff x \in A \text{ or } x \in B, \\ x \in A \cap B &\iff x \in A \text{ and } x \in B, \\ x \in A \setminus B &\iff x \in A \text{ and } x \notin B. \end{aligned}$$

If A^c denotes the complement of A relative to an ambient set U, then

$$x \in A^c \iff x \in U \text{ and } x \notin A.$$

Proof. Each equivalence is simply the corresponding definition rewritten in words. For example, by definition,

$$A \setminus B = \{x \in A \mid x \notin B\},$$

so to say $x \in A \setminus B$ is exactly to say that $x \in A$ and $x \notin B$. The other cases are analogous. □

Proposition 2.3.9 (First inclusions)**.** *For any sets A and B,*

(i) $A \cap B \subseteq A$,

(ii) $A \cap B \subseteq B$,

(iii) $A \subseteq A \cup B$,

(iv) $B \subseteq A \cup B$,

(v) $A \setminus B \subseteq A$.

Proof. We prove (i), (iii), and (v); the others are similar.

For (i), let $x \in A \cap B$. By Proposition 2.3.8, this means that $x \in A$ and $x \in B$. In particular, $x \in A$. Therefore $A \cap B \subseteq A$.

For (iii), let $x \in A$. Then $x \in A$ or $x \in B$, so $x \in A \cup B$. Hence $A \subseteq A \cup B$.

For (v), let $x \in A \setminus B$. Then $x \in A$ and $x \notin B$. In particular, $x \in A$. Hence $A \setminus B \subseteq A$. □

Theorem 2.3.10 (A set decomposes into overlap and non-overlap)**.** *For any sets A and B,*

$$A = (A \setminus B) \cup (A \cap B),$$

and the two sets on the right-hand side are disjoint.

Proof. We begin with the equality. Let $x \in A$. Either $x \in B$ or $x \notin B$. If $x \in B$, then $x \in A \cap B$. If $x \notin B$, then $x \in A \setminus B$. In either case, $x \in (A \setminus B) \cup (A \cap B)$. Thus $A \subseteq (A \setminus B) \cup (A \cap B)$.

Conversely, if $x \in (A \setminus B) \cup (A \cap B)$, then either $x \in A \setminus B$ or $x \in A \cap B$. In the first case, $x \in A$ by Proposition 2.3.9. In the second case, $x \in A$ again by the same proposition. Hence $(A \setminus B) \cup (A \cap B) \subseteq A$, so the equality follows from Theorem 2.2.9.

Now we show disjointness. Suppose $x \in (A \setminus B) \cap (A \cap B)$. Then $x \in A \setminus B$ and $x \in A \cap B$. The first condition implies $x \notin B$, while the second implies $x \in B$. This is impossible. Therefore $(A \setminus B) \cap (A \cap B) = \varnothing$, so the two sets are disjoint. □

Theorem 2.3.11 (Basic laws of complements)**.** *Let U be an ambient set and suppose $A \subseteq U$. Then:*

(i) $A \cup A^c = U$,

(ii) $A \cap A^c = \varnothing$,

(iii) $(A^c)^c = A$.

Proof. For (i), let $x \in U$. Either $x \in A$ or $x \notin A$. In the first case, $x \in A \cup A^c$. In the second case, $x \in A^c$ by definition of complement, so again $x \in A \cup A^c$. Thus $U \subseteq A \cup A^c$. Conversely, if $x \in A \cup A^c$, then either $x \in A$ or $x \in A^c$. In both cases, $x \in U$. Hence $A \cup A^c \subseteq U$, and so $A \cup A^c = U$.

For (ii), if $x \in A \cap A^c$, then $x \in A$ and $x \in A^c$. The latter means $x \notin A$, a contradiction. So $A \cap A^c$ has no elements and is therefore $\varnothing$.

For (iii), let $x \in (A^c)^c$. Then $x \in U$ and $x \notin A^c$. Since A^c consists of precisely the elements of U that are not in A, the statement $x \notin A^c$ forces $x \in A$. Thus $(A^c)^c \subseteq A$. Conversely, if $x \in A$, then certainly $x \in U$, and because $x \notin A^c$, we have $x \in (A^c)^c$. So $A \subseteq (A^c)^c$. The sets are equal by Theorem 2.2.9. □

Remark 2.3.12 (On Venn diagrams). Venn diagrams are valuable because they let the eye see overlap, separation, and inclusion at a glance. They are especially helpful when one is first learning the meaning of union, intersection, and complement. But a picture is not a proof. The equalities in this chapter must hold for *every* set, including infinite sets that cannot literally be drawn. That is why our formal proofs return to membership statements.

2.4 Power Sets

A set may contain numbers, points, functions, or even other sets. But sometimes we want to gather not just some subsets of a given set, but *all* of them at once. That construction is so important that it has its own name and notation. The power set is the first place in which a single set naturally generates a much larger universe of new objects.

The set of all subsets

Definition 2.4.1 (Power set)**.** Let A be a set. The *power set* of A, denoted by $\mathcal{P}(A)$, is the set of all subsets of A:

$$\mathcal{P}(A) = \{X \mid X \subseteq A\}.$$

Example 2.4.2. The smallest power sets can be written down explicitly:

$$\begin{aligned} \mathcal{P}(\varnothing) &= \{\varnothing\}, \\ \mathcal{P}(\{a\}) &= \{\varnothing, \{a\}\}, \\ \mathcal{P}(\{a, b\}) &= \{\varnothing, \{a\}, \{b\}, \{a, b\}\}. \end{aligned}$$

Even in the two-element case, the power set already has four members.

Example 2.4.3 (Nested sets inside a power set)**.** Let

$$A = \{\varnothing, \{\varnothing\}\}.$$

Then the subsets of A are

$$\varnothing, \qquad \{\varnothing\}, \qquad \{\{\varnothing\}\}, \qquad \{\varnothing, \{\varnothing\}\}.$$

Hence

$$\mathcal{P}(A) = \big\{\varnothing, \{\varnothing\}, \{\{\varnothing\}\}, \{\varnothing, \{\varnothing\}\}\big\}.$$

This example is a useful reminder that an element of $\mathcal{P}(A)$ is itself a set.

Proposition 2.4.4 (The smallest and largest members of a power set)**.** *For every set A, both $\varnothing \in \mathcal{P}(A)$ and $A \in \mathcal{P}(A)$.*

Proof. By Proposition 2.2.8, we know that $\varnothing \subseteq A$ and $A \subseteq A$. By the definition of $\mathcal{P}(A)$, this means exactly that $\varnothing \in \mathcal{P}(A)$ and $A \in \mathcal{P}(A)$. □

Theorem 2.4.5 (Power sets detect inclusion)**.** *For any sets A and B,*

$$A \subseteq B \quad \textit{if and only if} \quad \mathcal{P}(A) \subseteq \mathcal{P}(B).$$

Proof. First assume that $A \subseteq B$. Let $X \in \mathcal{P}(A)$. Then $X \subseteq A$. Since $A \subseteq B$, transitivity of inclusion (Proposition 2.2.8) gives $X \subseteq B$. Thus $X \in \mathcal{P}(B)$. We have shown that every member of $\mathcal{P}(A)$ belongs to $\mathcal{P}(B)$, so $\mathcal{P}(A) \subseteq \mathcal{P}(B)$.

Conversely, assume that $\mathcal{P}(A) \subseteq \mathcal{P}(B)$. By Proposition 2.4.4, we know that $A \in \mathcal{P}(A)$. Since $\mathcal{P}(A) \subseteq \mathcal{P}(B)$, it follows that $A \in \mathcal{P}(B)$. By definition of power set, this means $A \subseteq B$. □

Example 2.4.6. Let $A = \{1\}$ and $B = \{1, 2\}$. Then $A \subseteq B$, and indeed

$$\mathcal{P}(A) = \{\varnothing, \{1\}\} \subseteq \{\varnothing, \{1\}, \{2\}, \{1, 2\}\} = \mathcal{P}(B).$$

If we reverse the roles of A and B, the inclusion fails. This agrees with Theorem 2.4.5.

Why power sets grow quickly

One can already see from small examples that power sets become large very quickly. Even if the original set has only a few elements, the collection of all its subsets is noticeably larger.

Example 2.4.7 (The eight subsets of a three-element set)**.** Let $A = \{a, b, c\}$. Then

$$\begin{aligned}\mathcal{P}(A) = \{&\varnothing, \{a\}, \{b\}, \{c\},\\ &\{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}.\end{aligned}$$

So a set with three elements has eight subsets.

Remark 2.4.8 (How many subsets should a finite set have?). If a set has n distinct elements, then each subset is formed by making n independent yes-or-no decisions: for each element, either include it or do not include it. This leads to the expectation that there should be 2^n subsets altogether. We are appealing here to familiar finite counting intuition; Chapter 7 will return to such counting arguments in a more systematic set-theoretic way.

Remark 2.4.9 (A historical preview). The power-set operation became one of Cantor's central tools in the study of infinity. Much later in the book, in Chapter 9, we shall prove that every set has a power set that is larger than the original set in a precise sense. Cantor's 1891 paper [19] is a classic source for this turn in the subject.

2.5 The Algebra of Sets

We now come to one of the most useful habits in elementary set theory: proving identities among set expressions. At this stage the subject begins to feel less like a list of definitions and more like a small calculus. The guiding idea is simple. Because sets are determined by their elements, an equality of sets is proved by testing membership on both sides.

This method is sometimes called an *element-chasing argument*. One fixes an arbitrary object x and asks exactly when it belongs to each expression. The logical training of Chapter 1 now becomes visibly useful: conjunction corresponds to intersection, disjunction to union, and negation to complement.

How set identities are proved

Proposition 2.5.1 (Membership test for set identities)**.** *Let X and Y be sets.*

(i) To prove $X \subseteq Y$, it suffices to let x be an arbitrary object of X and prove that $x \in Y$.

(ii) To prove $X = Y$, it suffices to let x be an arbitrary object and prove that $x \in X \iff x \in Y$.

Proof. Statement (i) is simply the definition of subset. Statement (ii) follows from Definition 2.1.7. One may also prove (ii) by using (i) twice and appealing to Theorem 2.2.9. □

Proposition 2.5.2 (Difference as intersection with a complement)**.** *Let U be an ambient set, and suppose $A, B \subseteq U$. Then*

$$A \setminus B = A \cap B^c.$$

Proof. Let x be any object. Then

$$\begin{aligned} x \in A \setminus B &\iff x \in A \text{ and } x \notin B \\ &\iff x \in A \text{ and } x \in B^c \\ &\iff x \in A \cap B^c. \end{aligned}$$

By Proposition 2.5.1, the two sets are equal. □

Example 2.5.3. Let $U = \{1, 2, 3, 4, 5, 6\}$, $A = \{1, 2, 3, 4\}$, and $B = \{3, 4, 5\}$. Then $B^c = \{1, 2, 6\}$, so

$$A \cap B^c = \{1, 2\} = A \setminus B.$$

This is a concrete instance of Proposition 2.5.2.

Standard laws

The next results should be read exactly as one reads familiar algebraic identities. The important difference is that every proof takes place at the level of membership.

Theorem 2.5.4 (Commutative, associative, and idempotent laws)**.** *For all sets A, B, and C, the following hold:*

(i) $A \cup B = B \cup A$ *and* $A \cap B = B \cap A$ (commutative laws).

(ii) $A \cup (B \cup C) = (A \cup B) \cup C$ *and* $A \cap (B \cap C) = (A \cap B) \cap C$ (associative laws).

(iii) $A \cup A = A$ *and* $A \cap A = A$ (idempotent laws).

Proof. We prove one identity of each type; the others are analogous.

For commutativity of union, let x be any object. Then

$$x \in A \cup B \iff x \in A \text{ or } x \in B \iff x \in B \text{ or } x \in A \iff x \in B \cup A.$$

Hence $A \cup B = B \cup A$.

For associativity of intersection, let x be any object. Then

$$\begin{aligned} x \in A \cap (B \cap C) &\iff x \in A \text{ and } x \in (B \cap C) \\ &\iff x \in A \text{ and } x \in B \text{ and } x \in C \\ &\iff x \in (A \cap B) \text{ and } x \in C \\ &\iff x \in (A \cap B) \cap C. \end{aligned}$$

So $A \cap (B \cap C) = (A \cap B) \cap C$.

For idempotence of union, let x be any object. Then

$$x \in A \cup A \iff x \in A \text{ or } x \in A \iff x \in A.$$

Thus $A \cup A = A$. □

Corollary 2.5.5 (Identity laws)**.** *For every set A,*

$$A \cup \varnothing = A \qquad \textit{and} \qquad A \cap \varnothing = \varnothing.$$

If $A \subseteq U$, then also

$$A \cup U = U \qquad \textit{and} \qquad A \cap U = A.$$

Proof. For $A \cup \varnothing = A$, let x be any object. Then

$$x \in A \cup \varnothing \iff x \in A \text{ or } x \in \varnothing \iff x \in A,$$

since nothing belongs to $\varnothing$. Therefore $A \cup \varnothing = A$.

For $A \cap \varnothing = \varnothing$, if $x \in A \cap \varnothing$, then $x \in \varnothing$, which is impossible. So the intersection has no elements.

The identities involving U are proved similarly, using the fact that every element of A belongs to U when $A \subseteq U$. □

Corollary 2.5.6 (Monotonicity)**.** *If $A \subseteq B$, then for every set C,*

$$A \cup C \subseteq B \cup C, \qquad A \cap C \subseteq B \cap C, \qquad A \setminus C \subseteq B \setminus C.$$

Proof. Assume $A \subseteq B$. If $x \in A \cup C$, then either $x \in A$ or $x \in C$. In the first case, $x \in B$, so in both cases $x \in B \cup C$. Hence $A \cup C \subseteq B \cup C$.

If $x \in A \cap C$, then $x \in A$ and $x \in C$. Since $A \subseteq B$, we get $x \in B$, and therefore $x \in B \cap C$. So $A \cap C \subseteq B \cap C$.

Finally, if $x \in A \setminus C$, then $x \in A$ and $x \notin C$. Again $x \in B$, so $x \in B \setminus C$. Thus $A \setminus C \subseteq B \setminus C$. □

Theorem 2.5.7 (Distributive and absorption laws)**.** *For all sets A, B, and C, the following hold:*

(i) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

(ii) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

(iii) $A \cup (A \cap B) = A$.

(iv) $A \cap (A \cup B) = A$.

Proof. For (i), let x be any object. Then

$$\begin{aligned} x \in A \cap (B \cup C) &\iff x \in A \text{ and } x \in (B \cup C) \\ &\iff x \in A \text{ and } (x \in B \text{ or } x \in C) \\ &\iff (x \in A \text{ and } x \in B) \text{ or } (x \in A \text{ and } x \in C) \\ &\iff x \in (A \cap B) \cup (A \cap C). \end{aligned}$$

Hence $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

For (ii), let x be any object. Then

$$\begin{aligned} x \in A \cup (B \cap C) &\iff x \in A \text{ or } (x \in B \text{ and } x \in C) \\ &\iff (x \in A \text{ or } x \in B) \text{ and } (x \in A \text{ or } x \in C) \\ &\iff x \in (A \cup B) \cap (A \cup C). \end{aligned}$$

Therefore $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

For (iii), let x be any object. If $x \in A \cup (A \cap B)$, then either $x \in A$ or $x \in A \cap B$. In the second case, we still have $x \in A$. So every element of $A \cup (A \cap B)$ lies in A. Conversely, if $x \in A$, then certainly $x \in A \cup (A \cap B)$. Hence $A \cup (A \cap B) = A$.

The proof of (iv) is similar. □

Example 2.5.8 (Checking a distributive law concretely)**.** Let

$$A = \{1, 2, 3\}, \qquad B = \{2, 4\}, \qquad C = \{1, 4\}.$$

Then

$$B \cup C = \{1, 2, 4\}, \qquad A \cap (B \cup C) = \{1, 2\}.$$

On the other hand,

$$A \cap B = \{2\}, \qquad A \cap C = \{1\}, \qquad (A \cap B) \cup (A \cap C) = \{1, 2\}.$$

So in this concrete example,

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C).$$

The theorem says that this is not an accident of the chosen sets, but a general law.

Theorem 2.5.9 (De Morgan's laws)**.** *Let U be an ambient set, and let $A, B \subseteq U$. Then*

$$(A \cup B)^c = A^c \cap B^c \qquad \textit{and} \qquad (A \cap B)^c = A^c \cup B^c.$$

Proof. We prove the first identity. Let x be any object. Then

$$\begin{aligned} x \in (A \cup B)^c &\iff x \in U \text{ and } x \notin A \cup B \\ &\iff x \in U \text{ and not}(x \in A \text{ or } x \in B) \\ &\iff x \in U \text{ and } x \notin A \text{ and } x \notin B \\ &\iff x \in A^c \text{ and } x \in B^c \\ &\iff x \in A^c \cap B^c. \end{aligned}$$

Thus $(A \cup B)^c = A^c \cap B^c$.

The proof of $(A \cap B)^c = A^c \cup B^c$ is analogous:

$$\begin{aligned} x \in (A \cap B)^c &\iff x \in U \text{ and } x \notin A \cap B \\ &\iff x \in U \text{ and not}(x \in A \text{ and } x \in B) \\ &\iff x \in U \text{ and } (x \notin A \text{ or } x \notin B) \\ &\iff x \in A^c \cup B^c. \end{aligned}$$

□

Corollary 2.5.10 (Difference laws)**.** *Let U be an ambient set, and let $A, B, C \subseteq U$. Then*

$$A \setminus (B \cup C) = (A \setminus B) \cap (A \setminus C)$$

and

$$A \setminus (B \cap C) = (A \setminus B) \cup (A \setminus C).$$

Proof. Using Proposition 2.5.2 and Theorem 2.5.9, we compute

$$\begin{aligned} A \setminus (B \cup C) &= A \cap (B \cup C)^c \\ &= A \cap (B^c \cap C^c) \\ &= (A \cap B^c) \cap (A \cap C^c) \\ &= (A \setminus B) \cap (A \setminus C). \end{aligned}$$

The second identity is proved similarly:

$$\begin{aligned} A \setminus (B \cap C) &= A \cap (B \cap C)^c \\ &= A \cap (B^c \cup C^c) \\ &= (A \cap B^c) \cup (A \cap C^c) \\ &= (A \setminus B) \cup (A \setminus C). \end{aligned}$$

□

Proposition 2.5.11 (Predicates and set operations)**.** *Let U be a set, and let $P(x)$ and $Q(x)$ be predicates on U. Define*

$$A = \{x \in U \mid P(x)\} \qquad \text{and} \qquad B = \{x \in U \mid Q(x)\}.$$

Then:

$$\begin{aligned} A \cup B &= \{x \in U \mid P(x) \text{ or } Q(x)\}, \\ A \cap B &= \{x \in U \mid P(x) \text{ and } Q(x)\}, \\ A^c &= \{x \in U \mid \text{not } P(x)\}. \end{aligned}$$

Proof. For example, let $x \in U$. Then

$$x \in A \cup B \iff x \in A \text{ or } x \in B \iff P(x) \text{ or } Q(x).$$

So $A \cup B$ is exactly the set of elements of U for which $P(x)$ or $Q(x)$ holds. The other identities are proved in the same way. □

The previous proposition gives a useful dictionary between the logical connectives of Chapter 1 and the set operations of the present chapter.

Logic on predicates	*Set operation*	*Resulting subset of U*
$P(x)$ or $Q(x)$	union	$A \cup B$
$P(x)$ and $Q(x)$	intersection	$A \cap B$
not $P(x)$	complement	A^c

Remark 2.5.12 (A first glimpse of Boolean algebra). The laws proved in this section are not random conveniences. They are part of a coherent algebraic pattern. The subsets of a fixed ambient set U, equipped with union, intersection, and complement, form a prototype of what is called a *Boolean algebra*. We shall not formalize that notion here, but it is worth noticing that logical laws and set-theoretic laws mirror one another very closely. For a later, more abstract treatment, see Davey and Priestley [8].

Looking ahead

This chapter has taught us how to speak about membership, inclusion, and the basic operations that build new sets from old ones. Yet we have also seen a limitation: ordinary sets remember membership but not order. The set $\{a, b\}$ does not distinguish the first position from the second, and so it cannot by itself represent a point in the plane, an input-output rule, or a sequence of choices. In the next chapter we therefore enrich our language by introducing ordered pairs, Cartesian products, and functions. These constructions allow us to move from mere collections to structured arrangements, and they will prepare the way for relations, indexed families, and the comparison of sizes of sets.

Chapter 3

Ordered Pairs, Cartesian Products, and Functions

In Chapter 2 we learned that a set records membership but forgets order. The set $\{a, b\}$ says only that a and b are present; it does not say which one comes first. That is exactly what we want in many situations. But it is not what we want when we try to describe a point in the plane, a table entry in a grid, an input-output rule, or a sequence of choices. A point with coordinates $(2, 5)$ is not the same as the point with coordinates $(5, 2)$, even though the underlying two numbers are the same. The first coordinate and the second coordinate play different roles.

This chapter develops the set-theoretic language needed to remember such roles. We begin with *ordered pairs*, from which we form *Cartesian products*. These are the natural homes for coordinate-like data. We then turn to *functions*. At school one often meets functions through formulas such as $x \mapsto x^2$ or $x \mapsto \sin x$. That viewpoint remains important, but set theory encourages a broader one. A function need not be given by a formula. It may be described by a table, by a rule in words, or by a set of ordered pairs.

This shift in viewpoint is more than a matter of taste. Later chapters will compare sizes of sets by studying bijections, define natural numbers recursively by functions, and describe general products as sets of functions. For that reason the present chapter is one of the main bridges from elementary school mathematics to the broader structural language of modern mathematics. Our goal is not merely to accumulate new notation, but to see why that notation is natural and how it is used in proofs.

3.1 Ordered Pairs and Cartesian Products

Sets do not remember order, so if we want an object with a distinguished first position and a distinguished second position, we must build it. We begin with the simplest such object: an ordered pair.

Why order matters

Suppose we want to represent the point of the plane whose first coordinate is 2 and whose second coordinate is 5. The unordered pair $\{2, 5\}$ is not enough, because it is identical to $\{5, 2\}$. Likewise, a recipe that sends an input x to an output $f(x)$ should distinguish the place of the input from the place of the output. Set theory therefore needs a construction that retains the information of position.

Definition 3.1.1 (Ordered pair)**.** Given objects a and b, the *ordered pair* of a and b is denoted by $\langle a, b\rangle$. In this book we realize it set-theoretically by the formula

$$\langle a, b\rangle := \big\{\{a\}, \{a, b\}\big\}.$$

Thus an ordered pair is itself a set, but it is a specially designed set that remembers which entry is first and which is second.

Example 3.1.2 (Ordered pairs are not pair sets)**.** The ordered pair $\langle 1, 2\rangle$ is not the same object as the unordered pair $\{1, 2\}$. Indeed,

$$\langle 1, 2\rangle = \big\{\{1\}, \{1, 2\}\big\},$$

which has as elements the singleton $\{1\}$ and the pair set $\{1, 2\}$. The set $\{1, 2\}$ itself has elements 1 and 2. These are completely different kinds of objects.

Likewise,

$$\langle 1, 2\rangle = \big\{\{1\}, \{1, 2\}\big\} \qquad \text{but} \qquad \langle 2, 1\rangle = \big\{\{2\}, \{1, 2\}\big\}.$$

Since $\{1\} \neq \{2\}$, the two ordered pairs are different. This is exactly what we wanted: order has become visible.

Example 3.1.3 (The diagonal case)**.** If the two entries agree, then the construction collapses neatly:

$$\langle a, a\rangle = \big\{\{a\}, \{a, a\}\big\} = \big\{\{a\}, \{a\}\big\} = \{\{a\}\}.$$

So the ordered pair $\langle a, a\rangle$ becomes a singleton whose only element is the singleton $\{a\}$.

The previous example shows that the construction is a little indirect. Its virtue is not visual simplicity, but the fact that it behaves correctly.

Theorem 3.1.4 (Equality of ordered pairs)**.** *For any objects a, b, c, d,*

$$\langle a, b\rangle = \langle c, d\rangle \quad \textit{if and only if} \quad a = c \textit{ and } b = d.$$

Proof. If $a = c$ and $b = d$, then the two displayed definitions are literally the same, so the ordered pairs are equal.

Conversely, assume that

$$\langle a, b\rangle = \langle c, d\rangle.$$

By Definition 3.1.1, the set $\{a\}$ is an element of $\langle a, b\rangle$. Since the two ordered pairs are equal, $\{a\}$ is also an element of $\langle c, d\rangle = \{\{c\}, \{c, d\}\}$. Hence either $\{a\} = \{c\}$ or $\{a\} = \{c, d\}$.

In the first case, $a = c$. In the second case, both c and d are elements of $\{c, d\} = \{a\}$, so $c = a$ and $d = a$. In particular, again $a = c$. Thus we have proved that $a = c$.

Now use this information in the equality of ordered pairs:

$$\langle a, b\rangle = \langle a, d\rangle.$$

The set $\{a, b\}$ is an element of $\langle a, b\rangle$, so it is also an element of $\langle a, d\rangle = \{\{a\}, \{a, d\}\}$. Thus $\{a, b\} = \{a\}$ or $\{a, b\} = \{a, d\}$.

If $\{a, b\} = \{a\}$, then $b = a$. Also $\{a, d\}$ is an element of $\langle a, d\rangle = \langle a, b\rangle$, which in this case equals $\{\{a\}\}$. Therefore $\{a, d\} = \{a\}$, and so $d = a = b$. Hence $b = d$.

If instead $\{a, b\} = \{a, d\}$, then $b \in \{a, b\} = \{a, d\}$, so $b = a$ or $b = d$. If $b = d$, we are done. If $b = a$, then we are back in the previous paragraph, which again gives $d = a = b$. Thus in all cases $b = d$.

We have shown $a = c$ and $b = d$, as required. □

Remark 3.1.5 (Why this construction works). Theorem 3.1.4 is the entire point of the construction. An ordered pair may be built out of ordinary sets, but it behaves as though it had genuine first and second coordinates. This standard realization appears in many set theory texts; see, for instance, Enderton [3] or Pinter [6].

Products of sets

Once ordered pairs are available, it becomes natural to ask for the set of all ordered pairs whose first entry comes from one set and whose second entry comes from another.

Definition 3.1.6 (Cartesian product)**.** Let A and B be sets. Their *Cartesian product* is the set

$$A \times B = \{\langle a, b\rangle \mid a \in A \text{ and } b \in B\}.$$

Thus $A \times B$ consists of all ordered pairs whose first component lies in A and whose second component lies in B.

Example 3.1.7 (A finite product)**.** If

$$A = \{1, 2, 3\} \qquad \text{and} \qquad B = \{x, y\},$$

then

$$A \times B = \{\langle 1, x\rangle, \langle 1, y\rangle, \langle 2, x\rangle, \langle 2, y\rangle, \langle 3, x\rangle, \langle 3, y\rangle\}.$$

Even though A has three elements and B has two, the product contains six ordered pairs because each choice from A can be paired with each choice from B.

Example 3.1.8 (Squares and coordinate spaces)**.** The product $A \times A$ consists of all ordered pairs of elements of A. For example, if $A = \{0, 1\}$, then

$$A \times A = \{\langle 0, 0\rangle, \langle 0, 1\rangle, \langle 1, 0\rangle, \langle 1, 1\rangle\}.$$

More generally,

$$\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$$

may be viewed as the set of all ordered pairs of real numbers, that is, as the set of points of the ordinary coordinate plane.

Proposition 3.1.9 (Basic properties of Cartesian products)**.** *Let A, B, C, D be sets. Then:*

(i) $\langle a, b\rangle \in A \times B$ *if and only if* $a \in A$ *and* $b \in B$.

(ii) $A \times \varnothing = \varnothing = \varnothing \times A$.

(iii) If $A \subseteq C$ *and* $B \subseteq D$, *then* $A \times B \subseteq C \times D$.

(iv) $(A \cup C) \times B = (A \times B) \cup (C \times B)$.

(v) $(A \cap C) \times B = (A \times B) \cap (C \times B)$.

Proof. Statement (i) is just a restatement of Definition 3.1.6.

For (ii), let us prove that $A \times \varnothing = \varnothing$. An element of $A \times \varnothing$ would have to be an ordered pair $\langle a, b \rangle$ with $a \in A$ and $b \in \varnothing$. But nothing belongs to $\varnothing$. Therefore no such ordered pair exists, and so $A \times \varnothing$ has no elements. The proof that $\varnothing \times A = \varnothing$ is analogous.

For (iii), suppose $\langle a, b \rangle \in A \times B$. Then $a \in A$ and $b \in B$. Since $A \subseteq C$ and $B \subseteq D$, we have $a \in C$ and $b \in D$. Hence $\langle a, b \rangle \in C \times D$. Therefore $A \times B \subseteq C \times D$.

For (iv), let x be any object. Then

$$\begin{aligned} x \in (A \cup C) \times B &\iff x = \langle u, v \rangle \text{ for some } u \in A \cup C \text{ and } v \in B \\ &\iff x = \langle u, v \rangle \text{ for some } v \in B \text{ and } (u \in A \text{ or } u \in C) \\ &\iff x \in A \times B \text{ or } x \in C \times B \\ &\iff x \in (A \times B) \cup (C \times B). \end{aligned}$$

Thus $(A \cup C) \times B = (A \times B) \cup (C \times B)$.

For (v), let x be any object. Then

$$\begin{aligned} x \in (A \cap C) \times B &\iff x = \langle u, v \rangle \text{ for some } u \in A \cap C \text{ and } v \in B \\ &\iff x = \langle u, v \rangle \text{ for some } v \in B \text{ and } (u \in A \text{ and } u \in C) \\ &\iff x \in A \times B \text{ and } x \in C \times B \\ &\iff x \in (A \times B) \cap (C \times B). \end{aligned}$$

So $(A \cap C) \times B = (A \times B) \cap (C \times B)$. □

Remark 3.1.10 (Products are the natural home of relations). A subset of $A \times B$ is simply a collection of ordered pairs with first coordinate in A and second coordinate in B. In Chapter 5 such subsets will be called *relations*. Functions will soon emerge as particularly well behaved relations: they are the relations in which each first coordinate is matched with exactly one second coordinate.

3.2 Functions as Assignments and as Graphs

Most students first encounter functions in calculus or algebra through formulas. That is an excellent beginning, but it is not the full story. The modern notion of function is wider and more flexible. What matters is not the presence of a formula, but the existence of a unique output for each allowed input.

Functions as input-output rules

Definition 3.2.1 (Function as an assignment)**.** Let A and B be sets. A *function* from A to B, written $f : A \to B$, is a rule that assigns to each element $a \in A$ exactly one element $b \in B$. That unique output is written $f(a)$.

> *A function is not required to come from a formula. It is enough that each permitted input has one and only one output.*

Example 3.2.2 (First examples of functions)**.** (i) The rule $s : \mathbb{R} \to \mathbb{R}$ given by $s(x) = x^2$ is a function. Each real number has exactly one square.

(ii) Let $p : \mathbb{Z} \to \{0, 1\}$ be defined by

$$p(n) = \begin{cases} 0, & \text{if } n \text{ is even,} \\ 1, & \text{if } n \text{ is odd.} \end{cases}$$

This is a function, even though it is not given by a single algebraic formula.

(iii) Let $A = \{1, 2, 3, 4\}$ and $B = \{a, b\}$. Define $f : A \to B$ by

$$f(1) = a, \qquad f(2) = b, \qquad f(3) = a, \qquad f(4) = b.$$

This is again a perfectly good function.

Remark 3.2.3 (What is not a function). Not every informal rule defines a function. For example, the rule "send $x \geq 0$ to $\pm\sqrt{x}$" is *not* a function from $[0, \infty)$ to $\mathbb{R}$, because a positive input such as 4 is assigned two outputs, namely 2 and -2. The phrase "exactly one" in Definition 3.2.1 is essential.

Functions as special sets of ordered pairs

The language of assignments is intuitive, but set theory also offers a very concrete way to package a function: record every input together with its output.

Definition 3.2.4 (Graph of a function)**.** Let $f : A \to B$ be a function. The *graph* of f is the set

$$\Gamma_f = \{\langle a, f(a)\rangle \mid a \in A\}.$$

Thus Γ_f is a subset of $A \times B$.

Example 3.2.5 (Graphs in finite and infinite settings)**.** (i) For the finite function in Example 3.2.2(iii), the graph is

$$\Gamma_f = \{\langle 1, a\rangle, \langle 2, b\rangle, \langle 3, a\rangle, \langle 4, b\rangle\}.$$

(ii) If $s : \mathbb{R} \to \mathbb{R}$ is given by $s(x) = x^2$, then

$$\Gamma_s = \{\langle x, x^2\rangle \mid x \in \mathbb{R}\}.$$

Geometrically, this is the familiar parabola in the plane.

Proposition 3.2.6 (When a set of ordered pairs is the graph of a function)**.** *Let A and B be sets, and let $G \subseteq A \times B$. Then G is the graph of a function $f : A \to B$ if and only if the following condition holds:*

for every $a \in A$, there exists exactly one $b \in B$ such that $\langle a, b\rangle \in G$.

When this condition holds, the function is determined by the rule "$f(a)$ is the unique b with $\langle a, b\rangle \in G$."

Proof. First suppose that G is the graph of a function $f : A \to B$. Then

$$G = \Gamma_f = \{\langle a, f(a)\rangle \mid a \in A\}.$$

Fix $a \in A$. By construction, $\langle a, f(a)\rangle \in G$, so there is at least one suitable b, namely $b = f(a)$. If $\langle a, b\rangle \in G$, then this pair belongs to the graph of f, so it must be of the form $\langle a, f(a)\rangle$. By Theorem 3.1.4, we conclude that $b = f(a)$. Thus the suitable b is unique.

Conversely, suppose that the stated uniqueness condition holds. For $a \in A$, let $f(a)$ denote the unique element $b \in B$ such that $\langle a, b\rangle \in G$. This defines a function $f : A \to B$. By construction, every pair in G has the form $\langle a, f(a)\rangle$, and every pair $\langle a, f(a)\rangle$ lies in G. Therefore $G = \Gamma_f$, so G is indeed the graph of the function f. □

Remark 3.2.7 (Why set theorists identify functions with graphs). Proposition 3.2.6 shows that a function can be recovered completely from its graph. For that reason, set theory often identifies a function with its graph itself, that is, with the corresponding set of ordered pairs. Texts such as Halmos [2], Devlin [5], and Pinter [6] use this viewpoint extensively. We shall do the same whenever it is convenient, while continuing to speak in the more intuitive language of inputs and outputs.

3.3 Domain, Codomain, Image, and Preimage

Once functions are in place, we need vocabulary for talking about where they start, where they land, and which values they actually attain. These distinctions are simple, but they matter constantly. In particular, the *codomain* of a function is not always the same as its *image*.

The four basic notions

Definition 3.3.1 (Domain, codomain, and image)**.** Let $f : A \to B$ be a function.

(i) The set A is the *domain* of f. We also write $\mathrm{dom}(f) = A$. If we identify f with its graph, then

$$\mathrm{dom}(f) = \{a \mid \langle a, b\rangle \in f \text{ for some } b\}.$$

(ii) The set B is a chosen *codomain* of f.

(iii) The *image* (or *range*) of f is the set

$$\mathrm{ran}(f) = \{f(a) \mid a \in A\}.$$

If we identify f with its graph, then equally

$$\mathrm{ran}(f) = \{b \mid \langle a, b\rangle \in f \text{ for some } a\}.$$

Thus $\mathrm{ran}(f)$ is the set of actual outputs of f, and it is a subset of the codomain.

Example 3.3.2 (Codomain versus image)**.** Consider the squaring function

$$s : \mathbb{R} \to \mathbb{R}, \qquad s(x) = x^2.$$

Its domain is $\mathbb{R}$, and the chosen codomain is also $\mathbb{R}$. But its image is only

$$\operatorname{ran}(s) = [0, \infty),$$

because a square is never negative.

Now consider the same assignment with a different codomain:

$$\widetilde{s}\colon \mathbb{R} \to [0, \infty), \qquad \widetilde{s}(x) = x^2.$$

The actual output values have not changed at all, but the codomain has. In the second presentation the image equals the codomain, while in the first it does not. This simple example shows why codomain and image must not be confused.

Notation	*Name*	*Meaning*
$f : A \to B$	function	inputs from A, outputs in B
$\operatorname{dom}(f) = A$	domain	the allowed inputs
B	codomain	the target set in the notation $f : A \to B$
$\operatorname{ran}(f)$	image/range	the values actually attained

A great deal of confusion disappears once these three sets are kept separate.

Images and preimages of subsets

Functions act not only on individual elements but also on subsets. This leads to two constructions that behave differently and are both important.

Definition 3.3.3 (Image and preimage of a subset)**.** Let $f : A \to B$ be a function.

(i) If $S \subseteq A$, the *image of S under f* is the set

$$f[S] = \{f(a) \mid a \in S\}.$$

(ii) If $T \subseteq B$, the *preimage* (or *inverse image*) of T under f is the set

$$f^{-1}[T] = \{a \in A \mid f(a) \in T\}.$$

Example 3.3.4 (Images and preimages for the squaring map)**.** Let $s : \mathbb{R} \to \mathbb{R}$ be given by $s(x) = x^2$.

(i) If $S = [-1, 2]$, then

$$s[S] = [0, 4].$$

Indeed, every square of a number between -1 and 2 lies between 0 and 4, and every value in $[0, 4]$ occurs as the square of some number in $[-1, 2]$.

(ii) If $T = [0, 1]$, then

$$s^{-1}[T] = [-1, 1],$$

because $x^2 \in [0, 1]$ exactly when $-1 \le x \le 1$.

(iii) If $T = (-\infty, 0)$, then

$$s^{-1}[T] = \varnothing,$$

because no square is negative.

Theorem 3.3.5 (How images and preimages interact with set operations)**.** *Let $f : A \to B$ be a function. For all subsets $S, T \subseteq A$ and $U, V \subseteq B$, the following hold:*

(i) $f[S \cup T] = f[S] \cup f[T]$.

(ii) $f[S \cap T] \subseteq f[S] \cap f[T]$.

(iii) $f^{-1}[U \cup V] = f^{-1}[U] \cup f^{-1}[V]$.

(iv) $f^{-1}[U \cap V] = f^{-1}[U] \cap f^{-1}[V]$.

(v) $f^{-1}[B \setminus U] = A \setminus f^{-1}[U]$.

Proof. For (i), let y be any object. Then

$$\begin{aligned} y \in f[S \cup T] &\iff y = f(a) \text{ for some } a \in S \cup T \\ &\iff y = f(a) \text{ for some } a \in S \text{ or for some } a \in T \\ &\iff y \in f[S] \text{ or } y \in f[T] \\ &\iff y \in f[S] \cup f[T]. \end{aligned}$$

Hence $f[S \cup T] = f[S] \cup f[T]$.

For (ii), let $y \in f[S \cap T]$. Then $y = f(a)$ for some $a \in S \cap T$. In particular, $a \in S$ and $a \in T$, so $y \in f[S]$ and $y \in f[T]$. Thus $y \in f[S] \cap f[T]$. Therefore $f[S \cap T] \subseteq f[S] \cap f[T]$.

For (iii), let $a \in A$. Then

$$\begin{aligned} a \in f^{-1}[U \cup V] &\iff f(a) \in U \cup V \\ &\iff f(a) \in U \text{ or } f(a) \in V \\ &\iff a \in f^{-1}[U] \text{ or } a \in f^{-1}[V] \\ &\iff a \in f^{-1}[U] \cup f^{-1}[V]. \end{aligned}$$

So $f^{-1}[U \cup V] = f^{-1}[U] \cup f^{-1}[V]$.

For (iv), let $a \in A$. Then

$$\begin{aligned} a \in f^{-1}[U \cap V] &\iff f(a) \in U \cap V \\ &\iff f(a) \in U \text{ and } f(a) \in V \\ &\iff a \in f^{-1}[U] \text{ and } a \in f^{-1}[V] \\ &\iff a \in f^{-1}[U] \cap f^{-1}[V]. \end{aligned}$$

Hence $f^{-1}[U \cap V] = f^{-1}[U] \cap f^{-1}[V]$.

For (v), let $a \in A$. Then

$$\begin{aligned} a \in f^{-1}[B \setminus U] &\iff f(a) \in B \setminus U \\ &\iff f(a) \in B \text{ and } f(a) \notin U \\ &\iff a \in A \text{ and } a \notin f^{-1}[U] \\ &\iff a \in A \setminus f^{-1}[U]. \end{aligned}$$

Therefore $f^{-1}[B \setminus U] = A \setminus f^{-1}[U]$. □

Example 3.3.6 (Why the image of an intersection may be smaller)**.** Let $s : \mathbb{R} \to \mathbb{R}$ be given by $s(x) = x^2$, and let

$$S = [-1, 0], \qquad T = [0, 1].$$

Then $S \cap T = \{0\}$, so

$$s[S \cap T] = s[\{0\}] = \{0\}.$$

On the other hand,

$$s[S] = [0, 1] = s[T],$$

so

$$s[S] \cap s[T] = [0, 1].$$

Hence

$$s[S \cap T] \subsetneq s[S] \cap s[T].$$

This is one of the first places where image and preimage behave asymmetrically.

Remark 3.3.7 (Preimage is not the same as inverse function). The notation $f^{-1}[T]$ makes sense for *every* function f and every subset T of the codomain. It refers to a subset of the domain. By contrast, an *inverse function* need not exist at all. When we later speak of an inverse function, we will write it as $f^{-1} : B \to A$ without brackets, and it will exist only for bijections.

3.4 Composition, Identity, Inverse, Injection, Surjection, Bijection

Functions become mathematically powerful when we can combine them, compare them, and sometimes reverse them. Composition expresses the idea of "do one function, then another." Injectivity and surjectivity tell us whether information is lost or whether every possible target value is reached. Bijections, which are both injective and surjective, will later become the basic tool for comparing the sizes of sets.

The algebra of composition

Definition 3.4.1 (Composition and identity)**.** Let $f : A \to B$ and $g : B \to C$ be functions.

(i) Their *composition* is the function $g \circ f : A \to C$ defined by

$$(g \circ f)(a) = g(f(a)) \qquad \text{for all } a \in A.$$

Thus one first applies f and then applies g.

(ii) The *identity function* on a set A is the function $\mathrm{id}_A : A \to A$ given by

$$\mathrm{id}_A(a) = a \qquad \text{for all } a \in A.$$

Example 3.4.2 (Order matters in composition)**.** Let $f, g : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = x + 1, \qquad g(x) = x^2.$$

Then

$$(g \circ f)(x) = g(x+1) = (x+1)^2,$$

whereas

$$(f \circ g)(x) = f(x^2) = x^2 + 1.$$

These are different functions. Composition is therefore not commutative in general.

Proposition 3.4.3 (Associativity of composition and identity laws)**.** *Let $f : A \to B$, $g : B \to C$, and $h : C \to D$ be functions. Then:*

(i) $h \circ (g \circ f) = (h \circ g) \circ f$.

(ii) $\mathrm{id}_B \circ f = f$ *and* $f \circ \mathrm{id}_A = f$.

Proof. For (i), let $a \in A$. Then

$$(h \circ (g \circ f))(a) = h\big((g \circ f)(a)\big) = h(g(f(a))) = ((h \circ g) \circ f)(a).$$

Since the two functions have the same value at every element of A, they are equal.

For (ii), let $a \in A$. Then

$$(\mathrm{id}_B \circ f)(a) = \mathrm{id}_B(f(a)) = f(a)$$

and

$$(f \circ \mathrm{id}_A)(a) = f(\mathrm{id}_A(a)) = f(a).$$

Hence $\mathrm{id}_B \circ f = f$ and $f \circ \mathrm{id}_A = f$. □

Injective, surjective, bijective

The next definitions classify functions according to how faithfully they transport information from the domain to the codomain.

Definition 3.4.4 (Injective, surjective, bijective)**.** Let $f : A \to B$ be a function.

(i) f is *injective* (or *one-to-one*) if whenever $f(a_1) = f(a_2)$, we have $a_1 = a_2$.

(ii) f is *surjective* (or *onto*) if every element of B is actually attained, that is, if for every $b \in B$ there exists $a \in A$ such that $f(a) = b$.

(iii) f is *bijective* if it is both injective and surjective.

Example 3.4.5 (Examples of the three types)**.** (i) The inclusion map $i : \mathbb{N} \to \mathbb{Z}$, given by $i(n) = n$, is injective but not surjective, because negative integers do not lie in its image.

(ii) The absolute-value map $a : \mathbb{Z} \to \mathbb{N}_0$, given by $a(n) = |n|$, is surjective but not injective, because $a(3) = a(-3) = 3$.

(iii) The function $t : \mathbb{R} \to \mathbb{R}$, $t(x) = x + 3$, is bijective. Every real number has exactly one predecessor under this rule.

Proposition 3.4.6 (Composition preserves injectivity and surjectivity)**.** *Let $f : A \to B$ and $g : B \to C$ be functions.*

(i) If f and g are injective, then $g \circ f$ is injective.

(ii) If f and g are surjective, then $g \circ f$ is surjective.

(iii) Consequently, if f and g are bijective, then $g \circ f$ is bijective.

Proof. For (i), assume f and g are injective, and suppose that

$$(g \circ f)(a_1) = (g \circ f)(a_2).$$

Then $g(f(a_1)) = g(f(a_2))$. Since g is injective, $f(a_1) = f(a_2)$. Since f is injective, $a_1 = a_2$. Hence $g \circ f$ is injective.

For (ii), assume f and g are surjective, and let $c \in C$. Since g is surjective, there exists $b \in B$ with $g(b) = c$. Since f is surjective, there exists $a \in A$ with $f(a) = b$. Therefore

$$(g \circ f)(a) = g(f(a)) = g(b) = c.$$

So every element of C lies in the image of $g \circ f$, and thus $g \circ f$ is surjective.

Statement (iii) follows immediately from (i) and (ii). □

Inverse functions

An inverse function reverses the action of a bijection. It is the function-theoretic version of "undoing" a procedure.

Definition 3.4.7 (Inverse function)**.** Let $f : A \to B$ be a function. A function $g : B \to A$ is called an *inverse function* of f if

$$g \circ f = \mathrm{id}_A \qquad \text{and} \qquad f \circ g = \mathrm{id}_B .$$

When such a function exists, it is denoted by f^{-1}.

Theorem 3.4.8 (A function has an inverse exactly when it is bijective)**.** *Let $f : A \to B$ be a function. Then f has an inverse function if and only if f is bijective. When the inverse exists, it is unique.*

Proof. Suppose first that f has an inverse $g : B \to A$. We show that f is injective and surjective.

To prove injectivity, assume that $f(a_1) = f(a_2)$. Apply g to both sides. Then

$$a_1 = (g \circ f)(a_1) = (g \circ f)(a_2) = a_2.$$

Hence f is injective.

To prove surjectivity, let $b \in B$. Then

$$b = \mathrm{id}_B(b) = (f \circ g)(b) = f(g(b)).$$

So b lies in the image of f. Thus f is surjective. Therefore f is bijective.

Conversely, assume that f is bijective. Let $b \in B$. Because f is surjective, there exists at least one $a \in A$ with $f(a) = b$. Because f is injective, there is at most one such a. Hence there exists a unique $a \in A$ with $f(a) = b$. We may therefore define a function $g : B \to A$ by declaring $g(b)$ to be that unique element a.

Now let $a \in A$. Since $f(a)$ is an element of B, the value $g(f(a))$ is the unique element of A that maps to $f(a)$. But a itself maps to $f(a)$, so uniqueness forces $g(f(a)) = a$. Hence $g \circ f = \mathrm{id}_A$.

Similarly, let $b \in B$. By definition of $g(b)$, we have $f(g(b)) = b$. Hence $f \circ g = \mathrm{id}_B$. So g is indeed an inverse of f.

Finally, suppose that both g and h are inverses of f. Then using Proposition 3.4.3,

$$g = g \circ \mathrm{id}_B = g \circ (f \circ h) = (g \circ f) \circ h = \mathrm{id}_A \circ h = h.$$

Thus the inverse is unique. □

Example 3.4.9 (Inverse and non-inverse examples)**.** (i) For the translation $t : \mathbb{R} \to \mathbb{R}$, $t(x) = x + 3$, the inverse is the function $t^{-1}(y) = y - 3$.

(ii) Let $\tau : A \times B \to B \times A$ be defined by

$$\tau\big(\langle a, b\rangle\big) = \langle b, a\rangle.$$

Then τ is bijective, and its inverse is itself: $\tau^{-1} = \tau$.

(iii) The squaring map $s : \mathbb{R} \to \mathbb{R}$, $s(x) = x^2$, has no inverse function from $\mathbb{R}$ to $\mathbb{R}$, because it is neither injective nor surjective.

Corollary 3.4.10 (Inverse of a composite bijection)**.** *If $f : A \to B$ and $g : B \to C$ are bijections, then $g \circ f$ is a bijection and*

$$(g \circ f)^{-1} = f^{-1} \circ g^{-1}.$$

Proof. By Proposition 3.4.6, the composite $g \circ f$ is bijective. Therefore it has an inverse by Theorem 3.4.8. We compute

$$\begin{aligned}(f^{-1} \circ g^{-1}) \circ (g \circ f) &= f^{-1} \circ (g^{-1} \circ g) \circ f \\ &= f^{-1} \circ \mathrm{id}_B \circ f \\ &= f^{-1} \circ f \\ &= \mathrm{id}_A,\end{aligned}$$

and similarly

$$\begin{aligned}(g \circ f) \circ (f^{-1} \circ g^{-1}) &= g \circ (f \circ f^{-1}) \circ g^{-1} \\ &= g \circ \mathrm{id}_B \circ g^{-1} \\ &= g \circ g^{-1} \\ &= \mathrm{id}_C\,.\end{aligned}$$

Hence $f^{-1} \circ g^{-1}$ is the inverse of $g \circ f$, which proves the formula. □

Remark 3.4.11 (Why bijections matter). A bijection pairs the elements of one set with the elements of another without repetition and without omission. That is why bijections will later become our basic instrument for saying that two sets have the same size. Chapters 7–9 will return to this idea again and again.

3.5 Equality of Functions and Extensional Reasoning

By now we have several ways to describe a function: by a formula, by a table, by an arrow diagram, or by a graph. Different descriptions may in fact determine the same function. We therefore need a criterion for function equality analogous to extensionality for sets.

How do we know two functions are the same?

Definition 3.5.1 (Equality of functions)**.** When functions are identified with their graphs, two functions f and g are said to be *equal* if they are equal as sets of ordered pairs.

This definition is formally simple, but in practice one wants a more usable test. Just as two sets are equal when they have the same elements, two functions are equal when they have the same inputs and give the same output at each input.

Theorem 3.5.2 (Extensionality criterion for functions)**.** *Let f and g be functions. Then*

$$f = g \quad \textit{if and only if} \quad \operatorname{dom}(f) = \operatorname{dom}(g) \textit{ and } f(a) = g(a) \textit{ for every } a \in \operatorname{dom}(f).$$

Proof. Assume first that $f = g$. Since the two graphs are equal as sets of ordered pairs, they have the same first coordinates; therefore $\operatorname{dom}(f) = \operatorname{dom}(g)$. Now let $a \in \operatorname{dom}(f)$. Because a lies in the domain of f, the ordered pair $\langle a, f(a)\rangle$ belongs to f. Since $f = g$, it also belongs to g. But the unique pair in g with first coordinate a is $\langle a, g(a)\rangle$. By Theorem 3.1.4, we conclude that $f(a) = g(a)$.

Conversely, suppose that $\operatorname{dom}(f) = \operatorname{dom}(g)$ and that $f(a) = g(a)$ for every a in this common domain. We prove that the graphs are equal by extensionality. Let x be any object. If $x \in f$, then x is an ordered pair, say $x = \langle a, b\rangle$, with $a \in \operatorname{dom}(f)$ and $b = f(a)$. Since $\operatorname{dom}(f) = \operatorname{dom}(g)$, the element a lies in $\operatorname{dom}(g)$, and by assumption $g(a) = f(a) = b$. Hence $x = \langle a, g(a)\rangle \in g$. This shows that $f \subseteq g$. The reverse inclusion is proved in the same way, so $g \subseteq f$. By the double inclusion criterion from Theorem 2.2.9, we obtain $f = g$. □

Example 3.5.3 (Different formulas, same function)**.** (i) Let $f, g\colon \mathbb{R} \to \mathbb{R}$ be given by

$$f(x) = x^2, \qquad g(x) = |x|^2.$$

Since $|x|^2 = x^2$ for every real number x, Theorem 3.5.2 shows that $f = g$.

(ii) Let $h, k\colon \mathbb{R} \setminus \{1\} \to \mathbb{R}$ be given by

$$h(x) = \frac{x^2 - 1}{x - 1}, \qquad k(x) = x + 1.$$

For every $x \neq 1$, we have

$$\frac{x^2 - 1}{x - 1} = x + 1.$$

Hence $h = k$ as functions on $\mathbb{R} \setminus \{1\}$.

Definition 3.5.4 (Restriction)**.** Let $f : A \to B$ be a function, and let $S \subseteq A$. The *restriction* of f to S, written $f \restriction_S$, is the function $f \restriction_S : S \to B$ defined by

$$f \restriction_S (s) = f(s) \qquad \text{for all } s \in S.$$

Example 3.5.5 (Changing the domain changes the function)**.** Let $f : \mathbb{R} \to \mathbb{R}$ be given by $f(x) = x^2$. Its restriction $f \restriction_{[0,\infty)}$ is the function

$$f \restriction_{[0,\infty)} : [0, \infty) \to \mathbb{R}, \qquad f \restriction_{[0,\infty)} (x) = x^2.$$

The formula is the same, but the domain is smaller. Therefore $f \restriction_{[0,\infty)}$ is not equal to f. This is an instance of Theorem 3.5.2: equality of functions requires equality of domains as well as equality of output values.

Remark 3.5.6 (A note on codomain conventions). Different books handle the codomain of a function somewhat differently. Some regard $f : A \to B$ and $f : A \to C$ as different functions whenever $B \neq C$, even if the underlying graph is the same. Others, especially in set-theoretic contexts, identify the function with its graph and treat the codomain as extra information that accompanies a particular presentation. In this book we shall usually follow the second viewpoint when reasoning extensionally: to prove two functions equal, we check that they have the same domain and the same value at every input. At the same time, we will keep codomains visible whenever questions of surjectivity, inverse functions, or later product constructions depend on them.

Remark 3.5.7 (Functions are extensional objects). Theorem 3.5.2 is the function analogue of extensionality for sets. A function has no hidden inner life beyond its domain and its values. How it is described—by a formula, a table, a verbal rule, or an algebraic simplification—does not matter. What matters is what it does.

Looking ahead

This chapter has introduced the basic set-theoretic machinery for remembering order and describing rules. Ordered pairs let us tell first position from second position. Cartesian products assemble all possible pairs from two sets. Functions describe systematic assignments, either as input-output rules or as special sets of ordered pairs. The notions of domain, image, preimage, composition, and bijection now give us a small algebra of mappings.

The next chapter will push this point of view further. Instead of handling only one set or two sets at a time, we will study whole *families of sets*. An indexed family may itself be viewed as a function whose values are sets. From that perspective, general unions, general intersections, and especially general products become natural extensions of the ideas developed here. In particular, the general product $\prod_{i \in I} A_i$ will be the set of all functions choosing one element from each A_i, a construction that eventually leads to the axiom of choice.

Chapter 4

Families of Sets: General Unions, Intersections, and Products

In Chapters 2 and 3 we learned how to manipulate one set at a time and how to pass from two sets to a Cartesian product. We also learned that a function packages a rule into a single mathematical object. The present chapter combines those ideas. We shall study not merely a set, nor merely a pair of sets, but a whole *family of sets* indexed by another set.

At first this may look like a small change of notation. It is not. Once we can speak about a family $(A_i)_{i \in I}$, we can form the union of all the sets in the family, the intersection of all of them, and—most importantly for later chapters—the product of all of them. The first two constructions generalize the familiar operations of Chapter 2; the third generalizes the Cartesian product of Chapter 3 from two factors to arbitrarily many. The conceptual step is that an element of a general product is no longer best pictured as a short tuple. It is best pictured as a *function* whose value at each index chooses one coordinate.

This viewpoint is one of the turning points of elementary set theory. General unions and intersections translate naturally into the language of quantifiers: an element lies in a union when it lies in *at least one* member of the family, and it lies in an intersection when it lies in *every* member of the family. General products go one step further: an element of $\prod_{i \in I} A_i$ is exactly a coherent way of choosing one point from each set A_i. Once we see that, the later meaning of the axiom of choice becomes much less mysterious.

We continue to work in the informal but disciplined spirit explained in Section 1.5. We shall speak freely of indexed families and their products, but we will also watch carefully which set plays the role of the index set, which set supplies the ambient universe when complements or empty intersections are discussed, and which pieces of information are remembered by the notation. As before, the point is not formalism for its own sake, but clarity.

4.1 Indexed Families

Many constructions in mathematics begin with several objects of the same kind: a sequence of numbers, a list of vectors, a collection of intervals, a system of equations, a row of matrices. When the number of objects is small, we can simply write A_1, A_2, A_3. But once the number of objects varies, or becomes large, or is indexed by a set that is not merely $\{1, 2, \dots, n\}$, we need a more flexible language. That language is the language of indexed families.

From lists to indexed collections

Definition 4.1.1 (Indexed family)**.** Let I be a set. An *indexed family of sets* with *index set* I is an assignment that to each $i \in I$ associates a set A_i.

We denote such a family by

$$(A_i)_{i\in I}.$$

Thus the symbol i tells us *which position* in the family we are looking at, and A_i is the set attached to that position.

It is often helpful to think of a family as a function whose inputs are indices and whose outputs are sets. In the informal set-theoretic style of this book, we may therefore regard an indexed family as a set-valued function on I. The point is not that we need a new kind of function, but that ordinary functional language already knows how to handle indexing.

Example 4.1.2 (Finite and infinite families)**.** (i) The list

$$A_1 = \{0, 1\}, \qquad A_2 = \{a, b, c\}, \qquad A_3 = \varnothing$$

is a family indexed by $\{1, 2, 3\}$.

(ii) For each natural number $n \in \mathbb{N}$, let

$$T_n = \{m \in \mathbb{N} \mid m \geq n\}.$$

Then $(T_n)_{n\in\mathbb{N}}$ is an infinite family of subsets of $\mathbb{N}$.

(iii) For each positive real number r, let

$$I_r = (-r, r).$$

Then $(I_r)_{r\in(0,\infty)}$ is a family indexed by the set of positive real numbers. Here the index set is not countable, and not of the form $\{1, 2, \ldots, n\}$.

Remark 4.1.3 (A family is more than a mere set of values). The notation $(A_i)_{i\in I}$ remembers the indices. In particular, a family is not usually determined by the set $\{A_i \mid i \in I\}$ of its values. Two different indices may carry the same set, and that repetition can matter in later constructions. For example, the family

$$A_1 = \{0, 1\}, \qquad A_2 = \{0, 1\}$$

has two positions, even though its set of values has only one member, namely $\{0, 1\}$.

Example 4.1.4 (Repetition is allowed)**.** Let $I = \{1, 2, 3\}$, and define

$$A_1 = \{0, 1\}, \qquad A_2 = \{0, 1\}, \qquad A_3 = \{1, 2\}.$$

Then $(A_i)_{i\in I}$ is a family with three indexed entries, even though only two different sets actually occur among the values. This is harmless for unions and intersections, but it matters for products: the first and second coordinates are distinct coordinate positions even though the underlying sets coincide.

Proposition 4.1.5 (Equality of indexed families)**.** *Let $(A_i)_{i\in I}$ and $(B_i)_{i\in I}$ be two families indexed by the same set I. Then these families are equal if and only if*

$$A_i = B_i \qquad \text{for every } i \in I.$$

Proof. Thinking of the families as functions on I, this is exactly the extensionality criterion for functions from Theorem 3.5.2. Two functions with domain I are equal if and only if they have the same value at every input. Here the inputs are indices and the values are sets. □

Remark 4.1.6 (Why families fit naturally with the earlier chapters). Chapter 3 taught us to regard a function as an object in its own right. Indexed families are one of the first places where that viewpoint becomes genuinely useful. Rather than speaking vaguely of "many sets," we package them into a single function-like object and then apply set operations to its values. This functional viewpoint is standard in elementary set theory texts such as Halmos [2], Enderton [3], and Pinter [6].

4.2 General Unions and General Intersections

Union and intersection were first introduced in Section 2.3 for two sets at a time. Nothing in the underlying idea, however, depends on the number two. We may ask just as naturally whether an element lies in *some* member of a family or in *every* member of a family. The only new ingredient is the index set that tells us which members are under discussion.

> *General union is the set-theoretic form of the phrase "there exists an index for which ..."*
> *General intersection is the set-theoretic form of the phrase "for every index ..."*

That observation is more than a mnemonic. It is the reason that proofs about general unions and intersections almost always amount to a careful translation between membership statements and quantified sentences, just as in Chapter 1.

Definitions and first examples

Definition 4.2.1 (General union and general intersection)**.** Let $(A_i)_{i\in I}$ be a family of sets.

(i) The *general union* of the family is the set

$$\bigcup_{i\in I} A_i = \{x \mid \text{there exists } i \in I \text{ such that } x \in A_i\}.$$

(ii) If $I \neq \varnothing$, the *general intersection* of the family is the set

$$\bigcap_{i\in I} A_i = \{x \mid \text{for every } i \in I,\ x \in A_i\}.$$

Thus $x \in \bigcup_{i\in I} A_i$ means that x belongs to at least one member of the family, while $x \in \bigcap_{i\in I} A_i$ means that x belongs to all of them. Notice how directly this mirrors the logical distinction between existence and universality from Section 1.2.

Example 4.2.2 (Tails of the natural numbers)**.** For each $n \in \mathbb{N}$, let

$$T_n = \{m \in \mathbb{N} \mid m \geq n\}.$$

Then

$$\bigcup_{n \in \mathbb{N}} T_n = \mathbb{N},$$

because every natural number m lies in T_1. On the other hand,

$$\bigcap_{n \in \mathbb{N}} T_n = \varnothing,$$

because no natural number is at least *every* natural number. If m is given, then $m \notin T_{m+1}$.

Example 4.2.3 (A nested family of intervals)**.** For each $n \in \mathbb{N}$, let

$$I_n = \left[-\frac{1}{n}, \frac{1}{n}\right].$$

Then

$$\bigcup_{n \in \mathbb{N}} I_n = [-1, 1],$$

because $I_1 = [-1, 1]$ already contains all the others. But

$$\bigcap_{n \in \mathbb{N}} I_n = \{0\}.$$

Indeed, 0 belongs to every interval I_n. If $x \neq 0$, choose n so large that $1/n < |x|$. Then $x \notin I_n$, so $x \notin \bigcap_{n \in \mathbb{N}} I_n$.

Example 4.2.4 (The index set need not be $\mathbb{N}$)**.** For each positive rational number $q \in \mathbb{Q}$ with $q > 0$, let

$$J_q = (-q, q).$$

Then

$$\bigcup_{q \in \mathbb{Q},\ q>0} J_q = \mathbb{R} \qquad \text{and} \qquad \bigcap_{q \in \mathbb{Q},\ q>0} J_q = \{0\}.$$

The first equality holds because every real number x satisfies $|x| < q$ for some positive rational q. The second holds because only 0 lies in every interval centered at 0, no matter how small the positive rational radius is. This example is a good reminder that families are indexed by arbitrary sets, not only by the natural numbers.

Finite notation as a special case

The earlier notation $A \cup B$ and $A \cap B$ is contained inside these more general definitions. We are not replacing the old notation; we are extending it.

Proposition 4.2.5 (Finite unions and intersections are special cases)**.** *Let $A_1, \ldots, A_n$ be sets, where*

$n \geq 1$, and consider the family indexed by $\{1, \ldots, n\}$. Then

$$\bigcup_{i=1}^{n} A_i = A_1 \cup A_2 \cup \cdots \cup A_n$$

and

$$\bigcap_{i=1}^{n} A_i = A_1 \cap A_2 \cap \cdots \cap A_n.$$

In particular, when $n = 2$, we recover the binary operations of Definition 2.3.1.

Proof. We prove the statement for unions; the proof for intersections is parallel.

Let x be any object. Then

$$x \in \bigcup_{i=1}^{n} A_i$$

means that there exists an index $i \in \{1, \ldots, n\}$ such that $x \in A_i$. But that is exactly the same as saying

$$x \in A_1 \cup A_2 \cup \cdots \cup A_n.$$

Hence the two sets have the same elements, so they are equal by extensionality. □

Remark 4.2.6 (Indexing clarifies what the ellipsis means). The notation $A_1 \cup A_2 \cup \cdots \cup A_n$ is familiar, but it can hide the fact that an index set is already present. Writing $\bigcup_{i=1}^{n} A_i$ makes the indexing explicit and prepares us for families in which the indices are not consecutive integers. This is one reason the indexed notation is preferred in serious mathematical prose.

Basic laws for general unions and intersections

The finite laws from Chapter 2 have natural extensions to indexed families. Their proofs are still element-chasing proofs, but now the logical content is easier to see because the words "there exists" and "for every" appear explicitly.

Proposition 4.2.7 (Monotonicity)**.** *Let $(A_i)_{i \in I}$ and $(B_i)_{i \in I}$ be families indexed by the same set I, and suppose that $A_i \subseteq B_i$ for every $i \in I$. Then:*

(i) $\displaystyle\bigcup_{i \in I} A_i \subseteq \bigcup_{i \in I} B_i.$

(ii) If $I \neq \varnothing$, then $\displaystyle\bigcap_{i \in I} A_i \subseteq \bigcap_{i \in I} B_i.$

Proof. For (i), let $x \in \bigcup_{i \in I} A_i$. Then there exists $i \in I$ such that $x \in A_i$. Since $A_i \subseteq B_i$, we also have $x \in B_i$. Therefore $x \in \bigcup_{i \in I} B_i$.

For (ii), let $x \in \bigcap_{i \in I} A_i$. Then for every $i \in I$, we have $x \in A_i$. Since $A_i \subseteq B_i$ for each index, it follows that $x \in B_i$ for every $i \in I$. Hence $x \in \bigcap_{i \in I} B_i$. □

Proposition 4.2.8 (Distributing a fixed set across a family)**.** *Let $(A_i)_{i \in I}$ be a family of sets, and let C be a set. Then:*

(i)

$$C \cap \bigcup_{i \in I} A_i = \bigcup_{i \in I} (C \cap A_i).$$

(ii) If $I \neq \varnothing$, then

$$C \cup \bigcap_{i \in I} A_i = \bigcap_{i \in I} (C \cup A_i).$$

Proof. We prove (i). Let x be any object. Then

$$\begin{aligned} x \in C \cap \bigcup_{i \in I} A_i &\iff x \in C \text{ and } x \in \bigcup_{i \in I} A_i \\ &\iff x \in C \text{ and there exists } i \in I \text{ such that } x \in A_i \\ &\iff \text{there exists } i \in I \text{ such that } x \in C \text{ and } x \in A_i \\ &\iff \text{there exists } i \in I \text{ such that } x \in C \cap A_i \\ &\iff x \in \bigcup_{i \in I} (C \cap A_i). \end{aligned}$$

Hence the two sets are equal.

For (ii), let x be any object. Then

$$\begin{aligned} x \in C \cup \bigcap_{i \in I} A_i &\iff x \in C \text{ or } x \in \bigcap_{i \in I} A_i \\ &\iff x \in C \text{ or for every } i \in I,\ x \in A_i. \end{aligned}$$

If $x \in C$, then certainly $x \in C \cup A_i$ for every $i \in I$, so $x \in \bigcap_{i \in I} (C \cup A_i)$. If instead $x \notin C$, then the displayed equivalence reduces to the statement that $x \in A_i$ for every i, which is again equivalent to $x \in C \cup A_i$ for every i. Therefore $x \in \bigcap_{i \in I} (C \cup A_i)$.

Conversely, suppose $x \in \bigcap_{i \in I} (C \cup A_i)$. Then for every $i \in I$, we have $x \in C \cup A_i$. If $x \in C$, then $x \in C \cup \bigcap_{i \in I} A_i$. If $x \notin C$, then for each i the statement $x \in C \cup A_i$ forces $x \in A_i$, and so $x \in \bigcap_{i \in I} A_i$. Again $x \in C \cup \bigcap_{i \in I} A_i$. This proves (ii). □

Theorem 4.2.9 (General De Morgan laws)**.** *Let U be an ambient set, and let $(A_i)_{i \in I}$ be a family of subsets of U.*

(i)

$$\left(\bigcup_{i \in I} A_i \right)^c = \bigcap_{i \in I} A_i^c,$$

where the complements are taken in U, and if $I = \varnothing$ the right-hand side is understood relative to the ambient set U.

(ii) If $I \neq \varnothing$, then

$$\left(\bigcap_{i \in I} A_i \right)^c = \bigcup_{i \in I} A_i^c.$$

Proof. We prove (i). Let $x \in U$. Then

$$\begin{aligned} x \in \left(\bigcup_{i\in I} A_i\right)^c &\iff x \notin \bigcup_{i\in I} A_i \\ &\iff \text{it is not true that there exists } i \in I \text{ with } x \in A_i \\ &\iff \text{for every } i \in I,\ x \notin A_i \qquad \text{by Proposition 1.2.8} \\ &\iff \text{for every } i \in I,\ x \in A_i^c \\ &\iff x \in \bigcap_{i\in I} A_i^c. \end{aligned}$$

Therefore $(\bigcup_{i\in I} A_i)^c = \bigcap_{i\in I} A_i^c$.

For (ii), let $x \in U$. Then

$$\begin{aligned} x \in \left(\bigcap_{i\in I} A_i\right)^c &\iff x \notin \bigcap_{i\in I} A_i \\ &\iff \text{it is not true that for every } i \in I,\ x \in A_i \\ &\iff \text{there exists } i \in I \text{ such that } x \notin A_i \qquad \text{by Proposition 1.2.8} \\ &\iff \text{there exists } i \in I \text{ such that } x \in A_i^c \\ &\iff x \in \bigcup_{i\in I} A_i^c. \end{aligned}$$

Hence $(\bigcap_{i\in I} A_i)^c = \bigcup_{i\in I} A_i^c$. □

Remark 4.2.10 (Empty unions and empty intersections). The union over an empty index set is unproblematic:

$$\bigcup_{i\in\varnothing} A_i = \varnothing,$$

because there is no index witnessing membership. The empty intersection is subtler. If we write $\{x \mid \text{for every } i \in \varnothing,\ x \in A_i\}$, then every x satisfies the condition vacuously, so the result should be the whole ambient universe. But in our informal development there is no single absolute universal set available. For that reason we use empty intersections only relative to a previously fixed ambient set U, in which case the empty intersection is understood to be U. This is entirely analogous to the caution about complements in Remark 2.3.7. Discussions of this point appear in texts such as Devlin [5] and Moschovakis [7].

4.3 Disjoint Unions and Partitions of a Set

A general union forgets which set an element came from. If $x \in A_i \cup A_j$, then as an element of the union it is simply the object x; the union does not record whether x arrived through A_i, through A_j, or through both. Sometimes that forgetfulness is harmless. Sometimes it is exactly what we want. But when the member sets do not overlap, the union behaves especially cleanly, because each element comes from a unique place.

Pairwise disjoint families

Definition 4.3.1 (Pairwise disjoint family)**.** A family $(A_i)_{i\in I}$ of sets is said to be *pairwise disjoint* if

$$A_i \cap A_j = \varnothing \qquad \text{whenever } i, j \in I \text{ and } i \neq j.$$

Thus different members of the family have no elements in common. This condition is stronger than merely requiring the total intersection to be empty. For example, three sets may have empty common intersection even though two of them still overlap.

Example 4.3.2 (Pairwise disjoint and not pairwise disjoint)**.** (i) The family

$$\{2n \mid n \in \mathbb{Z}\}, \qquad \{2n+1 \mid n \in \mathbb{Z}\}$$

is pairwise disjoint: no integer is both even and odd.

(ii) Let

$$A_1 = \{1, 2\}, \qquad A_2 = \{2, 3\}, \qquad A_3 = \{1, 3\}.$$

Then $A_1 \cap A_2 \cap A_3 = \varnothing$, but the family is *not* pairwise disjoint, because $A_1 \cap A_2 = \{2\}$, $A_1 \cap A_3 = \{1\}$, and $A_2 \cap A_3 = \{3\}$.

Proposition 4.3.3 (Unique location inside a pairwise disjoint union)**.** *Let $(A_i)_{i\in I}$ be a pairwise disjoint family of sets. If $x \in \bigcup_{i\in I} A_i$, then there exists a unique index $i \in I$ such that $x \in A_i$.*

Proof. Because $x \in \bigcup_{i\in I} A_i$, there exists at least one index $i \in I$ with $x \in A_i$. We must show uniqueness.

Suppose that $x \in A_i$ and $x \in A_j$. Then $x \in A_i \cap A_j$. Since the family is pairwise disjoint, the intersection $A_i \cap A_j$ is empty whenever $i \neq j$. Therefore we cannot have $i \neq j$. Hence $i = j$, which proves uniqueness. □

Remark 4.3.4 (Disjoint unions as especially well-behaved unions). When a family is pairwise disjoint, its union is often called a *disjoint union*. In such a union, every element has a unique home. This is why disjoint unions are so useful in classification arguments: they allow us to decompose a set into pieces without ambiguity.

Tagged disjoint unions

Even if the original family overlaps, there is a simple set-theoretic trick that forces the pieces apart: attach the index to each element. This keeps track of where the element came from.

Remark 4.3.5 (Tagged disjoint unions). Given any family $(A_i)_{i\in I}$, the family $(\{i\}\times A_i)_{i\in I}$ is pairwise disjoint. Indeed, if $i \neq j$, then an element of $\{i\} \times A_i$ has first coordinate i, while an element of $\{j\} \times A_j$ has first coordinate j, so the two sets cannot overlap.

For that reason one often defines the *tagged disjoint union* by

$$\bigsqcup_{i\in I} A_i = \bigcup_{i\in I} (\{i\} \times A_i).$$

If the original sets are already pairwise disjoint, this tagging is not needed; if they overlap, the tag remembers which copy of an element we meant. For example, if $A_1 = A_2 = \{0, 1\}$, then

$$A_1 \cup A_2 = \{0, 1\}$$

has only two elements, but

$$\bigsqcup_{i=1}^{2} A_i = \{\langle 1, 0\rangle, \langle 1, 1\rangle, \langle 2, 0\rangle, \langle 2, 1\rangle\}$$

has four. The tags distinguish the first copy of 0 from the second copy of 0. This construction is standard in set-theoretic and algebraic contexts alike.

Partitions

The cleanest kind of disjoint union is one that exhausts a given set. Such decompositions will reappear in the next chapter when we study relations and equivalence classes.

Definition 4.3.6 (Partition)**.** Let X be a set. A family $(P_i)_{i\in I}$ of subsets of X is a *partition* of X if the following conditions hold:

(i) each P_i is nonempty;

(ii) the family is pairwise disjoint;

(iii) $X = \bigsqcup_{i\in I} P_i$.

In words, a partition breaks X into nonempty pieces that do not overlap and whose union is the whole set.

Example 4.3.7 (First examples of partitions)**.** (i) The sets of even and odd integers form a partition of $\mathbb{Z}$:

$$E = \{2n \mid n \in \mathbb{Z}\}, \qquad O = \{2n + 1 \mid n \in \mathbb{Z}\}.$$

(ii) For any set X, the family of all singleton sets $\big(\{x\}\big)_{x\in X}$ forms a partition of X.

(iii) On the set $\{1, 2, 3, 4, 5, 6\}$, the three subsets

$$\{1, 4\}, \qquad \{2, 5, 6\}, \qquad \{3\}$$

form a partition.

Proposition 4.3.8 (Every element lies in exactly one part)**.** *Let $(P_i)_{i\in I}$ be a partition of a set X. Then every $x \in X$ belongs to exactly one set P_i.*

Proof. Because $X = \bigcup_{i\in I} P_i$, each $x \in X$ belongs to at least one part. Because the family is pairwise disjoint, Proposition 4.3.3 shows that it belongs to at most one part. Therefore it belongs to exactly one part. □

Remark 4.3.9 (Why partitions matter). A partition is a controlled way of saying that a set has been broken into cases. In Chapter 5, partitions will arise from equivalence relations, and order-theoretic texts such as Davey and Priestley [8] use this language constantly. The key idea is already present here: sometimes understanding a set means understanding the pieces into which it naturally decomposes.

4.4 General Cartesian Products

A union gathers together elements that may come from different sets. A product does something very different: it records a simultaneous choice of one coordinate from each set in the family. For two sets we already know this construction: $A \times B$ consists of ordered pairs $\langle a, b \rangle$ with $a \in A$ and $b \in B$. For a family of many sets, especially an infinite family, the correct replacement for an ordered pair is a function on the index set.

This is one of the clearest examples of the usefulness of the function viewpoint from Chapter 3. A point in a general product is not best understood as a mysterious giant tuple. It is simply a function whose value at index i lies in the set A_i.

Definition and interpretation

Definition 4.4.1 (General Cartesian product)**.** Let $(A_i)_{i \in I}$ be a family of sets. The *general Cartesian product* of the family is the set

$$\prod_{i \in I} A_i = \big\{ f \mid \operatorname{dom}(f) = I \text{ and } f(i) \in A_i \text{ for every } i \in I \big\}.$$

An element of $\prod_{i \in I} A_i$ is therefore a function on the index set I that chooses one element from each factor.

To belong to a general product is to make one coherent choice at every index.

If all the factors are the same set A, we often write

$$A^I = \prod_{i \in I} A.$$

Thus A^I is the set of all functions from I to A. The notation suggests exponentiation, and later chapters on cardinality will explain why that suggestion is not accidental.

Example 4.4.2 (A finite three-fold product)**.** Let

$$A_1 = \{0, 1\}, \qquad A_2 = \{a, b\}, \qquad A_3 = \{x\}.$$

An element of $\prod_{i=1}^{3} A_i$ is a function $f : \{1, 2, 3\} \to A_1 \cup A_2 \cup A_3$ such that $f(1) \in A_1$, $f(2) \in A_2$, and $f(3) \in A_3$. There are four such functions. We usually display them suggestively as triples:

$$(0, a, x), \qquad (0, b, x), \qquad (1, a, x), \qquad (1, b, x).$$

Behind this notation, however, each triple is really a function on the index set $\{1, 2, 3\}$.

Proposition 4.4.3 (The general product extends the ordinary product)**.** *Let $I = \{1, 2\}$, and let $(A_i)_{i\in I}$ be a family of sets. Then the map*

$$\Phi\colon \prod_{i\in I} A_i \to A_1 \times A_2, \qquad \Phi(f) = \langle f(1), f(2)\rangle$$

is a bijection.

Proof. We first show that Φ is well defined. If $f \in \prod_{i\in I} A_i$, then $f(1) \in A_1$ and $f(2) \in A_2$ by Definition 4.4.1. Therefore $\langle f(1), f(2)\rangle \in A_1 \times A_2$ by Proposition 3.1.9(i).

To prove injectivity, suppose $\Phi(f) = \Phi(g)$. Then

$$\langle f(1), f(2)\rangle = \langle g(1), g(2)\rangle.$$

By Theorem 3.1.4, we have $f(1) = g(1)$ and $f(2) = g(2)$. Since both functions have the same domain $\{1, 2\}$, Theorem 3.5.2 gives $f = g$.

To prove surjectivity, let $\langle a, b\rangle \in A_1 \times A_2$. Define a function f on $\{1, 2\}$ by

$$f(1) = a, \qquad f(2) = b.$$

Then $f \in \prod_{i\in I} A_i$, and $\Phi(f) = \langle a, b\rangle$. Thus Φ is bijective. □

Example 4.4.4 (Constant products and sequences)**.** (i) If $A = \{0, 1\}$ and $I = \mathbb{N}$, then

$$A^{\mathbb{N}} = \prod_{n\in\mathbb{N}} \{0, 1\}$$

is the set of all infinite binary sequences $(a_1, a_2, a_3, \dots)$ with each $a_n \in \{0, 1\}$.

(ii) If $A = \mathbb{R}$ and $I = \mathbb{N}$, then

$$\mathbb{R}^{\mathbb{N}} = \prod_{n\in\mathbb{N}} \mathbb{R}$$

is the set of all real sequences.

(iii) If $I = X$ is any set, then $\{0, 1\}^X$ is the set of all 0-1-valued functions on X. In a moment we shall see that these functions encode subsets of X.

When are general products empty?

The general product behaves differently from general unions and intersections in one striking respect: it may fail to have any elements at all even when each factor is perfectly legitimate as a set. The reason is simple. To produce one point of the product, we must make a choice in every coordinate simultaneously.

Proposition 4.4.5 (An empty factor makes the whole product empty)**.** *Let $(A_i)_{i\in I}$ be a family of sets. If there exists an index $i_0 \in I$ such that $A_{i_0} = \varnothing$, then*

$$\prod_{i\in I} A_i = \varnothing.$$

Proof. Suppose, for contradiction, that there exists $f \in \prod_{i\in I} A_i$. By definition of the product, we must have $f(i_0) \in A_{i_0}$. But $A_{i_0} = \varnothing$, so no such membership statement is possible. This

contradiction shows that no function f can belong to the product. Hence the product is empty. □

The converse question is more delicate: if every factor is nonempty, must the whole product be nonempty? For finite families the answer is plainly yes. For arbitrary families the answer is precisely the content of the axiom of choice, which we will preview in Section 4.5.

Proposition 4.4.6 (The empty product has exactly one element)**.** *If the index set is empty, then*

$$\prod_{i\in\varnothing} A_i = \{\varnothing\}.$$

In words: the product over an empty family consists of the empty function alone.

Proof. A function belongs to $\prod_{i\in\varnothing} A_i$ exactly when its domain is $\varnothing$ and, for every $i \in \varnothing$, the value $f(i)$ belongs to A_i. The second condition imposes no further restriction, because there is no index i to check. Thus the only question is: which functions have domain $\varnothing$?

There is exactly one such function, namely the empty set viewed as a function with no ordered pairs. Indeed, if f and g both have empty domain, then by Theorem 3.5.2 they are equal, since they have the same domain and there are no input values at which they could differ. Therefore the product has exactly one element, namely $\varnothing$. □

Remark 4.4.7 (Why the empty product is not empty). The empty product can seem paradoxical at first. But the definition is perfectly natural: a point of the product is a rule choosing one element for each index, and when there are no indices, there is exactly one way to do that—namely, do nothing. This is analogous to the convention in ordinary arithmetic that the product of zero numbers is 1: there is one neutral way to multiply nothing at all.

Proposition 4.4.8 (Monotonicity of products)**.** *Let $(A_i)_{i\in I}$ and $(B_i)_{i\in I}$ be families indexed by the same set I, and suppose that $A_i \subseteq B_i$ for every $i \in I$. Then*

$$\prod_{i\in I} A_i \subseteq \prod_{i\in I} B_i.$$

Proof. Let $f \in \prod_{i\in I} A_i$. Then $\operatorname{dom}(f) = I$ and $f(i) \in A_i$ for every $i \in I$. Since $A_i \subseteq B_i$, we also have $f(i) \in B_i$ for every i. Therefore $f \in \prod_{i\in I} B_i$. □

Characteristic functions and subsets

One of the simplest constant products already contains a great deal of set-theoretic information.

Definition 4.4.9 (Characteristic function)**.** Let X be a set, and let $S \subseteq X$. The *characteristic function* of S is the function

$$\chi_S \colon X \to \{0,1\}$$

defined by

$$\chi_S(x) = \begin{cases} 1, & \text{if } x \in S, \\ 0, & \text{if } x \notin S. \end{cases}$$

Proposition 4.4.10 (Subsets correspond to 0-1-valued functions)**.** *For every set X, the map*

$$\Psi\colon \mathcal{P}(X) \to \{0,1\}^X, \qquad \Psi(S) = \chi_S$$

is a bijection.

Proof. We first show that Ψ is injective. Suppose $\Psi(S) = \Psi(T)$. Then $\chi_S = \chi_T$ as functions on X. Let $x \in X$. If $x \in S$, then $\chi_S(x) = 1$, so $\chi_T(x) = 1$ and hence $x \in T$. Thus $S \subseteq T$. By the same argument with S and T reversed, we get $T \subseteq S$. Therefore $S = T$ by the double inclusion criterion, Theorem 2.2.9.

Next we show surjectivity. Let $f \in \{0,1\}^X$. Define

$$S = \{x \in X \mid f(x) = 1\}.$$

We claim that $f = \chi_S$. Indeed, for each $x \in X$, either $f(x) = 1$ or $f(x) = 0$. In the first case $x \in S$, so $\chi_S(x) = 1 = f(x)$. In the second case $x \notin S$, so $\chi_S(x) = 0 = f(x)$. Thus f and χ_S agree at every point of X, and hence are equal by Theorem 3.5.2. Therefore Ψ is surjective.

So Ψ is bijective. □

Remark 4.4.11 (Why this correspondence matters). Proposition 4.4.10 translates a subset into a sequence of yes-or-no answers. This will be useful again when we study infinite cardinality: the power set $\mathcal{P}(X)$ can be viewed as a space of binary choices indexed by X. The theme already appears in introductory texts such as Halmos [2] and Enderton [3].

4.5 Choice Functions as a First Preview

The definition of the general product already contains the seed of one of the most famous principles in set theory. To specify an element of $\prod_{i\in I} A_i$ is to choose one element from each set A_i. When the family is finite, this feels completely routine. When the family is infinite and the sets have no distinguished elements, the situation becomes surprisingly subtle.

Products as simultaneous choices

Definition 4.5.1 (Choice function)**.** Let $(A_i)_{i\in I}$ be a family of nonempty sets. A *choice function* for the family is a function c with domain I such that

$$c(i) \in A_i \qquad \text{for every } i \in I.$$

Proposition 4.5.2 (Choice functions are exactly product elements)**.** *Let $(A_i)_{i\in I}$ be a family of nonempty sets. A function c is a choice function for this family if and only if*

$$c \in \prod_{i\in I} A_i.$$

Proof. This is an immediate restatement of Definition 4.4.1. Belonging to the product means having domain I and choosing one element from each factor; that is exactly what a choice function does. □

Example 4.5.3 (Concrete choice functions)**.** (i) For each $n \in \mathbb{N}$, let

$$A_n = \{m \in \mathbb{N} \mid m \geq n\}.$$

Then the function $c : \mathbb{N} \to \mathbb{N}$ defined by $c(n) = n$ is a choice function for the family $(A_n)_{n\in\mathbb{N}}$.

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

$$B_n = \left(0, \frac{1}{n}\right).$$

Then

$$c(n) = \frac{1}{n+1}$$

defines a choice function, because $0 < 1/(n+1) < 1/n$ for every n.

(iii) Suppose $(C_i)_{i\in I}$ is a family of nonempty subsets of $\mathbb{N}$. Then the rule

$$c(i) = \min(C_i)$$

defines a choice function. Here the natural order on $\mathbb{N}$ gives a distinguished element in each nonempty subset, namely its least member.

The last example is worth pausing over. It shows that many infinite families admit obvious choice functions because the member sets come with a built-in way to pick a distinguished element. The real difficulty arises when there is no preferred element in each set.

Finite families are easy

Proposition 4.5.4 (A finite nonempty product is nonempty)**.** *Let $n \geq 1$, and let $A_1, \ldots, A_n$ be nonempty sets. Then*

$$\prod_{i=1}^{n} A_i \neq \varnothing.$$

Proof. Because each A_k is nonempty, we may choose an element $a_k \in A_k$ for each $k = 1, \ldots, n$. Define a function $f : \{1, \ldots, n\} \to \bigcup_{i=1}^{n} A_i$ by

$$f(k) = a_k.$$

Then $f(k) \in A_k$ for every k, so $f \in \prod_{i=1}^{n} A_i$. Therefore the product is nonempty. □

Remark 4.5.5 (Why finite choice feels trivial)*.* Proposition 4.5.4 does not depend on any deep principle. When only finitely many sets are involved, we can simply choose one element from the first set, one from the second, and so on. The subtlety of choice begins only when the indexing set is arbitrary. In that case we are no longer describing a finite list of selections, but a single function that performs all the selections at once.

The axiom of choice on the horizon

The language of products gives an elegant way to say what is at stake. For a family of sets $(A_i)_{i\in I}$, the assertion

$$\prod_{i\in I} A_i \neq \varnothing$$

means exactly that there exists a function choosing one element from every factor. If some factor is empty, this is impossible by Proposition 4.4.5. The remarkable question is whether nonemptiness of every factor is already enough to guarantee a product element.

Remark 4.5.6 (Preview of the axiom of choice). A later chapter will study the following statement in depth:

For every indexed family $(A_i)_{i\in I}$ of nonempty sets, the product $\prod_{i\in I} A_i$ is nonempty.

This is one common form of the *axiom of choice*. At first sight it may look obvious, because in everyday life we are used to making choices. But mathematically it is far from trivial: for a completely arbitrary family there may be no explicit rule that singles out one element from each set. The axiom of choice asserts that a global choosing function exists anyway. See Jech [16] or Herrlich [17] for fuller discussions of its many equivalent forms and its role throughout mathematics.

Remark 4.5.7 (Products explain why the axiom of choice matters). Without the language of general products, the axiom of choice can sound like a philosophical slogan. With that language in hand, it becomes a concrete statement about the existence of elements in a certain set. In other words, the axiom of choice is not a separate topic floating above the rest of set theory; it is intimately tied to the product construction introduced in this chapter.

Looking ahead

This chapter has taken us from individual sets to whole indexed families. We learned that a family is best viewed as a set-valued assignment on an index set, that general unions and intersections turn membership questions into existential and universal statements, and that general products are sets of functions choosing one coordinate from each factor. Along the way we met pairwise disjoint families, partitions, characteristic functions, and the first meaningful appearance of choice functions.

The next chapter will use this language to study *relations*. A relation is a subset of a Cartesian product, so the product viewpoint of Chapter 3 and the family viewpoint of the present chapter both feed directly into it. In particular, the partitions introduced here will return in a central theorem: equivalence relations and partitions are two ways of describing the same kind of structure. We will also begin to study order relations, which prepare the ground for well-orderings, ordinals, and the transfinite viewpoint that lies ahead.

Chapter 5

Relations, Equivalence, and Order

In ordinary mathematical language we constantly compare objects in pairs. We say that one integer is less than another, that one number is divisible by another, that two triangles have the same shape, that a point lies on a line, or that two fractions represent the same rational number. Sentences of this sort all have the same broad form: they place some condition on two inputs at once. The set-theoretic language for such two-place conditions is the language of *relations*.

Relations are one of the first places where the earlier chapters begin to work together in a visible way. Chapter 3 introduced ordered pairs and Cartesian products, and a relation is simply a subset of a Cartesian product. Chapter 4 introduced partitions, and one of the central facts of the present chapter is that equivalence relations and partitions are really two ways of describing the same decomposition of a set. Functions told us how to go from each input to a single output; relations are more flexible, for they allow one input to be related to many outputs, or to none at all.

A second theme of this chapter is that not all relations express "sameness." Some express *comparison*. The usual order on the real numbers tells us which number comes first; divisibility on the natural numbers tells us which number is a factor of another; inclusion on $\mathcal{P}(A)$ tells us when one subset sits inside another. These are not equivalence relations, because comparison behaves differently from sameness. Instead they lead to the language of *partial orders* and *total orders*.

At first this may sound like a collection of separate topics, but there is a common question behind all of them:

> *How can a set carry internal structure before we even begin to add, multiply, or measure its elements?*

A relation is the simplest answer. It equips a set with a pattern of comparison. Sometimes that pattern groups elements into classes. Sometimes it arranges them in a line. Sometimes it produces only a partial hierarchy in which some elements can be compared and others cannot. Later chapters will build on all three possibilities. The quotient sets that arise from equivalence relations will reappear in many guises. Ordered sets will lead us toward well-orderings, and well-orderings in turn will lead toward ordinals, transfinite induction, and the axiom of choice.

For now we continue to work in the same informal spirit as before. We shall speak of relations, classes, and orders exactly as one does in an elementary set theory course, without trying yet to reduce everything to an axiomatic foundation. The point is to learn how these structures behave and why they matter.

5.1 Relations and Their Basic Properties

A relation is a convenient way to package any statement involving two places. For example, the sentence "$x < y$" becomes meaningful once we choose two real numbers x and y. Likewise "m divides n," "$A \subseteq B$," and "$x \in S$" are all conditions with two inputs. Since Chapter 3 showed us how to gather ordered pairs into a set, it is natural to encode a relation by collecting exactly those ordered pairs for which the condition holds.

Relations as subsets of Cartesian products

Definition 5.1.1 (Relation between two sets)**.** Let A and B be sets. A *relation from A to B* is a subset

$$R \subseteq A \times B.$$

If $\langle a, b\rangle \in R$, we often write

$$a \, R \, b$$

and say that a is *related* to b by R.

When $A = B$, we call R a *relation on A* .

Thus a relation is not mysterious. It is simply a set of ordered pairs, and the notation aRb is a convenient way to say that the ordered pair $\langle a, b\rangle$ belongs to that set. This is exactly the same set-theoretic move that allowed us to think of a function as a special set of ordered pairs in Chapter 3. The difference is that a general relation need not behave functionally: a single element a may be related to several different elements of B, or to none at all.

Example 5.1.2 (First examples of relations)**.** (i) On $\mathbb{R}$, the usual relation "less than" is

$$< \; = \{\langle x, y\rangle \in \mathbb{R} \times \mathbb{R} \mid x < y\}.$$

Writing $x < y$ is simply shorter than writing $\langle x, y\rangle \in <$.

(ii) On $\mathbb{N}$, divisibility is the relation

$$\mid \; = \{\langle m, n\rangle \in \mathbb{N} \times \mathbb{N} \mid \text{there exists } k \in \mathbb{N} \text{ with } n = mk\}.$$

We write $m \mid n$ and read this as "m divides n."

(iii) Between $\mathbb{N}$ and $\mathcal{P}(\mathbb{N})$, membership gives a relation

$$E = \{\langle n, S\rangle \in \mathbb{N} \times \mathcal{P}(\mathbb{N}) \mid n \in S\}.$$

Here the two coordinates are not the same kind of object: the first is a natural number, the second is a subset of $\mathbb{N}$.

(iv) On $\mathbb{R} \times \mathbb{R}$, define R by

$$\langle x, y\rangle \, R \, \langle u, v\rangle \quad \text{if and only if} \quad x + y = u + v.$$

Thus two ordered pairs are related when they have the same sum of coordinates.

Remark 5.1.3 (Relations need not be functional). If $R \subseteq A \times B$ is a function graph, then every $a \in A$ appears in exactly one pair of the form $\langle a, b \rangle$. A general relation is much looser. Under the membership relation from Example 5.1.2(iii), a fixed number n is related to many sets S, while the empty set is related to no natural number at all. So relations generalize functions, but they are not limited by the "exactly one output" rule.

Definition 5.1.4 (Domain, range, and inverse relation)**.** Let $R \subseteq A \times B$ be a relation from A to B.

(i) The *domain* of R is

$$\mathrm{dom}(R) = \{a \in A \mid \text{there exists } b \in B \text{ such that } aRb\}.$$

(ii) The *range* of R is

$$\mathrm{ran}(R) = \{b \in B \mid \text{there exists } a \in A \text{ such that } aRb\}.$$

(iii) The *inverse relation* R^{-1} is the relation from B to A defined by

$$b \, R^{-1} \, a \quad \text{if and only if} \quad aRb.$$

Equivalently,

$$R^{-1} = \{\langle b, a \rangle \in B \times A \mid \langle a, b \rangle \in R\}.$$

Example 5.1.5 (The membership relation revisited)**.** Let $E \subseteq \mathbb{N} \times \mathcal{P}(\mathbb{N})$ be the membership relation from Example 5.1.2(iii). Then

$$\mathrm{dom}(E) = \mathbb{N},$$

because every natural number n belongs to the singleton set $\{n\}$. On the other hand,

$$\mathrm{ran}(E) = \mathcal{P}(\mathbb{N}) \setminus \{\varnothing\},$$

because a subset of $\mathbb{N}$ is related to some natural number exactly when it is nonempty.

The inverse relation $E^{-1} \subseteq \mathcal{P}(\mathbb{N}) \times \mathbb{N}$ satisfies

$$S \, E^{-1} \, n \quad \text{if and only if} \quad n \in S.$$

So the inverse merely swaps the order in which we view the two inputs.

Proposition 5.1.6 (Extensionality for relations)**.** *Let $R, S \subseteq A \times B$ be relations from A to B. Then*

$$R = S \quad \textit{if and only if} \quad \textit{for all } a \in A \textit{ and } b \in B, \; aRb \textit{ exactly when } aSb.$$

Proof. By Definition 5.1.1, the statement aRb means precisely that $\langle a, b \rangle \in R$, and similarly for S. Thus R and S have the same elements if and only if they contain exactly the same ordered pairs. That is exactly the displayed condition. □

Remark 5.1.7 (A standard point of view). The idea that a relation is a subset of a Cartesian product is standard throughout elementary set theory; see, for example, Halmos [2], Enderton [3], and Pinter [6]. Once one becomes comfortable with this viewpoint, many apparently different notions—equivalence, ordering, divisibility, congruence, membership—become instances of the same underlying pattern.

Reflexivity, symmetry, antisymmetry, and transitivity

Some properties of relations appear so often that they receive special names. These names describe how a relation behaves when we compare an element with itself, swap the two inputs, or chain comparisons together.

Definition 5.1.8 (Basic properties of a relation)**.** Let R be a relation on a set A.

(i) R is *reflexive* if
$$aRa \qquad \text{for every } a \in A.$$

(ii) R is *symmetric* if
$$aRb \implies bRa \qquad \text{for all } a, b \in A.$$

(iii) R is *antisymmetric* if
$$aRb \text{ and } bRa \implies a = b \qquad \text{for all } a, b \in A.$$

(iv) R is *transitive* if
$$aRb \text{ and } bRc \implies aRc \qquad \text{for all } a, b, c \in A.$$

Example 5.1.9 (Testing the basic properties)**.** (i) Equality on any set A is reflexive, symmetric, antisymmetric, and transitive.

(ii) The relation $\leq$ on $\mathbb{R}$ is reflexive, antisymmetric, and transitive, but it is not symmetric. For example, $2 \leq 5$, but $5 \leq 2$ is false.

(iii) The relation $\neq$ on any set with at least two elements is symmetric, but it is neither reflexive nor transitive. Indeed, $a \neq a$ is always false, and if $1 \neq 2$ and $2 \neq 1$, then transitivity would force $1 \neq 1$, which is impossible.

(iv) Divisibility on $\mathbb{N}$ is reflexive, antisymmetric, and transitive. It is not symmetric: $2 \mid 6$, but $6 \nmid 2$.

(v) On $\mathbb{R}$, define $x \sim y$ if and only if $|x| = |y|$. Then $\sim$ is reflexive, symmetric, and transitive, but not antisymmetric, since $1 \sim -1$ while $1 \neq -1$.

Remark 5.1.10 (Antisymmetry is not the opposite of symmetry). Students often first read "antisymmetric" as if it meant "nonsymmetric," but that is not correct. A relation can be both symmetric and antisymmetric: equality is the most important example. A relation can also be antisymmetric without being symmetric, as with $\leq$ or $\subseteq$. Antisymmetry says only that whenever the relation goes in both directions, the two elements must already be the same.

Proposition 5.1.11 (Symmetry and inverse relations)**.** *Let R be a relation on a set A. Then R is symmetric if and only if*

$$R = R^{-1}.$$

Proof. Assume first that R is symmetric. To show $R \subseteq R^{-1}$, let $\langle a, b\rangle \in R$. Then aRb, so by symmetry bRa. Hence $\langle b, a\rangle \in R$, which means $\langle a, b\rangle \in R^{-1}$. Thus $R \subseteq R^{-1}$. The reverse inclusion is proved in exactly the same way, so $R = R^{-1}$.

Conversely, suppose $R = R^{-1}$. If aRb, then $\langle a, b\rangle \in R = R^{-1}$, so by the definition of inverse relation we have bRa. Therefore R is symmetric. □

Remark 5.1.12 (Finite relations as pictures). A finite relation on a set A may be pictured by drawing one dot for each element of A, and an arrow from a to b whenever aRb. Reflexive pairs become loops, symmetry means arrows come in opposite-direction pairs, and transitivity means that whenever there is an arrow from a to b and from b to c, there is also one from a to c. Such pictures are often helpful in small finite examples, although we will not rely on them heavily here.

5.2 Equivalence Relations and Partitions

Among all relations, some are designed to express that two objects are "the same for the purpose at hand." Of course they may not be literally equal. Two integers may have the same remainder upon division by 5; two fractions may represent the same rational number; two points in the plane may lie on the same circle centered at the origin. In each case we are not forgetting distinctions at random. We are organizing a set into classes whose elements should be treated as interchangeable in a particular context.

The idea of "same up to"

Definition 5.2.1 (Equivalence relation)**.** A relation $\sim$ on a set A is an *equivalence relation* if it is reflexive, symmetric, and transitive.

Thus an equivalence relation is exactly the right kind of relation for capturing a notion of "same up to some criterion." Reflexivity says every object is equivalent to itself. Symmetry says equivalence does not depend on the order in which we mention the two objects. Transitivity says that equivalent objects stay equivalent when we pass through an intermediate representative.

Example 5.2.2 (Congruence modulo n)**.** Fix a positive integer n. Define a relation on $\mathbb{Z}$ by

$$a \equiv b \pmod{n} \quad \text{if and only if} \quad n \mid (a - b).$$

This is an equivalence relation.

Indeed, $a - a = 0$, and every positive integer divides 0, so the relation is reflexive. If $n \mid (a-b)$, then $b - a = -(a - b)$, so $n \mid (b - a)$; hence the relation is symmetric. Finally, if $n \mid (a - b)$ and $n \mid (b - c)$, then n divides the sum

$$(a - b) + (b - c) = a - c,$$

so $a \equiv c \pmod{n}$. Therefore the relation is transitive.

For example, modulo 3 the integers split into the three classes

$$[0] = \{\dots, -6, -3, 0, 3, 6, \dots\}, \qquad [1] = \{\dots, -5, -2, 1, 4, 7, \dots\}, \qquad [2] = \{\dots, -4, -1, 2, 5, 8, \dots\}.$$

Example 5.2.3 (Fractions as equivalence classes)**.** Let

$$F = \{\langle a, b\rangle \in \mathbb{Z} \times \mathbb{Z} \mid b \neq 0\}.$$

Define a relation on F by

$$\langle a, b\rangle \sim \langle c, d\rangle \quad \text{if and only if} \quad ad = bc.$$

Then $\sim$ is an equivalence relation.

Reflexivity holds because $ab = ba$. Symmetry is immediate because the equation $ad = bc$ is unchanged when read in reverse as $bc = ad$. For transitivity, suppose

$$ad = bc \qquad \text{and} \qquad cf = de.$$

Then

$$adf = bcf = bde.$$

Since $d \neq 0$, we may cancel d in the integers and obtain $af = be$. Therefore

$$\langle a, b\rangle \sim \langle e, f\rangle.$$

So $\sim$ is transitive.

This relation identifies different fraction symbols that represent the same rational number. For instance,

$$\langle 1, 2\rangle \sim \langle 2, 4\rangle \sim \langle -3, -6\rangle.$$

In a later foundational construction one can define the rational numbers as the equivalence classes of this relation.

Example 5.2.4 (Same distance from the origin)**.** Define a relation on $\mathbb{R} \times \mathbb{R}$ by

$$\langle x, y\rangle \sim \langle u, v\rangle \quad \text{if and only if} \quad x^2 + y^2 = u^2 + v^2.$$

This is an equivalence relation. The equivalence class of a point $\langle x, y\rangle$ consists of all points lying on the circle centered at the origin with radius $\sqrt{x^2 + y^2}$, except that when $x = y = 0$ the class is just $\{\langle 0, 0\rangle\}$.

Equivalence classes

Definition 5.2.5 (Equivalence class and quotient set)**.** Let $\sim$ be an equivalence relation on a set A, and let $a \in A$.

(i) The *equivalence class* of a is the set

$$[a]_{\sim} = \{x \in A \mid x \sim a\}.$$

When the relation is clear from context, we simply write $[a]$.

(ii) The *quotient set* $A/\sim$ is the set of all equivalence classes:

$$A/\sim = \{[a]_\sim \mid a \in A\}.$$

The quotient set does not remember every element of A separately. Instead it remembers only the classes into which the relation groups the elements. One may think of it as the result of collapsing equivalent elements into a single block.

Proposition 5.2.6 (Basic properties of equivalence classes)**.** *Let $\sim$ be an equivalence relation on a set A, and let $a, b \in A$.*

(i) $a \in [a]$.

(ii) If $b \in [a]$, *then* $[b] = [a]$.

(iii) Either $[a] = [b]$ *or* $[a] \cap [b] = \emptyset$.

Proof. For (i), reflexivity gives $a \sim a$, so $a \in [a]$.

For (ii), assume $b \in [a]$. Then $b \sim a$. We show that $[b] \subseteq [a]$. If $x \in [b]$, then $x \sim b$. Since $b \sim a$ and $\sim$ is transitive, we have $x \sim a$. Hence $x \in [a]$. So $[b] \subseteq [a]$.

To prove the reverse inclusion, take $x \in [a]$. Then $x \sim a$. Because $b \sim a$ and $\sim$ is symmetric, we also have $a \sim b$. Transitivity now yields $x \sim b$, so $x \in [b]$. Thus $[a] \subseteq [b]$, and therefore $[a] = [b]$.

For (iii), suppose $[a] \cap [b] \neq \emptyset$. Choose $x \in [a] \cap [b]$. Then $x \in [a]$ and $x \in [b]$. By part (ii), $[x] = [a]$ and $[x] = [b]$. Hence $[a] = [b]$. Therefore if the two classes are not equal, they must be disjoint. □

Theorem 5.2.7 (Equivalence relations give partitions)**.** *Let $\sim$ be an equivalence relation on a set A. Then the family of equivalence classes*

$$A/\sim = \{[a] \mid a \in A\}$$

is a partition of A in the sense of Definition 4.3.6.

Proof. We verify the three conditions in Definition 4.3.6.

First, every equivalence class is nonempty by Proposition 5.2.6(i).

Second, the union of all equivalence classes is A. Indeed, every $a \in A$ belongs to its own class $[a]$, so

$$A \subseteq \bigcup_{x \in A} [x].$$

Conversely, each class $[x]$ is a subset of A, so the union is contained in A.

Third, distinct classes are disjoint by Proposition 5.2.6(iii). Therefore the equivalence classes form a partition of A. □

The theorem says that an equivalence relation carves a set into pieces, and those pieces are exactly the equivalence classes. No element is left out, no class is empty, and no element belongs to two different classes. This is precisely the content of the idea that equivalence means "belonging to the same block."

Theorem 5.2.8 (Partitions give equivalence relations)**.** *Let $\mathcal{P}$ be a partition of a set A. Define a relation $\sim_{\mathcal{P}}$ on A by declaring*

$$x \sim_{\mathcal{P}} y \quad \text{if and only if} \quad \text{there exists } P \in \mathcal{P} \text{ such that } x, y \in P.$$

Then $\sim_{\mathcal{P}}$ is an equivalence relation. Moreover, its equivalence classes are exactly the members of $\mathcal{P}$.

Proof. Because $\mathcal{P}$ is a partition, every element $x \in A$ lies in some block $P \in \mathcal{P}$. Hence $x \sim_{\mathcal{P}} x$, so the relation is reflexive.

If $x \sim_{\mathcal{P}} y$, then x and y lie in a common block $P \in \mathcal{P}$. Then of course y and x also lie in P, so $y \sim_{\mathcal{P}} x$. Thus the relation is symmetric.

Now suppose $x \sim_{\mathcal{P}} y$ and $y \sim_{\mathcal{P}} z$. Then there exist blocks $P, Q \in \mathcal{P}$ such that

$$x, y \in P \qquad \text{and} \qquad y, z \in Q.$$

So $y \in P \cap Q$. Since the members of a partition are pairwise disjoint unless equal, we must have $P = Q$. Therefore $x, z \in P$, and hence $x \sim_{\mathcal{P}} z$. So the relation is transitive.

It remains to show that the equivalence classes are exactly the blocks. Let $P \in \mathcal{P}$, and choose $a \in P$ (possible because every block of a partition is nonempty). We claim that $[a] = P$ for the relation $\sim_{\mathcal{P}}$.

If $x \in [a]$, then $x \sim_{\mathcal{P}} a$. Thus x and a lie in a common block. But $a \in P$, and by Proposition 4.3.8 the block containing a is unique. Hence that common block must be P, so $x \in P$. Therefore $[a] \subseteq P$.

Conversely, if $x \in P$, then x and a lie in the same block P, so $x \sim_{\mathcal{P}} a$. Hence $x \in [a]$. Therefore $P \subseteq [a]$, and thus $[a] = P$. □

Example 5.2.9 (The quotient set $\mathbb{Z}/3\mathbb{Z}$)**.** Under congruence modulo 3, the quotient set $\mathbb{Z}/\sim$ consists of three classes, usually denoted by

$$\mathbb{Z}/3\mathbb{Z} = \{[0], [1], [2]\}.$$

Every integer belongs to exactly one of these classes. For example,

$$[4] = [1], \qquad [-2] = [1], \qquad [8] = [2].$$

So $\mathbb{Z}/3\mathbb{Z}$ is not a new set of mysterious objects; it is simply the set of the three congruence classes modulo 3.

Proposition 5.2.10 (The canonical projection)**.** *Let $\sim$ be an equivalence relation on a set A. Define a function*

$$\pi \colon A \to A/\sim$$

by

$$\pi(a) = [a].$$

Then π is surjective, and for all $a, b \in A$,

$$\pi(a) = \pi(b) \quad \text{if and only if} \quad a \sim b.$$

Proof. The function is surjective because every element of $A/\sim$ is, by definition, some class $[a]$, and $\pi(a) = [a]$.

Now suppose $\pi(a) = \pi(b)$. Then $[a] = [b]$. Since $a \in [a]$, we have $a \in [b]$. By the definition of $[b]$, this means $a \sim b$.

Conversely, if $a \sim b$, then $b \in [a]$. By Proposition 5.2.6(ii), this implies $[a] = [b]$. Therefore $\pi(a) = \pi(b)$. □

Remark 5.2.11 (Equivalence means "same up to"). The correspondence between equivalence relations and partitions is one of the basic organizing facts of elementary set theory. It says that to specify a notion of "same up to" is exactly to specify a way of breaking a set into disjoint blocks. Texts such as Halmos [2], Enderton [3], and Pinter [6] all treat this correspondence as a central structural idea.

5.3 Partial Orders

An equivalence relation groups elements together. An order relation, by contrast, tries to arrange them. In some situations every pair of elements is comparable, as with the usual order on the real numbers. In other situations only some pairs can be compared. A subset of a set may or may not contain another subset; one integer may or may not divide another. These are naturally ordered worlds, but not linearly ordered ones. The correct general notion is that of a partial order.

Comparison without totality

Definition 5.3.1 (Partial order and poset)**.** A relation $\preceq$ on a set P is a *partial order* if it is reflexive, antisymmetric, and transitive. The pair $(P, \preceq)$ is called a *partially ordered set*, or *poset*.

The word "partial" is important. It means that the relation behaves like an order where it is defined, but it does not require that every two elements be comparable.

Definition 5.3.2 (Comparable and incomparable)**.** Let $(P, \preceq)$ be a poset, and let $x, y \in P$.

(i) We say that x and y are *comparable* if either $x \preceq y$ or $y \preceq x$.

(ii) We say that x and y are *incomparable* if neither $x \preceq y$ nor $y \preceq x$ holds.

Example 5.3.3 (Important partial orders)**.** (i) The usual relation $\leq$ on $\mathbb{R}$ is a partial order. Indeed, it is reflexive, antisymmetric, and transitive.

(ii) Divisibility on $\mathbb{N}$ is a partial order. Reflexivity and transitivity were discussed in Example 5.1.9. For antisymmetry, if $m \mid n$ and $n \mid m$, then $n = mk$ and $m = n\ell$ for some natural numbers k, ℓ. Hence $m = m(k\ell)$, so $k\ell = 1$, and therefore $k = \ell = 1$. Thus $m = n$.

(iii) Inclusion $\subseteq$ on $\mathcal{P}(A)$ is a partial order. Reflexivity and transitivity are immediate. Antisymmetry is exactly the double-inclusion criterion from Theorem 2.2.9.

(iv) Reverse inclusion $\supseteq$ on $\mathcal{P}(A)$ is also a partial order. This is a good reminder that the symbol used for an order need not have anything to do with numerical size.

(v) If $(P, \preceq_P)$ and $(Q, \preceq_Q)$ are posets, then the *product order* on $P \times Q$ is defined by

$$\langle p, q\rangle \preceq \langle p', q'\rangle \quad \text{if and only if} \quad p \preceq_P p' \text{ and } q \preceq_Q q'.$$

For instance, in $\mathbb{N} \times \mathbb{N}$ with the product order, $\langle 1, 3\rangle$ and $\langle 2, 2\rangle$ are incomparable.

Remark 5.3.4 (Partial orders compare some pairs, not all). In the divisibility order on $\mathbb{N}$, the numbers 2 and 3 are incomparable: neither divides the other. In the inclusion order on $\mathcal{P}(\{1, 2, 3\})$, the sets $\{1\}$ and $\{2\}$ are likewise incomparable. So a partial order need not arrange its elements in a single line. This failure of universal comparability is not a defect; it is exactly what makes partial orders useful in practice.

Least, greatest, minimal, and maximal elements

Definition 5.3.5 (Least, greatest, minimal, and maximal elements)**.** Let $(P, \preceq)$ be a poset, and let $B \subseteq P$.

(i) An element $\ell \in B$ is a *least element* of B if

$$\ell \preceq x \qquad \text{for every } x \in B.$$

(ii) An element $g \in B$ is a *greatest element* of B if

$$x \preceq g \qquad \text{for every } x \in B.$$

(iii) An element $m \in B$ is *minimal in B* if whenever $x \in B$ and $x \preceq m$, then $x = m$.

(iv) An element $M \in B$ is *maximal in B* if whenever $x \in B$ and $M \preceq x$, then $x = M$.

A least element is below every other element of B. A minimal element merely has nothing strictly smaller than it *inside* B. These are not the same notion, and distinguishing them is one of the first conceptual tests in order theory.

Example 5.3.6 (Minimal need not mean least)**.** (i) Let

$$B = \big\{\{1\}, \{2\}\big\} \subseteq \mathcal{P}(\{1, 2\}),$$

ordered by inclusion. Then both $\{1\}$ and $\{2\}$ are minimal in B, and both are maximal in B. But B has neither a least nor a greatest element, because neither set contains the other.

(ii) Let $D = \{2, 3, 6\} \subseteq \mathbb{N}$, ordered by divisibility. Then 2 and 3 are minimal elements of D, because nothing in D divides either of them except themselves. The element 6 is maximal. But there is no least element of D, since neither $2 \mid 3$ nor $3 \mid 2$.

(iii) In the set of positive divisors of 12, ordered by divisibility, the element 1 is least and 12 is greatest.

Proposition 5.3.7 (Least and greatest elements are unique)**.** *Let $(P, \preceq)$ be a poset, and let $B \subseteq P$.*

(i) If B has a least element, then it is unique.

(ii) If B has a greatest element, then it is unique.

Proof. We prove (i); the proof of (ii) is analogous. Suppose ℓ and ℓ' are both least elements of B. Since ℓ is least, $\ell \preceq \ell'$. Since ℓ' is least, $\ell' \preceq \ell$. By antisymmetry, $\ell = \ell'$. □

Proposition 5.3.8 (Least implies minimal, greatest implies maximal)**.** *Let $(P, \preceq)$ be a poset, and let $B \subseteq P$.*

(i) Every least element of B is minimal in B.

(ii) Every greatest element of B is maximal in B.

Proof. We prove (i). Let ℓ be a least element of B. Suppose $x \in B$ and $x \preceq \ell$. Since ℓ is least, we also have $\ell \preceq x$. By antisymmetry, $x = \ell$. Hence ℓ is minimal.

The proof of (ii) is the same argument with the order reversed. □

Remark 5.3.9 (Associated strict order). If $\preceq$ is a partial order, we often write

$$x \prec y \quad \text{to mean} \quad x \preceq y \text{ and } x \neq y.$$

Then an element $m \in B$ is minimal exactly when there is no $x \in B$ with $x \prec m$, and maximal exactly when there is no $x \in B$ with $m \prec x$. This notation is especially convenient when drawing Hasse diagrams later in the chapter.

Remark 5.3.10 (Order theory beyond this chapter). Even the basic notions introduced here already lead to a rich subject. Upper and lower bounds, suprema and infima, chains and antichains, and lattice operations all belong to the larger world of order theory. For an accessible introduction to that broader picture, see Davey and Priestley [8].

5.4 Total Orders, Lexicographic Orders, and Hasse Diagrams

Partial orders allow incomparability. But sometimes we do have a true linear arrangement: every pair of elements can be placed one before the other. That stronger situation leads to total orders. It is familiar from the usual order on numbers, but it also appears in less obvious settings such as dictionary order on words or lexicographic order on pairs.

Total orders

Definition 5.4.1 (Total order)**.** A partial order $\preceq$ on a set L is a *total order* if every two elements of L are comparable; that is, for all $x, y \in L$,

$$x \preceq y \quad \text{or} \quad y \preceq x.$$

A totally ordered set is also called a *linearly ordered set*.

Example 5.4.2 (Total and nontotal orders)**.** (i) The usual order $\leq$ on $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$ is total.

(ii) Inclusion on $\mathcal{P}(A)$ is usually not total when A has at least two elements. For example, if $A = \{1, 2\}$, then $\{1\}$ and $\{2\}$ are incomparable.

(iii) Divisibility on $\mathbb{N}$ is not total, since 2 and 3 are incomparable.

Proposition 5.4.3 (In a total order, minimal means least)**.** *Let $(L, \leq)$ be a totally ordered set, and let $B \subseteq L$.*

(i) If $m \in B$ is minimal in B, then m is the least element of B.

(ii) If $M \in B$ is maximal in B, then M is the greatest element of B.

Proof. We prove (i). Let $m \in B$ be minimal. Take any $x \in B$. Since L is totally ordered, either $x \leq m$ or $m \leq x$. The first possibility would force $x = m$, because m is minimal in B. Hence in either case $m \leq x$. Since $x \in B$ was arbitrary, m is the least element of B.

The proof of (ii) is analogous. □

Lexicographic order

The most familiar total orders come from numbers, but we can build new ones from old ones. The lexicographic order does this by declaring that the first coordinate has priority; the second coordinate matters only when the first coordinates agree. This is the same rule used in an ordinary dictionary.

Definition 5.4.4 (Lexicographic order on a product)**.** Let $(A, \leq_A)$ and $(B, \leq_B)$ be totally ordered sets. Define a relation $\leq_{\text{lex}}$ on $A \times B$ by

$$\langle a, b\rangle \leq_{\text{lex}} \langle a', b'\rangle$$

if and only if either

$$a <_A a',$$

or else

$$a = a' \quad \text{and} \quad b \leq_B b'.$$

Here $a <_A a'$ means that $a \leq_A a'$ and $a \neq a'$. This relation is called the *lexicographic order* on $A \times B$.

Proposition 5.4.5 (Lexicographic order is a total order)**.** *If $(A, \leq_A)$ and $(B, \leq_B)$ are totally ordered sets, then the lexicographic order on $A \times B$ is a total order.*

Proof. We verify the properties one by one.

Reflexivity. For any $\langle a, b\rangle \in A \times B$, we have $a = a$ and $b \leq_B b$. Hence

$$\langle a, b\rangle \leq_{\text{lex}} \langle a, b\rangle.$$

Antisymmetry. Suppose

$$\langle a, b\rangle \leq_{\text{lex}} \langle a', b'\rangle \qquad \text{and} \qquad \langle a', b'\rangle \leq_{\text{lex}} \langle a, b\rangle.$$

If $a <_A a'$, then the second inequality cannot hold, because it would require either $a' <_A a$, which is impossible, or $a' = a$, which also contradicts $a <_A a'$. So $a <_A a'$ is impossible. By the same reason $a' <_A a$ is impossible. Therefore $a = a'$. The two inequalities then reduce to $b \leq_B b'$ and $b' \leq_B b$, so by antisymmetry in B we obtain $b = b'$. Hence $\langle a, b\rangle = \langle a', b'\rangle$.

Transitivity. Suppose

$$\langle a, b\rangle \leq_{\text{lex}} \langle c, d\rangle \qquad \text{and} \qquad \langle c, d\rangle \leq_{\text{lex}} \langle e, f\rangle.$$

If $a <_A c$, then in either subcase of the second inequality we have $c \leq_A e$, and hence $a <_A e$. Therefore $\langle a, b\rangle \leq_{\text{lex}} \langle e, f\rangle$. If instead $a = c$ and $b \leq_B d$, then either $c <_A e$, in which case $a <_A e$, or $c = e$ and $d \leq_B f$, in which case $a = e$ and $b \leq_B f$. Again we conclude that $\langle a, b\rangle \leq_{\text{lex}} \langle e, f\rangle$. Thus the relation is transitive.

Totality. Let $\langle a, b\rangle, \langle a', b'\rangle \in A \times B$. Because A is totally ordered, either $a <_A a'$, or $a = a'$, or $a' <_A a$. In the first case, $\langle a, b\rangle \leq_{\text{lex}} \langle a', b'\rangle$. In the third case, $\langle a', b'\rangle \leq_{\text{lex}} \langle a, b\rangle$. If $a = a'$, then B is totally ordered, so either $b \leq_B b'$ or $b' \leq_B b$. Hence one of the two ordered pairs is lexicographically less than or equal to the other. Therefore the lexicographic order is total. □

Example 5.4.6 (Lexicographic order versus product order)**.** Consider the two points $\langle 1, 3\rangle$ and $\langle 2, 2\rangle$ in $\mathbb{N} \times \mathbb{N}$.

(i) In the product order from Example 5.3.3(v), these two points are incomparable, because $1 \leq 2$ but $3 \leq 2$ is false, and $2 \leq 1$ is also false.

(ii) In the lexicographic order, we have

$$\langle 1, 3\rangle <_{\text{lex}} \langle 2, 2\rangle,$$

because the first coordinate 1 is less than 2.

Thus the same underlying set $\mathbb{N} \times \mathbb{N}$ carries at least two very different natural orders.

Hasse diagrams

A finite partial order is often best understood visually. Instead of listing every ordered pair in the relation, we can draw only the most important comparisons and let transitivity fill in the rest.

Definition 5.4.7 (Cover relation)**.** Let $(P, \leq)$ be a poset. An element $y \in P$ *covers* an element $x \in P$ if

$$x \prec y$$

and there is no element $z \in P$ such that

$$x \prec z \prec y.$$

Remark 5.4.8 (Hasse diagrams). For a finite poset, the *Hasse diagram* is the picture obtained by placing elements as dots, drawing y above x when y covers x, and omitting arrowheads because "up" already indicates direction. Reflexive loops and transitive edges are omitted. The full order can then be read off from the upward paths in the diagram. This convention is standard in introductory order theory; see Davey and Priestley [8].

Example 5.4.9 (Divisors of 12 under divisibility)**.** Let

$$D_{12} = \{1, 2, 3, 4, 6, 12\},$$

ordered by divisibility. The cover relations are

$$1 \lessdot 2, \qquad 1 \lessdot 3, \qquad 2 \lessdot 4, \qquad 2 \lessdot 6, \qquad 3 \lessdot 6, \qquad 4 \lessdot 12, \qquad 6 \lessdot 12.$$

The Hasse diagram therefore has 1 at the bottom, 12 at the top, and the elements $2, 3, 4, 6$ arranged between them according to these cover relations. In particular, the fact that $1 \mid 12$ is not drawn as a direct edge, because it already follows transitively from the intermediate steps.

Remark 5.4.10 (Another useful picture). The set $\mathcal{P}(\{1, 2, 3\})$, ordered by inclusion, also has a familiar Hasse diagram. Its elements fall naturally into levels according to cardinality: the empty set at the bottom, the one-element subsets above it, then the two-element subsets, and finally $\{1, 2, 3\}$ at the top. This picture is often called the *subset lattice* of $\{1, 2, 3\}$.

5.5 Well-Ordered Sets as a Preview of the Transfinite

A total order puts every pair of elements in line. A well-order does something stronger: it ensures that every nonempty subset has a first element. This sounds like a modest extra condition, but it is one of the most powerful ideas in set theory. It underlies the familiar well-ordering principle for the natural numbers, and later it will lead us to ordinals, transfinite induction, and transfinite recursion.

The least-element property

Definition 5.5.1 (Well-order)**.** A total order $\leq$ on a set W is a *well-order* if every nonempty subset of W has a least element. A set equipped with a well-order is called a *well-ordered set*.

The key phrase is "every nonempty subset." It is not enough for the whole set W to have a least element. We require that the same property persist inside every smaller nonempty part of W.

Example 5.5.2 (First well-orders and nonexamples)**.** (i) Every finite totally ordered set is well-ordered. Indeed, any nonempty subset is finite, so one can compare its finitely many elements and choose the smallest.

(ii) The usual order on $\mathbb{N}$ is a well-order. This is the familiar well-ordering principle from elementary arithmetic.

(iii) The usual order on $\mathbb{Z}$ is *not* a well-order, because the subset $\mathbb{Z}$ itself has no least element.

(iv) The usual order on the interval $(0, 1) \subseteq \mathbb{R}$ is not a well-order, because $(0, 1)$ has no least element.

(v) If we reverse the usual order on $\mathbb{N}$ by declaring $m \preceq n$ when $m \geq n$ in the usual sense, then $(\mathbb{N}, \preceq)$ is not well-ordered: the whole set has no least element with respect to $\preceq$.

Example 5.5.3 (A less familiar well-order)**.** Consider $\mathbb{N} \times \mathbb{N}$ with the lexicographic order from Definition 5.4.4. Then this is a well-ordered set.

To see why, let $S \subseteq \mathbb{N} \times \mathbb{N}$ be nonempty. Look at the set of first coordinates that occur in S:

$$F = \{m \in \mathbb{N} \mid \text{there exists } n \in \mathbb{N} \text{ with } \langle m, n \rangle \in S\}.$$

Because S is nonempty, the set F is nonempty. Since $\mathbb{N}$ is well-ordered, F has a least element, say m_0. Now consider the set of second coordinates that occur together with this least first coordinate:

$$G = \{n \in \mathbb{N} \mid \langle m_0, n \rangle \in S\}.$$

Again G is nonempty, so it has a least element n_0. We claim that $\langle m_0, n_0 \rangle$ is the least element of S in the lexicographic order.

Indeed, if $\langle m, n \rangle \in S$, then by the choice of m_0 we have $m_0 \leq m$. If $m_0 < m$, then $\langle m_0, n_0 \rangle <_{\text{lex}} \langle m, n \rangle$. If $m_0 = m$, then by the choice of n_0 we have $n_0 \leq n$, so again $\langle m_0, n_0 \rangle \leq_{\text{lex}} \langle m, n \rangle$. Hence $\langle m_0, n_0 \rangle$ is least in S.

Proposition 5.5.4 (Subsets of well-ordered sets are well-ordered)**.** *Let $(W, \leq)$ be a well-ordered set, and let $B \subseteq W$. Then B, equipped with the restricted order, is also well-ordered.*

Proof. The restricted order on B is still a total order. Let $S \subseteq B$ be nonempty. Then S is also a nonempty subset of W. Since W is well-ordered, S has a least element in W, say s_0. Because $s_0 \in S \subseteq B$, the same element serves as the least element of S inside B. Therefore B is well-ordered. □

Proposition 5.5.5 (No infinite strictly descending sequence)**.** *Let $(W, \leq)$ be a well-ordered set. Then there is no sequence*

$$w_0 > w_1 > w_2 > \cdots$$

of elements of W that decreases forever.

Proof. Suppose such a sequence existed. Consider the set

$$S = \{w_n \mid n \in \mathbb{N}\}.$$

This set is nonempty, so it has a least element, say w_k, because W is well-ordered. But the sequence is strictly decreasing, so $w_{k+1} < w_k$. Also $w_{k+1} \in S$, contradicting the fact that w_k was least in S. Therefore no such infinite descending sequence can exist. □

Remark 5.5.6 (A useful intuition). Proposition 5.5.5 captures one of the basic intuitions behind well-ordering: one cannot keep moving strictly downward forever. In many concrete situations this "no endless descent" picture is an effective way to think about why well-ordered arguments work.

Well-orders and canonical choices

One reason well-orders matter is that they turn existence into a canonical choice. If every nonempty subset has a least element, then there is no ambiguity about which element to select.

Proposition 5.5.7 (Well-orders give canonical choice functions)**.** *Let $(W, \leq)$ be a well-ordered set, and let $(A_i)_{i \in I}$ be a family of nonempty subsets of W. Define a function c on I by letting $c(i)$ be the least element of A_i. Then c is a choice function for the family in the sense of Definition 4.5.1.*

Proof. Because each A_i is a nonempty subset of the well-ordered set W, it has a least element. So the rule defining $c(i)$ makes sense for every $i \in I$. By construction,

$$c(i) \in A_i \qquad \text{for every } i \in I.$$

This is exactly the defining property of a choice function. □

Remark 5.5.8 (A first glimpse of the axiom of choice). Proposition 5.5.7 shows that choice is easy inside a well-ordered set: one simply pick the least element each time. The real content of the axiom of choice, which we shall study later, is that for an arbitrary family of nonempty sets one may still be able to choose an element from each set even when no such canonical "least element" is available. This is one reason that well-orderings and the axiom of choice are so closely connected in more advanced set theory; see Jech [16] or Herrlich [17].

Remark 5.5.9 (Toward ordinals). Well-ordered sets are the doorway to the transfinite. Later we will ask when two well-ordered sets have the same ordered shape, how one well-order may sit as an initial segment of another, and how to assign a number-like object to each possible well-ordered type. Those questions lead to the theory of ordinals. Accessible introductions may be found in Moschovakis [7] and Jech [10].

Looking ahead

This chapter introduced relations as subsets of Cartesian products and showed how a single general notion can support several very different kinds of structure. Equivalence relations organize a set into classes, and those classes form partitions. Partial orders compare elements without forcing every pair to be comparable. Total orders line elements up completely, while well-orders do something stronger by guaranteeing a least element in every nonempty subset.

These ideas prepare the way for the next stage of the book. In the next chapter we will construct the natural numbers set-theoretically and use them as our first major example of a well-ordered set. The familiar facts that every natural number has a successor and that every nonempty subset of $\mathbb{N}$ has a least element will then be recast inside the language of sets. So Chapter 6 will not leave the present material behind; it will deepen it by turning one of our most familiar ordered worlds into a genuine set-theoretic construction.

Part II

Numbers Built from Sets and the First Infinite Worlds

Chapter 6

The Natural Numbers as Sets

Up to this point, the natural numbers have mostly stood outside the story. We have used them to count examples, to label terms of a sequence, to describe indexed families, and to talk about familiar relations such as divisibility and order. That is perfectly natural on first contact with set theory, because counting numbers are among the first mathematical objects anyone learns. But if set theory is meant to provide a common language for mathematics, then eventually the numbers should move from the background into the foreground. They should become objects *inside* the theory rather than tools borrowed from outside it.

This chapter makes that move. Its goal is not to convince the reader that children secretly count by manipulating sets. The goal is more structural. We want a single mathematical universe in which sets, functions, relations, ordered pairs, and the natural numbers all live side by side. Once that is done, counting can be studied set-theoretically, finite sets can later be defined in terms of bijections with natural numbers, and the methods of induction and recursion can be expressed in purely set-theoretic language.

There is an immediate question, however. Why should a number be a set at all? After all, when we count three apples, the number 3 does not look like a collection of collections. The answer is that there is no single sacred representation forced on us by nature. Mathematics often chooses a convenient model of an object rather than trying to discover its hidden physical essence. A complex number can be represented by an ordered pair of real numbers. A function can be represented by a set of ordered pairs. In the same spirit, the natural numbers can be represented by specially chosen sets.

Many such representations are possible. What makes the *von Neumann natural numbers* so useful is that each number is taken to be the set of all smaller numbers. In this model,

$$0 = \varnothing, \qquad 1 = \{0\}, \qquad 2 = \{0, 1\}, \qquad 3 = \{0, 1, 2\}, \quad \text{and so on.}$$

The number 3 therefore remembers its own predecessors automatically, and the order relation is built directly into membership. This is one of the reasons the construction is so elegant: arithmetic, induction, and order all become visible in the same picture.

This chapter is also a turning point in the philosophy of the book. Earlier chapters repeatedly said that we were proceeding informally and that an axiomatic treatment would come later. The present chapter keeps that promise honest. When we speak of "the set of all von Neumann natural numbers," we are still working at the intuitive level. In a fully axiomatic treatment one must justify the existence of such a set by an infinity axiom and related principles. Chapter 15 will return to that question. For now we focus on the mathematics that this construction makes possible.

One notational point matters immediately. In the notation table of Chapter 1, the symbol

$\mathbb{N}$ was reserved for $\{1, 2, 3, \dots\}$, while $\mathbb{N}_0$ denotes $\{0, 1, 2, 3, \dots\}$. Because the set-theoretic construction begins with 0, the natural home of this chapter is $\mathbb{N}_0$. We shall therefore build the full successor-generated system inside $\mathbb{N}_0$, and we can recover the positive integers later as $\mathbb{N} = \mathbb{N}_0 \setminus \{0\}$.

The guiding question is therefore this:

Can counting be carried out from inside set theory rather than merely alongside it?

The answer is yes. By the end of the chapter we will have an internal set-theoretic model of the natural numbers, the principle of induction, a precise theorem about recursive definitions, and recursive definitions of addition, multiplication, and exponentiation. Counting will no longer sit outside the theory as unexplained background. It will become one of the theory's own creations.

6.1 Why Define Numbers Inside Set Theory?

Before constructing numbers as sets, we should understand what problem that construction is trying to solve. In ordinary life the natural numbers seem so basic that they hardly need introduction. But foundational mathematics asks a different question from everyday life. It asks not merely how we *use* the numbers, but how they can be incorporated into a single mathematical framework with other objects.

From that viewpoint an internal model of the natural numbers should do at least four things. First, it should provide a distinguished starting point, which we call 0. Second, it should provide a rule that moves from one number to the next. Third, it should support the principle of mathematical induction, because induction is one of the basic ways in which arguments about counting numbers proceed. Fourth, it should allow functions to be defined recursively, step by step along the natural number sequence.

A model that does all this gives more than a mere dictionary between numbers and sets. It shows that the familiar structure of the counting numbers can be realized inside set theory itself. Once that has been done, later chapters can speak of finite sets, countable sets, and transfinite constructions using one common language. The payoff is not only philosophical but practical.

There is also a matter of canonicity. Suppose we want some set to play the role of 2. Many candidates would do: $\{a, b\}$, $\{\varnothing, \{\varnothing\}\}$, or any other set with two objects in it. But the moment we ask which of these candidates sits before 3, which contains 1, or which works naturally with recursion, most of these arbitrary choices begin to look awkward. We want a representation that organizes the whole number system at once, not a separate improvised encoding for each numeral.

The von Neumann idea answers that need beautifully. Instead of trying to decorate each number with some unrelated set, it builds the next number directly from the previous one and ensures that each stage keeps all earlier stages visible inside it. The result is a sequence of sets that is at once cumulative and ordered.

Remark 6.1.1 (Dedekind and von Neumann). Dedekind emphasized that the natural numbers should be understood through a starting point, a successor operation, and the induction principle [20]. Von Neumann later observed that sets themselves provide a particularly clean realization of this pattern: each number can be taken to be the set of all smaller numbers [24]. The construction in this chapter follows that now-standard viewpoint.

6.2 Successor and the von Neumann Naturals

The key step is to decide how one number should lead to the next. Since our objects are sets, the successor operation should be described in the language of unions and singletons developed in Chapter 2. The standard choice is simple and very natural: to move from a set x to its successor, we adjoin x itself as one new element.

The successor operation

Definition 6.2.1 (Successor)**.** For any set x, the *successor* of x is

$$S(x) = x \cup \{x\}.$$

Thus the successor of x contains everything that was already in x, and in addition it contains x itself as a new element. In that sense the operation S is cumulative: each stage keeps the previous stage and then adds it back in one level higher.

Example 6.2.2 (The first successor stages)**.** Starting with the empty set, we obtain

$$\begin{aligned} S(\varnothing) &= \varnothing \cup \{\varnothing\} = \{\varnothing\}, \\ S(\{\varnothing\}) &= \{\varnothing\} \cup \big\{\{\varnothing\}\big\} = \big\{\varnothing, \{\varnothing\}\big\}, \\ S(\{\varnothing, \{\varnothing\}\}) &= \big\{\varnothing, \{\varnothing\}\big\} \cup \big\{\{\varnothing, \{\varnothing\}\}\big\} \\ &= \big\{\varnothing, \{\varnothing\}, \{\varnothing, \{\varnothing\}\}\big\}. \end{aligned}$$

These are the set-theoretic objects that will become 1, 2, and 3.

The example already suggests an important pattern: each stage seems to contain all earlier stages. To describe that pattern cleanly, we need one more notion.

Transitive and inductive sets

Definition 6.2.3 (Transitive set)**.** A set A is called *transitive* if every element of an element of A is itself an element of A. In other words,

$$x \in y \in A \Longrightarrow x \in A.$$

Equivalently,

$$y \in A \Longrightarrow y \subseteq A.$$

A transitive set contains not only its elements but also everything that lies one step below those elements. Since the von Neumann numerals are supposed to contain their predecessors, transitivity is exactly the kind of closure property we expect to see.

Example 6.2.4 (Transitive and nontransitive examples)**.** (i) The empty set $\varnothing$ is transitive, because there are no elements to check.

(ii) The set $\{\varnothing\}$ is transitive. Its only element is $\varnothing$, and every element of $\varnothing$ is already in $\{\varnothing\}$ vacuously.

(iii) The set $\{\varnothing, \{\varnothing\}\}$ is transitive. The only nonempty element it contains is $\{\varnothing\}$, whose only element is $\varnothing$, and $\varnothing$ already belongs to the whole set.

(iv) The set $\{\{\varnothing\}\}$ is *not* transitive, because $\varnothing \in \{\varnothing\}$ but $\varnothing \notin \{\{\varnothing\}\}$.

The natural numbers will not be arbitrary transitive sets. They will be built by starting at 0 and closing under the successor operation. This leads to the next definition.

Definition 6.2.5 (Inductive set)**.** A set A is called *inductive* if

(i) $\varnothing \in A$, and

(ii) whenever $x \in A$, we also have $S(x) \in A$.

So an inductive set contains 0 and is closed under the operation that moves from one stage to the next. The intuitive natural number system should be the smallest such set.

Definition 6.2.6 (The von Neumann natural numbers)**.** We define

$$0 := \varnothing, \qquad 1 := S(0), \qquad 2 := S(1), \qquad 3 := S(2), \quad \dots$$

and, more generally, each new numeral is obtained by taking the successor of the previous one.

The set $\mathbb{N}_0$ of *von Neumann natural numbers* is the smallest inductive set. Its elements are precisely the stages obtained from 0 by finitely many applications of the successor operation.

Example 6.2.7 (The first von Neumann numerals)**.** The first few numerals are

$$0 = \varnothing, \qquad 1 = \{0\}, \qquad 2 = \{0, 1\}, \qquad 3 = \{0, 1, 2\}, \qquad 4 = \{0, 1, 2, 3\}.$$

Thus each numeral literally consists of the numerals that come before it. In particular,

$$2 \in 3, \qquad 1 \in 4, \qquad 0 \in 4,$$

which matches the ordinary intuition that $2 < 3$, $1 < 4$, and $0 < 4$.

Remark 6.2.8 (Why this construction is so useful). The von Neumann model is standard in elementary set theory; see, for example, Halmos [2], Enderton [3], and Moschovakis [7]. Its main advantage is that it makes several ideas coincide: the successor is built from union and singleton, the order relation is reflected by membership, and the induction principle becomes a statement about the smallest inductive set.

Remark 6.2.9 (Axiomatic caution). At the present intuitive stage, Definition 6.2.6 should be read as a mathematically guided description rather than as a fully axiomatic construction. In a formal treatment one first postulates an inductive set by an infinity axiom and then defines $\mathbb{N}_0$ as the intersection of all inductive sets. We postpone that axiomatic justification until Chapter 15.

A final notational remark will save confusion later. We often write $n + 1$ for $S(n)$. At the

moment this is only suggestive notation: it means "the successor of n." In Section 6.5 we will define addition recursively and prove that this notation is fully justified.

6.3 Induction

The phrase "and so on" appears constantly when one talks about the natural numbers. Induction is the theorem that makes that phrase mathematically legitimate. Once a property holds at the starting point and once we know that the property passes from each stage to its successor, then the property holds at every stage. This is not magic. It is simply a restatement of the fact that $\mathbb{N}_0$ was defined to be *the smallest inductive set*.

Ordinary induction

Theorem 6.3.1 (Principle of mathematical induction)**.** *Let $P(n)$ be a statement for each $n \in \mathbb{N}_0$. Suppose that*

(i) $P(0)$ is true, and

(ii) for every $n \in \mathbb{N}_0$, if $P(n)$ is true, then $P(S(n))$ is true.

Then $P(n)$ is true for every $n \in \mathbb{N}_0$.

Proof. Let

$$A = \{n \in \mathbb{N}_0 \mid P(n) \text{ is true}\}.$$

By assumption $P(0)$ is true, so $0 \in A$. Also, whenever $n \in A$, the statement $P(n)$ is true, so the induction step implies that $P(S(n))$ is true. Hence $S(n) \in A$. Therefore A is an inductive subset of $\mathbb{N}_0$.

But $\mathbb{N}_0$ was defined to be the smallest inductive set. Since $A \subseteq \mathbb{N}_0$ and A is itself inductive, we must have $A = \mathbb{N}_0$. That means exactly that $P(n)$ is true for every $n \in \mathbb{N}_0$. □

Remark 6.3.2 (Why induction is natural here). Theorem 6.3.1 is not an extra rule imported from outside set theory. It is a direct consequence of the way $\mathbb{N}_0$ was defined. In other words, induction is built into the very architecture of the natural numbers as the smallest inductive set. Dedekind already saw induction as one of the defining structural features of the natural numbers [20].

Once induction is available, we can prove the structural properties that make the von Neumann numerals behave the way we expect.

Proposition 6.3.3 (Every natural number is a subset of $\mathbb{N}_0$)**.** *For every $n \in \mathbb{N}_0$, we have*

$$n \subseteq \mathbb{N}_0.$$

Proof. We apply Theorem 6.3.1 to the statement $P(n)$: "$n \subseteq \mathbb{N}_0$."

For $n = 0$, the claim is true because $0 = \varnothing$, and the empty set is a subset of every set.

Now assume that $n \subseteq \mathbb{N}_0$. Then

$$S(n) = n \cup \{n\} \subseteq \mathbb{N}_0,$$

because every element of n lies in $\mathbb{N}_0$ by the inductive hypothesis, and the element n itself lies in $\mathbb{N}_0$ because n is a natural number. Thus $P(S(n))$ holds.

By induction, $P(n)$ holds for every $n \in \mathbb{N}_0$. □

Corollary 6.3.4 (Elements of a natural number are natural numbers)**.** *If $m \in n$ and $n \in \mathbb{N}_0$, then $m \in \mathbb{N}_0$.*

Proof. By Proposition 6.3.3, the set n is a subset of $\mathbb{N}_0$. Therefore every element of n belongs to $\mathbb{N}_0$. □

Proposition 6.3.5 (Every natural number is transitive)**.** *For every $n \in \mathbb{N}_0$, the set n is transitive.*

Proof. We again use Theorem 6.3.1.

For $n = 0$, the statement is true because the empty set is transitive.

Assume that n is transitive. We show that $S(n) = n \cup \{n\}$ is transitive. Let $x \in y \in S(n)$. Then either $y \in n$ or $y = n$.

If $y \in n$, then since n is transitive, we have $x \in n$, and therefore $x \in S(n)$.

If $y = n$, then $x \in n$, and again $x \in S(n)$.

So every element of an element of $S(n)$ lies in $S(n)$, which means that $S(n)$ is transitive.

By induction, every $n \in \mathbb{N}_0$ is transitive. □

Corollary 6.3.6 (Earlier numbers sit inside later ones)**.** *If $m \in n$ with $n \in \mathbb{N}_0$, then*

$$m \subseteq n.$$

Proof. Since n is transitive by Proposition 6.3.5, every element of n is a subset of n. In particular, $m \subseteq n$. □

The last corollary explains why the von Neumann picture is so useful: when one number lies inside another, it lies there as an *initial segment*. This is the set-theoretic shadow of ordinary numerical order.

Definition 6.3.7 (Order on $\mathbb{N}_0$)**.** For $m, n \in \mathbb{N}_0$, we define

$$m < n \quad \text{if and only if} \quad m \in n,$$

and

$$m \leq n \quad \text{if and only if} \quad m < n \text{ or } m = n.$$

Thus the elements of n are exactly the natural numbers smaller than n. In the von Neumann model, the order relation is therefore built into the membership relation itself.

Remark 6.3.8 (Order by membership). The definition above agrees with the ordinary picture suggested by the explicit numerals

$$0 = \varnothing, \qquad 1 = \{0\}, \qquad 2 = \{0, 1\}, \qquad 3 = \{0, 1, 2\}, \quad \ldots$$

For example, $2 < 5$ because $2 \in 5$, and $0 < 4$ because $0 \in 4$. Later, in Chapter 11, we will see that these objects are exactly the *finite ordinals*, which is why membership behaves so perfectly as an order relation here.

Proposition 6.3.9 (The successor map is injective)**.** *If $m, n \in \mathbb{N}_0$ and $S(m) = S(n)$, then $m = n$.*

Proof. Because m is transitive, we have $\bigcup m \subseteq m$. Hence

$$\bigcup S(m) = \bigcup (m \cup \{m\}) = \bigcup m \cup m = m.$$

The same argument gives $\bigcup S(n) = n$. Therefore, if $S(m) = S(n)$, then

$$m = \bigcup S(m) = \bigcup S(n) = n.$$

So the successor operation is injective on $\mathbb{N}_0$. □

Corollary 6.3.10 (Zero is not a successor)**.** *There is no $n \in \mathbb{N}_0$ such that $S(n) = 0$.*

Proof. For every n, the set $S(n) = n \cup \{n\}$ contains the element n. In particular, $S(n)$ is nonempty. But $0 = \varnothing$ is empty. Therefore $S(n) \neq 0$ for every n. □

Corollary 6.3.11 (Every nonzero natural has a unique predecessor)**.** *If $n \in \mathbb{N}_0$ and $n \neq 0$, then there exists a unique $m \in \mathbb{N}_0$ such that*

$$n = S(m).$$

Proof. By Definition 6.2.6, every nonzero natural number is obtained by one more successor step from an earlier stage. So some m exists with $n = S(m)$. If also $n = S(k)$, then Proposition 6.3.9 gives $m = k$. Thus the predecessor is unique. □

Remark 6.3.12 (The Peano picture inside set theory). We have now recovered several familiar structural facts about the natural numbers from set-theoretic definitions alone: there is a distinguished starting point 0, the successor map is injective, 0 is not a successor, and induction holds. These are among the central features of the classical Peano description of the natural numbers. The present chapter shows that they can be realized internally in set theory.

Strong induction and least elements

The ordinary induction principle is often reformulated in stronger-looking but equivalent ways. One of them allows the induction step for n to use *all* earlier values rather than only the immediately preceding one.

Theorem 6.3.13 (Strong induction)**.** *Let $P(n)$ be a statement for each $n \in \mathbb{N}_0$. Assume that for every $n \in \mathbb{N}_0$, whenever $P(m)$ is true for all $m < n$, the statement $P(n)$ is also true. Then $P(n)$ is true for every $n \in \mathbb{N}_0$.*

Proof. For each $n \in \mathbb{N}_0$, let $Q(n)$ be the statement

$$Q(n) : \text{“for every } m < n, \text{ the statement } P(m) \text{ is true.”}$$

We prove $Q(n)$ for all n by ordinary induction.

For $n = 0$, the statement $Q(0)$ is vacuously true, because there is no $m < 0$; equivalently, 0 has no elements.

Now assume that $Q(n)$ is true. We show that $Q(S(n))$ is true. Let $m < S(n)$. Then $m \in S(n) = n \cup \{n\}$, so either $m \in n$ or $m = n$.

If $m \in n$, then $m < n$, and $Q(n)$ tells us that $P(m)$ is true.

If $m = n$, then the hypothesis of strong induction applies to n, because $Q(n)$ asserts that $P(k)$ is true for every $k < n$. Hence $P(n)$ is true, so again $P(m)$ is true.

Thus $Q(S(n))$ holds. By ordinary induction, $Q(n)$ is true for all $n \in \mathbb{N}_0$.

Finally, fix any $n \in \mathbb{N}_0$. Then $n < S(n)$, because $n \in S(n)$ by definition of successor. So $Q(S(n))$ implies that $P(n)$ is true. Since n was arbitrary, $P(n)$ holds for every $n \in \mathbb{N}_0$. □

Remark 6.3.14 (Why strong induction is not really stronger). Theorem 6.3.13 looks more powerful than ordinary induction because the induction step is allowed to use all earlier cases. But the proof shows that it is only a reformulation of ordinary induction. The two principles express the same underlying structure of $\mathbb{N}_0$.

Corollary 6.3.15 (Least-element principle)**.** *Let $A \subseteq \mathbb{N}_0$ be nonempty. Then there exists an element $a \in A$ such that*

$$A \cap a = \varnothing.$$

Equivalently, A has an element with no smaller element of A below it.

Proof. Assume for contradiction that A is nonempty and that every $a \in A$ satisfies $A \cap a \neq \varnothing$.

We prove by induction that

$$P(n) : A \cap n = \varnothing$$

for every $n \in \mathbb{N}_0$.

For $n = 0$, the statement is true because $0 = \varnothing$, so $A \cap 0 = \varnothing$.

Now assume that $A \cap n = \varnothing$. We show that $A \cap S(n) = \varnothing$. Suppose instead that some $x \in A \cap S(n)$. Then either $x \in n$ or $x = n$.

If $x \in n$, then $x \in A \cap n$, contradicting the inductive hypothesis.

If $x = n$, then $n \in A$. By our contradiction assumption, $A \cap n \neq \varnothing$, again contradicting the inductive hypothesis.

So $A \cap S(n) = \varnothing$. By induction, $A \cap n = \varnothing$ for every $n \in \mathbb{N}_0$.

In particular, if $a \in A$, then $A \cap S(a) = \varnothing$. But $a \in S(a)$, so this would imply $a \notin A$, a contradiction. Therefore our assumption was false, and some $a \in A$ must satisfy $A \cap a = \varnothing$. □

Remark 6.3.16 (Well-ordering in the natural-number case). Because the elements of a are exactly the natural numbers smaller than a, Corollary 6.3.15 is the usual least-element principle for the natural numbers. In ordinary language, every nonempty subset of $\mathbb{N}_0$ has a least element. This is the first substantial example in the book of a well-ordered world, just as Chapter 5 anticipated.

Corollary 6.3.17 (No natural number belongs to itself)**.** *For every $n \in \mathbb{N}_0$,*

$$n \notin n.$$

Proof. Let

$$A = \{n \in \mathbb{N}_0 \mid n \in n\}.$$

Suppose that A were nonempty. By Corollary 6.3.15, there would be some $a \in A$ such that $A \cap a = \varnothing$. But $a \in A$ means precisely that $a \in a$. Since also $a \in A$, we would have $a \in A \cap a$, contradicting $A \cap a = \varnothing$. Therefore A is empty, so no natural number belongs to itself. □

6.4 Recursion on the Natural Numbers

Induction tells us how to *prove* statements about all natural numbers. Recursion tells us how to *define* objects step by step along the natural number sequence. This is one of the deepest reasons

why the natural numbers are so fundamental. We do not merely inspect them; we use them to generate new constructions stage by stage.

The everyday examples are familiar. We define the factorial by saying what happens at 0 and how to pass from n to $n+1$. We define partial sums by specifying the first value and then adding one more term at each step. We define many sequences in exactly this way. What set theory contributes is a clean theorem explaining why such definitions are legitimate and why they determine a unique function.

Recursive specifications

Definition 6.4.1 (Recursive specification on $\mathbb{N}_0$)**.** Let A be a set. A *recursive specification on $\mathbb{N}_0$ with values in A* consists of

(i) an initial value $a_0 \in A$, and

(ii) an update rule $G\colon A \to A$.

A function $f : \mathbb{N}_0 \to A$ is said to *satisfy* this specification if

$$f(0) = a_0 \qquad \text{and} \qquad f(S(n)) = G\big(f(n)\big) \quad \text{for every } n \in \mathbb{N}_0.$$

The definition says that once we know the current value $f(n)$, the rule G tells us the next value $f(S(n))$. The initial value gives the starting point.

Theorem 6.4.2 (Recursion theorem on $\mathbb{N}_0$)**.** *Let A be a set, let $a_0 \in A$, and let $G\colon A \to A$ be a function. Then there exists a unique function $f : \mathbb{N}_0 \to A$ such that*

$$f(0) = a_0 \qquad \textit{and} \qquad f(S(n)) = G\big(f(n)\big) \quad \textit{for every } n \in \mathbb{N}_0.$$

Proof. The proof proceeds in two stages. First we construct compatible finite approximations. Then we assemble them into a single function on all of $\mathbb{N}_0$.

For each $n \in \mathbb{N}_0$, call a function $u\colon n \to A$ an *n-approximation* if the following conditions hold:

(i) if $0 \in n$, then $u(0) = a_0$,

(ii) whenever $k \in n$ and $S(k) \in n$, we have $u(S(k)) = G(u(k))$.

Thus an n-approximation is a partial recursive definition carried out up to stage n.

We claim that for each $n \in \mathbb{N}_0$, there exists a unique n-approximation u_n. We prove this by induction on n.

For $n = 0$, the unique function $0 \to A$ is the empty function, and it vacuously satisfies the approximation conditions. So there is a unique 0-approximation u_0.

Now assume that there is a unique n-approximation u_n. We will construct the unique $S(n)$-approximation.

Because of Corollary 6.3.17, the new point n is not already in the domain n. So to pass from a function on n to a function on $S(n) = n \cup \{n\}$, it is enough to choose one new value at n.

Define an element $b_n \in A$ as follows:

(i) if $n = 0$, let $b_n = a_0$,

(ii) if $n \neq 0$, let k be the unique predecessor of n given by Corollary 6.3.11, so that $n = S(k)$, and define

$$b_n = G\big(u_n(k)\big).$$

Now set

$$u_{S(n)} = u_n \cup \{\langle n, b_n\rangle\}.$$

Because $n \notin n$, the point n is new, so this really is a function from $S(n)$ to A.

We check that $u_{S(n)}$ is an $S(n)$-approximation. First, if $0 \in S(n)$, then either $n = 0$, in which case $u_{S(0)}(0) = b_0 = a_0$, or else $0 \in n$, in which case $u_{S(n)}(0) = u_n(0) = a_0$.

Second, let $k \in S(n)$ with $S(k) \in S(n)$. If $S(k) \in n$, then both k and $S(k)$ lie in n, so the recursive condition already holds for u_n, and hence also for $u_{S(n)}$. If instead $S(k) = n$, then by definition of b_n we have

$$u_{S(n)}(S(k)) = u_{S(n)}(n) = b_n = G\big(u_n(k)\big) = G\big(u_{S(n)}(k)\big).$$

So $u_{S(n)}$ is indeed an $S(n)$-approximation.

To prove uniqueness, let v be any $S(n)$-approximation. Then its restriction $v \restriction_n$ is an n-approximation, so by uniqueness of u_n we have

$$v \restriction_n = u_n.$$

The value at the new point n is then forced. If $n = 0$, the initial-value condition gives $v(0) = a_0 = b_0$. If $n \neq 0$ and $n = S(k)$, then the recursive condition gives

$$v(n) = v(S(k)) = G(v(k)) = G(u_n(k)) = b_n.$$

Hence $v = u_{S(n)}$. So the $S(n)$-approximation is unique. This completes the induction and proves the existence and uniqueness of u_n for every $n \in \mathbb{N}_0$.

Next we show that these approximations are compatible. If $m \in n$, then the restriction $u_n \restriction_m$ is an m-approximation, so by uniqueness it must equal u_m:

$$u_n \restriction_m = u_m.$$

We can now define the desired global function $f : \mathbb{N}_0 \to A$ by

$$f(n) = u_{S(n)}(n).$$

This makes sense because $n \in S(n)$, so $u_{S(n)}$ is defined at n.

For the initial value, we have

$$f(0) = u_{S(0)}(0) = a_0.$$

For the recursive step, observe that $u_{S(S(n))} \restriction_{S(n)} = u_{S(n)}$ by compatibility. Since $S(n)$ is the new point added in the passage from $S(n)$ to $S(S(n))$, the defining property of the approximation gives

$$f(S(n)) = u_{S(S(n))}(S(n)) = G\big(u_{S(n)}(n)\big) = G\big(f(n)\big).$$

So f satisfies the required recursive specification.

Finally, we prove uniqueness of f. Suppose that $g, h : \mathbb{N}_0 \to A$ both satisfy the same recursion. Let

$$B = \{n \in \mathbb{N}_0 \mid g(n) = h(n)\}.$$

Then $0 \in B$, since $g(0) = a_0 = h(0)$. Also, if $n \in B$, then

$$g(S(n)) = G(g(n)) = G(h(n)) = h(S(n)),$$

so $S(n) \in B$. Thus B is an inductive subset of $\mathbb{N}_0$, and Theorem 6.3.1 implies that $B = \mathbb{N}_0$. Hence $g = h$. The recursive function is unique. □

Example 6.4.3 (The identity function as a recursive construction)**.** Take $A = \mathbb{N}_0$, let $a_0 = 0$, and let $G = S$. Then Theorem 6.4.2 gives a unique function $f : \mathbb{N}_0 \to \mathbb{N}_0$ satisfying

$$f(0) = 0, \qquad f(S(n)) = S(f(n)).$$

The identity function $\mathrm{id}_{\mathbb{N}_0}$ clearly satisfies the same two conditions, so uniqueness forces

$$f = \mathrm{id}_{\mathbb{N}_0}.$$

This tiny example is worth noticing: even the most familiar function on $\mathbb{N}_0$ is determined by a recursive rule.

Example 6.4.4 (A Fibonacci-type recursion by storing two values)**.** The recursion theorem above uses only the current value to determine the next one. But many familiar sequences, such as the Fibonacci sequence, seem to depend on more than one previous term. The cure is to enlarge the state space.

Define a function $T : \mathbb{N}_0 \times \mathbb{N}_0 \to \mathbb{N}_0 \times \mathbb{N}_0$ by

$$T\big(\langle a, b\rangle\big) = \langle b, a + b\rangle,$$

using the ordinary arithmetic intuition for the moment. Starting from $\langle 0, 1\rangle$, the recursion theorem gives a unique function $g : \mathbb{N}_0 \to \mathbb{N}_0 \times \mathbb{N}_0$ with

$$g(0) = \langle 0, 1\rangle, \qquad g(S(n)) = T\big(g(n)\big).$$

If we write $g(n) = \langle F_n, F_{n+1}\rangle$, then the first coordinates produce

$$0, 1, 1, 2, 3, 5, 8, \ldots$$

which is the Fibonacci sequence. So even a two-step recursion can be handled by an ordinary one-step recursion once we store enough information in the state.

Remark 6.4.5 (Recursion in standard texts). Theorem 6.4.2 is standard in introductory set-theory texts such as Enderton [3] and Moschovakis [7]. The philosophical point is simple but deep: recursive definitions are not vague appeals to "continue the pattern." They are justified by a precise existence-and-uniqueness theorem.

6.5 Addition, Multiplication, and Exponentiation on $\mathbb{N}_0$

The recursion theorem now lets us reconstruct the familiar arithmetic operations from the successor structure alone. Conceptually, the idea is beautifully simple:

> *Addition is repeated successor, multiplication is repeated addition, and exponentiation is repeated multiplication.*

What was formerly taken for granted in school arithmetic now becomes a consequence of recursive definitions.

Addition

Definition 6.5.1 (Addition on $\mathbb{N}_0$)**.** Fix $m \in \mathbb{N}_0$. By Theorem 6.4.2, there is a unique function $f_m : \mathbb{N}_0 \to \mathbb{N}_0$ such that

$$f_m(0) = m, \qquad f_m(S(n)) = S\big(f_m(n)\big) \quad \text{for every } n \in \mathbb{N}_0.$$

We define

$$m + n := f_m(n).$$

This is called *addition* on $\mathbb{N}_0$.

Thus, for fixed m, the function $n \mapsto m + n$ starts at m and moves forward by one successor step each time n advances by one.

Example 6.5.2 (Computing a small sum)**.** Let us compute $2 + 3$. By the recursive definition,

$$\begin{aligned} 2 + 0 &= 2, \\ 2 + 1 &= S(2 + 0) = S(2) = 3, \\ 2 + 2 &= S(2 + 1) = S(3) = 4, \\ 2 + 3 &= S(2 + 2) = S(4) = 5. \end{aligned}$$

So the familiar answer appears exactly as expected.

Proposition 6.5.3 (Basic addition identities)**.** *For all $m, n \in \mathbb{N}_0$,*

(i) $m + 0 = m$,

(ii) $m + S(n) = S(m + n)$,

(iii) $m + 1 = S(m)$.

Proof. Statements (i) and (ii) are exactly the defining properties of addition. For (iii), we use $1 = S(0)$:

$$m + 1 = m + S(0) = S(m + 0) = S(m).$$

□

Proposition 6.5.4 (Useful left-hand laws for addition)**.** *For all $m, n \in \mathbb{N}_0$,*

(i) $0 + n = n$,

(ii) $S(m) + n = S(m + n)$.

Proof. For (i), we use induction on n. When $n = 0$, we have $0 + 0 = 0$. Assume that $0 + n = n$. Then

$$0 + S(n) = S(0 + n) = S(n).$$

So $0 + n = n$ for all n.

For (ii), we again use induction on n. When $n = 0$,

$$S(m) + 0 = S(m) = S(m + 0).$$

Assume that $S(m) + n = S(m + n)$. Then

$$S(m) + S(n) = S\big(S(m) + n\big) = S\big(S(m + n)\big) = S\big(m + S(n)\big).$$

So $S(m) + n = S(m + n)$ for all n. □

Theorem 6.5.5 (Addition is associative and commutative)**.** *For all $m, n, k \in \mathbb{N}_0$,*

$$(m + n) + k = m + (n + k)$$

and

$$m + n = n + m.$$

Proof. We prove associativity first. Fix m and n, and use induction on k.

For $k = 0$,

$$(m + n) + 0 = m + n = m + (n + 0).$$

Assume that $(m + n) + k = m + (n + k)$. Then

$$\begin{aligned}(m + n) + S(k) &= S\big((m + n) + k\big)\\ &= S\big(m + (n + k)\big)\\ &= m + S(n + k)\\ &= m + (n + S(k)).\end{aligned}$$

So addition is associative.

Now we prove commutativity. For each $n \in \mathbb{N}_0$, let $P(n)$ be the statement "for every $m \in \mathbb{N}_0$, $m + n = n + m$."

For $n = 0$, Proposition 6.5.4(i) gives

$$m + 0 = m = 0 + m$$

for every m. So $P(0)$ holds.

Assume that $P(n)$ holds, and let $m \in \mathbb{N}_0$. Then

$$\begin{aligned}m + S(n) &= S(m + n)\\ &= S(n + m)\\ &= S(n) + m,\end{aligned}$$

where the first equality uses Proposition 6.5.3(ii), the second uses the inductive hypothesis $P(n)$, and the third uses Proposition 6.5.4(ii) with n and m interchanged. Thus $P(S(n))$ holds.

By induction, addition is commutative. □

Remark 6.5.6 (Addition as repeated succession). Definition 6.5.1 and Theorem 6.5.5 reconstruct the familiar additive structure of the natural numbers from the successor operation alone. Nothing essentially new was postulated; recursion built addition from the repeated act of moving one step forward.

Multiplication

Definition 6.5.7 (Multiplication on $\mathbb{N}_0$)**.** Fix $m \in \mathbb{N}_0$. By Theorem 6.4.2, there is a unique function $g_m : \mathbb{N}_0 \to \mathbb{N}_0$ such that

$$g_m(0) = 0, \qquad g_m(S(n)) = g_m(n) + m \quad \text{for every } n \in \mathbb{N}_0.$$

We define

$$m \cdot n := g_m(n).$$

This is called *multiplication* on $\mathbb{N}_0$.

So for fixed m, the function $n \mapsto m \cdot n$ starts at 0 and adds one more copy of m at each step.

Example 6.5.8 (Computing a small product)**.** Let us compute $2 \cdot 3$. By definition,

$$\begin{aligned} 2 \cdot 0 &= 0, \\ 2 \cdot 1 &= 2 \cdot 0 + 2 = 2, \\ 2 \cdot 2 &= 2 \cdot 1 + 2 = 4, \\ 2 \cdot 3 &= 2 \cdot 2 + 2 = 6. \end{aligned}$$

Thus multiplication is indeed repeated addition.

Proposition 6.5.9 (Basic multiplication identities)**.** *For all $m, n \in \mathbb{N}_0$,*

(i) $m \cdot 0 = 0$,

(ii) $m \cdot S(n) = m \cdot n + m$,

(iii) $m \cdot 1 = m$,

(iv) $0 \cdot n = 0$,

(v) $1 \cdot n = n$.

Proof. Statements (i) and (ii) are the defining properties of multiplication. For (iii),

$$m \cdot 1 = m \cdot S(0) = m \cdot 0 + m = 0 + m = m,$$

using Proposition 6.5.4(i).

For (iv), we use induction on n. When $n = 0$, we have $0 \cdot 0 = 0$. If $0 \cdot n = 0$, then

$$0 \cdot S(n) = 0 \cdot n + 0 = 0 + 0 = 0.$$

So $0 \cdot n = 0$ for all n.

For (v), we again use induction on n. When $n = 0$, $1 \cdot 0 = 0$. If $1 \cdot n = n$, then

$$1 \cdot S(n) = 1 \cdot n + 1 = n + 1 = S(n),$$

using Proposition 6.5.3(iii). Thus $1 \cdot n = n$ for all n. □

Proposition 6.5.10 (Distributivity of multiplication over addition)**.** *For all $m, n, k \in \mathbb{N}_0$,*

$$m \cdot (n + k) = m \cdot n + m \cdot k.$$

Proof. Fix m and n, and use induction on k.
For $k = 0$,

$$m \cdot (n + 0) = m \cdot n = m \cdot n + 0 = m \cdot n + m \cdot 0.$$

Assume that $m \cdot (n + k) = m \cdot n + m \cdot k$. Then

$$\begin{aligned} m \cdot (n + S(k)) &= m \cdot S(n + k) \\ &= m \cdot (n + k) + m \\ &= (m \cdot n + m \cdot k) + m \\ &= m \cdot n + (m \cdot k + m) \\ &= m \cdot n + m \cdot S(k). \end{aligned}$$

So the distributive law holds for all k. □

Proposition 6.5.11 (Multiplication is associative)**.** *For all $m, n, k \in \mathbb{N}_0$,*

$$(m \cdot n) \cdot k = m \cdot (n \cdot k).$$

Proof. Fix m and n, and use induction on k.
For $k = 0$,

$$(m \cdot n) \cdot 0 = 0 = m \cdot 0 = m \cdot (n \cdot 0).$$

Assume that $(m \cdot n) \cdot k = m \cdot (n \cdot k)$. Then

$$\begin{aligned} (m \cdot n) \cdot S(k) &= (m \cdot n) \cdot k + m \cdot n \\ &= m \cdot (n \cdot k) + m \cdot n \\ &= m \cdot \big((n \cdot k) + n\big) \\ &= m \cdot \big(n \cdot S(k)\big), \end{aligned}$$

where the third equality uses Proposition 6.5.10. Therefore multiplication is associative. □

Lemma 6.5.12 (Successor in the first factor)**.** *For all $m, n \in \mathbb{N}_0$,*

$$S(m) \cdot n = m \cdot n + n.$$

Proof. Fix m and use induction on n.
For $n = 0$,

$$S(m) \cdot 0 = 0 = m \cdot 0 + 0.$$

Assume that $S(m) \cdot n = m \cdot n + n$. Then

$$\begin{aligned} S(m) \cdot S(n) &= S(m) \cdot n + S(m) \\ &= (m \cdot n + n) + S(m) \\ &= S\big((m \cdot n + n) + m\big) \\ &= S\big((m \cdot n + m) + n\big) \\ &= (m \cdot n + m) + S(n) \\ &= m \cdot S(n) + S(n). \end{aligned}$$

So the lemma holds for all n. □

Theorem 6.5.13 (Multiplication is commutative)**.** *For all $m, n \in \mathbb{N}_0$,*

$$m \cdot n = n \cdot m.$$

Proof. For each $n \in \mathbb{N}_0$, let $P(n)$ be the statement "for every $m \in \mathbb{N}_0$, $m \cdot n = n \cdot m$."

For $n = 0$, Proposition 6.5.9 gives

$$m \cdot 0 = 0 = 0 \cdot m$$

for every m, so $P(0)$ holds.

Assume that $P(n)$ holds, and let $m \in \mathbb{N}_0$. Then

$$\begin{aligned} m \cdot S(n) &= m \cdot n + m \\ &= n \cdot m + m \\ &= S(n) \cdot m, \end{aligned}$$

where the first equality comes from the recursive definition of multiplication, the second from the inductive hypothesis, and the third from Lemma 6.5.12 with the names of the variables interchanged. Thus $P(S(n))$ holds.

By induction, multiplication is commutative. □

Remark 6.5.14 (Left distributivity now follows automatically). Proposition 6.5.10 proved a right-distributive law because multiplication was defined recursively in the second variable. Once multiplication is known to be commutative, left distributivity follows immediately:

$$(m + n) \cdot k = k \cdot (m + n) = k \cdot m + k \cdot n = m \cdot k + n \cdot k.$$

Exponentiation

Definition 6.5.15 (Exponentiation on $\mathbb{N}_0$)**.** Fix $m \in \mathbb{N}_0$. By Theorem 6.4.2, there is a unique function $h_m : \mathbb{N}_0 \to \mathbb{N}_0$ such that

$$h_m(0) = 1, \qquad h_m(S(n)) = h_m(n) \cdot m \quad \text{for every } n \in \mathbb{N}_0.$$

We define

$$m^n := h_m(n).$$

This is called *exponentiation* on $\mathbb{N}_0$.

So for fixed m, the function $n \mapsto m^n$ starts at 1 and multiplies by m at each step.

Example 6.5.16 (Computing a small power)**.** Let us compute 2^3. By definition,

$$\begin{aligned} 2^0 &= 1, \\ 2^1 &= 2^0 \cdot 2 = 2, \\ 2^2 &= 2^1 \cdot 2 = 4, \\ 2^3 &= 2^2 \cdot 2 = 8. \end{aligned}$$

So exponentiation is repeated multiplication, exactly as expected.

Proposition 6.5.17 (Basic exponent laws)**.** *For all $m, n, k \in \mathbb{N}_0$,*

(i) $m^0 = 1$,

(ii) $m^1 = m$,

(iii) $m^{n+k} = m^n \cdot m^k$.

Proof. Statement (i) is the defining property of exponentiation. For (ii),

$$m^1 = m^{S(0)} = m^0 \cdot m = 1 \cdot m = m.$$

For (iii), fix m and n, and use induction on k.
When $k = 0$,

$$m^{n+0} = m^n = m^n \cdot 1 = m^n \cdot m^0.$$

Assume that $m^{n+k} = m^n \cdot m^k$. Then

$$\begin{aligned} m^{n+S(k)} &= m^{S(n+k)} \\ &= m^{n+k} \cdot m \\ &= (m^n \cdot m^k) \cdot m \\ &= m^n \cdot (m^k \cdot m) \\ &= m^n \cdot m^{S(k)}. \end{aligned}$$

So the formula holds for all k. □

Remark 6.5.18 (Arithmetic rebuilt from set theory). We now have recursive definitions of addition, multiplication, and exponentiation on $\mathbb{N}_0$, together with the first familiar algebraic laws they satisfy. The point is not that school arithmetic was somehow wrong for taking these operations for granted. The point is that set-theoretic structure is rich enough to reconstruct them internally. Counting, adding, multiplying, and taking powers have all been expressed inside the same universe of sets.

Looking ahead

This chapter turned the natural numbers from background notation into set-theoretic objects. We defined the von Neumann numerals inside $\mathbb{N}_0$, saw how order is reflected by membership,

proved the principles of ordinary and strong induction, established a recursion theorem, and used recursion to define addition, multiplication, and exponentiation. In this way counting became internal to set theory rather than merely external bookkeeping.

The next chapter will build directly on this achievement. Once the natural numbers exist as sets, we can use them to make the everyday idea of a *finite set* precise. A set will be called finite when it can be matched with one of the numerals $0, 1, 2, \ldots$, and counting arguments will become statements about functions and bijections. So the transition to Chapter 7 is completely natural: now that we have constructed the numbers, we can begin to measure the sizes of sets by comparing them with those numbers.

Chapter 7

Finite Sets and Counting

In everyday life, counting feels almost primitive. We look at three books on a desk, five coins in a hand, or eight letters in a word, and we simply announce the number. But once sets become the basic objects of study, counting itself asks to be translated into set-theoretic language. What exactly does it mean for two collections to have the same number of elements? Why does the order in which we inspect the objects not matter? Why is a set with one more element than another set really larger? And why do arguments such as the pigeonhole principle feel so inevitable once one has seen the right definition?

The decisive idea is that counting is not primarily about reciting the symbols $0, 1, 2, 3, \ldots$. It is about *matching* objects. If the students in a classroom can be paired off perfectly with the chairs in that room, then there are exactly as many students as chairs, whether or not we know the actual number. A finite set has n elements when it can be matched, point for point, with the von Neumann numeral n constructed in Chapter 6. In that way the abstract natural numbers created in the previous chapter become standard models for finite sizes.

This chapter therefore performs a conceptual shift. Chapter 6 built the natural numbers inside set theory; the present chapter uses those numbers to measure other sets. The result is the first genuine theory of size in the book. At the finite level this theory agrees with everyday intuition, but it already uses the ideas that will later make infinite sets intelligible. In particular, the definition of "same number of elements" by means of bijections is not a temporary trick for finite sets. It is the central idea that Cantor used to compare sizes in general [18, 19].

There is also an important pedagogical reason for slowing down here. Many students first meet counting in school as a collection of formulas: $m + n$, mn, 2^n, and $n!$. Those formulas are correct, but if they are memorized without structural understanding, they feel like a bag of disconnected recipes. Set theory explains why they are natural. Addition counts disjoint unions, multiplication counts Cartesian products, exponentiation counts functions, and factorial counts permutations. Once these interpretations become clear, the formulas are not merely remembered; they are understood.

The chapter ends by showing that the finite world, although familiar, is not the whole story. Several principles that are unquestionable for finite sets fail dramatically for infinite ones. A finite set is never in bijection with a proper subset of itself, and every injective self-map of a finite set is automatically surjective. Later we will meet sets for which both statements fail. So the present chapter is both a culmination of the finite viewpoint and a bridge to the infinite one.

7.1 Bijections and the Meaning of "Same Number of Elements"

Before we define finite sets formally, we should clarify what it means for two sets to have the same size. In elementary life one often says "count the objects and compare the answers." But

that advice hides the real mathematical idea. The number obtained by counting is itself only a summary of a more primitive fact: the set can be paired with one of the natural numbers. And before even that step, two sets can be compared directly with each other by pairing their elements.

Pairing without counting

Definition 7.1.1 (Equinumerous sets)**.** Let A and B be sets. We say that A and B are *equinumerous* (or have the same number of elements) if there exists a bijection $f : A \to B$.

When this happens, we write

$$A \approx B.$$

This definition ignores the nature of the elements and remembers only whether they can be matched perfectly. That is exactly what we want from a concept of size. If every element of A can be paired with one and only one element of B, and every element of B is paired with exactly one element of A, then the two sets should count as equally large.

Example 7.1.2 (Three objects matched with three objects)**.** Let

$$A = \{\text{red}, \text{blue}, \text{green}\} \qquad \text{and} \qquad B = \{2, 5, 8\}.$$

The function $f : A \to B$ defined by

$$f(\text{red}) = 2, \qquad f(\text{blue}) = 5, \qquad f(\text{green}) = 8$$

is a bijection. Therefore $A \approx B$. The colors and the numbers have nothing in common as objects, but they still represent the same finite size.

Example 7.1.3 (Order does not matter)**.** Let

$$A = \{1, 2, 3, 4\} \qquad \text{and} \qquad B = \{4, 3, 2, 1\}.$$

As sets, $A = B$, because sets do not remember order. But even if we had written the second collection with different symbols,

$$C = \{a, b, c, d\},$$

we would still have $A \approx C$ by any bijection pairing the four numbers with the four letters. Counting concerns matching, not the way we happened to write the elements on paper.

The notation $A \approx B$ suggests that equinumerosity behaves like a kind of equality. It is not literal equality of sets, because the sets may have completely different elements. But it should at least have the basic formal properties of an equivalence relation. The next result confirms that expectation.

Proposition 7.1.4 (Equinumerosity is an equivalence relation)**.** *For all sets A, B, C, the relation $\approx$ satisfies:*

(i) $A \approx A$,

(ii) if $A \approx B$, *then* $B \approx A$,

(iii) if $A \approx B$ and $B \approx C$, then $A \approx C$.

Proof. For (i), the identity map

$$\mathrm{id}_A \colon A \to A, \qquad \mathrm{id}_A(a) = a,$$

is a bijection, so $A \approx A$.

For (ii), suppose $A \approx B$. Then there exists a bijection $f \colon A \to B$. By Theorem 3.4.8 from Chapter 3, the inverse function $f^{-1} \colon B \to A$ is also a bijection. Hence $B \approx A$.

For (iii), suppose $A \approx B$ and $B \approx C$. Then there are bijections $f \colon A \to B$ and $g \colon B \to C$. By Proposition 3.4.6 from Chapter 3, the composite $g \circ f \colon A \to C$ is again a bijection. Therefore $A \approx C$. □

Remark 7.1.5 (Cantor's viewpoint on size). The idea that size should be compared by one-to-one correspondence is one of Cantor's great insights [18, 19]. At the finite level the idea seems almost obvious. Its real power becomes visible only later, when the same definition continues to make sense for infinite sets and begins to distinguish different infinite sizes.

At this point we are finally ready to connect comparison by bijection with the natural numbers built in the previous chapter.

7.2 Finite Sets

The natural numbers $0, 1, 2, 3, \ldots$ now exist as specific sets inside our set-theoretic universe. We may therefore use them as standard models of finite sizes. Saying that a set has three elements should mean that it can be matched exactly with the numeral $3 = \{0, 1, 2\}$; saying that a set has no elements should mean that it can be matched with $0 = \varnothing$; and so on.

Finite sets as sets that match numerals

Definition 7.2.1 (Finite set)**.** Let A be a set and let $n \in \mathbb{N}_0$.

(i) We say that A is an *n-element set* if $A \approx n$, that is, if there exists a bijection $n \to A$.

(ii) We say that A is *finite* if it is an n-element set for some $n \in \mathbb{N}_0$.

So a finite set is one that can be counted out in finitely many steps and brought into bijection with one of the numerals produced in Chapter 6. The numeral itself serves as a canonical picture of that size.

Example 7.2.2 (The first finite sizes)**.** (i) The empty set $\varnothing$ is a 0-element set, because the empty function $0 \to \varnothing$ is a bijection.

(ii) Every singleton $\{a\}$ is a 1-element set. Indeed, $1 = \{0\}$, and the map $0 \mapsto a$ is a bijection $1 \to \{a\}$.

(iii) The set $\{2, 4, 6\}$ is a 3-element set. One bijection $3 \to \{2, 4, 6\}$ is given by

$$0 \mapsto 2, \qquad 1 \mapsto 4, \qquad 2 \mapsto 6.$$

The definition makes sense only if the natural number n that appears in it is uniquely determined by the set. Otherwise a set might be both three-element and four-element, which would destroy the point of the construction. The next theorem shows that the numerals are rigid enough to prevent that.

Theorem 7.2.3 (Bijections between numerals are rigid)**.** *Let $m, n \in \mathbb{N}_0$. If there exists a bijection $f : n \to m$, then $n = m$.*

Proof. We argue by induction on n.

If $n = 0$, and if $f : 0 \to m$ is a bijection, then m must be empty because the range of the empty function is empty. But the only empty natural number is 0. Indeed, if $m \neq 0$, then by Corollary 6.3.11 there exists $k \in \mathbb{N}_0$ with $m = S(k)$, and every successor is nonempty by Corollary 6.3.10. Hence $m = 0$.

Now assume the theorem is true for n, and suppose that $f : S(n) \to m$ is a bijection. Since $S(n)$ is nonempty, so is m. Therefore, by Corollary 6.3.11, there exists $k \in \mathbb{N}_0$ such that

$$m = S(k).$$

Let

$$a = f(n) \in S(k).$$

Define a function $\tau : S(k) \to S(k)$ by

$$\tau(x) = \begin{cases} k, & \text{if } x = a, \\ a, & \text{if } x = k, \\ x, & \text{otherwise.} \end{cases}$$

This function simply swaps a and k and leaves all other points fixed, so it is a bijection. Hence

$$g = \tau \circ f : S(n) \to S(k)$$

is also a bijection, and by construction

$$g(n) = k.$$

We claim that the restriction

$$g \restriction_n : n \to k$$

is a bijection. First, if $x \in n$, then $x \neq n$, so injectivity of g implies $g(x) \neq g(n) = k$. Since $g(x) \in S(k) = k \cup \{k\}$, it follows that $g(x) \in k$. Thus $g \restriction_n$ really maps n into k.

To see injectivity, note that a restriction of an injective function is again injective. For surjectivity, let $y \in k$. Since g is surjective onto $S(k)$, there exists $x \in S(n)$ such that $g(x) = y$. Because $y \neq k = g(n)$, injectivity of g forces $x \neq n$, so $x \in n$. Hence $g \restriction_n (x) = y$.

So $g \restriction_n : n \to k$ is a bijection. By the induction hypothesis, $n = k$. Therefore

$$m = S(k) = S(n).$$

This completes the induction. □

Corollary 7.2.4 (Finite cardinality is well defined)**.** *If a set A is finite, then there is a unique $n \in \mathbb{N}_0$ such that $A \approx n$.*

Proof. Suppose that $A \approx m$ and $A \approx n$. Then there are bijections $f : m \to A$ and $g : n \to A$. The composite

$$f^{-1} \circ g : n \to m$$

is a bijection. By Theorem 7.2.3, we must have $n = m$. □

Because of this corollary, it is now safe to introduce the usual notation for the size of a finite set.

Definition 7.2.5 (Finite cardinality notation)**.** If A is a finite set and $A \approx n$, where $n \in \mathbb{N}_0$, we write

$$|A| = n$$

and call n the *cardinality* or *size* of A.

The next two propositions explain how finite size changes when one point is added or removed.

Proposition 7.2.6 (Adding one new element increases the size by one)**.** *Let A be a finite set with $|A| = n$, and let $a \notin A$. Then*

$$|A \cup \{a\}| = S(n).$$

Proof. Choose a bijection $f : n \to A$. Define $g : S(n) \to A \cup \{a\}$ by

$$g(x) = \begin{cases} f(x), & \text{if } x \in n, \\ a, & \text{if } x = n. \end{cases}$$

Because $a \notin A$, the value at the new point n does not repeat any old value. Thus g is injective. It is also surjective, because every element of A is hit by f, and the element a is hit at n. Therefore g is a bijection, so $|A \cup \{a\}| = S(n)$. □

Example 7.2.7 (Adding one point to a three-element set)**.** The set $\{2, 4, 6\}$ has three elements, and 10 does not belong to it. Therefore

$$\big|\{2, 4, 6, 10\}\big| = 4.$$

This is exactly the set-theoretic form of the everyday rule that adding one genuinely new object increases the count by one.

Proposition 7.2.8 (Removing one element decreases the size by one)**.** *Let A be a finite set with $|A| = S(n)$, and let $a \in A$. Then*

$$|A \setminus \{a\}| = n.$$

Proof. Choose a bijection $f : S(n) \to A$, and let $b = f(n)$. Define a bijection $\tau : A \to A$ by swapping a and b and fixing all other elements:

$$\tau(x) = \begin{cases} a, & \text{if } x = b, \\ b, & \text{if } x = a, \\ x, & \text{otherwise.} \end{cases}$$

Then

$$g = \tau \circ f : S(n) \to A$$

is a bijection satisfying $g(n) = a$.

We claim that the restriction

$$g \restriction_n : n \to A \setminus \{a\}$$

is a bijection. If $x \in n$, then $x \neq n$, so injectivity of g gives $g(x) \neq g(n) = a$. Thus $g(x) \in A \setminus \{a\}$, so the restriction does map into $A \setminus \{a\}$.

It remains to check surjectivity. Let $y \in A \setminus \{a\}$. Since g is surjective onto A, some $x \in S(n)$ satisfies $g(x) = y$. Because $y \neq a = g(n)$, injectivity of g forces $x \neq n$, so $x \in n$. Hence $g \restriction_n (x) = y$.

Therefore $g \restriction_n$ is a bijection, and so $|A \setminus \{a\}| = n$. □

In the proofs that follow, it is useful to know that if one natural number is smaller than another, then its successor still does not leap past the larger number.

Proposition 7.2.9 (The successor of a smaller number is still at most the larger one)**.** *If $m < n$ in $\mathbb{N}_0$, then*

$$S(m) \leq n.$$

Proof. We prove the statement for all n by induction on n.

For $n = 0$, there is nothing to prove, because there is no $m < 0$.

Now assume that the statement holds for n, and let $m < S(n)$. Since $m \in S(n) = n \cup \{n\}$, either $m \in n$ or $m = n$.

If $m = n$, then $S(m) = S(n)$, so certainly $S(m) \leq S(n)$.

If $m \in n$, then $m < n$, so the induction hypothesis gives $S(m) \leq n$. Hence $S(m) \leq S(n)$ as well.

Thus the proposition holds for $S(n)$. By induction, it holds for all $n \in \mathbb{N}_0$. □

Corollary 7.2.10 (Successor preserves weak order)**.** *If $m \leq n$, then*

$$S(m) \leq S(n).$$

Proof. If $m = n$, then the conclusion is immediate. If $m < n$, then Proposition 7.2.9 gives $S(m) \leq n$, and therefore $S(m) \leq S(n)$. □

We are now ready for one of the decisive structural facts about finite sets: every subset of a finite set is finite, and every proper subset is strictly smaller.

Theorem 7.2.11 (Subsets of finite sets are finite, and proper subsets are smaller)**.** *Let A be a finite set, and let $B \subseteq A$.*

(i) The set B is finite.

(ii) If $B \subsetneq A$, then $|B| < |A|$.

Proof. Choose $n \in \mathbb{N}_0$ and a bijection $f : n \to A$. Let

$$C = f^{-1}[B] \subseteq n.$$

Because f is a bijection, the restriction

$$f \restriction_C : C \to B$$

is a bijection. So it is enough to prove the following claim:

If $C \subseteq n$, then C is finite; and if $C \subsetneq n$, then $|C| < n$.

We prove this claim by induction on n.

If $n = 0$, then the only subset of 0 is 0 itself, so the claim is trivial.

Now assume the claim has been proved for n, and let $C \subseteq S(n)$.

First suppose that $n \notin C$. Then $C \subseteq n$, so the induction hypothesis implies that C is finite and that, if $C \subsetneq n$, then $|C| < n$. In any case, C is finite and $|C| \leq n < S(n)$ whenever C is a proper subset of $S(n)$.

Now suppose that $n \in C$. Let

$$D = C \setminus \{n\}.$$

Then $D \subseteq n$. If $C = S(n)$, there is nothing to prove. Assume therefore that $C \subsetneq S(n)$. Then $D \subsetneq n$, so by the induction hypothesis D is finite and $|D| < n$. Write $|D| = m$. Since $m < n$, Proposition 7.2.9 gives $S(m) \leq n$. By Proposition 7.2.6, adding back the missing element n yields

$$|C| = S(m) \leq n < S(n).$$

So C is finite and, because it is a proper subset of $S(n)$, it is strictly smaller than $S(n)$.

This completes the inductive proof of the claim. Applying the claim to $C = f^{-1}[B]$, and then transporting the result across the bijection $f \restriction_C : C \to B$, proves the theorem. □

Corollary 7.2.12 (No proper subset of a finite set has the same size)**.** *If A is finite and $B \subsetneq A$, then there is no bijection $A \to B$.*

Proof. If there were a bijection $A \to B$, then $|A| = |B|$. But Theorem 7.2.11 shows that every proper subset of a finite set has strictly smaller size, so $|B| < |A|$. This is impossible. □

The previous theorem has an important conceptual consequence. For finite sets, "being smaller" really does mean "fitting inside as a proper subset." In the next chapter this principle will fail.

We also need one more general fact: the image of a finite set under any function is again finite. In the injective case the image has the same size; in the noninjective case it is strictly smaller.

Theorem 7.2.13 (Images of finite sets)**.** *Let A be a finite set with $|A| = n$, and let $f : A \to X$ be any function.*

(i) The image $f[A]$ is finite and satisfies $|f[A]| \leq n$.

(ii) If f is injective, then $|f[A]| = n$.

(iii) If f is not injective, then $|f[A]| < n$.

Proof. Choose a bijection $u : n \to A$ and consider the composite

$$g = f \circ u : n \to X.$$

Since u is surjective, we have

$$g[n] = f[A].$$

So it is enough to prove the theorem when the domain is the numeral n. We therefore prove by induction on n that every function $g : n \to X$ has finite image, with the stated refinements in the injective and noninjective cases.

If $n = 0$, then $g[0] = \varnothing$, so all three conclusions are clear.

Assume the statement has been proved for n, and let $g : S(n) \to X$. Write

$$C = g[n].$$

By the induction hypothesis, C is finite and $|C| \leq n$. We now compare $g(n)$ with C.

If $g(n) \in C$, then

$$g[S(n)] = C,$$

so the image has size at most n, hence strictly less than $S(n)$. In this case g is certainly not injective.

If $g(n) \notin C$, then

$$g[S(n)] = C \cup \{g(n)\}.$$

By Proposition 7.2.6, the size of $g[S(n)]$ is $S(|C|)$. Since $|C| \leq n$, Corollary 7.2.10 gives

$$|g[S(n)]| = S(|C|) \leq S(n).$$

Moreover, if g is injective, then $g \restriction_n$ is injective and $g(n) \notin C$, so the induction hypothesis yields $|C| = n$, and therefore

$$|g[S(n)]| = S(n).$$

If g is not injective but $g(n) \notin C$, then $g \restriction_n$ is not injective, so the induction hypothesis gives $|C| < n$. Hence

$$|g[S(n)]| = S(|C|) \leq n < S(n).$$

This proves all three statements for $S(n)$.

By induction, the theorem holds for numeral domains, and therefore it also holds for the original function $f : A \to X$. □

Corollary 7.2.14 (Injective and surjective maps between finite sets compare size)**.** *Let A and B be finite sets.*

(i) If there exists an injective map $A \to B$, then $|A| \leq |B|$.

(ii) If there exists a surjective map $A \to B$, then $|B| \leq |A|$.

Proof. For (i), let $f : A \to B$ be injective. Then Theorem 7.2.13 shows that $|f[A]| = |A|$. Since $f[A] \subseteq B$ and B is finite, Theorem 7.2.11 gives $|f[A]| \leq |B|$. Therefore $|A| \leq |B|$.

For (ii), let $f : A \to B$ be surjective. Then $f[A] = B$. By Theorem 7.2.13, we have $|B| = |f[A]| \leq |A|$. □

7.3 The Pigeonhole Principle

Counting arguments often feel more compelling than long chains of formal symbols. The pigeonhole principle is a perfect example. If we try to place more pigeons than there are pigeonholes, some hole must contain at least two pigeons. The statement is simple enough that one can believe it instantly, but it becomes much more powerful once it is translated into the language of functions.

In set theory, a placement of pigeons into holes is just a function. To each pigeon we assign the hole in which it lands. Saying that two pigeons land in the same hole means exactly that two different points of the domain receive the same value. So the pigeonhole principle is a statement about when a function cannot be injective.

Theorem 7.3.1 (Pigeonhole principle)**.** *Let A and B be finite sets. If*

$$|B| < |A|,$$

then no function $f : A \to B$ is injective. Equivalently, every function $f : A \to B$ identifies two distinct elements of A: there exist $a_1 \neq a_2$ in A such that

$$f(a_1) = f(a_2).$$

Proof. Suppose, for contradiction, that some function $f : A \to B$ is injective. Then Corollary 7.2.14 implies that

$$|A| \leq |B|.$$

But this contradicts the assumption $|B| < |A|$. Therefore no such injective function exists.

The second statement is just a reformulation of the failure of injectivity. □

Corollary 7.3.2 (Surjective form of the pigeonhole principle)**.** *Let A and B be finite sets. If $|B| < |A|$, then there is no surjective function $B \to A$.*

Proof. If there were a surjective function $g : B \to A$, then Corollary 7.2.14 would imply $|A| \leq |B|$, contradicting $|B| < |A|$. □

Example 7.3.3 (Two numbers with the same remainder)**.** Fix $n \in \mathbb{N}$, and choose $n + 1$ integers

$$a_0, a_1, \ldots, a_n.$$

Each integer has a remainder in the set $\{0, 1, \ldots, n-1\}$ when divided by n. So we obtain a function from an $(n+1)$-element set of indices to an n-element set of remainders. By Theorem 7.3.1, that function cannot be injective. Therefore two of the chosen integers have the same remainder modulo n.

Equivalently, among any $n + 1$ integers, two differ by a multiple of n.

The pigeonhole principle is often the first place where students feel that set-theoretic language is doing real work. A plain verbal fact about pigeons and holes becomes a general theorem about functions, and once stated in that form it applies immediately to remainders, birthdays, colorings, repeated values, and many other situations.

A second important finite principle is closely related: on a finite set, an injective self-map is automatically surjective, and a surjective self-map is automatically injective. This is false in infinite settings, as the next chapter will show.

Corollary 7.3.4 (Injective and surjective self-maps of a finite set)**.** *Let A be a finite set, and let $f : A \to A$. Then f is injective if and only if it is surjective.*

Proof. Suppose first that f is injective. By Theorem 7.2.13, the image $f[A]$ has the same size as A. Since $f[A] \subseteq A$ and no proper subset of a finite set has the same size as the whole set by Corollary 7.2.12, we must have $f[A] = A$. Hence f is surjective.

Conversely, suppose that f is surjective. Then $f[A] = A$. If f were not injective, Theorem 7.2.13 would imply $|f[A]| < |A|$, that is,

$$|A| < |A|,$$

which is impossible. Therefore f is injective. □

Remark 7.3.5 (A finite phenomenon). Corollary 7.3.4 is so familiar in school arithmetic that one may not notice how special it is. Later we will meet the shift map on $\mathbb{N}_0$, which is injective but not surjective. That single example will show that the finite world and the infinite world obey genuinely different size laws.

7.4 Counting Functions, Subsets, and Permutations of Finite Sets

Now that finite size is defined and its basic behavior understood, we can recover the familiar counting formulas of elementary combinatorics in a set-theoretic way. The formulas will no longer appear as isolated recipes. They will emerge from constructions we have already studied: disjoint union, Cartesian product, power set, and the set of all functions from one set to another.

Disjoint unions and products

Theorem 7.4.1 (Disjoint unions add)**.** *Let A and B be finite sets with*

$$A \cap B = \varnothing.$$

Then

$$|A \cup B| = |A| + |B|.$$

Proof. We prove the theorem by induction on $|B|$.

If $|B| = 0$, then $B = \varnothing$, so

$$|A \cup B| = |A| = |A| + 0.$$

Now assume the theorem has been proved whenever the second set has n elements, and suppose that $|B| = S(n)$. Choose an element $b \in B$. By Proposition 7.2.8, the set $B \setminus \{b\}$ has size n. Since A and B are disjoint, the element b does not belong to $A \cup (B \setminus \{b\})$. By the induction hypothesis,

$$\big|A \cup (B \setminus \{b\})\big| = |A| + n.$$

Adding the new element b back in and using Proposition 7.2.6 gives

$$|A \cup B| = S(|A| + n) = |A| + S(n) = |A| + |B|,$$

where the second equality is Proposition 6.5.3(ii) from Chapter 6. □

Example 7.4.2 (Why disjointness matters)**.** If

$$A = \{1, 2, 3\} \qquad \text{and} \qquad B = \{3, 4\},$$

then $|A| + |B| = 5$, but

$$|A \cup B| = 4.$$

The theorem does not fail here; its hypothesis fails. The two sets are not disjoint, so the element 3 is counted only once in the union.

Theorem 7.4.3 (Cartesian products multiply)**.** *Let A and B be finite sets. Then*

$$|A \times B| = |A| \cdot |B|.$$

Proof. We prove the theorem by induction on $|A|$.

If $|A| = 0$, then $A = \varnothing$, so $A \times B = \varnothing$ and the formula is true.

Now assume the theorem has been proved for all sets with m elements, and let $|A| = S(m)$. Choose $a \in A$, and set

$$A' = A \setminus \{a\}.$$

Then $|A'| = m$ by Proposition 7.2.8. Moreover,

$$A \times B = (A' \times B) \cup (\{a\} \times B),$$

and these two sets are disjoint because no ordered pair can have first coordinate both in A' and equal to a.

By the induction hypothesis,

$$|A' \times B| = m \cdot |B|.$$

The map

$$\phi \colon B \to \{a\} \times B, \qquad \phi(b) = \langle a, b \rangle,$$

is a bijection, so

$$|\{a\} \times B| = |B|.$$

Using Theorem 7.4.1, we obtain

$$|A \times B| = m \cdot |B| + |B|.$$

By Lemma 6.5.12 from Chapter 6, this equals

$$S(m) \cdot |B| = |A| \cdot |B|.$$

□

Example 7.4.4 (A rectangular array)**.** If $|A| = 2$ and $|B| = 3$, then $A \times B$ has $2 \cdot 3 = 6$ elements. Concretely, if

$$A = \{x, y\} \qquad \text{and} \qquad B = \{1, 2, 3\},$$

then

$$A \times B = \{\langle x, 1 \rangle, \langle x, 2 \rangle, \langle x, 3 \rangle, \langle y, 1 \rangle, \langle y, 2 \rangle, \langle y, 3 \rangle\}.$$

The theorem says that every such rectangular arrangement is counted by multiplication.

Functions and exponentiation

Chapter 4 already showed that a constant product can be viewed as a set of functions. If A and B are sets, we write

$$B^A = \{f \mid f \colon A \to B\}$$

for the set of all functions from A to B. At the finite level, this construction is counted by exponentiation.

Theorem 7.4.5 (Functions are counted by exponentiation)**.** *Let A and B be finite sets. Then*

$$|B^A| = |B|^{|A|}.$$

Proof. We first prove the statement for numeral sets. Fix $n \in \mathbb{N}_0$, and for each $m \in \mathbb{N}_0$ let

$$F_{m,n} = \{f \mid f : m \to n\}.$$

We claim that

$$|F_{m,n}| = n^m$$

for every m.

We use induction on m. If $m = 0$, there is exactly one function $0 \to n$, namely the empty function. Therefore

$$|F_{0,n}| = 1 = n^0.$$

Now assume that $|F_{m,n}| = n^m$. Every function $f : S(m) \to n$ is determined uniquely by two pieces of data:

(i) its restriction $f \restriction_m : m \to n$, and

(ii) the value $f(m) \in n$.

Conversely, if $g : m \to n$ is a function and $b \in n$, then

$$g \cup \{\langle m, b\rangle\}$$

is a function $S(m) \to n$. Thus the map

$$\Phi : F_{S(m),n} \to F_{m,n} \times n, \qquad \Phi(f) = \langle f \restriction_m, f(m)\rangle$$

is a bijection. Therefore

$$|F_{S(m),n}| = |F_{m,n} \times n| = |F_{m,n}| \cdot n = n^m \cdot n = n^{S(m)},$$

where the last equality is Proposition 6.5.17(iii) from Chapter 6. This proves the numeral case.

Now let $|A| = m$ and $|B| = n$, and choose bijections $u : m \to A$ and $v : n \to B$. Define

$$\Psi : B^A \to F_{m,n}$$

by

$$\Psi(f) = v^{-1} \circ f \circ u.$$

This is a bijection, with inverse

$$h \longmapsto v \circ h \circ u^{-1}.$$

Hence

$$|B^A| = |F_{m,n}| = n^m = |B|^{|A|}.$$

□

Example 7.4.6 (Binary strings and functions)**.** A function $f : 5 \to 2$ assigns to each of the five inputs either 0 or 1. So it is exactly the same thing as a binary string of length five. By

Theorem 7.4.5, there are $2^5 = 32$ such functions. This is the familiar count of length-five binary strings.

Subsets and the power set

The power set construction from Chapter 2 becomes particularly vivid when the underlying set is finite. We already know from Proposition 4.4.10 that subsets correspond exactly to characteristic functions with values in $\{0, 1\}$. Combining that fact with the previous theorem gives the standard counting formula for power sets.

Theorem 7.4.7 (The power set of an n-element set has 2^n elements)**.** *If A is a finite set with $|A| = n$, then*

$$|\mathcal{P}(A)| = 2^n.$$

Proof. By Proposition 4.4.10 from Chapter 4, the power set $\mathcal{P}(A)$ is in bijection with the set $\{0, 1\}^A$ of all characteristic functions on A. Since $|\{0, 1\}| = 2$, Theorem 7.4.5 gives

$$|\mathcal{P}(A)| = |\{0, 1\}^A| = 2^{|A|} = 2^n.$$

□

Example 7.4.8 (The subsets of a three-element set)**.** If $A = \{a, b, c\}$, then A has three elements, so its power set has $2^3 = 8$ elements. Indeed,

$$\mathcal{P}(A) = \{\varnothing, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}.$$

The theorem says that this list of eight subsets is not a coincidence; it is the general pattern for every three-element set.

Permutations and factorial

The last classical counting formula in this chapter concerns *permutations*. Intuitively, a permutation of a set is a way of rearranging its elements. Formally, it is simply a bijection from the set to itself.

Definition 7.4.9 (Permutation and factorial)**.** (i) A *permutation* of a set A is a bijection $A \to A$.

(ii) The *factorial* function on $\mathbb{N}_0$ is defined recursively by

$$0! = 1, \qquad S(n)! = S(n) \cdot n!.$$

The recursive definition of factorial is justified by the recursion machinery of Chapter 6; compare Theorem 6.4.2. The next theorem explains why this particular recursive rule is natural.

Theorem 7.4.10 (Bijections from n count as $n!$)**.** *If A is a finite set with $|A| = n$, then the set of all bijections $n \to A$ has exactly $n!$ elements.*

Proof. We prove the theorem by induction on n.

If $n = 0$, then $A = \varnothing$, and there is exactly one bijection $0 \to \varnothing$, namely the empty function. Thus the set of all such bijections has one element, which is $0!$.

Now assume the theorem has been proved for n, and let A be a set with $|A| = S(n)$. For each $a \in A$, let E_a denote the set of all bijections $f : S(n) \to A$ satisfying $f(n) = a$.

The sets E_a are pairwise disjoint, and their union is exactly the set of all bijections $S(n) \to A$. Fix $a \in A$. By Proposition 7.2.8, the set $A \setminus \{a\}$ has size n. Restriction gives a bijection from E_a to the set of bijections $n \to A \setminus \{a\}$:

$$f \longmapsto f \restriction_n .$$

The inverse map extends a bijection $g : n \to A \setminus \{a\}$ by declaring $g(n) = a$. By the induction hypothesis, each set E_a therefore has exactly $n!$ elements.

Now choose a bijection $u : S(n) \to A$. For each $i \in S(n)$, choose a bijection

$$\phi_i : n! \to E_{u(i)}.$$

Since the sets $E_{u(i)}$ are pairwise disjoint, the map

$$\Phi : S(n) \times n! \to \bigcup_{a \in A} E_a, \qquad \Phi(i, t) = \phi_i(t)$$

is a bijection. Therefore

$$\left| \bigcup_{a \in A} E_a \right| = |S(n) \times n!| = S(n) \cdot n! = S(n)!.$$

But the union on the left is precisely the set of all bijections $S(n) \to A$. Hence that set has $S(n)!$ elements. □

Corollary 7.4.11 (A finite set with n elements has $n!$ permutations)**.** *If A is a finite set with $|A| = n$, then the set of permutations of A has exactly $n!$ elements.*

Proof. Choose a bijection $u : n \to A$. Then composition with u induces a bijection from the set of bijections $n \to A$ to the set of permutations of A:

$$f \longmapsto f \circ u^{-1}.$$

By Theorem 7.4.10, the first set has $n!$ elements. Therefore the second one does as well. □

Example 7.4.12 (Six permutations of a three-element set)**.** A set with three elements has $3! = 6$ permutations. For example, the set $\{a, b, c\}$ can be rearranged in exactly six ways:

$$abc,\ acb,\ bac,\ bca,\ cab,\ cba.$$

The factorial formula says that this familiar school fact is really a statement about bijections of finite sets.

7.5 Where Finite Intuition Begins to Fail

The results of this chapter line up beautifully with everyday expectation. Proper subsets are smaller than the sets that contain them. Injective self-maps of finite sets are automatically surjective. Surjective self-maps are automatically injective. A set with n elements has 2^n subsets and $n!$ permutations. Everything feels stable, familiar, and finite.

Precisely for that reason, this is the right moment to ask a dangerous question:

Which of these truths are really truths about all sets, and which are truths only about finite sets?

The next chapter will formalize the idea of infinity. But even before we give the official definitions, we can already see that some finite principles are going to break.

Proposition 7.5.1 (The shift map on $\mathbb{N}_0$)**.** *Define* $s : \mathbb{N}_0 \to \mathbb{N}_0$ *by*

$$s(n) = S(n) = n + 1.$$

Then s *is injective but not surjective.*

Proof. The function s is injective by Proposition 6.3.9. It is not surjective because 0 is not in its range: by Corollary 6.3.10, no natural number has successor 0. □

Example 7.5.2 (A proper subset that matches the whole)**.** The image of the shift map is

$$s[\mathbb{N}_0] = \mathbb{N} = \{1, 2, 3, \dots\}.$$

So s is a bijection from $\mathbb{N}_0$ onto the proper subset $\mathbb{N}$ of $\mathbb{N}_0$. By Corollary 7.2.12, this could never happen for a finite set. The example therefore announces, before the formal definitions arrive, that $\mathbb{N}_0$ does not behave like a finite set.

Remark 7.5.3 (Dedekind's idea of infinity). One of Dedekind's great observations is that a set should count as infinite when it can be put in bijection with a proper subset of itself [20]. We have not yet adopted that as a formal definition, but the shift map above shows exactly why the idea is natural. The finite world is characterized by the impossibility of such a bijection; the infinite world begins where that impossibility breaks.

So the chapter has done two things at once. It has made finite counting precise, and it has prepared us to distrust finite intuition once we leave the finite setting. The definition of size by bijection was powerful enough to explain finite counting, and it will be powerful enough to describe the first infinite sets as well.

Looking ahead

This chapter translated ordinary counting into set-theoretic language. We defined equinumerosity by bijections, used the natural numbers as standard models of finite size, proved that finite cardinality is well defined, and established the basic structural facts of the finite world: subsets of finite sets are finite, proper subsets are smaller, the pigeonhole principle holds, and finite counting formulas arise from unions, products, function sets, power sets, and permutations.

The next chapter begins the real study of infinity. We will define what it means for a set to be infinite, countably infinite, and countable, and we will discover that several of the laws just proved for finite sets no longer remain true. In particular, the shift map from Proposition 7.5.1 will become our first model of a new phenomenon: a set that can be matched with a proper subset of itself. So Chapter 8 begins exactly where the present chapter leaves off: at the point where finite intuition starts to fail and a richer theory of size is needed.

Chapter 8

Infinite, Countable, and Countably Infinite Sets

In the previous chapter we translated ordinary counting into the language of bijections. That point of view made the finite world look almost perfectly orderly. A finite set could not be matched with a proper subset of itself; injective self-maps of finite sets were automatically surjective; and the familiar counting formulas for unions, products, functions, and permutations fell naturally out of the set-theoretic picture. But the last pages of Chapter 7 already warned us that this comfortable world does not extend unchanged to all sets. The shift map on the natural numbers behaved in a way that no finite set could imitate.

So this chapter begins the first serious study of infinity. At first, that may sound like a philosophical rather than mathematical topic. After all, everyone has some informal feeling for what "infinite" means: endless, unbounded, never finishing. Yet one of Cantor's great insights was that infinity becomes mathematically manageable when we do not ask vague questions such as "How large is the infinite?" but rather ask precise structural questions. Can one infinite set be matched with another? Can it be arranged in a sequence? Can it be listed, one object after another, so that every element eventually appears?

The first surprise is that some infinite sets are still, in a precise sense, no larger than the natural numbers. The set of even numbers, the set of all integers, the set of ordered pairs of natural numbers, and even the set of rational numbers can all be listed in a sequence. This is not obvious when one first hears it. The rational numbers, for example, seem vastly more complicated than the natural numbers. Between any two distinct real numbers there are infinitely many rationals, and yet the rationals can still be arranged in a list.

The second surprise is that the possibility of such a list is the right borderline. A set is *countable* when it can be brought into line, in principle, as first, second, third, and so on. Some infinite sets have this property and some do not. The present chapter develops the countable side of the story. The next chapter will show, by Cantor's diagonal argument, that not every infinite set can be listed.

There is also an important foundational theme running quietly in the background. For standard sets such as $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{Q}$, we will be able to write down explicit enumerations and explicit functions. That makes the arguments feel concrete and trustworthy. But when we move from one set to whole families of sets, especially infinite families, the act of choosing an enumeration for each member of the family begins to raise deeper questions. Those questions will lead us, in Chapter 10, to the axiom of choice. So the present chapter does two jobs at once: it introduces the first infinite worlds, and it prepares us to see why choice enters set theory at all.

8.1 Infinite Sets and Dedekind's Idea

The word "infinite" is used so casually in ordinary speech that it is worth pausing to decide what it should mean mathematically. Since Chapter 7 gave a precise meaning to "finite," the most direct approach is immediate: a set is infinite when it is not finite. That simple definition is enough to start the subject.

But there is another idea, emphasized by Dedekind [20], that captures a special kind of self-similarity. Finite sets cannot be matched with proper subsets of themselves. So if a set *can* be matched with a proper subset, then the set must already have crossed the border into the infinite world. For the basic sets studied in this chapter, that idea gives an especially vivid picture of infinity.

Infinity as the failure of finiteness

Definition 8.1.1 (Infinite set)**.** A set A is called *infinite* if it is not finite.

So the empty set is not infinite, any singleton is not infinite, and in general every n-element set from Chapter 7 is not infinite. The definition is extremely short, but it becomes useful only when we discover workable ways to recognize infinite sets in practice.

Definition 8.1.2 (Dedekind-infinite set)**.** A set A is called *Dedekind-infinite* if there exists a proper subset $B \subsetneq A$ such that $A \approx B$.

The point of this definition is not that it replaces Definition 8.1.1. Rather, it gives a concrete test for one especially important kind of infinite behavior. A Dedekind-infinite set contains a proper copy of itself.

The next proposition shows that two natural formulations of this idea say exactly the same thing.

Proposition 8.1.3 (Two forms of Dedekind's idea)**.** *For a set A, the following are equivalent:*

(i) A is Dedekind-infinite.

(ii) There exists an injective function $f : A \to A$ that is not surjective.

Proof. Assume first that A is Dedekind-infinite. Then there exists a proper subset $B \subsetneq A$ and a bijection $g : A \to B$. Let $i : B \to A$ be the inclusion map, $i(b) = b$. Then the composite

$$i \circ g : A \to A$$

is injective, because both g and i are injective by Proposition 3.4.6. It is not surjective, because its range is exactly B, and B is a proper subset of A.

Conversely, suppose there exists an injective function $f : A \to A$ that is not surjective. Let $B = f[A]$ be its range. Then $B \subsetneq A$, because f is not surjective. The function $f : A \to B$ is surjective by definition of the range, and it is injective because we assumed it. Hence $f : A \to B$ is a bijection, so $A \approx B$. Therefore A is Dedekind-infinite. □

Theorem 7.2.12 from the previous chapter already tells us what happens in the finite world: no finite set can be equinumerous with a proper subset of itself. So Dedekind's idea really does witness infinity.

Theorem 8.1.4 (Dedekind-infinite sets are infinite)**.** *Every Dedekind-infinite set is infinite.*

Proof. Let A be Dedekind-infinite. Then by Definition 8.1.2, there exists a proper subset $B \subsetneq A$ such that $A \approx B$. If A were finite, Corollary 7.2.12 would say that no proper subset of A can be equinumerous with A. That is a contradiction. Therefore A is not finite, hence infinite by Definition 8.1.1. □

Example 8.1.5 (The natural numbers are Dedekind-infinite)**.** Consider the function $s : \mathbb{N} \to \mathbb{N}$ given by

$$s(n) = n + 1.$$

It is injective by Proposition 6.3.9, because $n + 1 = S(n)$. It is not surjective, because 1 is not in its range. So by Proposition 8.1.3, the set $\mathbb{N}$ is Dedekind-infinite. By Theorem 8.1.4, $\mathbb{N}$ is infinite.

This is the same phenomenon that already appeared in Proposition 7.5.1. The point is that infinity allows a set to "shift over" and still cover almost all of itself.

Example 8.1.6 (The even numbers are as numerous as all natural numbers)**.** Let

$$2\mathbb{N} = \{2n \in \mathbb{N} \mid n \in \mathbb{N}\}$$

be the set of even positive integers. The function $d : \mathbb{N} \to 2\mathbb{N}$ defined by

$$d(n) = 2n$$

is a bijection. So $\mathbb{N} \approx 2\mathbb{N}$, even though $2\mathbb{N}$ is a proper subset of $\mathbb{N}$.

For a student meeting set theory for the first time, this is often the moment when infinity stops being a vague adjective and becomes a genuine mathematical structure. The even numbers are not "half as many" as the natural numbers in the sense relevant to set theory, because they can be matched perfectly.

Remark 8.1.7 (A caution about converses). At the intuitive level it is tempting to say that a set is infinite *exactly when* it is Dedekind-infinite. For the standard sets of this chapter that intuition works beautifully. In later axiomatic set theory, however, the exact relationship between different formulations of infinity becomes subtler than it first appears. So in this book we keep Definition 8.1.1 as the official meaning of "infinite" and use Dedekind's criterion as a powerful and vivid way to recognize many important examples.

8.2 Countably Infinite and Countable Sets

Once we know that infinite sets exist, the next question is not merely whether a set is infinite, but what kind of infinity it has. The most basic infinite size is the size of the natural numbers themselves. A set with that size is large enough to be infinite, but still orderly enough to be listed one item after another.

This idea leads to the central definition of the chapter.

The first infinite size

Definition 8.2.1 (Countably infinite and countable)**.** Let A be a set.

(i) We say that A is *countably infinite* if $A \approx \mathbb{N}$.

(ii) We say that A is *countable* if either A is finite or A is countably infinite.

So a countable set is one whose elements can be exhausted by a finite list or by an infinite list of the form first, second, third, and so on. The distinction between "countably infinite" and "countable" is worth remembering. Countably infinite means *the same size as* $\mathbb{N}$. Countable allows the finite case as well.

Remark 8.2.2 (Why we use $\mathbb{N}$ here). Chapter 6 used $\mathbb{N}_0$ for the von Neumann natural numbers $0, 1, 2, \ldots$, while the present chapter uses $\mathbb{N}$ for the positive integers $1, 2, 3, \ldots$. This is only a matter of notation. Since Proposition 7.5.1 shows that $\mathbb{N}_0 \approx \mathbb{N}$, a set is countably infinite if and only if it is in bijection with $\mathbb{N}_0$. We choose $\mathbb{N}$ here because it fits the usual notation for lists $a_1, a_2, a_3, \ldots$.

Example 8.2.3 (The first countably infinite sets)**.** Each of the following sets is countably infinite:

(i) $\mathbb{N}$ itself, by the identity map.

(ii) $\mathbb{N}_0$, because $\mathbb{N}_0 \approx \mathbb{N}$ by Proposition 7.5.1.

(iii) $2\mathbb{N}$, by Example 8.1.6.

(iv) The set of odd positive integers, $2\mathbb{N} - 1 = \{2n - 1 \mid n \in \mathbb{N}\}$, by the bijection $n \mapsto 2n - 1$.

The definition says that countable sets are exactly those that behave, from the viewpoint of size, like finite initial segments of counting or like the full counting process itself. To use the definition well, however, we need practical tests. The next theorem begins with the most basic case: subsets of the natural numbers.

Theorem 8.2.4 (Every subset of $\mathbb{N}$ is countable)**.** *If $S \subseteq \mathbb{N}$, then S is countable.*

Proof. If S is finite, there is nothing to prove. So suppose that S is infinite.

Because S is a nonempty subset of $\mathbb{N}_0$, Corollary 6.3.15 gives a least element of S. Call it a_1.

Now assume that a_n has been chosen. Consider the set

$$S_{n+1} = \{s \in S \mid s > a_n\}.$$

We claim that S_{n+1} is nonempty. Indeed, the set $S \cap \{1, 2, \ldots, a_n\}$ is a subset of the finite set $\{1, 2, \ldots, a_n\}$, so it is finite by Theorem 7.2.11. If S_{n+1} were empty, then every element of S would lie in $S \cap \{1, 2, \ldots, a_n\}$, which would make S finite. That contradicts our assumption. Since S_{n+1} is a nonempty subset of $\mathbb{N}_0$, it has a least element by Corollary 6.3.15. Call that element a_{n+1}.

In this way we obtain an increasing sequence

$$a_1 < a_2 < a_3 < \cdots$$

of elements of S. Define $e : \mathbb{N} \to S$ by

$$e(n) = a_n.$$

The function e is injective because the sequence is strictly increasing.

It remains to prove that e is surjective. Let $s \in S$. The set

$$F = S \cap \{1, 2, \dots, s\}$$

is finite, nonempty, and contained in $\mathbb{N}$. We may therefore list its elements in increasing order:

$$b_1 < b_2 < \cdots < b_k = s.$$

We claim that $a_i = b_i$ for all $1 \le i \le k$. For $i = 1$, both a_1 and b_1 are the least element of S, so $a_1 = b_1$. Now assume $a_i = b_i$ for some $i < k$. By construction, a_{i+1} is the least element of S larger than $a_i = b_i$. But among the elements of S that are larger than b_i and at most s, the least one is precisely b_{i+1}. Hence $a_{i+1} = b_{i+1}$. By induction, $a_k = b_k = s$.

Thus every element $s \in S$ appears as $e(k)$ for some k, so e is surjective. Therefore e is a bijection $\mathbb{N} \to S$, and S is countably infinite. In particular, S is countable. □

The same idea immediately extends from subsets of $\mathbb{N}$ to subsets of any countable set.

Corollary 8.2.5 (Subsets of countable sets are countable)**.** *If A is countable and $B \subseteq A$, then B is countable.*

Proof. If A is finite, then B is finite by Theorem 7.2.11, hence countable.

Now suppose that A is countably infinite. Choose a bijection $f : \mathbb{N} \to A$. Let

$$S = f^{-1}[B] = \{n \in \mathbb{N} \mid f(n) \in B\}.$$

Then $S \subseteq \mathbb{N}$, so S is countable by Theorem 8.2.4. The restriction $f \restriction_S : S \to B$ is a bijection. Hence $B \approx S$, so B is countable. □

The next theorem gathers together the three most useful ways to think about countable sets. The official definition uses bijections with $\mathbb{N}$ in the infinite case, but in practice one often proves countability by finding an injection into $\mathbb{N}$ or a surjection from $\mathbb{N}$.

Theorem 8.2.6 (Equivalent tests for countability)**.** *Let A be a nonempty set. Then the following are equivalent:*

(i) A is countable.

(ii) There exists an injective function $f : A \to \mathbb{N}$.

(iii) There exists a surjective function $g : \mathbb{N} \to A$.

Proof. We prove $(i) \Rightarrow (ii) \Rightarrow (i)$ and $(i) \Rightarrow (iii) \Rightarrow (i)$.

$(i) \Rightarrow (ii)$. If A is finite, say it has n elements, then there is a bijection from A to the finite subset $\{1, 2, \dots, n\}$ of $\mathbb{N}$, and hence an injection $A \to \mathbb{N}$. If A is countably infinite, then by Definition 8.2.1 there is a bijection $A \to \mathbb{N}$, which in particular is injective.

$(ii) \Rightarrow (i)$. Suppose $f : A \to \mathbb{N}$ is injective. Then $f[A] \subseteq \mathbb{N}$, so $f[A]$ is countable by Theorem 8.2.4. Because f is injective, the function $f : A \to f[A]$ is a bijection. Thus $A \approx f[A]$, and therefore A is countable.

$(i) \Rightarrow (iii)$. If A is countably infinite, then by definition there exists a bijection $g\colon \mathbb{N} \to A$. In particular, g is surjective. If A is a finite nonempty set, list its elements as $a_1, \ldots, a_n$, and define $g\colon \mathbb{N} \to A$ by

$$g(i) = a_i \quad \text{for } 1 \le i \le n, \qquad \text{and} \qquad g(m) = a_n \quad \text{for } m > n.$$

Then g is surjective.

$(iii) \Rightarrow (i)$. Assume now that $g\colon \mathbb{N} \to A$ is surjective. Define

$$S = \{n \in \mathbb{N} \mid (\forall m < n)\, g(m) \neq g(n)\}.$$

Thus S is the set of *first occurrences* of the values of g. Since $S \subseteq \mathbb{N}$, Theorem 8.2.4 shows that S is countable.

We claim that the restriction $g \restriction_S\colon S \to A$ is a bijection. It is injective because if $m < n$ are in S, then the condition defining $n \in S$ says in particular that $g(m) \neq g(n)$. It is surjective because for any $a \in A$, the set

$$E_a = \{n \in \mathbb{N} \mid g(n) = a\}$$

is nonempty, and therefore has a least element by Corollary 6.3.15. That least element belongs to S by definition, and its image under g is a.

So $A \approx S$. Since S is countable, the set A is also countable. □

Theorem 8.2.7 (Images of countable sets are countable)**.** *Let A be a countable set, and let $f\colon A \to B$ be a function. Then the image $f[A]$ is countable.*

Proof. If A is finite, then $f[A]$ is finite by Theorem 7.2.13, hence countable.

Now suppose that A is countably infinite. Choose a bijection $e\colon \mathbb{N} \to A$. Then the composite

$$f \circ e\colon \mathbb{N} \to f[A]$$

is surjective. By Theorem 8.2.6, the set $f[A]$ is countable. □

Remark 8.2.8 (Countable and infinite means countably infinite). Because Definition 8.2.1 says that a countable set is either finite or countably infinite, any set that is *both* countable and infinite must automatically be countably infinite. We will use this observation repeatedly in what follows.

8.3 Standard Countable Sets

The general criteria of the previous section become convincing only when we see them at work on concrete examples. The point is not merely to collect a few clever tricks. Each example teaches a different strategy. For $\mathbb{Z}$ we use a zigzag list that alternates signs. For $\mathbb{N} \times \mathbb{N}$ we sweep the grid diagonally. For $\mathbb{Q}$ we use the fact that rationals can be represented by pairs of integers, though not uniquely. These are the basic patterns that reappear throughout set theory.

From lines to grids to fractions

Theorem 8.3.1 (The integers are countably infinite)**.** *The set $\mathbb{Z}$ is countably infinite.*

Proof. Define $\eta\colon \mathbb{N} \to \mathbb{Z}$ by

$$\eta(n) = \begin{cases} \dfrac{n-1}{2}, & \text{if } n \text{ is odd,} \\ -\dfrac{n}{2}, & \text{if } n \text{ is even.} \end{cases}$$

Thus

$$\eta(1) = 0, \quad \eta(2) = -1, \quad \eta(3) = 1, \quad \eta(4) = -2, \quad \eta(5) = 2, \ldots$$

We claim that η is a bijection.

It is surjective: if $z = 0$, then $\eta(1) = 0$. If $z = k > 0$, then $\eta(2k+1) = k = z$. If $z = -k < 0$ with $k \in \mathbb{N}$, then $\eta(2k) = -k = z$.

It is injective because every nonnegative integer occurs at a unique odd index and every negative integer occurs at a unique even index. More explicitly, if $\eta(m) = \eta(n)$, then either both values are nonnegative, forcing both indices to be odd and then equal, or both are negative, forcing both indices to be even and then equal. A positive integer can never equal a negative one, so odd and even indices cannot produce the same value.

Therefore η is a bijection $\mathbb{N} \to \mathbb{Z}$, and $\mathbb{Z}$ is countably infinite. □

Example 8.3.2 (A zigzag listing of the integers)**.** The theorem says that the integers may be listed as

$$0, -1, 1, -2, 2, -3, 3, -4, 4, \ldots$$

The list does not respect the usual order on $\mathbb{Z}$, but that is not required. Countability asks only that every integer appear somewhere at a finite stage.

The next example is more surprising, because $\mathbb{N} \times \mathbb{N}$ looks like a two-dimensional array rather than a one-dimensional list. But the array can still be traversed systematically.

Theorem 8.3.3 (The grid $\mathbb{N} \times \mathbb{N}$ is countably infinite)**.** *The set $\mathbb{N} \times \mathbb{N}$ is countably infinite.*

Proof. We define a bijection $\pi\colon \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ by the diagonal formula

$$\pi(m, n) = \frac{(m+n-2)(m+n-1)}{2} + m.$$

It is helpful to think of pairs (m, n) as lying on diagonals of constant sum $m + n$. The pairs on the diagonal $m + n = s$ are

$$(1, s-1), (2, s-2), \ldots, (s-1, 1),$$

and there are exactly $s - 1$ such pairs. The formula above says that we first count all pairs on earlier diagonals, and then count forward by m places along the diagonal $m + n = s$.

To prove injectivity, suppose that $\pi(m, n) = \pi(m', n')$. Set $s = m + n$ and $s' = m' + n'$. The values taken on the diagonal $m + n = s$ are precisely the integers in the interval

$$\left\{ \frac{(s-2)(s-1)}{2} + 1, \ldots, \frac{(s-1)s}{2} \right\}.$$

Different diagonals give disjoint intervals, so equality of the values forces $s = s'$. Once the diagonal is fixed, the formula $\pi(m, n) = \frac{(s-2)(s-1)}{2} + m$ shows that the value of π determines m, and then $n = s - m$. Hence $(m, n) = (m', n')$, so π is injective.

To prove surjectivity, let $k \in \mathbb{N}$. There is a unique $s \geq 2$ such that

$$\frac{(s-2)(s-1)}{2} < k \leq \frac{(s-1)s}{2}.$$

Then define

$$m = k - \frac{(s-2)(s-1)}{2}, \qquad n = s - m.$$

The inequality shows that $1 \leq m \leq s-1$, so both m and n lie in $\mathbb{N}$. A direct substitution gives $\pi(m,n) = k$. Thus every natural number lies in the range of π, and π is surjective.

So π is a bijection, and $\mathbb{N} \times \mathbb{N}$ is countably infinite. □

Example 8.3.4 (The diagonal sweep)**.** The proof may be visualized as the list

$$\begin{aligned}
&(1,1),\\
&(1,2),(2,1),\\
&(1,3),(2,2),(3,1),\\
&(1,4),(2,3),(3,2),(4,1),\\
&\quad\vdots
\end{aligned}$$

Each diagonal is finite, so we can finish one diagonal and then move to the next. Although the grid looks two-dimensional, it can still be threaded into a single sequence.

Now we turn to the rational numbers. They seem much more complicated than the integers, because they include fractions of every possible shape. But each rational number is still encoded by finitely many integers, and that is enough to keep the whole set countable.

Theorem 8.3.5 (The rational numbers are countably infinite)**.** *The set $\mathbb{Q}$ is countably infinite.*

Proof. From Theorem 8.3.1 we have a bijection $\eta\colon \mathbb{N} \to \mathbb{Z}$. Define a function $q\colon \mathbb{N} \times \mathbb{N} \to \mathbb{Q}$ by

$$q(m,n) = \frac{\eta(m)}{n}.$$

This function is surjective. Indeed, every rational number r can be written in the form $r = a/b$ with $a \in \mathbb{Z}$ and $b \in \mathbb{N}$. Since η is surjective, there exists $m \in \mathbb{N}$ with $\eta(m) = a$, and then $q(m,b) = r$.

By Theorem 8.3.3, the set $\mathbb{N} \times \mathbb{N}$ is countably infinite, hence countable. Therefore its image $q[\mathbb{N} \times \mathbb{N}] = \mathbb{Q}$ is countable by Theorem 8.2.7.

Finally, $\mathbb{Q}$ is infinite because $\mathbb{N} \subseteq \mathbb{Q}$. By Remark 8.2.8, a set that is both countable and infinite is countably infinite. Hence $\mathbb{Q}$ is countably infinite. □

Remark 8.3.6 (Density is not the same as size). The countability of $\mathbb{Q}$ is one of Cantor's early and most famous results [18]. It teaches an important lesson: being "everywhere" is not the same as being uncountable. There are rational numbers between any two distinct real numbers, and yet the rationals can still be listed. Order-theoretic richness and set-theoretic size are different ideas.

8.4 Countable Unions and Products

By now we know that many familiar sets are countable. The next natural question is structural: when can we build new countable sets out of old ones? For finite sets the answer was clean. Finite unions and finite products stayed finite. In the countable world the situation is more interesting. A countable union of countable sets is often countable, and a finite product of countable sets is countable, but the proofs now carry the flavor of infinite organization rather than finite counting.

This is also the place where a first shadow of the axiom of choice appears. To prove that a union of countably many countable sets is countable, we usually need an enumeration of each member of the family. If those enumerations are given, the proof is explicit. If they are not given, the question of choosing them simultaneously becomes a separate issue. For the moment we work constructively, with the enumerations in hand.

Countable unions with chosen enumerations

Theorem 8.4.1 (A constructive countable-union theorem)**.** *Let $(A_n)_{n\in\mathbb{N}}$ be a family of nonempty sets. Suppose that for each $n \in \mathbb{N}$ we are given a surjection $e_n : \mathbb{N} \to A_n$. Then*

$$\bigcup_{n\in\mathbb{N}} A_n$$

is countable.

Proof. Define $E : \mathbb{N} \times \mathbb{N} \to \bigcup_{n\in\mathbb{N}} A_n$ by

$$E(n, m) = e_n(m).$$

This is well defined, because $e_n(m) \in A_n$ for every pair (n, m).

The function E is surjective. Indeed, if $x \in \bigcup_{n\in\mathbb{N}} A_n$, then $x \in A_k$ for some $k \in \mathbb{N}$. Since $e_k : \mathbb{N} \to A_k$ is surjective, there exists $m \in \mathbb{N}$ with $e_k(m) = x$. Therefore $E(k, m) = x$.

By Theorem 8.3.3, the set $\mathbb{N} \times \mathbb{N}$ is countably infinite, hence countable. So its image under the surjection E is countable by Theorem 8.2.7. That image is exactly $\bigcup_{n\in\mathbb{N}} A_n$. □

Remark 8.4.2 (Where choice begins to whisper). The theorem proves countability of the union once the maps $e_n : \mathbb{N} \to A_n$ have been specified. That is the constructive heart of the argument. But notice what has quietly been assumed: for each set A_n, we have already selected an enumeration. When the family is infinite, the question of whether one may always choose such a whole sequence of enumerations belongs to the landscape of choice. Chapter 10 will return to this point in a more systematic way.

Example 8.4.3 (Finite binary strings are countable)**.** For each $n \in \mathbb{N}_0$, let B_n be the set of binary strings of length n. Thus

$$B_0 = \{\epsilon\}, \qquad B_1 = \{0, 1\}, \qquad B_2 = \{00, 01, 10, 11\},$$

and so on. Each B_n is finite, hence countable. Because $\mathbb{N}_0$ is countably infinite by Example 8.2.3, we may reindex the family as $C_n = B_{n-1}$ for $n \in \mathbb{N}$. Each length comes with an obvious explicit list, so Theorem 8.4.1 shows that the set

$$\bigcup_{n\in\mathbb{N}_0} B_n = \bigcup_{n\in\mathbb{N}} C_n$$

of all finite binary strings is countable.

This example is worth remembering. In the next chapter we will meet the set of *infinite* binary sequences, and the situation there will be completely different.

The countable-union theorem says that countability survives stacking sets one above another in countably many rows. The next result says that countability also survives a fixed finite number of coordinates.

Theorem 8.4.4 (The product of two countable sets is countable)**.** *If A and B are countable, then $A \times B$ is countable.*

Proof. If either A or B is empty, then $A \times B = \varnothing$, which is finite and therefore countable. So assume both sets are nonempty.

By Theorem 8.2.6, there exist injections $i\colon A \to \mathbb{N}$ and $j\colon B \to \mathbb{N}$. Let $\pi\colon \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ be the bijection from the proof of Theorem 8.3.3. Define $h\colon A \times B \to \mathbb{N}$ by

$$h\big(\langle a, b\rangle\big) = \pi\big(i(a), j(b)\big).$$

Because i, j, and π are all injective, Proposition 3.4.6 shows that h is injective. Hence $A \times B$ injects into $\mathbb{N}$. By Theorem 8.2.6, $A \times B$ is countable. □

Corollary 8.4.5 (Finite products of countable sets are countable)**.** *If $A_1, \ldots, A_k$ are countable sets, then the finite product*

$$A_1 \times A_2 \times \cdots \times A_k$$

is countable.

Proof. We argue by induction on k.

For $k = 1$, the statement is immediate. Assume it holds for some $k \in \mathbb{N}$, and let $A_1, \ldots, A_k, A_{k+1}$ be countable sets. Set

$$P = A_1 \times \cdots \times A_k.$$

By the induction hypothesis, the set P is countable. Applying Theorem 8.4.4 to P and to A_{k+1}, we conclude that

$$(A_1 \times \cdots \times A_k) \times A_{k+1}$$

is countable. This product is canonically identified with $A_1 \times \cdots \times A_k \times A_{k+1}$, so the induction step is complete. □

Remark 8.4.6 (Why countable products are another story). The corollary concerns a *fixed finite* number of factors. That is very different from an infinite product. The set $\{0,1\}^{\mathbb{N}}$ of all infinite binary sequences is a countable product of a two-element set with itself, and Chapter 9 will show that this set is not countable at all. So one should not read "countable products" here as "products of countably many factors." The finite number of coordinates is essential.

8.5 Lists, Enumerations, and the Shape of Countability

The formal results of the chapter can now be compressed into a single image: a countable set is a set that can be listed. But because the word "list" is so ordinary, it is easy to underestimate how much it is saying. A list may be finite or endless. It may repeat elements. It need not respect the geometric or algebraic shape of the set being listed. And it need not be practical for

human memory. The point is not that we can finish the list, but that every individual element has a finite address in it.

This is the right place to make the informal picture explicit.

What it means to be listable in principle

Definition 8.5.1 (Enumeration)**.** Let A be a set.

(i) An *enumeration* of A is a surjective function $e : \mathbb{N} \to A$.

(ii) A *repetition-free enumeration* of A is a bijection $e : \mathbb{N} \to A$.

The first notion allows repeated appearances of the same element. The second does not. For finite nonempty sets, one may still have an enumeration in the first sense by repeating the final element forever. Under this convention the empty set has no enumeration, because there is no surjection from the nonempty set $\mathbb{N}$ onto $\varnothing$. For countably infinite sets, the most informative enumerations are usually the repetition-free ones.

Proposition 8.5.2 (Countability as listability)**.** *Let A be a set.*

(i) The set A is nonempty and countable if and only if it admits an enumeration.

(ii) The set A is countably infinite if and only if it admits a repetition-free enumeration.

Proof. Part (i) is exactly the equivalence between Definition 8.2.1 and the surjective criterion from Theorem 8.2.6 for nonempty sets.

For part (ii), if A is countably infinite, then by definition there exists a bijection $\mathbb{N} \to A$, which is a repetition-free enumeration. Conversely, a repetition-free enumeration is a bijection $\mathbb{N} \to A$, so $A \approx \mathbb{N}$. Hence A is countably infinite. □

Example 8.5.3 (An enumeration may have repetitions)**.** Let $A = \{a, b, c\}$. The function $e : \mathbb{N} \to A$ given by

$$e(3k+1) = a, \qquad e(3k+2) = b, \qquad e(3k+3) = c$$

for $k \in \mathbb{N}_0$ is an enumeration of A. It is not injective, because the same three values repeat forever, but it is surjective. So finiteness is compatible with the existence of an infinite list.

Remark 8.5.4 (What an enumeration does *not* require). An enumeration does not have to respect the "natural" order already present on a set. The zigzag list of $\mathbb{Z}$ from Example 8.3.2 is not increasing. The diagonal list of $\mathbb{N} \times \mathbb{N}$ from Example 8.3.4 ignores the usual geometry of the grid. And the proof that $\mathbb{Q}$ is countable does not list the rationals by size, denominator, or distance from zero. Countability is about the existence of a sequence that eventually reaches every element, not about preserving some pre-existing arrangement.

Remark 8.5.5 (Every element appears at a finite stage). When a set is countable, we should imagine not a completed infinite list sitting all at once before us, but a process with the following feature: for each particular element, there is some finite stage at which that element appears. There is no last stage containing everything, but there is a finite place for every individual object. That is the mathematical content of listability.

> *A countable set is one whose elements can, in principle, be lined up as first, second, third, and so on, so that no element is lost forever.*

Remark 8.5.6 (Countability is not the end of infinity). By the end of this chapter we have seen that many sets which look larger than $\mathbb{N}$ are still countable: the integers, the grid $\mathbb{N} \times \mathbb{N}$, and the rationals. That success could easily lead to a wrong guess: perhaps every infinite set can be listed if one is clever enough. The next chapter shows that this guess is false. Cantor's diagonal argument will prove that the real numbers, and more generally power sets, escape all possible enumerations.

Looking ahead

This chapter began the real mathematics of infinite size. We defined infinite sets as nonfinite sets, introduced Dedekind's idea of a set that matches a proper subset of itself, and used that idea to recognize familiar infinite examples such as $\mathbb{N}$ and the even numbers. We then defined countably infinite and countable sets, proved practical criteria for countability in terms of injections into $\mathbb{N}$ and surjections from $\mathbb{N}$, and established that many standard sets are countably infinite: $\mathbb{Z}$, $\mathbb{N} \times \mathbb{N}$, and $\mathbb{Q}$.

We also saw the first closure properties of the countable world. With explicit enumerations in hand, countable unions of countable sets remain countable, and finite products of countable sets remain countable. At the same time, the proofs already hinted that infinite families require careful choices of representatives and lists. That hint will become a major theme later in the book.

The next chapter asks the question that naturally follows all this success: are *all* infinite sets countable? Cantor's answer was no. We will prove that no list can exhaust the real numbers, and more generally that no set can ever be put into bijection with its power set. That is the moment when infinity splits into genuinely different sizes.

Chapter 9

Uncountability and the Power Set

Chapter 8 produced a string of pleasant surprises. The integers are countable. The grid $\mathbb{N} \times \mathbb{N}$ is countable. Even the rational numbers are countable. After seeing one successful enumeration after another, it is very natural to begin suspecting that perhaps every infinite set can be listed if only one is clever enough.

This chapter is where that suspicion breaks. Cantor's great discovery was that countability is not the whole story of infinity [18, 19]. Some sets really do escape every possible list. The real numbers do. More generally, for every set X, the power set $\mathcal{P}(X)$ does. This is not merely a technical curiosity. It means that infinity does not come in a single size. There are genuinely different infinite worlds.

The central method is one of the most beautiful and durable ideas in all of mathematics: the *diagonal argument*. Whenever someone claims to have listed every object of a certain kind, diagonalization tries to build a new object by changing the list at its nth place in the nth step. The resulting object is guaranteed to disagree with the nth entry at least somewhere, and so it escapes the whole enumeration. The method looks almost playful when first seen, but its consequences are profound.

There is also a foundational lesson here. In Chapter 1 we warned that unrestricted set formation leads to paradox. The present chapter shows that diagonal reasoning itself is not the problem. On the contrary, diagonal reasoning becomes mathematically safe and powerful when it is carried out *inside* an already given set, such as $\mathbb{N}$ or an arbitrary ambient set X. This point will return when we prove Cantor's theorem for power sets and compare it with Russell's paradox.

By the end of the chapter we will know far more than simply that $\mathbb{R}$ is uncountable. We will see that subsets of $\mathbb{N}$, binary sequences, and certain families of real numbers are tightly connected. We will also see that the gap between countable and uncountable sets has concrete consequences: for example, there are not merely some but uncountably many transcendental real numbers. The familiar continuum of the real line will no longer look like an anonymous endless collection of points; it will acquire a clear structural profile.

9.1 Cantor's Diagonal Argument

Chapter 8 taught us to interpret countability as listability in principle. A countable set is one whose elements can be placed in a sequence

$$a_1, a_2, a_3, \ldots$$

so that every element eventually appears. To prove that a set is *not* countable, we therefore try to show that every such claimed list must fail.

That is the setting in which diagonalization first appears.

When a list cannot be complete

Definition 9.1.1 (Uncountable set)**.** A set A is called *uncountable* if it is not countable.

Since every finite set is countable by Definition 8.2.1, every uncountable set is automatically infinite. But the point of the new word is stronger: an uncountable set is not merely endless. It is too large to be captured by any enumeration.

The following easy observation is often useful.

Proposition 9.1.2 (A set containing an uncountable subset is uncountable)**.** *Let $B \subseteq A$. If B is uncountable, then A is uncountable.*

Proof. We argue by contrapositive. Suppose that A is countable. Then every subset of A is countable by Corollary 8.2.5. In particular, B is countable. Therefore if B is uncountable, A cannot be countable. □

Our first major theorem concerns the power set of the natural numbers. This is a natural testing ground because Chapter 8 already encouraged us to think of countability in terms of lists, and Chapter 4 already taught us to think of $\mathcal{P}(\mathbb{N})$ as the set of all subsets of $\mathbb{N}$. The question is therefore concrete and simple to state:

Can all subsets of $\mathbb{N}$ be listed as $A_1, A_2, A_3, \ldots$?

Cantor's answer is no.

Theorem 9.1.3 (The power set of the natural numbers is uncountable)**.** *The set $\mathcal{P}(\mathbb{N})$ is uncountable.*

Proof. Suppose, for contradiction, that $\mathcal{P}(\mathbb{N})$ is countable. Since it is certainly nonempty, Proposition 8.5.2 shows that it admits an enumeration. So there exists a surjective function

$$e : \mathbb{N} \to \mathcal{P}(\mathbb{N}).$$

For each $n \in \mathbb{N}$, write

$$A_n = e(n).$$

Thus

$$A_1, A_2, A_3, \ldots$$

is a list that is supposed to contain every subset of $\mathbb{N}$.

Now define a new subset $D \subseteq \mathbb{N}$ by

$$D = \{n \in \mathbb{N} \mid n \notin A_n\}.$$

Because D is a subset of $\mathbb{N}$, we have $D \in \mathcal{P}(\mathbb{N})$. Since e is surjective, there exists some $m \in \mathbb{N}$ such that

$$e(m) = D.$$

In other words,

$$A_m = D.$$

We now ask whether $m \in D$. By the definition of D,

$$m \in D \quad \text{if and only if} \quad m \notin A_m.$$

But $A_m = D$, so this becomes

$$m \in D \quad \text{if and only if} \quad m \notin D,$$

which is impossible.

The contradiction shows that our original assumption was false. Therefore $\mathcal{P}(\mathbb{N})$ is uncountable. □

Example 9.1.4 (How the diagonal set escapes the list)**.** The set D from the proof is built to disagree with the list at the diagonal places:

$$1 \text{ versus } A_1, \qquad 2 \text{ versus } A_2, \qquad 3 \text{ versus } A_3, \quad \ldots$$

If $1 \in A_1$, then $1 \notin D$; if $1 \notin A_1$, then $1 \in D$. So D differs from A_1 at the element 1. Likewise D differs from A_2 at the element 2, from A_3 at the element 3, and so on.

Thus D is not merely missing from the list for some mysterious reason. It is constructed to miss the nth listed set at the nth place. That is the heart of diagonalization.

Remark 9.1.5 (Why the method is called "diagonal"). If one writes the list $A_1, A_2, A_3, \ldots$ as a table of 0's and 1's, with rows indexed by the sets and columns indexed by the natural numbers, then the entries

$$(1,1),\ (2,2),\ (3,3),\ \ldots$$

lie on the diagonal of the table. The set D is obtained by changing the membership decision at exactly those diagonal positions. That geometric picture is so vivid that the entire argument is now known as a diagonal argument or a diagonalization argument .

Remark 9.1.6 (Cantor's turning point). Theorem 9.1.3 is one of the decisive moments in the history of mathematics. It shows, in a single short argument, that the set-theoretic universe already contains more subsets of $\mathbb{N}$ than there are natural numbers themselves. Cantor's papers [18, 19] turned this from a philosophical puzzle into a precise mathematical fact.

9.2 The Real Numbers Are Uncountable

The theorem just proved tells us that the collection of all subsets of $\mathbb{N}$ cannot be listed. But students often first meet uncountability through the real numbers. That is natural: the real line is the most familiar continuous object in elementary mathematics, and it feels far more crowded than the discrete sequence $1, 2, 3, \ldots$.

To prove that $\mathbb{R}$ is uncountable, we again use diagonalization. This time the objects to be listed are not subsets of $\mathbb{N}$, but decimal expansions of real numbers in the unit interval.

A diagonal construction among decimal expansions

Remark 9.2.1 (Our convention on decimal expansions). Every real number $r \in (0,1)$ has at least one decimal expansion

$$r = 0.d_1d_2d_3\ldots.$$

Some numbers have two, for example

$$0.5000\cdots = 0.4999\ldots.$$

Whenever this happens, we choose the decimal expansion that does *not* end in an endless string of 9's. With this convention, each real number in $(0,1)$ is represented by a single chosen decimal expansion.

This is only a bookkeeping device. It allows us to talk about the nth digit of the nth listed real number without ambiguity.

Theorem 9.2.2 (The unit interval is uncountable)**.** *The interval* $(0,1)$ *is uncountable.*

Proof. Suppose, for contradiction, that $(0,1)$ is countable. Since it is nonempty, Proposition 8.5.2 provides an enumeration

$$r_1, r_2, r_3, \ldots$$

of all numbers in $(0,1)$.

Write the chosen decimal expansion of each r_n as

$$r_n = 0.d_{n1}d_{n2}d_{n3}\ldots,$$

where each d_{nk} is one of the digits $0, 1, \ldots, 9$, and where by Remark 9.2.1 no chosen expansion ends in an endless string of 9's.

Now define a new decimal expansion

$$x = 0.c_1c_2c_3\ldots$$

by setting

$$c_n = \begin{cases} 1, & \text{if } d_{nn} \neq 1, \\ 2, & \text{if } d_{nn} = 1. \end{cases}$$

Thus each c_n is either 1 or 2, and in particular $x \in (0,1)$. Also, the decimal expansion of x certainly does not end in an endless string of 9's.

For each $n \in \mathbb{N}$, the number x differs from r_n in the nth decimal place, because $c_n \neq d_{nn}$. Since both decimal expansions are written in our chosen form, this implies that

$$x \neq r_n$$

for every n.

But the sequence $r_1, r_2, r_3, \ldots$ was supposed to enumerate all of $(0,1)$, and $x \in (0,1)$. So x should appear somewhere on the list. We have just proved that it does not. This contradiction shows that $(0,1)$ is uncountable. □

Corollary 9.2.3 (The real numbers are uncountable)**.** *The set* $\mathbb{R}$ *of real numbers is uncountable.*

Proof. We have $(0,1) \subseteq \mathbb{R}$. Since $(0,1)$ is uncountable by Theorem 9.2.2, Proposition 9.1.2 implies that $\mathbb{R}$ is uncountable. □

Corollary 9.2.4 (The irrational numbers are uncountable)**.** *The set $\mathbb{R} \setminus \mathbb{Q}$ of irrational real numbers is uncountable.*

Proof. The rational numbers $\mathbb{Q}$ are countably infinite by Theorem 8.3.5. Suppose, for contradiction, that $\mathbb{R} \setminus \mathbb{Q}$ were countable. Then

$$\mathbb{R} = \mathbb{Q} \cup (\mathbb{R} \setminus \mathbb{Q})$$

would be a union of two countable sets. Since the index set $\{1, 2\}$ is countable, Theorem 8.4.1 would imply that $\mathbb{R}$ is countable. This contradicts Corollary 9.2.3. Therefore $\mathbb{R} \setminus \mathbb{Q}$ is uncountable. □

Remark 9.2.5 (Density is not the same as size). The rational numbers are dense in $\mathbb{R}$: between any two distinct real numbers there lies a rational number. Yet $\mathbb{Q}$ is countable and $\mathbb{R}$ is uncountable. This is a good reminder that "spread throughout" and "having the same number of elements" are very different notions. Density concerns the order structure of a set inside the line; countability concerns the possibility of listing its elements.

Remark 9.2.6 (Diagonalization is a method of escape). The proof of Theorem 9.2.2 does not depend on any deep property of decimal notation. The essential idea is again the same as before: from the nth listed object, look at its nth piece of data and change it. That one simple recipe forces the newly constructed object to escape the entire list.

9.3 Cantor's Theorem for Power Sets

The uncountability of $\mathcal{P}(\mathbb{N})$ and of $(0, 1)$ may at first look like two separate miracles. In fact they are instances of a far more general phenomenon. The power-set argument does not really depend on anything special about $\mathbb{N}$. The same diagonal construction works for *every* set.

This result is usually called *Cantor's theorem*.

The general diagonal theorem

Theorem 9.3.1 (Cantor's theorem)**.** *For every set X, there is no surjective function*

$$f : X \to \mathcal{P}(X).$$

In particular, there is no bijection between X and $\mathcal{P}(X)$.

Proof. Suppose, for contradiction, that there exists a surjective function

$$f : X \to \mathcal{P}(X).$$

Define

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

Since $D \subseteq X$, we have $D \in \mathcal{P}(X)$.

Now f is surjective, so there exists some $d \in X$ such that

$$f(d) = D.$$

We now ask whether $d \in D$. By the definition of D,

$$d \in D \quad \text{if and only if} \quad d \notin f(d).$$

Since $f(d) = D$, this becomes

$$d \in D \quad \text{if and only if} \quad d \notin D,$$

which is impossible.

Therefore no such surjective function can exist. □

Example 9.3.2 (The special case $X = \mathbb{N}$)**.** If we take $X = \mathbb{N}$ in Theorem 9.3.1, we recover Theorem 9.1.3. Indeed, if $\mathcal{P}(\mathbb{N})$ were countable, then it would admit a surjection from $\mathbb{N}$ by Proposition 8.5.2, and Cantor's theorem would rule that out.

So the concrete diagonal argument from Section 9.1 was already the first example of a completely general principle.

Remark 9.3.3 (Cantor's theorem and Russell's paradox). The set D in the proof of Theorem 9.3.1 looks very much like the set that appeared in Russell's paradox in Chapter 1. The resemblance is real, but so is the difference.

In Russell's paradox, one tries to form a set of *all* sets having a certain property. That unrestricted step is exactly what creates the contradiction; see Proposition 1.5.4 and Remark 1.5.5. In Cantor's theorem, by contrast, the diagonal set is formed only as a subset of a previously given set X. So the construction stays safely inside the ambient power set $\mathcal{P}(X)$. The same diagonal idea is therefore dangerous in an unrestricted setting but perfectly legitimate in a bounded one [2, 3, 21].

Corollary 9.3.4 (Power sets create a strictly larger world)**.** *For every set X, there is an injective function*

$$i \colon X \to \mathcal{P}(X)$$

but no surjective function

$$X \to \mathcal{P}(X).$$

Proof. Define

$$i(x) = \{x\}.$$

If $i(x) = i(y)$, then $\{x\} = \{y\}$, and therefore $x = y$. So i is injective.

On the other hand, Theorem 9.3.1 says that no surjection $X \to \mathcal{P}(X)$ exists. □

Remark 9.3.5 (There is no final stage of size). Corollary 9.3.4 may be read as a recipe that never stops. Starting from any set X, one can pass to $\mathcal{P}(X)$, then to $\mathcal{P}(\mathcal{P}(X))$, then to $\mathcal{P}(\mathcal{P}(\mathcal{P}(X)))$, and so on. At each step one obtains a new set that cannot be exhausted by any surjection from the previous one. Long before we introduce cardinal numbers formally, this already tells us that the universe of sets cannot have a largest size.

Remark 9.3.6 (The finite case revisited). For a finite n-element set, Chapter 7 showed that the power set has 2^n elements. Cantor's theorem says that a similar strict increase happens for every set, finite or infinite. In the finite world the jump is measured by the familiar formula $n \mapsto 2^n$; in the general world the jump is measured by the impossibility of surjecting X onto $\mathcal{P}(X)$.

9.4 Binary Sequences, Intervals, and the Continuum

So far we have met uncountability in two guises: as the impossibility of listing all subsets of $\mathbb{N}$, and as the impossibility of listing all real numbers in an interval. This section explains why those two worlds are so closely connected.

The key idea is that a subset of $\mathbb{N}$ can be recorded by an infinite string of 0's and 1's, and such strings can in turn be encoded inside the real line. In that sense, the continuum already contains a concrete shadow of the whole power set of $\mathbb{N}$.

Binary sequences as characteristic functions

Definition 9.4.1 (Binary sequence)**.** A *binary sequence* is a function

$$b : \mathbb{N} \to \{0, 1\}.$$

In more informal language, it is an infinite sequence

$$b(1), b(2), b(3), \ldots$$

whose terms are all either 0 or 1.

Example 9.4.2 (A subset of $\mathbb{N}$ as a binary sequence)**.** Let

$$A = \{1, 3, 4, 7, 9, \ldots\} \subseteq \mathbb{N}.$$

Its characteristic function

$$\chi_A : \mathbb{N} \to \{0, 1\}$$

is a binary sequence. The first several values are

$$\chi_A(1) = 1, \quad \chi_A(2) = 0, \quad \chi_A(3) = 1, \quad \chi_A(4) = 1, \quad \chi_A(5) = 0.$$

So the set A may be encoded as the binary string

$$1\,0\,1\,1\,0\,0\,1\,0\,1\ldots.$$

The nth entry tells us whether n belongs to A.

Proposition 9.4.3 (Binary sequences correspond to subsets of $\mathbb{N}$)**.** *The map*

$$A \mapsto \chi_A$$

is a bijection from $\mathcal{P}(\mathbb{N})$ to the set $\{0, 1\}^{\mathbb{N}}$ of all binary sequences.

Proof. This is exactly Proposition 4.4.10 from Chapter 4 in the special case $X = \mathbb{N}$. □

Corollary 9.4.4 (The set of binary sequences is uncountable)**.** *The set $\{0, 1\}^{\mathbb{N}}$ of all binary sequences is uncountable.*

Proof. By Proposition 9.4.3, $\{0, 1\}^{\mathbb{N}}$ is in bijection with $\mathcal{P}(\mathbb{N})$. Since $\mathcal{P}(\mathbb{N})$ is uncountable by Theorem 9.1.3, the same is true of $\{0, 1\}^{\mathbb{N}}$. □

The next result places a concrete copy of $\mathcal{P}(\mathbb{N})$ inside the unit interval. To avoid the small ambiguity caused by binary expansions such as

$$0.01111\ldots_2 = 0.10000\ldots_2,$$

we will use decimal expansions with only the digits 1 and 2.

Proposition 9.4.5 (A copy of $\mathcal{P}(\mathbb{N})$ inside $(0,1)$)**.** *Let E be the set of all real numbers in $(0,1)$ whose decimal expansions have the form*

$$0.a_1a_2a_3\ldots$$

with each $a_n \in \{1,2\}$. Then $\mathcal{P}(\mathbb{N})$ is in bijection with E.

Proof. Define

$$\Phi\colon \mathcal{P}(\mathbb{N}) \to E$$

as follows. For $A \subseteq \mathbb{N}$, let

$$\Phi(A) = 0.a_1a_2a_3\ldots,$$

where

$$a_n = \begin{cases} 2, & \text{if } n \in A, \\ 1, & \text{if } n \notin A. \end{cases}$$

Then $\Phi(A)$ indeed belongs to E.

We first show that Φ is injective. Suppose $A \neq B$. Then there exists some $n \in \mathbb{N}$ such that $n \in A$ and $n \notin B$, or vice versa. Hence the nth digits of $\Phi(A)$ and $\Phi(B)$ are different. Since both decimal expansions use only the digits 1 and 2, the corresponding real numbers are different. Therefore $\Phi(A) \neq \Phi(B)$, so Φ is injective.

Now let $x \in E$. Write

$$x = 0.a_1a_2a_3\ldots$$

with each $a_n \in \{1,2\}$. Define

$$A = \{n \in \mathbb{N} \mid a_n = 2\}.$$

Then by construction $\Phi(A) = x$. So Φ is surjective.

Therefore Φ is a bijection. □

Remark 9.4.6 (Binary expansions and a small ambiguity). The most direct way to encode a subset $A \subseteq \mathbb{N}$ is often to use the binary real

$$0.\chi_A(1)\chi_A(2)\chi_A(3)\ldots_2.$$

This works beautifully most of the time, but binary expansions have the same kind of ambiguity as decimal expansions:

$$0.01111\ldots_2 = 0.10000\ldots_2.$$

Our decimal 1-2 encoding avoids that difficulty while keeping the same underlying idea. The important point is not the particular base, but the fact that infinite digit strings can carry the information of subsets of $\mathbb{N}$.

We now turn from special subsets of the unit interval to the interval itself.

Proposition 9.4.7 (All nondegenerate open intervals have the same size)**.** *Let $a, b \in \mathbb{R}$ with $a < b$. Then the open interval (a,b) is in bijection with $(0,1)$.*

Proof. Define

$$f\colon (0,1) \to (a,b)$$

by

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

This is the familiar affine map from school algebra. It is injective because $b-a > 0$, so $f(x) = f(y)$ implies $x = y$. It is surjective because if $y \in (a, b)$, then

$$x = \frac{y-a}{b-a}$$

lies in $(0, 1)$ and satisfies $f(x) = y$. Therefore f is a bijection. □

Theorem 9.4.8 (The real line has the same size as the unit interval)**.** *The set $\mathbb{R}$ is in bijection with $(0, 1)$.*

Proof. First define

$$g : \mathbb{R} \to (-1, 1)$$

by

$$g(x) = \frac{x}{1+|x|}.$$

We claim that g is a bijection.

To see injectivity, suppose $g(x) = g(y)$. If x and y are both nonnegative, then

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

and cross-multiplication gives $x = y$. The same argument works if both are nonpositive, because then $|x| = -x$ and $|y| = -y$. If one of x, y is nonnegative and the other is negative, then $g(x) \geq 0 > g(y)$, so equality is impossible. Hence g is injective.

To see surjectivity, let $t \in (-1, 1)$. Define

$$x = \frac{t}{1-|t|}.$$

Then $1 - |t| > 0$, so x is well defined. If $t \geq 0$, then $x \geq 0$, so $|x| = x$, and

$$g(x) = \frac{x}{1+x} = \frac{\frac{t}{1-t}}{1+\frac{t}{1-t}} = t.$$

If $t < 0$, then $x < 0$, so $|x| = -x$, and

$$g(x) = \frac{x}{1-x} = \frac{\frac{t}{1+t}}{1-\frac{t}{1+t}} = t.$$

Thus g is surjective.

So g is a bijection from $\mathbb{R}$ onto $(-1, 1)$. The affine map

$$h(t) = \frac{t+1}{2}$$

is a bijection from $(-1, 1)$ onto $(0, 1)$. Therefore

$$h \circ g : \mathbb{R} \to (0, 1)$$

is a bijection. □

Definition 9.4.9 (The continuum)**.** We will use the word *continuum* for the real line $\mathbb{R}$ and, equivalently, for the unit interval $(0, 1)$. By Theorem 9.4.8 and Proposition 9.4.7, every nondegenerate open interval represents the same set-theoretic size.

Later, when cardinal numbers are introduced, the word "continuum" will also refer to that common cardinality.

Remark 9.4.10 (Intervals are much alike from the viewpoint of size). The interval $(0, 1)$ may look much smaller than the whole line $\mathbb{R}$, just as the interval $(2, 3)$ may look much smaller than $(0, 100)$. Geometrically, those impressions are reasonable: the sets sit in different places and have different lengths. Set-theoretically, however, they have the same number of points. The difference between countable and uncountable is therefore much coarser than ordinary geometric measurement.

9.5 The New Landscape of Infinite Size

By now the picture is dramatically richer than it was at the end of Chapter 8. We no longer have a single vague idea called "infinity." We have at least two sharply different possibilities. Some sets, such as $\mathbb{Z}$ and $\mathbb{Q}$, are infinite yet listable. Others, such as $\mathcal{P}(\mathbb{N})$, $(0, 1)$, and $\mathbb{R}$, escape every list.

This new landscape is not just a matter of classification. It has surprising mathematical consequences. One of the most famous is that the transcendental numbers are unavoidable.

A countable family inside an uncountable world

Definition 9.5.1 (Algebraic and transcendental real numbers)**.** A real number x is called *algebraic* if there exists a nonzero polynomial

$$p(t) = a_0 + a_1 t + \cdots + a_n t^n$$

with integer coefficients $a_0, a_1, \ldots, a_n \in \mathbb{Z}$ such that

$$p(x) = 0.$$

A real number that is not algebraic is called *transcendental*.

The examples $\sqrt{2}$ and $\sqrt[3]{5}$ are algebraic because they satisfy the equations

$$x^2 - 2 = 0 \qquad \text{and} \qquad x^3 - 5 = 0.$$

By contrast, numbers such as π and e are transcendental, but that fact is far deeper than anything we need here. For the moment, we ask only whether transcendental numbers must exist at all.

Theorem 9.5.2 (The algebraic real numbers form a countable set)**.** *The set of all algebraic real numbers is countable.*

Proof. For each $n \in \mathbb{N}_0$, let P_n be the set of all nonzero polynomials

$$p(t) = a_0 + a_1 t + \cdots + a_n t^n$$

with integer coefficients. Such a polynomial is completely determined by its coefficient tuple

$$(a_0, a_1, \ldots, a_n) \in \mathbb{Z}^{n+1}.$$

So P_n may be viewed as a subset of $\mathbb{Z}^{n+1}$. Since $\mathbb{Z}$ is countably infinite by Theorem 8.3.1, finite products of countable sets are countable by Corollary 8.4.5. Hence $\mathbb{Z}^{n+1}$ is countable, and therefore P_n is countable by Corollary 8.2.5.

Now let

$$P = \bigcup_{n \in \mathbb{N}_0} P_n.$$

Because $\mathbb{N}_0$ is countable and each P_n is countable, Theorem 8.4.1 implies that P is countable. So there are only countably many nonzero polynomials with integer coefficients.

For each polynomial $p \in P$, let

$$R_p = \{x \in \mathbb{R} \mid p(x) = 0\}.$$

A nonzero polynomial of degree n has at most n real roots, so each R_p is finite. In particular, each R_p is countable.

The set A of algebraic real numbers is exactly

$$A = \bigcup_{p \in P} R_p.$$

This is a countable union of countable sets, because P is countable and each R_p is countable. Therefore A is countable by Theorem 8.4.1. □

Corollary 9.5.3 (There are uncountably many transcendental real numbers)**.** *The set of transcendental real numbers is uncountable.*

Proof. Let A be the set of algebraic real numbers and let T be the set of transcendental real numbers. Then

$$\mathbb{R} = A \cup T.$$

By Theorem 9.5.2, the set A is countable. If T were also countable, then $\mathbb{R}$ would be a union of two countable sets, and hence countable by Theorem 8.4.1. This contradicts Corollary 9.2.3. Therefore T is uncountable. □

Remark 9.5.4 (Most real numbers are transcendental). Set-theoretically speaking, algebraic real numbers occupy only a countable part of the uncountable continuum. So transcendental numbers do not merely exist; they are overwhelmingly abundant. The familiar algebraic numbers form only a thin countable layer inside the real line.

> *Infinity does not merely continue the counting process forever; it branches into genuinely different sizes.*

Remark 9.5.5 (What has changed since the previous chapter). At the end of Chapter 8, one might still hope that ingenious enough lists could eventually tame every infinite set. This chapter

proves that such hope is false. The power set of $\mathbb{N}$ cannot be listed. The continuum cannot be listed. And from any set at all, passing to its power set creates a new realm that no surjection from the original set can exhaust.

We therefore enter the next phase of the book with a new problem. Once different infinite sizes exist, how should we compare them systematically? One especially ambitious answer is to try to place sets into well-ordered form. That attempt leads directly to the axiom of choice.

Looking ahead

This chapter showed that countability is not the end of the infinite story. We defined uncountable sets, proved by Cantor's diagonal argument that $\mathcal{P}(\mathbb{N})$ is uncountable, and then adapted the same idea to decimal expansions to prove that $(0, 1)$ and therefore $\mathbb{R}$ are uncountable. We also saw that the diagonal construction is not tied to the natural numbers: Cantor's theorem says that no set can ever be put into bijection with its own power set.

Along the way we made the continuum more concrete. Subsets of $\mathbb{N}$ correspond to binary sequences, and those sequences can be encoded inside the unit interval. We also saw that every nondegenerate open interval has the same size as $(0, 1)$, and that the whole real line has that same size as well. Finally, as a first application of the countable/ uncountable divide, we proved that the algebraic real numbers are countable while the transcendental real numbers are uncountable.

The next chapter turns to a different but closely related theme. Up to now we have often chosen elements from sets or families of sets when a clear rule was available. But what if no such rule is visible? Can one still choose one element from each set in an arbitrary family? That question leads to the axiom of choice, one of the central principles of modern set theory and a decisive tool for comparing and organizing large collections.

Part III

Choice and the Transfinite Viewpoint

Chapter 10

The Axiom of Choice

By the time a student first hears the phrase "the axiom of choice," it is often presented as if it were a mysterious decree dropped from the sky: somehow we are supposed to accept that one may choose an element from each set in an arbitrary family, even when no rule is visible. Presented that way, the subject can feel needlessly dramatic. But by now we have already met the real mathematical sources of the problem.

In Chapter 4 we learned that an element of a general product

$$\prod_{i \in I} A_i$$

is exactly a simultaneous choice of one element from each factor. In Chapter 5 we saw that well-orders make choice trivial inside a fixed ordered universe: one simply picks the least element of each nonempty subset. In Chapters 8 and 9 we also learned that infinite collections come in sharply different forms. Some can be listed explicitly; others cannot. So the real question is no longer whether "choice" is a familiar everyday action. The question is whether arbitrary set-valued families admit a single global selector.

That is what the axiom of choice asserts. It does not tell us *how* to choose; it says that a choosing function exists. This distinction is crucial. When a family comes with a built-in recipe—a least element, a midpoint, a smallest integer, a first term of an explicit enumeration—we do not feel that any special principle is being used. The interest of the axiom begins precisely when no such recipe is visible.

One of the remarkable features of modern set theory is that the same principle reappears in apparently different guises. The assertion that every family of nonempty sets has a choice function turns out to be equivalent to the claim that every set can be well-ordered. It is also equivalent to Zorn's lemma, a maximal-principle statement about partially ordered sets. At first sight these theorems do not even look as though they belong to the same subject. By the end of this chapter we will see that they are three faces of one idea.

There is also an important philosophical point. Up to this stage of the book we have generally preferred explicit constructions: we listed the rationals, built the natural numbers as sets, and used Cantor's diagonal argument to exhibit missing elements. The axiom of choice marks a shift in tone. It allows us to prove strong existence theorems even when no concrete formula for the desired object is available. Some mathematicians regard this as perfectly natural; others regard it as a genuine extra commitment. Our aim in this chapter is not to settle that philosophical debate, but to understand clearly what the axiom says, why it appears, and what it changes.

10.1 Why Choice Appears Naturally

When we first choose one element from a set, nothing seems subtle. If A is nonempty, we may simply say "let $a \in A$." That is just ordinary mathematical language. The difference between one choice and an arbitrary family of choices is easy to overlook because the sentences look so similar. But there is a real shift in logical strength between

> *for this particular nonempty set, choose an element,*

and

> *for every family of nonempty sets, there exists a single function that chooses one element from each member of the family.*

The first statement concerns one set at a time. The second concerns a global selector for a whole family, possibly with no common structure and no visible pattern.

Where choice is automatic

Many familiar families come equipped with a canonical rule. In those situations the existence of a choice function is not mysterious at all.

Example 10.1.1 (Easy families with visible selectors)**.** (i) If $(A_i)_{i\in I}$ is a family of nonempty subsets of $\mathbb{N}$, then the function

$$c(i) = \min(A_i)$$

is a choice function. This works because $\mathbb{N}$ is well-ordered.

(ii) If $I = (0, 1)$ and $A_r = (0, r)$ for each $r \in I$, then

$$c(r) = \frac{r}{2}$$

defines a choice function.

(iii) If $(B_i)_{i\in I}$ is a family of nonempty closed intervals $B_i = [a_i, b_i] \subseteq \mathbb{R}$, then the midpoint rule

$$c(i) = \frac{a_i + b_i}{2}$$

gives a choice function.

These examples already contain the two main ways in which choice can be easy: either each set has a distinguished least element, or the family comes with an explicit formula that picks one point from each member. When such a rule is present, the axiom of choice is not needed.

Remark 10.1.2 (Well-ordering makes choice canonical). Chapter 5 already isolated the decisive point in Proposition 5.5.7: inside a well-ordered set, every nonempty subset has a least element, so a choice function arises automatically. This is our first hint that the axiom of choice and well-ordering should be closely connected.

Where the difficulty lies

The interest of the subject begins when there is no common recipe. We may have a family of nonempty sets, each of which contains elements, and yet have no obvious way to pick one element from each set in a coherent manner.

Example 10.1.3 (A family with no evident selector)**.** Let $\mathcal{F}$ be the family of all nonempty subsets of $\mathbb{R}$ that have no least element in the usual order. Every member of $\mathcal{F}$ certainly contains elements, but the most obvious choice rule—"take the least element"—is unavailable by design. Nor is there any single elementary formula that clearly selects one point from *every* member of $\mathcal{F}$.

The point is not that a choice function is impossible. The point is that its existence is not automatic from the description of the family.

Example 10.1.4 (A product viewpoint)**.** Suppose $(A_i)_{i\in I}$ is a family of nonempty sets. To specify an element of the product

$$\prod_{i\in I} A_i$$

is to specify, all at once, one value $a_i \in A_i$ for every $i \in I$. So the question "does the product have an element?" is exactly the question "can we choose one element from each set in the family?"

This translation is important because it shows that choice is not a separate slogan floating above set theory. It is built into the product construction itself.

The next proposition is only a restatement of what we already learned in Chapter 4, but it is so central to the present chapter that it is worth placing again in the foreground.

Proposition 10.1.5 (Choice functions and product elements)**.** *Let $(A_i)_{i\in I}$ be a family of nonempty sets. The following are equivalent:*

(i) There exists a choice function for the family.

(ii) The product $\prod_{i\in I} A_i$ is nonempty.

Proof. By Proposition 4.5.2, a function c is a choice function for the family if and only if $c \in \prod_{i\in I} A_i$. So a choice function exists exactly when that product contains at least one element. □

Remark 10.1.6 (Why finite choice feels different). For finite families, the existence of a choice function was already proved in Proposition 4.5.4. We simply choose one element from the first set, then one from the second, and so on. The subtlety begins only when there is no final step. The axiom of choice is therefore not about whether individual choices are possible; it is about whether an arbitrary family admits one global choosing function.

> *Choice becomes mathematically interesting not when one set is nonempty, but when an arbitrary family of nonempty sets asks for one selector at every index simultaneously.*

10.2 The Axiom of Choice and Choice Functions

The preceding section explained why a special principle might be needed. We now state that principle formally.

The central statement

Definition 10.2.1 (The axiom of choice)**.** The *axiom of choice* is the statement:

> *For every family $(A_i)_{i\in I}$ of nonempty sets, there exists a choice function c with domain I such that $c(i) \in A_i$ for every $i \in I$.*

By Proposition 10.1.5, we may express the same statement in product language.

Proposition 10.2.2 (Product form of the axiom of choice)**.** *The axiom of choice is equivalent to the statement that for every family $(A_i)_{i\in I}$ of nonempty sets,*

$$\prod_{i\in I} A_i \neq \varnothing.$$

Proof. This follows immediately from Proposition 10.1.5. □

Remark 10.2.3 (An existence statement, not an algorithm). The axiom of choice does *not* promise a formula for the selector. It does not tell us how to compute $c(i)$ from i. It says only that a function with the required property exists. This is one reason choice is often described as a nonconstructive principle.

Finite, countable, and arbitrary choice

Because finite families are easy, mathematicians often distinguish between weaker and stronger forms of choice.

Definition 10.2.4 (Countable choice)**.** The *axiom of countable choice* is the statement:

> *Every countable family of nonempty sets has a choice function.*

Full choice implies countable choice, because a countable family is a special case of an arbitrary family. The converse, however, is not known to hold in axiomatic set theory and in fact fails in general: the full axiom of choice is strictly stronger than countable choice.

Example 10.2.5 (Countable families may still lack a visible rule)**.** Suppose $(A_n)_{n\in\mathbb{N}}$ is a sequence of nonempty sets. If each A_n is a nonempty subset of $\mathbb{N}$, then the least-element rule $c(n) = \min(A_n)$ gives a selector. But if the sets are arbitrary, there need not be any evident common recipe.

So even countable choice is already more than a triviality. The point is not that each individual A_n contains elements; the point is that one wants a single function defined on all of $\mathbb{N}$ at once.

Remark 10.2.6 (Finite choice is a theorem). What is often called "finite choice" is not really an extra axiom in our setting. Proposition 4.5.4 already proved that every finite family of nonempty sets admits a choice function. So the genuine new content begins only at infinite index sets.

Why mathematicians care

The importance of the axiom of choice lies partly in the fact that it keeps reappearing in different parts of mathematics. Sometimes it is used openly, as in the existence of a well-order on an arbitrary set. Sometimes it appears in disguised form, as a maximality principle or as a theorem about selecting representatives from equivalence classes. One reason this chapter matters is that, after seeing several equivalent versions, we will be better able to recognize when a proof is really a choice argument even if the word "choice" never appears.

Remark 10.2.7 (A historical note). Zermelo's 1904 paper [22] gave the first proof that the axiom of choice implies the well-ordering theorem. The result quickly became controversial because it produced strong existence theorems without explicit constructions. Zermelo's 1908 axiomatization of set theory [23] was one of the decisive steps toward the modern axiomatic viewpoint.

10.3 Equivalent Forms: Well-Ordering and Zorn's Lemma

One of the great surprises of set theory is that the axiom of choice can be reformulated in ways that look entirely different. Two of the most important are the well-ordering theorem and Zorn's lemma. The first is a statement about arranging sets in a least-element order. The second is a statement about the existence of maximal elements in partially ordered sets. Their equivalence with the axiom of choice is one of the central organizing facts of modern mathematics.

The easy direction is that well-ordering gives choice: if the union of a family is well-ordered, we simply choose the least element of each set. The reverse direction is deeper. There we must show that a global choice function can be used to organize an arbitrary set into a well-order. After that, Zorn's lemma can be brought into the picture.

Initial segments

Before turning to the main equivalences, we need one small piece of order-theoretic language.

Definition 10.3.1 (Initial segment)**.** Let $(L, \leq)$ be a linearly ordered set. A subset $I \subseteq L$ is called an *initial segment* of L if whenever $y \in I$ and $x < y$, then $x \in I$.

Thus an initial segment contains everything that lies earlier than any of its members. In the ordered set $\mathbb{N}$, the subsets $\varnothing$, $\{1\}$, $\{1, 2\}$, and more generally $\{1, 2, \ldots, n\}$ are initial segments, as is the whole set $\mathbb{N}$ itself.

From well-ordering to choice

Theorem 10.3.2 (Well-ordering theorem)**.** *Every set can be well-ordered.*

Proposition 10.3.3 (The well-ordering theorem implies the axiom of choice)**.** *Assume Theorem 10.3.2. Then the axiom of choice holds.*

Proof. Let $(A_i)_{i \in I}$ be a family of nonempty sets. Put

$$U = \bigcup_{i \in I} A_i.$$

By the well-ordering theorem, there exists a well-order $\leq$ on U. Since each A_i is a nonempty subset of the well-ordered set $(U, \leq)$, it has a least element. Define

$$c(i) = \text{the least element of } A_i \text{ with respect to } \leq .$$

Then $c(i) \in A_i$ for every $i \in I$, so c is a choice function. Since the family was arbitrary, the axiom of choice holds. □

The converse is the heart of the section.

Theorem 10.3.4 (The axiom of choice implies the well-ordering theorem)**.** *Assume the axiom of choice. Then every set can be well-ordered.*

Proof. Let X be an arbitrary set. We must construct a well-order on X. The proof is somewhat technical on first reading, so it helps to keep one idea in mind: we would like to build X in stages by repeatedly choosing one new element from what remains.

Because we are assuming the axiom of choice, the family of all nonempty subsets of X has a choice function. So there exists a function

$$c : \mathcal{P}(X) \setminus \{\varnothing\} \to X$$

such that $c(A) \in A$ for every nonempty subset $A \subseteq X$. For every proper subset $A \subsetneq X$, define

$$\sigma(A) = c(X \setminus A).$$

Thus $\sigma(A)$ chooses one element of X that is not already in A.

The problem is to organize these choices coherently. To do so, we work not directly with individual elements but with families of previously constructed stages.

Definition 10.3.5 (Regular family)**.** A family $\mathcal{R} \subseteq \mathcal{P}(X)$ is called *regular* if the following conditions hold:

(i) $\mathcal{R}$ is linearly ordered by inclusion.

(ii) $\mathcal{R}$ is well-ordered by inclusion.

(iii) $\varnothing \in \mathcal{R}$.

(iv) For every nonempty $A \in \mathcal{R}$, if

$$A^- = \bigcup \{B \in \mathcal{R} \mid B \subsetneq A\},$$

then

$$A = A^- \cup \{\sigma(A^-)\}.$$

Condition (iv) says that each nonempty stage is obtained from the union of all earlier stages by adjoining exactly one new element, namely the element chosen by σ from what remains.

The smallest example is the family $\{\varnothing\}$, which is regular. So regular families do exist.

We next show that any two regular families agree as long as they are both defined. In other words, one must be an initial segment of the other. This is the key uniqueness feature that makes the whole construction work.

Lemma 10.3.6 (Regular families are comparable)**.** *If $\mathcal{R}$ and $\mathcal{S}$ are regular families, then one of them is an initial segment of the other under the inclusion order.*

Proof. Let C be the set of all members $A \in \mathcal{R} \cap \mathcal{S}$ such that the predecessors of A are the same in both families; that is,

$$\{B \in \mathcal{R} \mid B \subsetneq A\} = \{B \in \mathcal{S} \mid B \subsetneq A\}.$$

By construction, C is an initial segment of both $\mathcal{R}$ and $\mathcal{S}$.

Suppose, toward a contradiction, that C is a proper initial segment of both families. Since $\mathcal{R}$ and $\mathcal{S}$ are well-ordered by inclusion, there exists a least element $R_0 \in \mathcal{R} \setminus C$ and a least element $S_0 \in \mathcal{S} \setminus C$.

Because R_0 is the first element of $\mathcal{R}$ beyond C, its predecessors in $\mathcal{R}$ are exactly the members of C. Therefore, by Definition 10.3.5(iv),

$$R_0 = U \cup \{\sigma(U)\}, \qquad U = \bigcup C.$$

The same reasoning applies to S_0, because its predecessors in $\mathcal{S}$ are also exactly the members of C. Hence

$$S_0 = U \cup \{\sigma(U)\}.$$

So $R_0 = S_0$.

But then R_0 belongs to both families and has the same predecessors in both, which means $R_0 \in C$, contradicting the choice of R_0. Therefore C cannot be a proper initial segment of both families. Hence one regular family is an initial segment of the other. □

Now let $\mathcal{Q}$ be the union of all regular families. Because regular families are comparable by Lemma 10.3.6, this union is itself linearly ordered by inclusion. In fact it is again regular.

Lemma 10.3.7 (The union of all regular families is regular)**.** *The family $\mathcal{Q}$ is regular.*

Proof. We verify the four conditions of Definition 10.3.5.

Condition (i): let $A, B \in \mathcal{Q}$. Then A belongs to some regular family $\mathcal{R}$, and B belongs to some regular family $\mathcal{S}$. By Lemma 10.3.6, one of these families is an initial segment of the other. In particular, both A and B belong to the larger family, where they are comparable by inclusion. Hence $\mathcal{Q}$ is linearly ordered by inclusion.

Condition (ii): let $\Gamma \subseteq \mathcal{Q}$ be nonempty. Choose some $A \in \Gamma$, and let $\mathcal{R}$ be a regular family containing A. We claim that every member of Γ that is contained in A already lies in $\mathcal{R}$. Indeed, if $B \in \Gamma$ and $B \subseteq A$, choose a regular family $\mathcal{S}$ containing B. By Lemma 10.3.6, either $\mathcal{S}$ is an initial segment of $\mathcal{R}$, in which case $B \in \mathcal{R}$, or $\mathcal{R}$ is an initial segment of $\mathcal{S}$, in which case $A \in \mathcal{S}$ and the initial-segment property forces $B \in \mathcal{R}$. Therefore the set

$$\Gamma_A = \{B \in \Gamma \mid B \subseteq A\}$$

is a nonempty subset of the well-ordered family $\mathcal{R}$, so it has a least element B_0 under inclusion.

We claim that B_0 is the least element of Γ. Let $C \in \Gamma$. Since $\mathcal{Q}$ is linearly ordered by inclusion, either $C \subseteq A$ or $A \subseteq C$. In the first case, $C \in \Gamma_A$, so $B_0 \subseteq C$ by minimality of B_0. In the second case, we have $B_0 \subseteq A \subseteq C$. Thus $B_0 \subseteq C$ for every $C \in \Gamma$. So Γ has a least element. Therefore $\mathcal{Q}$ is well-ordered by inclusion.

Condition (iii): every regular family contains $\varnothing$, so $\varnothing \in \mathcal{Q}$.

Condition (iv): let $A \in \mathcal{Q}$ be nonempty. Choose a regular family $\mathcal{R}$ containing A. We claim that every predecessor of A in $\mathcal{Q}$ already belongs to $\mathcal{R}$. Indeed, if $B \in \mathcal{Q}$ and $B \subsetneq A$, choose a regular family $\mathcal{S}$ containing B. As above, Lemma 10.3.6 shows that both A and B lie in a regular family in which $\mathcal{R}$ is an initial segment; therefore the fact that $B \subsetneq A$ forces $B \in \mathcal{R}$. So the predecessors of A in $\mathcal{Q}$ are exactly the predecessors of A in $\mathcal{R}$. Since $\mathcal{R}$ is regular, condition (iv) follows for A inside $\mathcal{Q}$ as well. □

Let

$$M = \bigcup \mathcal{Q}$$

be the union of all stages appearing in $\mathcal{Q}$. We now show that in fact $M = X$.

Lemma 10.3.8 (The regular construction covers the whole set)**.** *We have $M = X$.*

Proof. Suppose, toward a contradiction, that $M \subsetneq X$. Then $\sigma(M)$ is defined and belongs to $X \setminus M$. Consider the larger family

$$\mathcal{Q}^+ = \mathcal{Q} \cup \{M \cup \{\sigma(M)\}\}.$$

We claim that $\mathcal{Q}^+$ is regular. Conditions (i), (ii), and (iii) are clear: we have simply added one new stage at the top of the well-ordered inclusion chain $\mathcal{Q}$. Condition (iv) also holds, because the predecessors of the new stage are exactly the members of $\mathcal{Q}$, whose union is M, and the new stage is $M \cup \{\sigma(M)\}$.

So $\mathcal{Q}^+$ is a regular family. But $\mathcal{Q}$ was the union of *all* regular families, so every member of any regular family already belongs to $\mathcal{Q}$. In particular, the new stage $M \cup \{\sigma(M)\}$ should belong to $\mathcal{Q}$, which is impossible because it contains the element $\sigma(M) \notin M$. This contradiction shows that $M = X$. □

We now read a well-order off the regular family $\mathcal{Q}$. Since $\mathcal{Q}$ is well-ordered by inclusion and its union is all of X, every element of X enters the construction at a unique first stage.

For each $x \in X$, let A_x be the least member of $\mathcal{Q}$ containing x. This is well-defined because Lemma 10.3.8 says that $x \in M = X$, so at least one stage contains x, and $\mathcal{Q}$ is well-ordered. Define a relation $\triangleleft$ on X by

$$x \triangleleft y \quad \text{if and only if} \quad A_x \subsetneq A_y.$$

We claim that $\triangleleft$ is a well-order on X.

First we show that the map $x \mapsto A_x$ is injective. Suppose that $A_x = A_y = A$. Let

$$A^- = \bigcup \{B \in \mathcal{Q} \mid B \subsetneq A\}.$$

By regularity,

$$A = A^- \cup \{\sigma(A^-)\}.$$

Since A is the first stage containing x, we have $x \notin A^-$, so $x = \sigma(A^-)$. Likewise $y \notin A^-$, so $y = \sigma(A^-)$. Hence $x = y$. Thus different elements of X appear at different first stages.

Because $\mathcal{Q}$ is linearly ordered by inclusion, the relation $\triangleleft$ is a total order on X. It remains to prove the least-element property. Let $S \subseteq X$ be nonempty. Consider the set of stages

$$\mathcal{A}_S = \{A_x \mid x \in S\}.$$

This is a nonempty subset of the well-ordered family $\mathcal{Q}$, so it has a least element, say A_s with $s \in S$. For any $x \in S$, we have $A_s \subseteq A_x$; by injectivity of the stage map, this means either $s = x$ or $A_s \subsetneq A_x$, that is, $s \triangleleft x$. So s is the least element of S.

Therefore $\triangleleft$ is a well-order on X. Since X was arbitrary, every set can be well-ordered. □

Remark 10.3.9 (What the proof really did). The preceding proof may look elaborate, but its core idea is simple. A choice function tells us how to pick one element from whatever part of X remains. The technical work is devoted to making sure that these successive picks are compatible and can be assembled into one global ordering. Zermelo's original argument [22] was the first demonstration that arbitrary choice leads to a well-order of every set.

Chains and Zorn's lemma

We now turn to the other famous form of choice.

Definition 10.3.10 (Chain and upper bound)**.** Let $(P, \leq)$ be a partially ordered set.

(i) A subset $C \subseteq P$ is a *chain* if it is totally ordered by $\leq$; that is, for every $x, y \in C$, either $x \leq y$ or $y \leq x$.

(ii) An element $u \in P$ is an *upper bound* of C if $x \leq u$ for every $x \in C$.

Thus a chain is a subset in which no incomparability remains, and an upper bound is a point that lies above the whole chain.

Definition 10.3.11 (Zorn's lemma)**.** *Zorn's lemma* says:

> *If $(P, \leq)$ is a partially ordered set in which every chain has an upper bound in P, then P has a maximal element.*

At first sight this looks far removed from choice functions. There is no family of sets, no product, and no visible selecting rule. Yet the lemma turns out to be equivalent to the axiom of choice.

We first derive Zorn's lemma from the well-ordering theorem.

Theorem 10.3.12 (The well-ordering theorem implies Zorn's lemma)**.** *Assume the well-ordering theorem. Then Zorn's lemma holds.*

Proof. Let $(P, \leq)$ be a partially ordered set such that every chain in P has an upper bound. By the well-ordering theorem, there exists a well-order $\prec$ on the underlying set P.

We now inspect the elements of P one at a time in $\prec$-order and build a subset $C \subseteq P$ stage by stage. When we arrive at an element $x \in P$, we have already decided which earlier elements $y \prec x$ belong to C. Let

$$C_{\prec x} = \{y \in P \mid y \prec x \text{ and } y \in C\}$$

be the set of previously admitted elements. We declare that x is to belong to C exactly when every element of $C_{\prec x}$ lies below x; that is, exactly when

$$y \leq x \qquad \text{for every } y \in C_{\prec x}.$$

Because the decision at stage x depends only on what happened at strictly earlier stages, the well-order $\prec$ determines a unique subset C by this rule. (Chapter 12 will later develop this kind of stage-by-stage construction systematically.)

By construction, the defining property of C is:

> *for each $x \in P$, the element x belongs to C if and only if it lies above every earlier element of C with respect to $\prec$.*

We first show that C is a chain in P. Let $x, z \in C$. Since $\prec$ is a total order, either $x \prec z$, $z \prec x$, or $x = z$. If $x \prec z$, then $x \in C_{\prec z}$, so the defining rule for stage z gives $x \leq z$. If $z \prec x$, then similarly $z \leq x$. Thus every two elements of C are comparable, so C is a chain.

By hypothesis, C has an upper bound $u \in P$. We claim first that $u \in C$. Indeed, if $y \prec u$ and $y \in C$, then $y \leq u$ because u is an upper bound of the chain C. So the admission rule at stage u says that u belongs to C.

We next show that u is maximal in P. Suppose $u \leq v$ for some $v \in P$. Let $y \prec v$ and assume $y \in C$. Since u is an upper bound of C, we have $y \leq u \leq v$. Therefore stage v also passes the admission test, so $v \in C$. But u is an upper bound of the whole chain C, hence $v \leq u$. Together with $u \leq v$, antisymmetry gives $u = v$.

Thus no element of P lies strictly above u, and u is a maximal element. Therefore Zorn's lemma holds. □

The converse uses the idea that a partial well-order on a subset of a set should be extendable unless it already covers the whole set.

Theorem 10.3.13 (Zorn's lemma implies the well-ordering theorem)**.** *Assume Zorn's lemma. Then every set can be well-ordered.*

Proof. Let X be a set. Consider the collection $\mathcal{W}$ of all pairs $\langle A, \leq_A \rangle$, where $A \subseteq X$ and $\leq_A$ is a well-order on A.

We partially order $\mathcal{W}$ by declaring

$$\langle A, \leq_A \rangle \preceq \langle B, \leq_B \rangle$$

if and only if all of the following hold:

(i) $A \subseteq B$,

(ii) $\leq_B$ restricts to $\leq_A$ on A, and

(iii) A is an initial segment of $(B, \leq_B)$.

One checks directly that $\preceq$ is a partial order.

Let $\mathcal{C} \subseteq \mathcal{W}$ be a chain under $\preceq$. We must show that $\mathcal{C}$ has an upper bound in $\mathcal{W}$. Let

$$U = \bigcup \{A \mid \langle A, \leq_A \rangle \in \mathcal{C}\}.$$

We define an order $\leq_U$ on U by saying that for $x, y \in U$,

$$x \leq_U y$$

if and only if for some member $\langle A, \leq_A \rangle \in \mathcal{C}$ containing both x and y, we have $x \leq_A y$.

This is well-defined. Indeed, if two members of the chain contain both x and y, then one extends the other and therefore agrees with it on the smaller domain.

We claim that $(U, \leq_U)$ is well-ordered. It is linearly ordered because every two elements lie together in some member of the chain, where they are comparable. Now let $S \subseteq U$ be nonempty. Choose $x \in S$, and choose a member $\langle A, \leq_A \rangle \in C$ with $x \in A$. Then $S \cap A$ is nonempty, so it has a least $\leq_A$-element, say a.

We show that a is least in all of S. Let $y \in S$, and choose $\langle B, \leq_B \rangle \in C$ with $y \in B$. Because C is a chain under $\preceq$, either $\langle A, \leq_A \rangle \preceq \langle B, \leq_B \rangle$ or the reverse. If $\langle B, \leq_B \rangle \preceq \langle A, \leq_A \rangle$, then $y \in A$, so $a \leq_A y$, hence $a \leq_U y$. If $\langle A, \leq_A \rangle \preceq \langle B, \leq_B \rangle$, then A is an initial segment of $(B, \leq_B)$. If we had $y <_B a$, the initial-segment property would force $y \in A$, contradicting the minimality of a in $S \cap A$. Therefore $a \leq_B y$, hence again $a \leq_U y$. So every nonempty subset of U has a least element, and $(U, \leq_U)$ is well-ordered.

Thus $\langle U, \leq_U \rangle$ is an upper bound of the chain C in $\mathcal{W}$. By Zorn's lemma, $\mathcal{W}$ has a maximal element, say $\langle M, \leq_M \rangle$.

If $M \neq X$, choose some $x \in X \setminus M$. Extend $\leq_M$ to an order $\leq'$ on $M \cup \{x\}$ by declaring

$$m <' x \qquad \text{for every } m \in M,$$

and leaving the old order on M unchanged. Then $(M \cup \{x\}, \leq')$ is well-ordered, and M is an initial segment of it. So

$$\langle M, \leq_M \rangle \prec \langle M \cup \{x\}, \leq' \rangle$$

in the partial order $\mathcal{W}$, contradicting maximality. Therefore $M = X$, and X carries a well-order. □

Corollary 10.3.14 (Choice, well-ordering, and Zorn are equivalent)**.** *The following statements are equivalent:*

(i) The axiom of choice.

(ii) The well-ordering theorem.

(iii) Zorn's lemma.

Proof. By Proposition 10.3.3, Theorem 10.3.4, Theorem 10.3.12, and Theorem 10.3.13, we have

$$(i) \Rightarrow (ii) \Rightarrow (i), \qquad (ii) \Rightarrow (iii), \qquad (iii) \Rightarrow (ii).$$

So all three statements are equivalent. □

Remark 10.3.15 (Why the equivalence matters). Different branches of mathematics often prefer different faces of the same principle. Set theory often uses the language of well-orders, algebra frequently uses Zorn's lemma, and product constructions naturally suggest the axiom of choice itself. Corollary 10.3.14 explains why these very different-looking arguments are, at bottom, using one and the same resource.

10.4 Representatives, Products, and Maximal Principles

Once the equivalent forms are in place, the axiom of choice begins to show up in concrete ways. The same idea of choosing one element from each set appears in selecting representatives from equivalence classes, in splitting surjective maps, and in extracting maximal chains from partially

ordered sets. These are not separate miracles; they are all expressions of the same underlying principle.

Choosing representatives

One of the most common uses of choice is to select exactly one element from each member of a partition.

Definition 10.4.1 (Transversal or set of representatives)**.** Let $\mathcal{F}$ be a family of pairwise disjoint nonempty sets. A subset $R \subseteq \bigcup \mathcal{F}$ is called a *transversal* (or a *set of representatives*) for $\mathcal{F}$ if every $A \in \mathcal{F}$ contains exactly one element of R.

Proposition 10.4.2 (Choice functions and transversals)**.** *Let $\mathcal{F}$ be a family of pairwise disjoint nonempty sets. The following are equivalent:*

(i) $\mathcal{F}$ has a choice function.

(ii) $\mathcal{F}$ has a transversal.

Proof. Assume first that c is a choice function on $\mathcal{F}$. Set

$$R = \{c(A) \mid A \in \mathcal{F}\}.$$

Since $c(A) \in A$ for each $A \in \mathcal{F}$, the set R meets every member of $\mathcal{F}$. Because the family is pairwise disjoint, two distinct members of $\mathcal{F}$ cannot contribute the same chosen element. Hence R meets each member of $\mathcal{F}$ in exactly one point, so R is a transversal.

Conversely, assume that R is a transversal. For each $A \in \mathcal{F}$, the set $A \cap R$ contains exactly one element. Define $c(A)$ to be that unique element. Then $c(A) \in A$ for every $A \in \mathcal{F}$, so c is a choice function. □

Corollary 10.4.3 (Equivalence classes admit representatives under choice)**.** *Assume the axiom of choice. Let E be an equivalence relation on a set X. Then there exists a subset $R \subseteq X$ containing exactly one element from each equivalence class.*

Proof. By Definition 5.2.5, the quotient set X/E is a family of pairwise disjoint nonempty subsets of X, namely the equivalence classes. By the axiom of choice, X/E has a choice function. Hence by Proposition 10.4.2, it has a transversal $R \subseteq X$. That transversal contains exactly one element from each equivalence class. □

Remark 10.4.4 (A famous example). On $\mathbb{R}$, define an equivalence relation by

$$x \sim y \quad \text{if and only if} \quad x - y \in \mathbb{Q}.$$

The classes are the rational translates of one another. By Corollary 10.4.3, the axiom of choice yields a set containing exactly one representative from each class. Such sets are among the classic sources of counterintuitive constructions in analysis. We will not pursue that direction here, but it is one reason the axiom of choice has such a large mathematical shadow.

Sections of surjective maps

Another useful reformulation of choice arises from the fibers of a surjective function.

Definition 10.4.5 (Right inverse or section)**.** Let $f : X \to Y$ be a function. A function $s : Y \to X$ is called a *right inverse* or a *section* of f if

$$f \circ s = \mathrm{id}_Y .$$

Equivalently, $s(y)$ is an element of the fiber $f^{-1}(\{y\})$ for every $y \in Y$.

Theorem 10.4.6 (Choice and sections of surjections)**.** *The axiom of choice is equivalent to the following statement:*

Every surjective function has a right inverse.

Proof. Assume first that the axiom of choice holds, and let $f : X \to Y$ be surjective. For each $y \in Y$, the fiber

$$F_y = f^{-1}(\{y\})$$

is nonempty because f is onto. By the axiom of choice, the family $(F_y)_{y \in Y}$ has a choice function s with $s(y) \in F_y$ for all $y \in Y$. Then

$$f(s(y)) = y \qquad \text{for every } y \in Y,$$

so $f \circ s = \mathrm{id}_Y$. Thus s is a right inverse of f.

Conversely, assume that every surjective function has a right inverse. Let $(A_i)_{i \in I}$ be a family of nonempty sets. Form the disjoint union

$$D = \{\langle i, a \rangle \mid i \in I,\ a \in A_i\}.$$

Define

$$\pi : D \to I, \qquad \pi(\langle i, a \rangle) = i.$$

Because each A_i is nonempty, π is surjective. By assumption, there exists a right inverse $s : I \to D$ with $\pi \circ s = \mathrm{id}_I$. Write

$$s(i) = \langle i, c(i) \rangle.$$

Then $c(i) \in A_i$ for every $i \in I$. So c is a choice function for the family. Therefore the axiom of choice holds. □

Example 10.4.7 (Why sections are selectors)**.** If $f : X \to Y$ is surjective, then each fiber $f^{-1}(\{y\})$ is a nonempty set of possible preimages of y. A section chooses one preimage in each fiber. Thus the theorem above says that the ability to split every surjection is exactly the ability to make arbitrary choices across fibers.

Maximal chains

Zorn's lemma is often used through a simpler-looking consequence.

Theorem 10.4.8 (Hausdorff maximal principle)**.** *Assume Zorn's lemma. Then every partially ordered set has a maximal chain.*

Proof. Let $(P, \leq)$ be a partially ordered set. Consider the set C of all chains in P, ordered by inclusion.

Let $\mathcal{D} \subseteq C$ be a chain of chains. We claim that

$$U = \bigcup \mathcal{D}$$

is itself a chain in P. Indeed, if $x, y \in U$, then $x \in C_1$ and $y \in C_2$ for some $C_1, C_2 \in \mathcal{D}$. Since $\mathcal{D}$ is linearly ordered by inclusion, either $C_1 \subseteq C_2$ or $C_2 \subseteq C_1$. Hence both x and y lie in one common chain, where they are comparable. So U is a chain.

Thus every chain in C has an upper bound in $(C, \subseteq)$, namely its union. By Zorn's lemma, C has a maximal element. That maximal element is a maximal chain in P. □

Remark 10.4.9 (A typical use of Zorn). In practice, one often does not use Zorn's lemma to find a maximal *element* directly. Instead one defines an auxiliary partially ordered set of partial constructions, checks that every chain has an upper bound obtained by taking a union, and then applies Zorn's lemma to obtain a maximal object. The Hausdorff maximal principle is the simplest model of this pattern.

10.5 What Choice Changes—and What It Does Not

At this point it is worth stepping back. The axiom of choice has now appeared in three equivalent forms and several applications. But what, practically speaking, does it change? And what does it leave unchanged? The answer is subtler than a simple yes-or-no slogan.

What choice does not do

The axiom of choice does not suddenly turn every proof into a mystery. Most of the mathematics developed earlier in this book did not require it. The construction of the natural numbers, the study of finite sets, countability arguments for $\mathbb{Z}$ and $\mathbb{Q}$, and Cantor's diagonal arguments are all concrete and explicit. Their force comes from clear functions, explicit encodings, or direct contradictions, not from an arbitrary global selector.

Remark 10.5.1 (Choice is not needed everywhere). Many statements about familiar countable or explicitly described objects can be proved without the full axiom of choice. For example, when the sets involved are nonempty subsets of $\mathbb{N}$, a least-element rule is already available. Likewise, finite products, finite unions, and the basic countability results of Chapters 8 and 9 were proved without invoking the axiom of choice.

Remark 10.5.2 (Choice does not produce formulas). Even when the axiom of choice gives existence, it may give no explicit description of the object obtained. A well-order of $\mathbb{R}$, for instance, is guaranteed by choice, but the theorem does not hand us a simple formula for such an order. Likewise, a set of representatives for the relation $x \sim y \iff x - y \in \mathbb{Q}$ is guaranteed to exist, but the proof does not tell us how to write one down concretely.

What choice does change

What choice does provide is a powerful existence principle for large and complicated structures.

Remark 10.5.3 (Three major consequences already seen). By this chapter alone, the axiom of choice has already been shown to imply three substantial facts:

(i) every set can be well-ordered;

(ii) every surjection has a section;

(iii) every partially ordered set has a maximal chain.

These are not cosmetic reformulations. They change the kinds of existence arguments one can make throughout mathematics.

Remark 10.5.4 (Weaker forms of choice). Full choice is not the only selection principle studied in set theory. Countable choice, introduced in Definition 10.2.4, asks only for selectors on countable families. Other principles, such as dependent choice, are strong enough for many parts of analysis while still weaker than the full axiom of choice. We will not pursue this hierarchy here, but it is useful to know that "choice" is not a single all-or-nothing topic.

Axiomatic status

Finally, from the foundational point of view, the axiom of choice is not something that can simply be proved from the intuitive material already developed. In modern axiomatic set theory it is treated as an extra axiom, usually added to the Zermelo–Fraenkel axioms to form ZFC. Later, in Chapter 15, we will step back and see where that larger axiomatic picture comes from.

Remark 10.5.5 (Independence in the modern axiomatic picture). One of the deepest twentieth-century discoveries is that the axiom of choice is independent of the remaining standard axioms of set theory. Very roughly speaking, Gödel showed in 1940 that if the usual axioms of set theory are consistent, then they remain consistent after adding the axiom of choice [25]. Cohen later showed that if those axioms are consistent, then they also remain consistent after adding the negation of the axiom of choice [26, 27]. Thus the axiom of choice cannot be settled from the other axioms alone.

> *The axiom of choice does not replace explicit mathematics; it enlarges the range of existence statements that explicit mathematics alone can prove.*

Looking ahead

This chapter was the book's first sustained encounter with a genuinely nonconstructive principle. We began by returning to the product view of families of sets and explaining why the existence of a global selector is not automatic for arbitrary families. We then stated the axiom of choice, distinguished it from finite and countable choice, and stressed that it is an existence statement rather than an algorithm.

The central result of the chapter was that three apparently different statements are equivalent: the axiom of choice, the well-ordering theorem, and Zorn's lemma. In proving these equivalences we saw that well-orders provide canonical choices, that choice can be used to build well-orders on arbitrary sets, and that maximal principles in partially ordered sets hide the same logical content. We also studied several applications: choosing representatives from equivalence classes, splitting surjective maps by sections, and deriving the Hausdorff maximal principle.

The next chapter takes the well-ordering side of the story seriously. Once a set has been well-ordered, we may ask when two well-ordered sets have the same ordered shape, how one sits as an initial segment of another, and how to turn well-ordered types into number-like objects. Those questions lead to ordinal numbers, the first true inhabitants of the transfinite world.

Chapter 11

Well-Ordered Sets and Ordinal Numbers

In Chapter 5 we met well-ordered sets as the ordered worlds in which every nonempty subset has a first element. In Chapter 10 we saw that the axiom of choice is equivalent to the statement that every set can be well-ordered. So by this point well-ordering is no longer a side topic. It has become one of the central structural ideas of the book.

The next natural question is not merely whether a set can be well-ordered, but how different well-orders compare. Are two given well-ordered sets "the same" from the order-theoretic point of view? Can one sit as an initial segment of another? And if well-ordered sets really behave like number systems, can we attach to each such order a canonical number-like object that records its ordered shape?

For finite sets, these questions are so familiar that we hardly notice them. A linearly ordered set with one point has the same ordered shape as any other one-point ordered set; a linearly ordered set with three points has the same ordered shape as any other three-point line. The natural number measuring the size of the initial segment $\{0, 1, \dots, n-1\}$ is both a counting number and an order type. Ordinals extend this idea beyond the finite.

The decisive step is to stop thinking of ordinals as vague labels and to realize them as actual sets. In the von Neumann picture, an ordinal is not merely a symbol standing for a well-ordered type; it is the set of all smaller ordinals. Thus

$$0 = \varnothing, \qquad 1 = \{0\}, \qquad 2 = \{0, 1\}, \qquad 3 = \{0, 1, 2\}, \quad \dots$$

This is already the pattern we used in Chapter 6 for the natural numbers. The present chapter explains why that pattern is not an accident, but the beginning of a transfinite hierarchy.

The chapter has two intertwined aims. The first is order-theoretic: we study initial segments and prove that well-ordered sets are rigidly comparable. The second is set-theoretic: we define ordinals as transitive sets well-ordered by membership and develop their first basic properties. By the end of the chapter we will have reached ω, the first infinite ordinal, and we will see how unions of sets of ordinals produce suprema and carry us farther upward. In the next chapter these ordinals will become the domain on which transfinite induction and transfinite recursion operate.

11.1 Order Isomorphism and Order Type

When we compare two sets as sets, the right notion of sameness is a bijection. When we compare two ordered sets, a plain bijection is not enough. We want a bijection that respects the order. That is the first step toward understanding ordinals.

Order-preserving maps and isomorphisms

Definition 11.1.1 (Order-preserving map)**.** Let $(A, \leq_A)$ and $(B, \leq_B)$ be linearly ordered sets. A function $f : A \to B$ is called *order-preserving* if for all $x, y \in A$,

$$x <_A y \quad \Longrightarrow \quad f(x) <_B f(y).$$

An order-preserving map sends earlier points to earlier points. It may collapse several elements to the same value, so it does not by itself capture "sameness" of ordered shape. For that we need a bijective version.

Definition 11.1.2 (Order isomorphism)**.** Let $(A, \leq_A)$ and $(B, \leq_B)$ be linearly ordered sets. An *order isomorphism* from A to B is a bijection $f : A \to B$ such that for all $x, y \in A$,

$$x <_A y \quad \Longleftrightarrow \quad f(x) <_B f(y).$$

If such a map exists, we say that A and B are *order-isomorphic*. In symbols, one often writes

$$(A, \leq_A) \cong (B, \leq_B).$$

Thus two ordered sets are order-isomorphic exactly when one can relabel the points of one by the points of the other without changing the order relations.

Example 11.1.3 (First examples of order isomorphism)**.** (i) Let

$$A = \{a_1, a_2, a_3\} \qquad \text{with} \qquad a_1 < a_2 < a_3,$$

and let

$$B = \{2, 5, 9\} \qquad \text{with the usual order.}$$

Then the map

$$f(a_1) = 2, \qquad f(a_2) = 5, \qquad f(a_3) = 9$$

is an order isomorphism. So the two ordered sets have the same ordered shape.

(ii) The usual order on $\mathbb{N}$ is order-isomorphic to the usual order on the set of even positive integers

$$2\mathbb{N} = \{2, 4, 6, \dots\}.$$

The order isomorphism is the map $n \mapsto 2n$.

(iii) The usual order on $\mathbb{Z}$ is *not* order-isomorphic to the usual order on $\mathbb{N}$. The reason is not merely that one set has negative numbers and the other does not. The deeper point is that $\mathbb{N}$ has a least element, while $\mathbb{Z}$ does not.

Remark 11.1.4 (Order-theoretic invariants). Any property that can be described purely in terms of order must be preserved by order isomorphism. Having a least element, having a greatest element, being dense, being well-ordered, or containing an endless descending sequence are all examples of such order-theoretic properties. This is why ordered sets are often studied "up to order isomorphism" rather than by the names of their underlying points.

Two quick observations are worth recording explicitly.

Proposition 11.1.5 (Order isomorphism is an equivalence relation)**.** *Among linearly ordered sets, order isomorphism is an equivalence relation.*

Proof. Reflexivity holds because the identity map on a linearly ordered set is an order isomorphism. Symmetry holds because if $f : A \to B$ is an order isomorphism, then its inverse $f^{-1} : B \to A$ is also an order isomorphism. Transitivity holds because the composition of two order isomorphisms is again an order isomorphism. □

Proposition 11.1.6 (Order isomorphisms preserve least elements and initial segments)**.** *Let $f : (A, \leq_A) \to (B, \leq_B)$ be an order isomorphism. Then:*

(i) if $S \subseteq A$ has a least element s_0, then $f(S)$ has least element $f(s_0)$;

(ii) for every $a \in A$,

$$f\big[\{x \in A \mid x <_A a\}\big] = \{y \in B \mid y <_B f(a)\}.$$

Proof. For (i), let s_0 be least in S. If $t \in f(S)$, then $t = f(s)$ for some $s \in S$. Since $s_0 \leq_A s$, order preservation gives $f(s_0) \leq_B t$. Thus $f(s_0)$ is least in $f(S)$.

For (ii), let

$$A_{<a} = \{x \in A \mid x <_A a\} \qquad \text{and} \qquad B_{<f(a)} = \{y \in B \mid y <_B f(a)\}.$$

If $x \in A_{<a}$, then $x <_A a$, so $f(x) <_B f(a)$, which means $f(x) \in B_{<f(a)}$. Thus $f[A_{<a}] \subseteq B_{<f(a)}$. Conversely, let $y \in B_{<f(a)}$. Since f is surjective, there exists $x \in A$ with $f(x) = y$. Then $f(x) = y <_B f(a)$, so $x <_A a$. Hence $x \in A_{<a}$, and therefore $y = f(x) \in f[A_{<a}]$. So the two sets are equal. □

Part (ii) says that an order isomorphism does not merely preserve the global order. It also preserves the entire local picture below each point.

Order type as ordered shape

Definition 11.1.7 (Order type)**.** The *order type* of a well-ordered set is its ordered shape up to order isomorphism.

This definition is intentionally informal. At the moment, "order type" means an equivalence class under order isomorphism. The idea is clear, but equivalence classes are not the most convenient objects to work with. We would like a canonical representative for each well-ordered shape, just as the natural number 3 is a canonical representative of every three-element well-ordered set.

Remark 11.1.8 (Why canonical representatives matter). In ordinary mathematics we often replace an abstract class of equivalent objects by one especially convenient model. In linear algebra, every finite-dimensional real vector space of dimension n is isomorphic to $\mathbb{R}^n$, so $\mathbb{R}^n$ serves as a standard model. For well-ordered sets, ordinals play exactly this role: they provide concrete set-theoretic representatives of well-ordered types.

Remark 11.1.9 (Cantor and von Neumann). Cantor introduced transfinite numbers in order to compare infinite ordered structures. The set-theoretic realization of ordinals that we use in this book is due to von Neumann [24]. It has one especially elegant feature: an ordinal is literally the set of all smaller ordinals. Standard modern treatments include Moschovakis [7], Jech [10], Kunen [11], and Potter [14].

The next section develops the rigidity of well-orders. Once we know how well-ordered sets compare by initial segments, ordinals will arise quite naturally.

11.2 Well-Ordered Sets and Initial Segments

For arbitrary linear orders, initial segments can be complicated and poorly behaved. In a well-ordered set they are much more rigid. Every proper initial segment is determined by a single cut point, and no well-ordered set can hide inside a proper initial segment of itself. These facts are the technical heart of the transition from well-orders to ordinals.

Principal initial segments

Chapter 10 introduced the general notion of an initial segment in Definition 10.3.1. For a well-ordered set it is useful to name the most obvious examples.

Definition 11.2.1 (Principal initial segment)**.** Let $(W, \leq)$ be a well-ordered set, and let $a \in W$. The set

$$W_{<a} = \{x \in W \mid x < a\}$$

is called the *principal initial segment* of W determined by a.

Thus $W_{<a}$ is simply the part of the well-order that lies strictly before a.

Example 11.2.2 (Principal initial segments)**.** (i) In $(\mathbb{N}, \leq)$, the principal initial segment below 5 is

$$\mathbb{N}_{<5} = \{1, 2, 3, 4\}.$$

(ii) In $(\mathbb{N}_0, \leq)$, the principal initial segment below 4 is

$$\{n \in \mathbb{N}_0 \mid n < 4\} = \{0, 1, 2, 3\} = 4$$

in the von Neumann notation of Chapter 6.

(iii) In the lexicographically ordered set $\mathbb{N} \times \mathbb{N}$, the principal initial segment below $\langle 2, 3 \rangle$ consists of all pairs whose first coordinate is less than 2, together with the pairs $\langle 2, 1 \rangle$ and $\langle 2, 2 \rangle$.

The first structural fact is that in a well-ordered set there are no mysterious proper initial segments.

Proposition 11.2.3 (Proper initial segments are principal)**.** *Let $(W, \leq)$ be a well-ordered set, and let $I \subsetneq W$ be a proper initial segment. Then there exists a unique element $a \in W$ such that*

$$I = W_{<a}.$$

Proof. Because I is a proper subset of W, the complement $W \setminus I$ is nonempty. Since W is well-ordered, $W \setminus I$ has a least element; call it a.

We claim that $I = W_{<a}$. First let $x \in I$. If we had $a \leq x$, then either $a = x$ or $a < x$. The equality $a = x$ is impossible because $a \notin I$. If $a < x$, then since I is an initial segment and $x \in I$, we would also get $a \in I$, again a contradiction. Therefore $x < a$, so $x \in W_{<a}$. Hence $I \subseteq W_{<a}$.

Conversely, let $x < a$. If $x \notin I$, then $x \in W \setminus I$ and $x < a$, contradicting the choice of a as the least element of $W \setminus I$. So $x \in I$. Thus $W_{<a} \subseteq I$, and the two sets are equal.

For uniqueness, suppose also that $I = W_{<b}$. If $a < b$, then $a \in W_{<b} = I = W_{<a}$, impossible. Similarly $b < a$ is impossible. Therefore $a = b$. □

Corollary 11.2.4 (A point is determined by what lies before it)**.** *Let $(W, \leq)$ be a well-ordered set. If $a, b \in W$ satisfy*

$$W_{<a} = W_{<b},$$

then $a = b$.

Proof. The set $W_{<a}$ is a proper initial segment of W. By Proposition 11.2.3, it determines a unique point of W. Since both a and b determine it, we must have $a = b$. □

Why a well-order cannot resemble a proper part of itself

The finite case suggests that a well-ordered set should never be order-isomorphic to a proper initial segment of itself. Unlike mere set-theoretic equinumerosity, well-ordered shape is rigid enough to prevent this.

Proposition 11.2.5 (No well-ordered set is isomorphic to a proper initial segment of itself)**.** *Let $(W, \leq)$ be a well-ordered set. Then $(W, \leq)$ is not order-isomorphic to any proper initial segment of itself.*

Proof. Suppose, toward a contradiction, that $I \subsetneq W$ is a proper initial segment and that

$$f : W \to I$$

is an order isomorphism. Since $I \neq W$, the map f cannot be the identity map on W. Thus the set

$$S = \{x \in W \mid f(x) \neq x\}$$

is nonempty. Because W is well-ordered, S has a least element; call it s.

By the choice of s, we have $f(x) = x$ for every $x < s$. Now compare $f(s)$ with s.

If $f(s) < s$, then $f(s) \notin S$ by minimality of s, so

$$f(f(s)) = f(s).$$

But f is injective, and also $f(f(s)) = f(s) = f(s)$. Injectivity therefore gives $f(s) = s$, contradicting $s \in S$.

If $s < f(s)$, then $f(s) \in I$, and because I is an initial segment of W, the inequality $s < f(s)$ forces $s \in I$ as well. Since f is surjective onto I, there exists $t \in W$ such that $f(t) = s$. Now

$$f(t) = s < f(s),$$

so order preservation gives $t < s$. By minimality of s, we have $f(t) = t$. Hence $t = s$, which contradicts $f(t) = s \neq f(s)$.

The two strict inequalities are both impossible, and equality $f(s) = s$ is impossible because $s \in S$. This contradiction shows that no such order isomorphism can exist. □

Remark 11.2.6 (A first sign of transfinite rigidity). Infinite sets can be equinumerous with proper subsets of themselves, as Chapter 8 emphasized. Proposition 11.2.5 shows that well-orders behave very differently. Their order type is too rigid to fit into a proper beginning segment of itself.

Comparing two well-orders

The next theorem is one of the basic structural results of ordinal thinking. Any two well-ordered sets are comparable by initial segment. There is never a genuinely tangled situation in which each order has a piece that looks unlike anything in the other.

Theorem 11.2.7 (Comparability of well-ordered sets)**.** *Let $(A, \leq_A)$ and $(B, \leq_B)$ be well-ordered sets. Then one of them is order-isomorphic to an initial segment of the other.*

Proof. For each $a \in A$ and $b \in B$, let

$$A_{<a} = \{x \in A \mid x <_A a\}, \qquad B_{<b} = \{y \in B \mid y <_B b\}.$$

Consider the relation $R \subseteq A \times B$ defined by

$$(a, b) \in R \qquad \Longleftrightarrow \qquad A_{<a} \text{ and } B_{<b} \text{ are order-isomorphic.}$$

We will show that R is the graph of an order isomorphism between an initial segment of A and an initial segment of B.

Step 1: for each a, there is at most one b with $(a, b) \in R$, and vice versa.

Suppose $(a, b) \in R$ and $(a, b') \in R$. Then $B_{<b}$ and $B_{<b'}$ are order-isomorphic. If $b < b'$, then $B_{<b}$ is a proper initial segment of $B_{<b'}$, contradicting Proposition 11.2.5. The case $b' < b$ is symmetric. Therefore $b = b'$. The same argument, with A and B exchanged, shows that for each b there is at most one a with $(a, b) \in R$.

Thus R is the graph of a partial function from A to B. We write

$$f(a) = b \qquad \text{whenever} \qquad (a, b) \in R.$$

Let

$$D = \operatorname{dom}(f) \qquad \text{and} \qquad E = \operatorname{ran}(f).$$

Step 2: D is an initial segment of A, and E is an initial segment of B.

Let $a \in D$, say $f(a) = b$, and let $x <_A a$. Choose an order isomorphism

$$h \colon A_{<a} \to B_{<b}.$$

Because $x \in A_{<a}$, the value $h(x)$ lies in $B_{<b}$. By Proposition 11.1.6, the restriction of h maps $A_{<x}$ order-isomorphically onto $B_{<h(x)}$. Thus $(x, h(x)) \in R$, so $x \in D$. Therefore D is an initial segment of A.

By symmetry, E is an initial segment of B.

Step 3: $f \colon D \to E$ is an order isomorphism.

Let $a, a' \in D$ with $a <_A a'$. Put $f(a) = b$ and $f(a') = b'$. Choose an order isomorphism

$$h\colon A_{<a'} \to B_{<b'}.$$

Since $a \in A_{<a'}$, let $c = h(a)$. Again by Proposition 11.1.6, the initial segments $A_{<a}$ and $B_{<c}$ are order-isomorphic. Thus $(a, c) \in R$. By uniqueness from Step 1, we must have $c = f(a) = b$. Because $c \in B_{<b'}$, this gives $b <_B b'$.

So f is order-preserving. Since the inverse relation of R satisfies the same construction with A and B interchanged, $f^{-1}\colon E \to D$ is also order-preserving. Hence f is an order isomorphism from D to E.

Step 4: one of D or E is the whole ambient set.

Suppose, toward a contradiction, that both $D \subsetneq A$ and $E \subsetneq B$ are proper initial segments. By Proposition 11.2.3, there exist unique points $a_0 \in A$ and $b_0 \in B$ such that

$$D = A_{<a_0} \qquad \text{and} \qquad E = B_{<b_0}.$$

But Step 3 tells us that D and E are order-isomorphic, so $A_{<a_0}$ and $B_{<b_0}$ are order-isomorphic. Hence $(a_0, b_0) \in R$, which means $a_0 \in D$ and $b_0 \in E$. That contradicts the definitions of a_0 and b_0.

Therefore at least one of the initial segments D and E is the whole of its ambient well-ordered set. If $D = A$, then A is order-isomorphic to the initial segment E of B. If $E = B$, then B is order-isomorphic to the initial segment D of A. This is exactly the desired conclusion. □

Corollary 11.2.8 (Trichotomy of well-ordered types)**.** *Let $(A, \leq_A)$ and $(B, \leq_B)$ be well-ordered sets. Then exactly one of the following holds:*

(i) A and B are order-isomorphic;

(ii) A is order-isomorphic to a proper initial segment of B;

(iii) B is order-isomorphic to a proper initial segment of A.

Proof. Existence follows from Theorem 11.2.7. Two of the alternatives cannot hold at once, because by Proposition 11.2.5 no well-ordered set is isomorphic to a proper initial segment of itself. □

Remark 11.2.9 (The hidden order among all well-orders). Corollary 11.2.8 says that the order types of well-ordered sets are themselves arranged in a linear order by initial-segment comparison. This is the conceptual bridge from well-orders to ordinals. Instead of dealing with arbitrary equivalence classes of well-ordered sets, we are about to replace them with actual set-theoretic objects that already come arranged in that same linear hierarchy.

11.3 Ordinals as Transitive Well-Ordered Sets

The previous section showed that well-ordered sets are comparable in an extremely rigid way. We now turn that order-theoretic picture into a set-theoretic one.

The guiding idea is simple enough to state in one sentence:

> *An ordinal should be a well-ordered set whose elements are exactly the ordinals that come before it.*

This is the von Neumann viewpoint. It is one of the cleanest ideas in all of set theory, because it makes order and membership coincide.

The definition

Definition 11.3.1 (Ordinal)**.** A set α is called an *ordinal* if the following two conditions hold:

(i) α is transitive in the sense of Definition 6.2.3;

(ii) the membership relation $\in$ well-orders α, that is, among elements of α it is a strict total order and every nonempty subset of α has an $\in$-least element.

The second condition means that if $x, y \in \alpha$, then exactly one of $x \in y$, $x = y$, or $y \in x$ holds, and every nonempty subset of α has a first element with respect to membership. The first condition ensures that whenever an ordinal contains something, it also contains everything below that thing.

Example 11.3.2 (The first few ordinals)**.** The empty set $0 = \varnothing$ is an ordinal. It is transitive, and there are no elements to check for well-ordering.

Next,

$$1 = \{0\}$$

is an ordinal: it is transitive, and its only nonempty subset is $\{0\}$, whose least element is 0.

Similarly,

$$2 = \{0, 1\} \qquad \text{and} \qquad 3 = \{0, 1, 2\}$$

are ordinals. In each case membership is exactly the usual order of the previous numerals:

$$0 \in 1 \in 2 \in 3.$$

The example shows the pattern, but the real strength of the definition comes from the general consequences below.

Proposition 11.3.3 (Elements of ordinals are ordinals)**.** *Let α be an ordinal, and let $\beta \in \alpha$. Then β is also an ordinal.*

Proof. We first show that β is transitive. Let $x \in y \in \beta$. Since $\beta \in \alpha$ and α is transitive, we have $y \in \alpha$. Because α is transitive again, the relation $x \in y \in \alpha$ implies $x \in \alpha$. Now x, y, β are all elements of the well-ordered set α, and in that ordered set the relation $\in$ is transitive. From $x \in y$ and $y \in \beta$ we obtain $x \in \beta$. So β is transitive.

Next we show that membership well-orders β. Let $B \subseteq \beta$ be nonempty. Then B is also a nonempty subset of α, so it has an $\in$-least element because α is well-ordered by membership. That least element lies in β, so it is also the $\in$-least element of B inside β. The linearity of membership on β is inherited from its linearity on α.

Therefore β is an ordinal. □

Proposition 11.3.4 (An element is exactly its predecessor segment)**.** *Let α be an ordinal and let $\beta \in \alpha$. Then the initial segment of α lying below β is exactly β:*

$$\{\gamma \in \alpha \mid \gamma \in \beta\} = \beta.$$

Proof. Because α is transitive and $\beta \in \alpha$, every element of β belongs to α. Thus

$$\beta \subseteq \{\gamma \in \alpha \mid \gamma \in \beta\}.$$

The reverse inclusion is immediate from the defining condition $\gamma \in \beta$. So the two sets are equal. □

This proposition is where the von Neumann picture truly becomes visible: in an ordinal, each element is literally the set of all earlier ordinals.

Corollary 11.3.5 (Proper initial segments of an ordinal are its elements)**.** *Let α be an ordinal. Every proper initial segment of $(\alpha, \in)$ is equal to some $\beta \in \alpha$.*

Proof. By Proposition 11.2.3, every proper initial segment of the well-ordered set $(\alpha, \in)$ is the segment below some element $\beta \in \alpha$. By Proposition 11.3.4, that segment is exactly β. □

Ordinals are rigid

For arbitrary well-ordered sets, order isomorphism is a relation between different objects. For ordinals something much stronger happens: order-isomorphic ordinals are not merely similar, they are equal.

Theorem 11.3.6 (Any isomorphism between ordinals is the identity)**.** *Let α and β be ordinals, and let*

$$f : \alpha \to \beta$$

be an order isomorphism with respect to membership. Then $\alpha = \beta$ and f is the identity map.

Proof. Suppose, toward a contradiction, that f is not the identity. Then the set

$$S = \{\xi \in \alpha \mid f(\xi) \neq \xi\}$$

is nonempty. Since α is well-ordered by membership, S has an $\in$-least element; call it ξ.

For every $\eta \in \xi$, we have $\eta \notin S$, so $f(\eta) = \eta$. Because f is an order isomorphism, Proposition 11.1.6 shows that f maps the initial segment below ξ onto the initial segment below $f(\xi)$. In ordinal language,

$$f[\xi] = f(\xi).$$

But the left-hand side is

$$f[\xi] = \{f(\eta) \mid \eta \in \xi\} = \{\eta \mid \eta \in \xi\} = \xi,$$

by the choice of ξ. Therefore $f(\xi) = \xi$, contradicting the fact that $\xi \in S$.

So f is the identity on α. Since its range is β, we must also have $\alpha = \beta$. □

Corollary 11.3.7 (Order-isomorphic ordinals are equal)**.** *If two ordinals are order-isomorphic, then they are equal.*

Proof. Any order isomorphism between them is the identity by Theorem 11.3.6. □

Remark 11.3.8 (Canonical order types). Theorem 11.3.6 is the reason ordinals can serve as canonical representatives of well-ordered types. Once we know that a well-ordered set W is order-isomorphic to some ordinal, that ordinal is automatically unique. It is then natural

to write $\operatorname{otp}(W)$ for this ordinal. The general existence theorem that every well-ordered set is isomorphic to a unique ordinal is most naturally proved by transfinite recursion, so we postpone that theorem until Chapter 12.

Trichotomy and comparison among ordinals

The comparability theorem for well-ordered sets becomes especially sharp for ordinals, because proper initial segments of an ordinal are elements of that ordinal.

Theorem 11.3.9 (Trichotomy of ordinals)**.** *Let α and β be ordinals. Then exactly one of the following holds:*

$$\alpha \in \beta, \qquad \alpha = \beta, \qquad \beta \in \alpha.$$

Proof. By Theorem 11.2.7, one of the well-ordered sets $(\alpha, \in)$ and $(\beta, \in)$ is order-isomorphic to an initial segment of the other.

If $(\alpha, \in)$ is order-isomorphic to $(\beta, \in)$, then $\alpha = \beta$ by Corollary 11.3.7.

If $(\alpha, \in)$ is order-isomorphic to a proper initial segment of $(\beta, \in)$, then by Corollary 11.3.5 that initial segment is some element $\gamma \in \beta$. Since α and γ are order-isomorphic ordinals, they are equal. Hence $\alpha = \gamma \in \beta$.

The remaining case similarly yields $\beta \in \alpha$.

These alternatives are mutually exclusive. Equality excludes the other two. Also, if $\alpha \in \beta$ and $\beta \in \alpha$, then the ordinals α and β would be distinct elements of the well-ordered set $(\alpha, \in)$ or $(\beta, \in)$ each lying below the other, which is impossible in a strict order. □

Corollary 11.3.10 (No ordinal belongs to itself)**.** *If α is an ordinal, then*

$$\alpha \notin \alpha.$$

Proof. If $\alpha \in \alpha$, then α would be an element of the well-ordered set $(\alpha, \in)$ that is strictly below itself, which is impossible. □

Definition 11.3.11 (Order on the ordinals)**.** For ordinals α and β, we write

$$\alpha < \beta \qquad \text{to mean} \qquad \alpha \in \beta,$$

and

$$\alpha \leq \beta \qquad \text{to mean} \qquad \alpha \in \beta \text{ or } \alpha = \beta.$$

By Theorem 11.3.9, this really is a linear order. The notation finally matches ordinary intuition: ordinals smaller than β are exactly the elements of β.

Corollary 11.3.12 (Membership and inclusion agree on ordinals)**.** *For ordinals α and β, the following are equivalent:*

(i) $\alpha < \beta$;

(ii) $\alpha \subsetneq \beta$.

Consequently,

$$\alpha \leq \beta \qquad \Longleftrightarrow \qquad \alpha \subseteq \beta.$$

Proof. Assume first that $\alpha < \beta$, that is, $\alpha \in \beta$. Since β is transitive, every element of α belongs to β, so $\alpha \subseteq \beta$. Also $\alpha \neq \beta$ because no ordinal belongs to itself by Corollary 11.3.10. Thus $\alpha \subsetneq \beta$.

Conversely, suppose $\alpha \subsetneq \beta$. By Theorem 11.3.9, exactly one of $\alpha < \beta$, $\alpha = \beta$, or $\beta < \alpha$ holds. Equality is impossible because the inclusion is proper. If $\beta < \alpha$, then $\beta \in \alpha$, and since $\alpha \subseteq \beta$, this would imply $\beta \in \beta$, contradicting Corollary 11.3.10. Hence $\alpha < \beta$.

The final equivalence follows immediately. □

Corollary 11.3.13 (An ordinal is the set of all smaller ordinals)**.** *For every ordinal α,*

$$\alpha = \{\beta \mid \beta \text{ is an ordinal and } \beta < \alpha\}.$$

Proof. If $\beta \in \alpha$, then β is an ordinal by Proposition 11.3.3, and $\beta < \alpha$ by Definition 11.3.11. Conversely, if β is an ordinal and $\beta < \alpha$, then by definition $\beta \in \alpha$. □

Corollary 11.3.14 (Sets of ordinals are well-ordered by membership)**.** *Let A be a set of ordinals. Then membership linearly orders A, and every nonempty subset of A has an $\in$-least element.*

Proof. Linearity follows from Theorem 11.3.9.

Let $B \subseteq A$ be nonempty. Choose some $\beta \in B$. If $B \cap \beta = \varnothing$, then no element of B lies below β, so β is the $\in$-least element of B.

If $B \cap \beta \neq \varnothing$, then because β is an ordinal, the nonempty subset $B \cap \beta$ of β has an $\in$-least element; call it γ. We claim that γ is least in all of B. Let $\delta \in B$. By Theorem 11.3.9, either $\delta \in \beta$, $\delta = \beta$, or $\beta \in \delta$.

If $\delta \in \beta$, then $\delta \in B \cap \beta$, so by the choice of γ we have $\gamma \leq \delta$. If $\delta = \beta$, then $\gamma \in \beta = \delta$, so again $\gamma < \delta$. If $\beta \in \delta$, then $\gamma \in \beta \in \delta$, and because membership is transitive inside the ordinal δ, we again have $\gamma \in \delta$, that is, $\gamma < \delta$. Thus $\gamma \leq \delta$ for every $\delta \in B$, so γ is the least element of B. □

Remark 11.3.15 (Why the definition is so efficient). The ordinals are now doing several jobs at once. They are sets, well-orders, and canonical order types. The same relation $\in$ records membership, "comes before," and "is smaller than." That is what makes the theory so economical. Later, when we define operations and recursion on ordinals, this economy will become even more visible.

11.4 Finite Ordinals, ω, Successor Ordinals, and Limit Ordinals

The ordinals begin exactly where our earlier construction of the natural numbers began. The objects $0, 1, 2, 3, \ldots$ from Chapter 6 were already hinting that the finite natural numbers should be the first ordinals. We now make that identification explicit and then take the first step beyond the finite.

Successors and the finite ordinals

The set-theoretic successor operation from Definition 6.2.1 works perfectly well for ordinals.

Proposition 11.4.1 (The successor of an ordinal is an ordinal)**.** *If α is an ordinal, then its successor*

$$S(\alpha) = \alpha \cup \{\alpha\}$$

is also an ordinal.

Proof. We first show that $S(\alpha)$ is transitive. Let $x \in y \in S(\alpha)$. There are two cases.

If $y \in \alpha$, then $x \in \alpha$ because α is transitive. Hence $x \in S(\alpha)$.

If $y = \alpha$, then again $x \in \alpha \subseteq S(\alpha)$. Thus $S(\alpha)$ is transitive.

Next we show that membership well-orders $S(\alpha)$. Let $B \subseteq S(\alpha)$ be nonempty. If $B \cap \alpha$ is nonempty, then because α is an ordinal, the set $B \cap \alpha$ has an $\in$-least element b_0. This b_0 is also least in all of B, because any element of B outside α must be α itself, and every element of α lies below α.

If $B \cap \alpha = \varnothing$, then B can only be the singleton $\{\alpha\}$, so α is its least element.

Finally, membership is a strict total order on $S(\alpha)$: any two distinct elements inside α are comparable because α is an ordinal, and every element of α lies below the new point α.

Therefore $S(\alpha)$ is an ordinal. □

Corollary 11.4.2 (Every natural number is an ordinal)**.** *Every $n \in \mathbb{N}_0$ is an ordinal.*

Proof. We use induction on n; see Theorem 6.3.1. The base case is $0 = \varnothing$, which is an ordinal by Example 11.3.2. For the inductive step, assume that n is an ordinal. Then $S(n) = n \cup \{n\}$ is an ordinal by Proposition 11.4.1. Therefore every natural number in $\mathbb{N}_0$ is an ordinal. □

Definition 11.4.3 (Finite ordinal)**.** An ordinal is called *finite* if it belongs to $\mathbb{N}_0$, equivalently if it is one of the ordinals

$$0, 1, 2, 3, \dots$$

built from 0 by finitely many successor steps.

So the natural numbers constructed in Chapter 6 are not merely like finite ordinals; they literally are the finite ordinals.

Remark 11.4.4 (A small notational shift). Earlier chapters used $\mathbb{N}$ for the positive integers and $\mathbb{N}_0$ for the natural numbers including 0. In ordinal theory, starting at 0 is much more natural than starting at 1, because 0 is the first stage of the successor construction. This is why the finite ordinals line up with $\mathbb{N}_0$, not with $\mathbb{N}$.

The first infinite ordinal

Definition 11.4.5 (ω)**.** We write

$$\omega = \mathbb{N}_0 = \{0, 1, 2, 3, \dots\}$$

and call it *omega* or the *first infinite ordinal*.

Proposition 11.4.6 (ω is an ordinal)**.** *The set ω is an ordinal.*

Proof. We first show that ω is transitive. Let $m \in n \in \omega$. Since $n \in \mathbb{N}_0$, Corollary 6.3.4 shows that $m \in \mathbb{N}_0$ as well. Hence $m \in \omega$.

Next we show that membership well-orders ω. Every element of ω is an ordinal by Corollary 11.4.2. Therefore Corollary 11.3.14 implies that every nonempty subset of ω has an $\in$-least element and that membership linearly orders ω.

Thus ω is an ordinal. □

Example 11.4.7 (What ω looks like)**.** The ordinal ω is the infinite well-order

$$0 \in 1 \in 2 \in 3 \in \cdots$$

with no last element. As a set it contains every finite ordinal and nothing else. Any nonempty subset of ω still has a first member, exactly as the least-element principle in Chapter 6 promised.

Proposition 11.4.8 (ω is the least infinite ordinal)**.** *Every ordinal $\alpha < \omega$ is finite. In particular, ω is the smallest ordinal that is not finite.*

Proof. If $\alpha < \omega$, then by Definition 11.3.11 we have $\alpha \in \omega = \mathbb{N}_0$. So α is a natural number, and hence a finite ordinal by Definition 11.4.3. Therefore no smaller ordinal than ω is infinite. Since ω contains all finite ordinals, it is the first ordinal beyond the finite ones. □

Successor and limit ordinals

Definition 11.4.9 (Successor ordinal and limit ordinal)**.** An ordinal β is called a *successor ordinal* if there exists an ordinal α such that

$$\beta = S(\alpha) = \alpha \cup \{\alpha\}.$$

An ordinal λ is called a *limit ordinal* if

(i) $\lambda \neq 0$, and

(ii) λ is not a successor ordinal.

So every nonzero ordinal is either a successor or a limit. The successor ordinals are obtained by adjoining one new last point. The limit ordinals are the stages that have no immediate predecessor.

Example 11.4.10 (First successor and limit ordinals)**.** (i) Every positive finite ordinal is a successor ordinal:

$$1 = S(0), \qquad 2 = S(1), \qquad 3 = S(2), \quad \ldots$$

(ii) The ordinal ω is not a successor. Intuitively, no finite stage sits immediately before all finite stages at once. Therefore ω is a limit ordinal.

(iii) The successor $S(\omega) = \omega \cup \{\omega\}$ is a new ordinal strictly larger than ω. It looks like the usual infinite sequence of finite ordinals followed by one extra last point. In the next chapter, when ordinal addition is introduced, it will be natural to denote this ordinal by $\omega + 1$.

The difference between successor and limit ordinals is visible in the presence or absence of a greatest element.

Proposition 11.4.11 (Limit ordinals are exactly the nonzero ordinals with no greatest element)**.** *Let λ be an ordinal. Then λ is a limit ordinal if and only if $\lambda \neq 0$ and λ has no greatest element.*

Proof. Suppose first that λ is a successor ordinal, say $\lambda = S(\alpha)$. Then $\alpha \in \lambda$, and every element of λ is either α itself or lies in α. Hence α is the greatest element of λ.

Conversely, suppose $\lambda \neq 0$ and that λ has a greatest element α. We claim that $\lambda = S(\alpha)$. Because $\alpha \in \lambda$ and λ is transitive, we have $\alpha \subseteq \lambda$. Therefore $S(\alpha) = \alpha \cup \{\alpha\} \subseteq \lambda$.

For the reverse inclusion, let $x \in \lambda$. Since α is the greatest element of λ, either $x = \alpha$ or $x \in \alpha$. Thus $x \in S(\alpha)$. So $\lambda \subseteq S(\alpha)$, and hence $\lambda = S(\alpha)$.

We have proved that a nonzero ordinal is a successor exactly when it has a greatest element. Therefore a nonzero ordinal is a limit ordinal exactly when it has no greatest element. □

Corollary 11.4.12 (ω is a limit ordinal)**.** *The ordinal ω is a limit ordinal.*

Proof. The ordinal ω is nonzero because $0 \in \omega$. It has no greatest element: if $n \in \omega$, then $S(n)$ is also in ω and satisfies $n < S(n)$. Therefore Proposition 11.4.11 shows that ω is a limit ordinal. □

Remark 11.4.13 (Successor stages versus accumulation stages). It is helpful to picture successor ordinals and limit ordinals as two kinds of stage. At a successor stage, one more point is appended to an earlier well-order. At a limit stage, nothing new appears "all at once" as a last element; instead the ordinal is built from everything that came earlier without ending in a greatest point. This successor/ limit distinction will guide nearly every transfinite definition in the next chapter.

11.5 Suprema of Sets of Ordinals

The ordinals form an endlessly increasing hierarchy, but not a chaotic one. Given a set of ordinals, we can gather together everything that appears in any of them simply by taking a union. Remarkably, the result is again an ordinal, and in fact it is the least ordinal lying above all of them.

This is the first genuinely transfinite closure property of the ordinal world.

Unions of ordinals

Proposition 11.5.1 (The union of a set of ordinals is an ordinal)**.** *Let A be a set of ordinals. Then*

$$U = \bigcup A$$

is an ordinal.

Proof. We first show that U is transitive. Let $x \in y \in U$. Since $y \in U$, there exists an ordinal $\alpha \in A$ with $y \in \alpha$. Because α is transitive and $x \in y$, we have $x \in \alpha$. Therefore $x \in U$. So U is transitive.

Next we show that membership well-orders U. Let $B \subseteq U$ be nonempty. Every element of B is an element of some ordinal in A, and therefore is itself an ordinal by Proposition 11.3.3. Thus B is a nonempty set of ordinals. By Corollary 11.3.14, membership linearly orders B and gives it an $\in$-least element. Since B was an arbitrary nonempty subset of U, membership well-orders U.

Therefore U is an ordinal. □

Definition 11.5.2 (Upper bound and supremum for ordinals)**.** Let A be a set of ordinals.

(i) An ordinal β is an *upper bound* for A if $\alpha \leq \beta$ for every $\alpha \in A$.

(ii) An ordinal σ is the *supremum* or *least upper bound* of A if σ is an upper bound for A, and whenever β is any upper bound for A, we have $\sigma \leq \beta$.

The next theorem shows that unions and suprema are the same thing for sets of ordinals.

Theorem 11.5.3 (The union is the supremum)**.** *Let A be a set of ordinals. Then $\bigcup A$ is the supremum of A. In symbols,*

$$\sup(A) = \bigcup A.$$

In particular, when $A = \varnothing$, we have

$$\sup(\varnothing) = 0.$$

Proof. By Proposition 11.5.1, the union $U = \bigcup A$ is an ordinal.

We first show that U is an upper bound for A. Let $\alpha \in A$. Since every element of α belongs to the union U, we have $\alpha \subseteq U$. By Corollary 11.3.12, this implies $\alpha \leq U$.

Now let β be any upper bound for A. Then $\alpha \leq \beta$ for every $\alpha \in A$, so again by Corollary 11.3.12, each $\alpha \in A$ satisfies $\alpha \subseteq \beta$. Therefore every element of every $\alpha \in A$ belongs to β, which means that

$$U = \bigcup A \subseteq \beta.$$

Applying Corollary 11.3.12 once more, we obtain $U \leq \beta$. Thus U is the least upper bound of A.

If $A = \varnothing$, then $\bigcup A = \varnothing = 0$, so $\sup(\varnothing) = 0$. □

Example 11.5.4 (Basic suprema)**.** (i) For the finite set $A = \{1, 4, 7\}$,

$$\sup(A) = 7.$$

For a finite set of ordinals, the supremum is simply the largest member.

(ii) Let

$$A = \{0, 2, 4, 6, \dots\}$$

be the set of even finite ordinals. Then

$$\sup(A) = \omega.$$

No finite ordinal can dominate all even finite ordinals, but ω lies above all of them.

(iii) The set of all finite ordinals has supremum ω:

$$\sup(\omega) = \omega.$$

This is the first example in which a set of ordinals has a supremum that is not already a largest member of the set.

Proposition 11.5.5 (When the supremum is already present)**.** *Let A be a nonempty set of ordinals.*

(i) If A has a largest element δ, then $\sup(A) = \delta$.

(ii) If A has no largest element, then $\sup(A)$ is a limit ordinal.

Proof. For (i), if δ is largest in A, then every $\alpha \in A$ satisfies $\alpha \leq \delta$, so δ is an upper bound. Since $\delta \in A$, every upper bound for A must be at least δ. Therefore δ is the least upper bound, that is, $\sup(A) = \delta$.

For (ii), let $\lambda = \sup(A) = \bigcup A$. Since A has no largest element, it is nonempty and $\lambda \neq 0$. We show that λ has no greatest element.

Let $\beta \in \lambda$. Then $\beta \in \alpha$ for some $\alpha \in A$. Because A has no largest element, there exists $\gamma \in A$ with $\alpha < \gamma$, that is, $\alpha \in \gamma$. Since γ is transitive and $\beta \in \alpha \in \gamma$, we have $\beta \in \gamma$. Also $\alpha \in \gamma \subseteq \lambda$, so $\alpha \in \lambda$ and $\beta < \alpha$. Thus every element $\beta \in \lambda$ has a strictly larger element $\alpha \in \lambda$ above it.

Therefore λ has no greatest element. By Proposition 11.4.11, λ is a limit ordinal. □

There is no set of all ordinals

The supremum theorem has an important conceptual consequence. However far up the ordinal hierarchy a set takes us, there is always a strictly larger ordinal beyond it.

Theorem 11.5.6 (There is no set of all ordinals)**.** *For every set A of ordinals, there exists an ordinal strictly larger than every member of A. Consequently, there is no set whose members are exactly all ordinals.*

Proof. Let A be a set of ordinals, and put

$$\sigma = \sup(A) = \bigcup A.$$

Then σ is an ordinal by Proposition 11.5.1. Its successor

$$S(\sigma) = \sigma \cup \{\sigma\}$$

is also an ordinal by Proposition 11.4.1.

For each $\alpha \in A$, Theorem 11.5.3 gives $\alpha \leq \sigma$, and therefore $\alpha < S(\sigma)$. So $S(\sigma)$ is strictly larger than every member of A.

If there were a set containing all ordinals, then applying the previous paragraph to that set would produce an ordinal larger than every ordinal in it, which is impossible. Hence there is no set of all ordinals. □

Remark 11.5.7 (The collection Ord). One often writes Ord for the total collection of all ordinals. By Theorem 11.5.6, this collection is too large to be a set. In more formal language, it is a *proper class*. We will postpone a systematic discussion of classes until the later chapters on axioms and foundations, but it is already worth keeping the picture in mind: the ordinal hierarchy never stops.

Remark 11.5.8 (A first glimpse of Burali-Forti's idea). Theorem 11.5.6 is a beginner-friendly form of a classical idea often associated with Burali-Forti. If one naively tried to gather all ordinals into a single set, then the supremum of that set would itself be an ordinal larger than every member of the set. The hierarchy defeats any attempt to package all of it into one further ordinal set. See Potter [14] or Jech [10] for fuller discussions.

Looking ahead

This chapter turned well-ordering into transfinite number. We began by asking when two ordered sets should count as the same from the order-theoretic point of view, and we introduced order isomorphism and order type. The crucial structural result was the comparability of well-ordered sets: one always fits as an initial segment of the other. That rigidity prepared the way for the von Neumann definition of an ordinal.

Once ordinals were introduced as transitive sets well-ordered by membership, several remarkable simplifications appeared. Elements of an ordinal turned out to be ordinals themselves, order-isomorphic ordinals turned out to be literally equal, and the three relations "is smaller than," "is an element of," and "is an initial segment of" collapsed into a single picture. We then identified the finite ordinals with the natural numbers, introduced ω as the first infinite ordinal, and distinguished successor ordinals from limit ordinals.

The final section showed that sets of ordinals have canonical suprema, given simply by unions, and that the ordinal hierarchy climbs forever: there is no largest ordinal and no set of all ordinals. The next chapter explains how to use this hierarchy as a domain of definition. There we will prove transfinite induction, develop transfinite recursion, and use ordinals not just as objects to compare, but as the stages along which genuinely new mathematics is built.

Chapter 12

Transfinite Induction, Recursion, and Ordinal Arithmetic

In Chapter 6 we learned two methods that are so familiar on the natural numbers that they can easily seem almost mechanical: ordinary induction and recursive definition. Induction says that if a property holds at the beginning and survives passage from one natural number to the next, then it holds forever. Recursion says that if we know how to specify the value at the beginning and how to pass from one stage to the next, then we obtain a function defined on all of $\mathbb{N}_0$.

Chapter 11 showed that $\mathbb{N}_0$ is only the first part of a much larger ordinal hierarchy. Beyond the finite ordinals lie ω, $\omega+1$, $\omega+2$, and far more exotic stages. Once that hierarchy is available, the natural question is whether the methods from Chapter 6 also extend beyond the finite. The answer is yes, but with an important twist. On the natural numbers we move only from n to $n+1$. On the ordinals we must also understand *limit stages*, where there is no immediate predecessor and the next value depends on an entire earlier history.

That is why the present chapter is one of the conceptual turning points of the book. Ordinals stop being merely objects that can be compared, and become the stages along which mathematics can be built. We begin with transfinite induction, which is the correct extension of ordinary induction from $\mathbb{N}_0$ to arbitrary ordinals. We then prove the transfinite recursion theorem, which allows functions to be defined one stage at a time along any well-order.

Once these methods are available, ordinal arithmetic becomes natural. Ordinal addition, multiplication, and exponentiation are not designed to measure size; they are designed to measure *ordered shape*. For finite ordinals they agree with the ordinary arithmetic of Chapter 6. For infinite ordinals, however, they reveal new phenomena. The simple-looking equalities

$$1+\omega=\omega \qquad \text{and} \qquad \omega+1>\omega$$

show at once that infinite order behaves very differently from infinite cardinality.

The chapter closes with countable ordinals and the first uncountable ordinal ω_1. This is the bridge from the transfinite ordering of sets to the next chapter's systematic study of cardinality. The basic message is worth keeping in mind from the start:

> *Ordinals are not only objects to be classified; they are stages of construction, and arithmetic on them records the geometry of those stages.*

12.1 Transfinite Induction

Ordinary induction on $\mathbb{N}_0$ rests on one decisive fact: every nonempty subset of $\mathbb{N}_0$ has a least element. The same idea works for every ordinal, because every ordinal is well-ordered by membership. So the real engine behind induction is not the special form of the natural numbers, but the general principle of *least counterexample*. Transfinite induction is simply that principle applied on arbitrary ordinals.

The least-counterexample principle beyond the finite

Theorem 12.1.1 (Transfinite induction below a fixed ordinal)**.** *Let θ be an ordinal, and let $A \subseteq \theta$. Suppose that for every $\alpha < \theta$,*

$$\big(\forall \beta < \alpha,\ \beta \in A\big) \Longrightarrow \alpha \in A.$$

Then $A = \theta$.

Proof. Assume for contradiction that $A \neq \theta$. Then the set

$$B = \theta \setminus A$$

is nonempty. Because θ is an ordinal, every element of θ is itself an ordinal, so B is a set of ordinals. By Corollary 11.3.14, the set B is well-ordered by membership and therefore has a least element; call it α_0.

Since α_0 is least in B, no smaller ordinal belongs to B. In other words, if $\beta < \alpha_0$, then $\beta \notin B$, so $\beta \in A$. Thus every ordinal below α_0 lies in A.

Applying the hypothesis to α_0, we conclude that $\alpha_0 \in A$. But by construction $\alpha_0 \in B$, hence $\alpha_0 \notin A$. This contradiction shows that B must have been empty. Therefore $A = \theta$. □

The theorem is the exact analogue of induction on $\mathbb{N}_0$, except that the "earlier stages" below α may now form a large infinite initial segment rather than a single predecessor chain.

Corollary 12.1.2 (Global transfinite induction)**.** *Let $P(\alpha)$ be a statement about ordinals. Suppose that for every ordinal α,*

$$\big(\forall \beta < \alpha,\ P(\beta)\big) \Longrightarrow P(\alpha).$$

Then $P(\alpha)$ holds for every ordinal α.

Proof. Fix an arbitrary ordinal θ, and define

$$A = \{\alpha \in \theta \mid P(\alpha)\}.$$

If $\alpha < \theta$ and every $\beta < \alpha$ belongs to A, then $P(\beta)$ holds for every $\beta < \alpha$, so the hypothesis gives $P(\alpha)$. Hence $\alpha \in A$. By Theorem 12.1.1, we obtain $A = \theta$. So $P(\alpha)$ holds for every $\alpha < \theta$.

Because θ was arbitrary, $P(\alpha)$ holds for every ordinal α. □

The corollary is the form of transfinite induction most often used in practice. To prove a statement for every ordinal, it is enough to show that whenever the statement holds at all earlier stages, it also holds at the current stage.

Corollary 12.1.3 (Successor-limit form of transfinite induction)**.** *Let $P(\alpha)$ be a statement about ordinals. Assume that:*

(i) $P(0)$ holds;

(ii) whenever $P(\alpha)$ holds, $P(S(\alpha))$ also holds;

(iii) for every limit ordinal λ, if $P(\beta)$ holds for all $\beta < \lambda$, then $P(\lambda)$ holds.

Then $P(\alpha)$ holds for every ordinal α.

Proof. We verify the hypothesis of Corollary 12.1.2. Let α be an ordinal, and assume that $P(\beta)$ holds for every $\beta < \alpha$. We must show that $P(\alpha)$ holds.

If $\alpha = 0$, then this is part (i). If α is a successor, say $\alpha = S(\gamma)$, then $\gamma < \alpha$, so $P(\gamma)$ holds by hypothesis; part (ii) now gives $P(\alpha)$. If α is neither 0 nor a successor, then by Definition 11.4.9 it is a limit ordinal, and part (iii) applies directly.

So the hypothesis of Corollary 12.1.2 is satisfied, and the conclusion follows. □

This successor-limit form is often the most intuitive one, because it separates the genuinely new feature of transfinite reasoning: besides a base case and a successor step, one must also understand limit stages.

Example 12.1.4 (Ordinary induction is a special case)**.** Take $P(n)$ to be a statement about the finite ordinals $n \in \mathbb{N}_0$. If we apply Corollary 12.1.3 only below ω, then there are no nonzero limit ordinals below ω other than ω itself, which lies outside the range of the proof. So the transfinite principle reduces to the ordinary induction principle from Chapter 6. In this sense, ordinary induction is simply the first finite fragment of a much larger method.

Remark 12.1.5 (Why limit stages matter). On $\mathbb{N}_0$, every nonzero stage has an immediate predecessor. On the ordinals, ω does not. Neither does $\omega \cdot 2$, nor ω^2, nor any other limit ordinal. So a transfinite proof can never rely only on "take the previous stage and add one." At limit ordinals we must often look at the whole earlier pattern at once.

Remark 12.1.6 (A standard set-theoretic method). Transfinite induction is one of the basic tools of set theory, topology, logic, and parts of algebra. Standard discussions may be found in Moschovakis [7], Potter [14], Levy [12], Jech [10], and Kunen [11].

12.2 Transfinite Recursion

Induction proves that something is true at every stage of a well-order. Recursion does something more constructive: it defines an object at each stage. On $\mathbb{N}_0$, a recursive definition usually has the form "start with an initial value, then explain how to get the next value from the current one." On the ordinals, that is no longer enough, because limit ordinals do not have immediate predecessors. A genuinely transfinite recursive rule must be allowed to depend on the entire initial segment built so far.

This is the guiding pattern:

$$\text{value at stage } \alpha = \text{a rule applied to the function on all stages below } \alpha.$$

The transfinite recursion theorem says that such definitions are not only intuitive; they actually determine unique functions.

Approximation functions

Definition 12.2.1 (Approximation for a transfinite recursion)**.** Let θ be an ordinal, let X be a set, and let F be a rule that assigns to every function h whose domain is an ordinal $< \theta$ and whose values lie in X, an element $F(h) \in X$.

If $\beta \leq \theta$, a function g is called a *β-approximation* for the rule F if:

(i) $\operatorname{dom}(g) = \beta$;

(ii) $\operatorname{ran}(g) \subseteq X$;

(iii) for every $\gamma < \beta$,

$$g(\gamma) = F\big(g \restriction_\gamma\big).$$

So a β-approximation is a partial solution defined on the first β stages. The crucial point is that at each stage γ, the value $g(\gamma)$ is determined from the earlier part $g \restriction_\gamma$.

Lemma 12.2.2 (Restrictions of approximations are approximations)**.** *Let $\beta \leq \gamma \leq \theta$. If g is a γ-approximation for the rule F, then $g \restriction_\beta$ is a β-approximation for the same rule.*

Proof. Because g is a function with domain γ, its restriction $g \restriction_\beta$ is a function with domain β, and clearly its range still lies in X.

Now let $\delta < \beta$. Since $\delta < \gamma$ as well, the defining property of the γ-approximation g gives

$$g(\delta) = F\big(g \restriction_\delta\big).$$

But restricting first to β and then to δ gives the same function as restricting directly to δ. Thus

$$\big(g \restriction_\beta\big)(\delta) = g(\delta) = F\big(g \restriction_\delta\big) = F\Big(\big(g \restriction_\beta\big) \restriction_\delta\Big).$$

So $g \restriction_\beta$ satisfies the defining recursion at every stage $< \beta$. Hence it is a β-approximation. □

The lemma expresses the basic coherence that a stage-by-stage definition ought to have: if we have built a correct solution up to a later stage, then the earlier part must already have been a correct solution on its own.

Existence and uniqueness of recursively defined functions

Theorem 12.2.3 (Transfinite recursion theorem)**.** *Let θ be an ordinal, let X be a set, and let F be a rule that assigns to every function h whose domain is an ordinal $< \theta$ and whose values lie in X, an element $F(h) \in X$. Then there exists a unique function $g \colon \theta \to X$ such that for every $\beta < \theta$,*

$$g(\beta) = F\big(g \restriction_\beta\big).$$

Proof. We prove the stronger statement that for every ordinal $\beta \leq \theta$, there exists a unique β-approximation for the rule F. Once this is done, the case $\beta = \theta$ gives exactly the desired function.

We argue by transfinite induction on $\beta \leq \theta$.

Base case: $\beta = 0$. The empty function $\varnothing$ has domain 0 and range contained in X. Since there are no ordinals below 0, the recursive condition is vacuous. Thus $\varnothing$ is a 0-approximation. It is the only function with domain 0, so the 0-approximation is unique.

Successor step. Assume that for some $\delta < \theta$ there exists a unique δ-approximation, say g_δ. We construct a $S(\delta)$-approximation by adjoining one new ordered pair:

$$g = g_\delta \cup \{\langle \delta, F(g_\delta)\rangle\}.$$

Because $F(g_\delta) \in X$, this is a function whose domain is $S(\delta) = \delta \cup \{\delta\}$, and its range lies in X.

If $\gamma < \delta$, then $g \restriction_\gamma = g_\delta \restriction_\gamma$, so

$$g(\gamma) = g_\delta(\gamma) = F\big(g_\delta \restriction_\gamma\big) = F\big(g \restriction_\gamma\big).$$

At the new stage δ, the defining equation holds by construction:

$$g(\delta) = F(g_\delta) = F\big(g \restriction_\delta\big).$$

So g is indeed an $S(\delta)$-approximation.

To prove uniqueness, let h be any $S(\delta)$-approximation. By Lemma 12.2.2, the restriction $h \restriction_\delta$ is a δ-approximation. By the inductive uniqueness hypothesis,

$$h \restriction_\delta = g_\delta.$$

Therefore

$$h(\delta) = F\big(h \restriction_\delta\big) = F(g_\delta) = g(\delta).$$

Since h and g agree on δ and also at the point δ, they are equal. So the $S(\delta)$-approximation is unique.

Limit step. Let $\lambda \leq \theta$ be a nonzero limit ordinal, and assume that for every $\delta < \lambda$ there exists a unique δ-approximation g_δ.

We first show that these approximations fit together coherently. If $\delta < \varepsilon < \lambda$, then Lemma 12.2.2 says that $g_\varepsilon \restriction_\delta$ is a δ-approximation. By the uniqueness of the δ-approximation,

$$g_\varepsilon \restriction_\delta = g_\delta.$$

So later approximations extend earlier ones.

Now define

$$g = \bigcup_{\delta<\lambda} g_\delta.$$

Because the family is coherent, the union is again a function. Its domain is

$$\operatorname{dom}(g) = \bigcup_{\delta<\lambda} \operatorname{dom}(g_\delta) = \bigcup_{\delta<\lambda} \delta = \lambda,$$

where the last equality is Theorem 11.5.3 applied to the set of ordinals below λ. Its range lies in X, because each g_δ has range in X.

Let $\gamma < \lambda$. Since λ is a limit ordinal, $S(\gamma) < \lambda$. Choose $\delta = S(\gamma)$. Then $\gamma < \delta < \lambda$, and by coherence $g \restriction_\gamma = g_\delta \restriction_\gamma$. Therefore

$$g(\gamma) = g_\delta(\gamma) = F\big(g_\delta \restriction_\gamma\big) = F\big(g \restriction_\gamma\big).$$

So g is a λ-approximation.

For uniqueness, let h be another λ-approximation. For any $\delta < \lambda$, Lemma 12.2.2 shows that $h \restriction_\delta$ is a δ-approximation. By uniqueness,

$$h \restriction_\delta = g_\delta.$$

Hence every ordered pair of h belongs to some g_δ, and so $h \subseteq g$. Conversely, each g_δ equals $h \restriction_\delta$, so every ordered pair of every g_δ belongs to h. Thus $g \subseteq h$. Therefore $g = h$.

This completes the induction. In particular, there exists a unique θ-approximation. That is exactly a unique function $g : \theta \to X$ satisfying $g(\beta) = F(g \restriction_\beta)$ for all $\beta < \theta$. □

Corollary 12.2.4 (Coherence of bounded recursive constructions)**.** *Fix a recursive rule F as in Theorem 12.2.3. If $\theta < \eta$, and if $g_\theta : \theta \to X$ and $g_\eta : \eta \to X$ are the unique functions produced by the theorem on domains θ and η, then*

$$g_\eta \restriction_\theta = g_\theta.$$

Proof. By Lemma 12.2.2, the restriction $g_\eta \restriction_\theta$ is a θ-approximation for the same recursive rule. By the uniqueness part of Theorem 12.2.3, it must equal g_θ. □

The corollary is what allows us to speak informally of a single recursive construction "on all ordinals," even though the theorem itself is stated only for a fixed domain θ. The bounded solutions agree on overlaps, so they fit into one consistent global picture.

Example 12.2.5 (The identity function from recursion)**.** Fix an ordinal θ, let $X = \theta$, and define a rule F by

$$F(h) = \operatorname{dom}(h).$$

This makes sense because the domain of any admissible h is an ordinal $< \theta$, hence an element of θ.

The transfinite recursion theorem gives a unique function $g : \theta \to \theta$ such that

$$g(\beta) = \operatorname{dom}\big(g \restriction_\beta\big)$$

for all $\beta < \theta$. But $\operatorname{dom}(g \restriction_\beta) = \beta$, so the defining equation becomes

$$g(\beta) = \beta.$$

Thus the recursively defined function is simply the identity map on θ.

Remark 12.2.6 (Recursion and class-size bookkeeping). In a fully axiomatic development, global transfinite recursion is often phrased in the language of class functions. We have not yet developed that formal framework, and we deliberately postpone most foundational bookkeeping until the later chapter on axioms. For our current purposes, the bounded theorem and Corollary 12.2.4 contain the essential mathematics. They justify the recursive definitions that follow and explain why they are unambiguous.

12.3 Ordinal Addition and Multiplication

For finite ordinals, arithmetic is so familiar that it is easy to think of $+$ and $\cdot$ as operations of pure quantity. For ordinals, however, the primary issue is not quantity but order. Ordinal addition records what happens when one well-ordered block is placed after another. Ordinal

multiplication records what happens when one block is repeated in well-ordered many copies. Once infinite blocks enter the picture, the order in which pieces are arranged becomes decisive.

A first warning sign is already visible in the pair

$$1 + \omega \qquad \text{and} \qquad \omega + 1.$$

If we place one point in front of an ω-sequence, we still get a well-order of type ω. But if we place one point *after* an ω-sequence, we obtain a genuinely new last point. The same finite set of pieces can therefore yield different ordinal sums, depending on how the order is assembled.

Ordinal addition

Definition 12.3.1 (Ordinal addition)**.** Fix an ordinal α. By transfinite recursion on the second variable, we define the *ordinal sum* $\alpha + \beta$ by the clauses

$$\alpha + 0 = \alpha,$$

$$\alpha + S(\beta) = S(\alpha + \beta),$$

and for every nonzero limit ordinal λ,

$$\alpha + \lambda = \sup_{\beta<\lambda}(\alpha + \beta).$$

Intuitively, $\alpha + \beta$ is the order type obtained by placing a copy of β after a copy of α. The recursive definition is the set-theoretic way of encoding that picture.

Proposition 12.3.2 (Agreement with finite addition)**.** *If $m, n \in \mathbb{N}_0$ are viewed as finite ordinals, then the ordinal sum $m + n$ agrees with the addition on $\mathbb{N}_0$ defined in Definition 6.5.1.*

Proof. Fix a finite ordinal m. Chapter 6 defines a function $a_m : \mathbb{N}_0 \to \mathbb{N}_0$ by the recursion

$$a_m(0) = m, \qquad a_m(S(n)) = S(a_m(n)).$$

By Definition 12.3.1, the function $b_m : \mathbb{N}_0 \to \mathbb{N}_0$ given by $b_m(n) = m + n$ satisfies the same recursion. By the uniqueness part of the ordinary recursion Theorem 6.4.2, we must have $a_m = b_m$. Thus the finite ordinal sum agrees with the already defined natural-number sum. □

Example 12.3.3 (First ordinal sums)**.** (i) Because $\alpha + 1 = \alpha + S(0) = S(\alpha + 0) = S(\alpha)$, adding 1 on the right simply appends one new last point.

(ii) In particular,

$$\omega + 1 = S(\omega),$$

so $\omega + 1$ is strictly larger than ω.

(iii) On the other hand,

$$1 + \omega = \sup_{n<\omega}(1 + n).$$

By Proposition 12.3.2, the finite values $1 + n$ are just the ordinary successors $1, 2, 3, \ldots$. Their supremum is ω. Hence

$$1 + \omega = \omega.$$

(iv) The same reasoning shows that for every nonzero finite ordinal k,

$$k + \omega = \omega.$$

A finite block added in front of an ω-sequence does not change the eventual order type.

(v) By contrast,

$$\omega + 2 = S(S(\omega))$$

has two new points at the end, and

$$\omega + \omega = \sup_{n<\omega}(\omega + n)$$

is two copies of ω placed one after the other.

The last example is the first vivid sign that ordinal arithmetic records ordered position rather than size. The sets underlying ω, $\omega + 1$, and $\omega + \omega$ are all countably infinite, but their order types are different.

Proposition 12.3.4 (Addition is strictly increasing in the right variable)**.** *For every ordinal α, if $\beta < \gamma$, then*

$$\alpha + \beta < \alpha + \gamma.$$

Proof. Fix α. We prove by transfinite induction on γ that for every $\beta < \gamma$, one has $\alpha + \beta < \alpha + \gamma$.

If $\gamma = 0$, there is nothing to prove.

Suppose the statement holds for γ, and consider $S(\gamma)$. Let $\beta < S(\gamma)$. If $\beta = \gamma$, then

$$\alpha + \beta = \alpha + \gamma < S(\alpha + \gamma) = \alpha + S(\gamma).$$

If $\beta < \gamma$, then by the inductive hypothesis,

$$\alpha + \beta < \alpha + \gamma < \alpha + S(\gamma).$$

So the claim holds at the successor stage.

Now let λ be a limit ordinal, and assume the claim holds for all smaller ordinals. If $\beta < \lambda$, then $\alpha + \lambda$ is the supremum of the set $\{\alpha + \xi \mid \xi < \lambda\}$. In particular, $\alpha + \beta \leq \alpha + \lambda$. To see that the inequality is strict, note that $S(\beta) < \lambda$, so

$$\alpha + \beta < \alpha + S(\beta) \leq \alpha + \lambda.$$

Hence $\alpha + \beta < \alpha + \lambda$.

This completes the transfinite induction. □

Lemma 12.3.5 (Adding a limit on the right gives a limit)**.** *If λ is a nonzero limit ordinal, then for every ordinal β, the ordinal $\beta + \lambda$ is also a limit ordinal. Moreover, every ordinal below $\beta + \lambda$ is below some $\beta + \delta$ with $\delta < \lambda$.*

Proof. By Definition 12.3.1,

$$\beta + \lambda = \sup_{\delta<\lambda}(\beta + \delta).$$

Each $\beta + \delta$ is below this supremum, so the family $\{\beta + \delta \mid \delta < \lambda\}$ is bounded above by $\beta + \lambda$. By the defining property of supremum, $\beta + \lambda$ is the least such upper bound, so every $\xi < \beta + \lambda$ must lie below some $\beta + \delta$ with $\delta < \lambda$; otherwise ξ itself would be an upper bound smaller than the supremum.

If $\beta + \lambda$ were a successor ordinal, say $S(\eta)$, then $\eta < \beta + \lambda$. By the previous paragraph there would exist $\delta < \lambda$ such that $\eta < \beta + \delta$. But then

$$S(\eta) \leq \beta + \delta < \beta + \lambda = S(\eta),$$

which is impossible. Therefore $\beta + \lambda$ is not a successor. Since $\lambda \neq 0$, the ordinal $\beta + \lambda$ is nonzero, and so it is a limit ordinal. □

Proposition 12.3.6 (Ordinal addition is associative)**.** *For all ordinals α, β, γ,*

$$(\alpha + \beta) + \gamma = \alpha + (\beta + \gamma).$$

Proof. Fix ordinals α and β. We prove by transfinite induction on γ that

$$(\alpha + \beta) + \gamma = \alpha + (\beta + \gamma).$$

If $\gamma = 0$, then both sides equal $\alpha + \beta$.

Assume the identity holds for γ. Then

$$(\alpha + \beta) + S(\gamma) = S\big((\alpha + \beta) + \gamma\big) = S\big(\alpha + (\beta + \gamma)\big) = \alpha + S(\beta + \gamma) = \alpha + (\beta + S(\gamma)).$$

So the identity holds at successor stages.

Now let λ be a limit ordinal, and assume the identity holds for all $\gamma < \lambda$. By the definition of addition,

$$(\alpha + \beta) + \lambda = \sup_{\gamma<\lambda}\big((\alpha + \beta) + \gamma\big) = \sup_{\gamma<\lambda}\big(\alpha + (\beta + \gamma)\big).$$

By Lemma 12.3.5, the ordinal $\beta + \lambda$ is a limit ordinal and every ordinal below it is below some $\beta + \gamma$ with $\gamma < \lambda$. Since Proposition 12.3.4 tells us that the map $\xi \mapsto \alpha + \xi$ is increasing, the supremum of $\{\alpha + (\beta + \gamma) \mid \gamma < \lambda\}$ is exactly $\alpha + (\beta + \lambda)$. Therefore

$$(\alpha + \beta) + \lambda = \alpha + (\beta + \lambda).$$

This completes the induction. □

Remark 12.3.7 (Addition is not commutative). Example 12.3.3 already shows that

$$1 + \omega = \omega \qquad \text{but} \qquad \omega + 1 > \omega.$$

So ordinal addition is not commutative. This is one of the most useful first lessons of transfinite arithmetic: infinite order is sensitive to where the new pieces are attached.

Ordinal multiplication

Definition 12.3.8 (Ordinal multiplication)**.** Fix an ordinal α. By transfinite recursion on the second variable, we define the *ordinal product* $\alpha \cdot \beta$ by

$$\alpha \cdot 0 = 0,$$

$$\alpha \cdot S(\beta) = (\alpha \cdot \beta) + \alpha,$$

and for every nonzero limit ordinal λ,

$$\alpha \cdot \lambda = \sup_{\beta<\lambda}(\alpha \cdot \beta).$$

Intuitively, $\alpha \cdot \beta$ is obtained by arranging β copies of α in order. The second variable tells us how many blocks appear and in what order they are stacked.

Proposition 12.3.9 (Agreement with finite multiplication)**.** *If $m, n \in \mathbb{N}_0$ are viewed as finite ordinals, then the ordinal product $m \cdot n$ agrees with the multiplication on $\mathbb{N}_0$ defined in Definition 6.5.7.*

Proof. Fix a finite ordinal m. Chapter 6 defines a function $\mu_m : \mathbb{N}_0 \to \mathbb{N}_0$ by the recursion

$$\mu_m(0) = 0, \qquad \mu_m(S(n)) = \mu_m(n) + m.$$

By Definition 12.3.8, the function $\nu_m(n) = m \cdot n$ satisfies exactly the same recursion. By the uniqueness part of Theorem 6.4.2, we conclude that $\mu_m = \nu_m$. Thus finite ordinal multiplication agrees with the multiplication already constructed on $\mathbb{N}_0$. □

Example 12.3.10 (First ordinal products)**.** (i) For every ordinal α,

$$\alpha \cdot 1 = \alpha \cdot S(0) = (\alpha \cdot 0) + \alpha = \alpha.$$

(ii) We have

$$\omega \cdot 2 = \omega + \omega,$$

two copies of ω placed one after the other.

(iii) By contrast,

$$2 \cdot \omega = \sup_{n<\omega}(2 \cdot n).$$

By Proposition 12.3.9, the values $2 \cdot n$ are the even finite ordinals. Their supremum is ω. Hence

$$2 \cdot \omega = \omega.$$

(iv) More generally, if k is any nonzero finite ordinal, then

$$k \cdot \omega = \omega.$$

No matter how many finite points occur in each block, stacking countably many finite blocks in order still produces an ω-type sequence.

(v) But

$$\omega \cdot 3 = \omega + \omega + \omega$$

really is three ω-blocks in succession.

Proposition 12.3.11 (Multiplication is strictly increasing in the right variable for positive left factor)**.** *If $0 < \alpha$ and $\beta < \gamma$, then*

$$\alpha \cdot \beta < \alpha \cdot \gamma.$$

Proof. Fix $0 < \alpha$. We prove by transfinite induction on γ that for every $\beta < \gamma$, one has $\alpha \cdot \beta < \alpha \cdot \gamma$.

If $\gamma = 0$, there is nothing to prove.

Suppose the statement holds for γ, and let $\beta < S(\gamma)$. If $\beta = \gamma$, then

$$\alpha \cdot \beta = \alpha \cdot \gamma < (\alpha \cdot \gamma) + \alpha = \alpha \cdot S(\gamma),$$

because $0 < \alpha$ and Proposition 12.3.4 shows that adding α on the right makes the ordinal strictly larger. If $\beta < \gamma$, then by the inductive hypothesis,

$$\alpha \cdot \beta < \alpha \cdot \gamma < \alpha \cdot S(\gamma).$$

So the claim holds at successor stages.

Now let λ be a limit ordinal, and assume the claim holds below λ. If $\beta < \lambda$, then $S(\beta) < \lambda$, so

$$\alpha \cdot \beta < \alpha \cdot S(\beta) \leq \alpha \cdot \lambda,$$

where the first inequality comes from the successor case and the second from the fact that $\alpha \cdot \lambda$ is the supremum of the values $\alpha \cdot \xi$ with $\xi < \lambda$.

This completes the induction. □

Remark 12.3.12 (Multiplication is not commutative). Example 12.3.10 shows that

$$2 \cdot \omega = \omega \qquad \text{but} \qquad \omega \cdot 2 = \omega + \omega > \omega.$$

So ordinal multiplication is not commutative.

Remark 12.3.13 (Left distributivity can fail). Ordinary finite arithmetic satisfies both distributive laws, but ordinal arithmetic is more delicate. For example,

$$(1 + 1) \cdot \omega = 2 \cdot \omega = \omega,$$

while

$$1 \cdot \omega + 1 \cdot \omega = \omega + \omega > \omega.$$

So

$$(1 + 1) \cdot \omega \neq 1 \cdot \omega + 1 \cdot \omega.$$

This failure is another reminder that ordinal multiplication is about ordered placement of blocks, not about size alone.

Remark 12.3.14 (Cantor's transfinite arithmetic). Cantor introduced transfinite numbers and their arithmetic in order to study infinite order types. One of the striking discoveries was that operations familiar from finite arithmetic cease to be commutative in this setting. Historically, this was not a defect but a revelation: it showed that order carries structure invisible to mere

counting. See Cantor's classical paper [19] and modern expositions such as Devlin [5] or Potter [14].

12.4 Ordinal Exponentiation

Exponentiation on the natural numbers is repeated multiplication. The same idea extends to ordinals, but once again the infinite case behaves in a distinctive way. The simplest surprise is already this:

$$2^{\omega} = \omega.$$

Why? Because the earlier values are the finite powers $1, 2, 4, 8, \ldots$, and their supremum is just ω. In the world of ordinal exponentiation, growth is governed by the ordered way in which earlier stages accumulate.

To keep the limit clause transparent, we define exponentiation only for positive bases. The exceptional base 0 can be treated separately, but it plays no essential role in the developments that follow.

Recursive definition and first examples

Definition 12.4.1 (Ordinal exponentiation for positive base)**.** Fix an ordinal $\alpha > 0$. By transfinite recursion on the exponent, we define the *ordinal power* α^{β} by

$$\alpha^{0} = 1,$$

$$\alpha^{S(\beta)} = \alpha^{\beta} \cdot \alpha,$$

and for every nonzero limit ordinal λ,

$$\alpha^{\lambda} = \sup_{\beta<\lambda}(\alpha^{\beta}).$$

Proposition 12.4.2 (Agreement with finite exponentiation)**.** *If $m \in \mathbb{N}_0 \setminus \{0\}$ and $n \in \mathbb{N}_0$ are viewed as finite ordinals, then the ordinal power m^n agrees with the exponentiation on $\mathbb{N}_0$ defined in Definition 6.5.15.*

Proof. Fix a positive finite ordinal m. Chapter 6 defines a function $e_m : \mathbb{N}_0 \to \mathbb{N}_0$ by

$$e_m(0) = 1, \qquad e_m(S(n)) = e_m(n) \cdot m.$$

By Definition 12.4.1, the function $p_m(n) = m^n$ satisfies the same recursion on $\mathbb{N}_0$. By the uniqueness part of Theorem 6.4.2, the two functions are equal. So finite ordinal exponentiation agrees with the usual exponentiation on $\mathbb{N}_0$. □

Example 12.4.3 (First ordinal powers)**.** (i) For every $\alpha > 0$,

$$\alpha^{1} = \alpha^{0} \cdot \alpha = 1 \cdot \alpha = \alpha.$$

(ii) For finite exponents, the values are the familiar ones. For example,

$$2^{3} = 8, \qquad 3^{2} = 9.$$

(iii) We have

$$2^{\omega} = \sup_{n<\omega} 2^n.$$

The earlier values 2^n are finite ordinals, and they are unbounded in ω. So their supremum is ω. Hence

$$2^{\omega} = \omega.$$

(iv) Since $\omega^2 = \omega \cdot \omega$, we have

$$\omega^2 = \omega \cdot \omega = \sup_{n<\omega}(\omega \cdot n).$$

Thus ω^2 may be pictured as ω many successive ω-blocks.

(v) Similarly,

$$\omega^{\omega} = \sup_{n<\omega} \omega^n.$$

So ω^{ω} is the first stage reached by taking all finite powers $\omega, \omega^2, \omega^3, \ldots$ and then passing to their supremum.

Proposition 12.4.4 (Basic identities for positive-base exponentiation)**.** *Let $\alpha > 0$ be an ordinal. Then:*

(i) $\alpha^0 = 1$;

(ii) $\alpha^1 = \alpha$;

(iii) if β is any ordinal, then $\alpha^{S(\beta)} = \alpha^{\beta} \cdot \alpha$;

(iv) $1^{\beta} = 1$ *for every ordinal β.*

Proof. Parts (i) and (iii) are exactly the defining clauses in Definition 12.4.1. Part (ii) follows from part (iii) with $\beta = 0$:

$$\alpha^1 = \alpha^{S(0)} = \alpha^0 \cdot \alpha = 1 \cdot \alpha = \alpha.$$

For part (iv), we use transfinite induction on β. The base case $1^0 = 1$ is immediate. If $1^{\beta} = 1$, then

$$1^{S(\beta)} = 1^{\beta} \cdot 1 = 1 \cdot 1 = 1.$$

If λ is a nonzero limit ordinal and $1^{\beta} = 1$ for all $\beta < \lambda$, then

$$1^{\lambda} = \sup_{\beta<\lambda} 1^{\beta} = \sup_{\beta<\lambda} 1 = 1.$$

So $1^{\beta} = 1$ for every ordinal β. □

Remark 12.4.5 (Exponentiation is not commutative). The examples above show that

$$2^{\omega} = \omega \qquad \text{while} \qquad \omega^2 > \omega.$$

So ordinal exponentiation is certainly not commutative. More generally, one should be cautious about importing familiar finite laws into the transfinite setting without proof.

Remark 12.4.6 (What the limit clause is really saying). In ordinary arithmetic, a power such as a^n is reached after finitely many multiplication steps. For a limit exponent like ω, there is no last

multiplication step. The definition therefore says that α^ω is the least ordinal above all the earlier finite powers α^n. This "take the supremum at limit stages" rule is one of the characteristic features of transfinite arithmetic.

12.5 Countable Ordinals and the First Uncountable Ordinal

Up to this point we have developed ordinals mainly as ordered objects. Now size returns to the picture. Some ordinals are finite, some are countably infinite, and some are larger than every countable order. This distinction leads to one of the most important landmarks of the transfinite hierarchy: the first uncountable ordinal ω_1.

The point of ω_1 is subtle and fundamental. It is not merely an uncountable set. It is the *least* ordinal whose underlying set is uncountable. Equivalently, every smaller ordinal is countable. So ω_1 marks the boundary between countable well-orders and truly uncountable order types.

Countable ordinals

Definition 12.5.1 (Countable ordinal)**.** An ordinal α is called a *countable ordinal* if the underlying set α is countable, that is, if α is either finite or countably infinite in the sense of Chapter 8.

Example 12.5.2 (First countable ordinals)**.** Every finite ordinal is countable, and ω is countable because its underlying set is $\mathbb{N}_0$. The ordinals $\omega+1$, $\omega+2$, and $\omega+\omega$ are also countable, because they are obtained from countable stages by finitely many further constructions. Later in this section we will prove a more systematic closure theorem.

Proposition 12.5.3 (The successor of a countable ordinal is countable)**.** *If α is a countable ordinal, then $S(\alpha) = \alpha \cup \{\alpha\}$ is also countable.*

Proof. Because α is countable, there exists an injection $i\colon \alpha \to \mathbb{N}$ by Theorem 8.2.6. By Corollary 11.3.10, the point α is not an element of α, so it really is one new point. Define $j\colon S(\alpha) \to \mathbb{N}$ by

$$j(x) = \begin{cases} 1, & \text{if } x = \alpha, \\ 2i(x) + 2, & \text{if } x \in \alpha. \end{cases}$$

This function is injective: the new point α is sent to 1, while every old point of α is sent to an even integer at least 2. Hence $S(\alpha)$ injects into $\mathbb{N}$, and so it is countable by Theorem 8.2.6. □

Theorem 12.5.4 (Countable suprema of countable ordinals are countable)**.** *Let A be a countable set of countable ordinals. Then* $\sup(A)$ *is countable.*

Proof. If $A = \varnothing$, then $\sup(A) = 0$, which is finite and therefore countable. So assume $A \neq \varnothing$.

Because A is countable, Proposition 8.5.2 gives an enumeration $e\colon \mathbb{N} \to A$. For each $n \in \mathbb{N}$, the ordinal $e(n)$ is countable, so its successor

$$B_n = S(e(n))$$

is nonempty and countable by Proposition 12.5.3. Therefore the set

$$E_n = \{f\colon \mathbb{N} \to B_n \mid f \text{ is surjective}\}$$

is nonempty for each n, again by Proposition 8.5.2.

Now apply countable choice (Definition 10.2.4) to the family $(E_n)_{n\in\mathbb{N}}$. We obtain surjections $s_n : \mathbb{N} \to B_n$ for all $n \in \mathbb{N}$. The constructive countable-union theorem, Theorem 8.4.1, now shows that

$$\bigcup_{n\in\mathbb{N}} B_n$$

is countable.

Finally,

$$\sup(A) = \bigcup A$$

by Theorem 11.5.3, and each $e(n) \subseteq B_n$. Hence

$$\sup(A) = \bigcup A \subseteq \bigcup_{n\in\mathbb{N}} B_n.$$

Since a subset of a countable set is countable by Corollary 8.2.5, the supremum $\sup(A)$ is countable. □

Remark 12.5.5 (Where choice reappears). The theorem is one of the places where choice quietly enters ordinary transfinite mathematics. To show that a countable union of countable ordinals is countable, we needed to choose enumerations for infinitely many countable sets at once. Countable choice is enough for that step, and full choice of course implies it. This is a good example of how the transfinite methods of this chapter interact with the choice principles of Chapter 10.

The first uncountable ordinal

Proposition 12.5.6 (There exist uncountable ordinals)**.** *There exists at least one uncountable ordinal.*

Proof. By Corollary 9.2.3, the set $\mathbb{R}$ is uncountable. By the well-ordering theorem, Theorem 10.3.2, there exists a well-order on $\mathbb{R}$. Chapter 11 explained that every well-ordered set is order-isomorphic to a unique ordinal. Let α be the ordinal representing the order type of this well-order on $\mathbb{R}$.

If α were countable as a set, then any set order-isomorphic to it would also be countable. In particular, $\mathbb{R}$ would be countable, contradicting Corollary 9.2.3. So α is an uncountable ordinal. □

Definition 12.5.7 (The first uncountable ordinal)**.** The *first uncountable ordinal* is the least uncountable ordinal. It is denoted by

$$\omega_1.$$

The definition makes sense because Proposition 12.5.6 shows that uncountable ordinals exist, and the ordinals are well-ordered by membership.

Theorem 12.5.8 (Characterization of ω_1)**.** *For an ordinal β, the following are equivalent:*

(i) $\beta < \omega_1$;

(ii) β *is countable.*

Proof. Assume first that $\beta < \omega_1$. Since ω_1 is the least uncountable ordinal, no smaller ordinal can be uncountable. Therefore β is countable.

Conversely, assume that β is countable. If $\beta \not< \omega_1$, then by the trichotomy of ordinals (Theorem 11.3.9) we must have either $\beta = \omega_1$ or $\omega_1 < \beta$. In either case, Corollary 11.3.12 implies that $\omega_1 \subseteq \beta$. But a subset of a countable set is countable by Corollary 8.2.5, so ω_1 would be countable. This contradicts the definition of ω_1 as an uncountable ordinal. Hence $\beta < \omega_1$. □

Corollary 12.5.9 (ω_1 is a limit ordinal)**.** *The ordinal ω_1 is a limit ordinal.*

Proof. The ordinal ω_1 is nonzero, because ω is countable and therefore smaller than ω_1 by Theorem 12.5.8.

If ω_1 were a successor ordinal, say $\omega_1 = S(\beta)$, then $\beta < \omega_1$. By Theorem 12.5.8, the ordinal β would be countable. Proposition 12.5.3 would then imply that $S(\beta) = \omega_1$ is countable, a contradiction. Hence ω_1 is not a successor. Therefore it is a limit ordinal. □

Proposition 12.5.10 (Countable ordinals are closed under the basic operations)**.** *Let $\alpha, \beta < \omega_1$. Then:*

(i) $\alpha + \beta < \omega_1$;

(ii) $\alpha \cdot \beta < \omega_1$;

(iii) if $0 < \alpha$, *then* $\alpha^\beta < \omega_1$.

Proof. By Theorem 12.5.8, both α and β are countable ordinals.

Part (i). Fix α, and prove by transfinite induction on $\beta < \omega_1$ that $\alpha + \beta$ is countable.

For $\beta = 0$, we have $\alpha + 0 = \alpha$, which is countable. If $\alpha + \beta$ is countable, then $\alpha + S(\beta) = S(\alpha + \beta)$ is countable by Proposition 12.5.3. Now let $\lambda < \omega_1$ be a limit ordinal and assume that $\alpha + \gamma$ is countable for every $\gamma < \lambda$. Since $\lambda < \omega_1$, the ordinal λ is countable by Theorem 12.5.8. The image

$$A = \{\alpha + \gamma \mid \gamma < \lambda\}$$

of the countable set λ under the function $\gamma \mapsto \alpha + \gamma$ is countable by Theorem 8.2.7. Every element of A is countable by the inductive hypothesis. Therefore Theorem 12.5.4 shows that

$$\alpha + \lambda = \sup A$$

is countable. This completes the induction. Hence $\alpha + \beta < \omega_1$.

Part (ii). Fix α, and prove by transfinite induction on $\beta < \omega_1$ that $\alpha \cdot \beta$ is countable.

For $\beta = 0$, we have $\alpha \cdot 0 = 0$. If $\alpha \cdot \beta$ is countable, then

$$\alpha \cdot S(\beta) = (\alpha \cdot \beta) + \alpha$$

is countable by part (i). If $\lambda < \omega_1$ is a limit ordinal and $\alpha \cdot \gamma$ is countable for all $\gamma < \lambda$, then as above the set

$$B = \{\alpha \cdot \gamma \mid \gamma < \lambda\}$$

is a countable set of countable ordinals, so

$$\alpha \cdot \lambda = \sup B$$

is countable by Theorem 12.5.4. Thus $\alpha \cdot \beta < \omega_1$.

Part (iii). Assume also that $0 < \alpha$. We prove by transfinite induction on $\beta < \omega_1$ that α^β is countable.

For $\beta = 0$, we have $\alpha^0 = 1$, which is finite. If α^β is countable, then

$$\alpha^{S(\beta)} = \alpha^\beta \cdot \alpha$$

is countable by part (ii). If $\lambda < \omega_1$ is a limit ordinal and α^γ is countable for all $\gamma < \lambda$, then the set

$$C = \{\alpha^\gamma \mid \gamma < \lambda\}$$

is a countable set of countable ordinals, so

$$\alpha^\lambda = \sup C$$

is countable by Theorem 12.5.4. Hence $\alpha^\beta < \omega_1$. □

Example 12.5.11 (Countable arithmetic stays below ω_1)**.** Because $\omega < \omega_1$, Proposition 12.5.10 shows that all of the ordinals

$$\omega + \omega, \qquad \omega \cdot \omega = \omega^2, \qquad \omega^\omega$$

are still below ω_1. So ω_1 lies far above the usual countable ordinal arithmetic of the first transfinite stages.

Corollary 12.5.12 (No countable sequence reaches ω_1)**.** *If $f : \mathbb{N} \to \omega_1$ is any function, then*

$$\sup(\mathrm{ran}(f)) < \omega_1.$$

Proof. The range $\mathrm{ran}(f)$ is countable by Theorem 8.2.7. Every element of $\mathrm{ran}(f)$ is an ordinal below ω_1, hence is countable by Theorem 12.5.8. Therefore Theorem 12.5.4 shows that $\sup(\mathrm{ran}(f))$ is countable. Applying Theorem 12.5.8 once more, we obtain $\sup(\mathrm{ran}(f)) < \omega_1$. □

Remark 12.5.13 (What ω_1 represents). The ordinal ω_1 is not just one more infinite stage after $\omega, \omega + 1, \omega \cdot 2$, and ω^2. It is the point at which countable well-orders run out. Every ordinal below it is the order type of some countable well-ordered set, and no countable process can climb cofinally to it by Corollary 12.5.12. That is why ω_1 is the natural bridge from ordinal theory to cardinal theory.

Remark 12.5.14 (A foundational caution). In a fully axiomatic treatment, one can define ω_1 without first well-ordering $\mathbb{R}$: it is the set of all countable ordinals. We have not yet introduced the axioms needed to justify that formal set existence argument, so for the moment we used a more intuitive route via well-ordering and leastness. Later, in the chapter on axioms, we will return to this kind of foundational bookkeeping.

Looking ahead

This chapter turned ordinals from static objects into working tools. The first section generalized ordinary induction to arbitrary ordinals and showed that the true source of induction is not finiteness, but the well-ordering principle. The second section then proved the transfinite recursion theorem, which allows us to define functions stage by stage along any ordinal. Together, these two methods form the basic technique of transfinite mathematics.

We then used recursion to define ordinal addition, multiplication, and positive-base exponentiation. For finite ordinals these operations agree with the arithmetic of Chapter 6, but the infinite examples quickly revealed a new landscape. Equalities such as $1 + \omega = \omega$ and inequalities such as $\omega + 1 > \omega$ showed that ordinal arithmetic measures order type, not cardinal size. The operations became richer precisely because the transfinite hierarchy contains limit stages.

The final section introduced countable ordinals and the first uncountable ordinal ω_1. The key point is that ω_1 is not merely an uncountable set: it is the least ordinal beyond all countable order types. This prepares the way for the next chapter, where we shift the emphasis from ordered shape to size itself. There we will study cardinal numbers, compare sets by injections and bijections, and see how the language of cardinals reorganizes many of the phenomena that have appeared so far.

Part IV

Cardinality Beyond Countability

Chapter 13

Cardinal Numbers

By the time we reach this chapter, we have already met several ways of saying that one set is "as large as" another. In Chapter 7 we defined equinumerosity by bijection and used the natural numbers to measure finite sizes. In Chapters 8 and 9 we discovered that the same idea continues to make sense for infinite sets, and that it separates countable sets from uncountable ones. In Chapters 10, 11, and 12 we developed the well-ordered world of ordinals and saw that choice lets us compare arbitrary sets with well-orders.

The present chapter brings those strands together. A *cardinal number* is meant to capture size while ignoring order. Two sets may have completely different elements and completely different internal orderings, yet still determine the same cardinal if they can be matched by a bijection. Ordinals, by contrast, remember the shape of a well-order. One of the main themes of this chapter is that cardinals can be represented by special ordinals — the *initial ordinals* — once we accept the axiom of choice.

There are therefore two points of view running in parallel. First, one can compare arbitrary sets directly by injections, surjections, and bijections. This is a very concrete and flexible language. Second, once choice is available, one can replace a set by a canonical well-ordered representative of its size. That second point of view lets us speak of $\aleph_0$, $\aleph_1$, and more generally the transfinite hierarchy of infinite cardinals.

The bridge between the two viewpoints is the Cantor–Bernstein theorem. It says that if each of two sets injects into the other, then in fact they have exactly the same size. This theorem turns comparison by injection into a robust structural tool. The chapter closes with Hartogs' theorem, which shows that every set has some strictly larger well-ordered size. In that sense, no set exhausts the transfinite hierarchy.

Ordinal theory remembers order. Cardinal theory forgets order and keeps only size.

13.1 Equinumerosity, Injections, Surjections, and Comparison

Before we introduce cardinals as objects in their own right, we should clarify the language in which sets are compared. A bijection says that two sets have exactly the same size. An injection says that one set can be fitted inside another. A surjection says that one set is large enough to cover another. All three ideas are familiar from earlier chapters, but here they become the basic grammar of cardinality.

At the finite level, comparison by injection agrees perfectly with ordinary intuition: if there is an injection from a set with seven elements into a set with ten elements, then seven is at most ten. At the infinite level the same idea still works, but with one important warning. A proper subset of an infinite set may still have exactly the same size as the whole set. So injections give

us a notion of "at most as many elements," but not yet a complete theorem saying when two comparisons force equality. That will be the content of the next section.

Size comparison by fitting one set into another

Definition 13.1.1 (Comparison by injection)**.** Let A and B be sets.

(i) We write
$$A \preceq B$$
and say that A is *at most as large as* B if there exists an injective function $f : A \to B$.

(ii) We write
$$A \prec B$$
and say that A is *strictly smaller than* B if $A \preceq B$ but $A \not\approx B$.

The symbol $A \preceq B$ is deliberately weaker than $A \approx B$. A bijection gives a perfect pairing between the two sets. An injection only says that every element of A can be assigned a distinct place inside B. In everyday language, A can be packed into B without collisions.

Example 13.1.2 (Subsets are at most as large as the sets that contain them)**.** If $A \subseteq B$, then the inclusion map
$$\iota : A \to B, \qquad \iota(a) = a,$$
is injective. Hence $A \preceq B$.

For finite sets this often means that A is genuinely smaller than B. For infinite sets it may or may not.

Example 13.1.3 (A proper subset of $\mathbb{N}$ with the same size)**.** Let
$$E = \{2, 4, 6, 8, \dots\} \subsetneq \mathbb{N}.$$
The map
$$f : \mathbb{N} \to E, \qquad f(n) = 2n,$$
is a bijection. So $E \approx \mathbb{N}$, even though E is a proper subset of $\mathbb{N}$.

This is one of the first unmistakable signs that infinite size behaves very differently from finite size.

Proposition 13.1.4 (Basic properties of $\preceq$)**.** *For all sets A, B, C:*

(i) $A \preceq A$;

(ii) if $A \preceq B$ *and* $B \preceq C$, *then* $A \preceq C$;

(iii) if $A \approx B$, *then* $A \preceq B$ *and* $B \preceq A$.

Proof. Part (i) is immediate from the identity map $\mathrm{id}_A : A \to A$, which is injective.

For part (ii), choose injections $f : A \to B$ and $g : B \to C$. By Proposition 3.4.6, the composite $g \circ f : A \to C$ is injective. Therefore $A \preceq C$.

For part (iii), if $A \approx B$, then there exists a bijection $h : A \to B$. Every bijection is in particular injective, so $A \preceq B$. By Theorem 3.4.8, the inverse map $h^{-1} : B \to A$ is also a bijection, hence injective. Thus $B \preceq A$ as well. □

So $\preceq$ is reflexive and transitive, and equality of size implies comparison in both directions. What is not yet obvious is the converse: if $A \preceq B$ and $B \preceq A$, must A and B already be equinumerous? The next section will answer yes.

There is also a useful way to read $A \preceq B$ geometrically: it means that a copy of A can be found inside B.

Proposition 13.1.5 (Injection means "copy inside")**.** *For sets A and B, the following are equivalent:*

(i) $A \preceq B$;

(ii) there exists a subset $C \subseteq B$ such that $A \approx C$.

Proof. Assume first that $A \preceq B$. Choose an injection $f : A \to B$. Then its range $f[A]$ is a subset of B, and because f is injective, the map

$$f : A \to f[A]$$

is a bijection. Hence $A \approx f[A]$.

Conversely, if some subset $C \subseteq B$ satisfies $A \approx C$, choose a bijection $g : A \to C$. Composing with the inclusion $C \hookrightarrow B$ gives an injection $A \to B$. So $A \preceq B$. □

Remark 13.1.6 (Finite and infinite subsets behave differently). If A is finite and $B \subsetneq A$, then Corollary 7.2.12 shows that $A \not\approx B$. Thus every proper subset of a finite set is strictly smaller.

By Example 13.1.3, that statement fails for infinite sets. So the finite world teaches us to expect "proper subset means smaller," while the infinite world forces us to separate inclusion from cardinality.

Surjections as covering maps

An injection says that A fits inside B. A surjection says that B is rich enough to hit every point of A. In many concrete arguments, surjections feel easier to construct than injections. For example, Chapter 8 characterized countable sets in terms of surjections from $\mathbb{N}$ in Theorem 8.2.6. So it is natural to ask how surjections and injections are related in general.

One direction uses the axiom of choice in a very transparent way. Suppose $g : B \to A$ is surjective. For each $a \in A$, the fiber

$$g^{-1}[\{a\}] = \{b \in B \mid g(b) = a\}$$

is nonempty. To build an injection back from A into B, we would like to choose one representative from each fiber. That is exactly a choice problem.

Proposition 13.1.7 (Surjections yield injections under choice)**.** *Assume the axiom of choice. Let $g : B \to A$ be a surjective function. Then $A \preceq B$.*

Proof. For each $a \in A$, define the fiber

$$F_a = g^{-1}[\{a\}] = \{b \in B \mid g(b) = a\}.$$

Because g is surjective, each F_a is nonempty. Consider the family

$$\mathcal{F} = \{F_a \mid a \in A\}.$$

By the axiom of choice, there exists a choice function $c : \mathcal{F} \to B$ such that $c(F) \in F$ for every $F \in \mathcal{F}$.

Now define $s : A \to B$ by

$$s(a) = c(F_a).$$

Then $s(a) \in F_a$, so $g(s(a)) = a$ for every $a \in A$.

We claim that s is injective. Suppose that $s(a) = s(a')$. Apply g to both sides. Since $g(s(a)) = a$ and $g(s(a')) = a'$, we obtain $a = a'$. Thus s is injective, and therefore $A \preceq B$. □

The opposite direction is much easier. If A injects into B and A is nonempty, then B can be collapsed onto A by sending all points outside the embedded copy to one fixed element of A.

Proposition 13.1.8 (Injections yield surjections onto nonempty sets)**.** *Let A and B be sets, and assume that A is nonempty. If $A \preceq B$, then there exists a surjective function $h : B \to A$.*

Proof. Choose an injection $f : A \to B$. Because A is nonempty, select one element $a_0 \in A$. Define $h : B \to A$ by

$$h(b) = \begin{cases} f^{-1}(b), & \text{if } b \in f[A], \\ a_0, & \text{if } b \notin f[A]. \end{cases}$$

The formula is well defined because f is injective, so each $b \in f[A]$ has a unique preimage under f.

To see that h is surjective, let $a \in A$. Then $f(a) \in B$, and by the first clause of the definition,

$$h(f(a)) = a.$$

So every element of A is hit. □

Corollary 13.1.9 (Comparison by surjections under choice)**.** *Assume the axiom of choice, and let A and B be sets with $A \neq \varnothing$. Then the following are equivalent:*

(i) $A \preceq B$;

(ii) there exists a surjective function $B \to A$.

Proof. Proposition 13.1.8 gives $(i) \Rightarrow (ii)$, and Proposition 13.1.7 gives $(ii) \Rightarrow (i)$. □

Remark 13.1.10 (Why this matters). At the finite level, injections, surjections, and bijections are tightly connected by elementary counting. At the infinite level their relation is more subtle. Proposition 13.1.7 shows exactly where the axiom of choice enters: a surjection covers the target, but to turn that cover into an injection back, one must choose a single representative from each fiber.

This is one of the reasons Chapter 10 was placed before the present chapter. Cardinal theory becomes especially clean once we know that arbitrary sets can be well-ordered and that fibers of surjections can be thinned out by choice.

Example 13.1.11 (Every power set is strictly larger than the original set)**.** For every set A, the map

$$x \longmapsto \{x\}$$

is an injection from A into $\mathcal{P}(A)$. Hence $A \preceq \mathcal{P}(A)$. On the other hand, Cantor's theorem (Theorem 9.3.1) says that there is no surjection $A \to \mathcal{P}(A)$, and in particular no bijection $A \to \mathcal{P}(A)$. Therefore

$$A \prec \mathcal{P}(A).$$

So the power-set operation always produces a genuinely larger size.

13.2 The Cantor–Bernstein Theorem

Comparison by injection would be much less useful if mutual comparison left us permanently uncertain. Suppose we can embed A inside B and also embed B inside A. Does that mean that one set is really larger, or do the two embeddings force equality of size?

At first sight the question is surprisingly delicate. The existence of two injections does not literally hand us a bijection. We are not told that the image of the first injection lines up nicely with the domain of the second, and in the infinite case one cannot simply subtract the "unused parts" and count what is left. The remarkable fact is that the right bijection nevertheless exists.

This theorem is one of the central structural principles of set theory. It turns comparison by injection into a genuine order relation on cardinalities. Philosophically, it says that size is determined by mutual embeddability.

Mutual injections force equal size

Theorem 13.2.1 (Cantor–Bernstein)**.** *Let A and B be sets. If $A \preceq B$ and $B \preceq A$, then $A \approx B$.*

Proof. Choose injections

$$f : A \to B, \qquad g : B \to A.$$

We will build a bijection $h : A \to B$ by deciding, for each point of A, whether to use f or the inverse of g.

The construction begins by isolating the points of A that do *not* lie in the image of g:

$$A_0 = A \setminus g[B].$$

Then define recursively

$$A_{n+1} = g\big(f[A_n]\big) \qquad \text{for } n \in \mathbb{N}_0.$$

Finally set

$$A_* = \bigcup_{n \in \mathbb{N}_0} A_n.$$

So A_* is the part of A generated by repeatedly moving forward with f and back with g, starting from the points not covered by g.

We now define $h : A \to B$ by

$$h(a) = \begin{cases} f(a), & \text{if } a \in A_*, \\ g^{-1}(a), & \text{if } a \notin A_*. \end{cases}$$

This is well defined. Indeed, if $a \notin A_*$, then in particular $a \notin A_0 = A \setminus g[B]$, so $a \in g[B]$. Since g is injective, there is a unique element of B mapped to a, namely $g^{-1}(a)$.

We show that h is injective.

Case 1: both points lie in A_*. If $a, a' \in A_*$ and $h(a) = h(a')$, then $f(a) = f(a')$. Since f is injective, $a = a'$.

Case 2: both points lie outside A_*. If $a, a' \notin A_*$ and $h(a) = h(a')$, then $g^{-1}(a) = g^{-1}(a')$. Apply g to both sides to obtain $a = a'$.

Case 3: one point lies in A_* and the other does not. Suppose that $a \in A_*$, $a' \notin A_*$, and $h(a) = h(a')$. Then

$$f(a) = g^{-1}(a').$$

Applying g, we get

$$g(f(a)) = a'.$$

Because $a \in A_*$, there exists $n \in \mathbb{N}_0$ such that $a \in A_n$. Therefore

$$a' = g(f(a)) \in g(f[A_n]) = A_{n+1} \subseteq A_*.$$

This contradicts $a' \notin A_*$. So the mixed case cannot occur. Hence h is injective.

It remains to prove that h is surjective. Let $b \in B$. We must find $a \in A$ with $h(a) = b$.

If $b \in f[A_*]$, then there exists $a \in A_*$ with $f(a) = b$, and therefore $h(a) = b$.

Now suppose that $b \notin f[A_*]$. We claim that $g(b) \notin A_*$. If this were false, then $g(b) \in A_n$ for some least $n \in \mathbb{N}_0$. The case $n = 0$ is impossible because A_0 is disjoint from $g[B]$, whereas $g(b) \in g[B]$. Thus $n \geq 1$. By definition of A_n, we then have

$$g(b) \in A_n = g\big(f[A_{n-1}]\big).$$

So there exists $a \in A_{n-1} \subseteq A_*$ such that $g(b) = g(f(a))$. Because g is injective, this implies $b = f(a)$, contradicting $b \notin f[A_*]$.

So indeed $g(b) \notin A_*$. For that element of A we use the second clause in the definition of h:

$$h(g(b)) = g^{-1}(g(b)) = b.$$

Thus every $b \in B$ lies in the image of h, and h is surjective.

Therefore h is a bijection from A to B, so $A \approx B$. □

Corollary 13.2.2 (Comparison by injection is antisymmetric up to bijection)**.** *For sets A and B, we have*

$$A \approx B \qquad \Longleftrightarrow \qquad A \preceq B \text{ and } B \preceq A.$$

Proof. The forward implication is part (iii) of Proposition 13.1.4. The reverse implication is Theorem 13.2.1. □

Example 13.2.3 (The real line and the unit interval have the same size)**.** The inclusion map

$$(0, 1) \hookrightarrow \mathbb{R}$$

is injective, so $(0, 1) \preceq \mathbb{R}$.

For the reverse comparison, define

$$\varphi : \mathbb{R} \to (0, 1), \qquad \varphi(x) = \frac{1}{2} + \frac{x}{2(1 + |x|)}.$$

Because the function $x \mapsto x/(1 + |x|)$ takes values in $(-1, 1)$ and is strictly increasing, φ is injective. Hence $\mathbb{R} \preceq (0, 1)$.

By Theorem 13.2.1, we conclude that

$$\mathbb{R} \approx (0, 1).$$

Chapter 9 already exhibited a direct bijection in another way. The present proof shows how the theorem converts mutual injections into equality of size.

Corollary 13.2.4 (A cardinality sandwich principle)**.** *Let $A \subseteq B \subseteq C$. If $A \approx C$, then $A \approx B \approx C$.*

Proof. Because $A \subseteq B$, Example 13.1.2 gives $A \preceq B$. Because $B \subseteq C$, we also have $B \preceq C$. Choose a bijection $u \colon C \to A$. Then the composite

$$B \hookrightarrow C \xrightarrow{u} A$$

is an injection, so $B \preceq A$. By Theorem 13.2.1, we obtain $A \approx B$.

Now choose a bijection $v \colon B \to A$. Composing with the inclusion $A \hookrightarrow C$ gives an injection $B \to C$, which we already had, and composing the inclusion $C \hookleftarrow B$ with the bijection $u \colon C \to A \approx B$ gives an injection $C \to B$. Another application of Theorem 13.2.1 yields $B \approx C$. □

Remark 13.2.5 (What the theorem accomplishes). The Cantor–Bernstein theorem does not tell us *which* bijection between A and B is the most natural one. In concrete examples there may be many different bijections, each illuminating a different feature of the sets involved. The theorem says something more basic: size comparison by injection is logically complete. Once we know that neither set is larger than the other in the sense of Definition 13.1.1, we know that they have the same cardinality.

Remark 13.2.6 (Historical and conceptual background). Cantor's way of comparing sets by correspondence goes back to the birth of set theory itself [18, 19]. Modern textbook discussions of the Cantor–Bernstein theorem may be found in Halmos [2], Enderton [3], Devlin [5], and Moschovakis [7].

13.3 Cardinals as Initial Ordinals

So far, cardinality has been expressed by a relation between sets: $A \approx B$ means that A and B have the same size. This is perfectly adequate for many purposes, but it leaves one natural desire unfulfilled. We would like to speak of a set's size as an object in its own right, just as we speak of an ordinal as the order type of a well-order.

One tempting answer is to define a cardinal as an equivalence class of sets under $\approx$. That is conceptually correct, but it leads quickly to foundational questions about classes of all sets equinumerous to a given set. In the present book we take a more concrete route. Because Chapter 10 gave us the well-ordering theorem, every set can be well-ordered and therefore matched with an ordinal. Among the ordinals equinumerous with a given set, there is a least one. That least ordinal will serve as the canonical representative of the set's cardinality.

The price of this concrete representation is that it depends on choice. Without the axiom of choice, some sets may fail to be well-orderable, and then not every cardinality can be represented by an initial ordinal. In this chapter we work explicitly in the choice-friendly setting opened up by Chapter 10.

The least ordinal of a given size

Definition 13.3.1 (Initial ordinal)**.** An ordinal κ is called an *initial ordinal* if there is no ordinal $\alpha < \kappa$ such that $\alpha \approx \kappa$.

A *cardinal number* will be represented in this chapter by an initial ordinal.

The definition says exactly what one would hope. An initial ordinal is the first place at which a certain size appears in the ordinal hierarchy. It may have many larger ordinals equinumerous with it, but none smaller.

Example 13.3.2 (The first initial ordinals)**.** The finite ordinals

$$0, 1, 2, 3, \ldots$$

are all initial.

Indeed, if $m < n$ are finite ordinals, then m is a proper subset of n. By Corollary 7.2.12, there is no bijection $m \to n$. So no smaller ordinal is equinumerous with n.

The ordinal ω is also initial. Every smaller ordinal $\beta < \omega$ is finite, whereas ω is countably infinite: it is in bijection with $\mathbb{N}$. So β cannot be equinumerous with ω.

Not every ordinal is initial. For example, $\omega + 1$, $\omega + 2$, and $\omega \cdot 2$ are all countable ordinals, so each of them will turn out to have the same cardinality as ω. They are different ordinals because they encode different order types, but they are not new cardinals because they do not represent new sizes.

Theorem 13.3.3 (Every set has a unique cardinal representative under choice)**.** *Assume the axiom of choice. For every set A, there exists a unique initial ordinal κ such that*

$$A \approx \kappa.$$

Proof. By the well-ordering theorem (Theorem 10.3.2), the set A can be well-ordered. Every well-ordered set is order-isomorphic to a unique ordinal, so there exists an ordinal α with $A \approx \alpha$.

Consider the set

$$S = \{\beta \in S(\alpha) \mid \beta \approx A\}.$$

This set is nonempty because $\alpha \in S$. Since S is a nonempty set of ordinals, Corollary 11.3.14 shows that S has a least element; call it κ.

Then $\kappa \approx A$ by construction. We claim that κ is initial. Suppose instead that some ordinal $\gamma < \kappa$ satisfies $\gamma \approx \kappa$. Since $\kappa \approx A$, transitivity of $\approx$ from Proposition 7.1.4 gives $\gamma \approx A$. Hence $\gamma \in S$, contradicting the leastness of κ. So κ is initial.

To prove uniqueness, suppose that λ is another initial ordinal with $\lambda \approx A$. Then $\kappa \approx \lambda$. By the trichotomy of ordinals (Theorem 11.3.9), exactly one of the following holds:

$$\kappa < \lambda, \qquad \kappa = \lambda, \qquad \lambda < \kappa.$$

If $\kappa < \lambda$, then λ is equinumerous with a smaller ordinal, contradicting that λ is initial. Similarly, $\lambda < \kappa$ contradicts that κ is initial. Therefore $\kappa = \lambda$. □

Definition 13.3.4 (Cardinality as an initial ordinal)**.** Assume the axiom of choice. For a set A, the unique initial ordinal provided by Theorem 13.3.3 is called the *cardinality* of A and is denoted by

$$\mathrm{Card}(A).$$

Many books write $|A|$ for the cardinality of A. We will continue to use bar notation informally when no ambiguity is possible, but the notation $\mathrm{Card}(A)$ is helpful when we want to emphasize that the cardinality is being treated as a particular ordinal representative.

Example 13.3.5 (Finite and countably infinite cardinals)**.** If A has exactly n elements, then $A \approx n$, so $\mathrm{Card}(A) = n$.

If A is countably infinite, then $A \approx \mathbb{N}$, and because $\mathbb{N} \approx \omega$, we have $\mathrm{Card}(A) = \omega$.

Once cardinals are represented by initial ordinals, comparison by injection becomes extremely clean.

Theorem 13.3.6 (Comparison of cardinals by injections)**.** *Assume the axiom of choice, and let κ and λ be cardinals (that is, initial ordinals). Then:*

(i) $\kappa \leq \lambda$ *if and only if* $\kappa \preceq \lambda$*;*

(ii) $\kappa < \lambda$ *if and only if* $\kappa \preceq \lambda$ *and* $\lambda \not\preceq \kappa$*.*

Proof. For part (i), if $\kappa \leq \lambda$, then because ordinals are transitive sets, we have $\kappa \subseteq \lambda$. So the inclusion map is an injection $\kappa \to \lambda$, and hence $\kappa \preceq \lambda$.

Conversely, suppose that $\kappa \preceq \lambda$. If we had $\lambda < \kappa$, then the inclusion $\lambda \hookrightarrow \kappa$ would give $\lambda \preceq \kappa$ as well. By Theorem 13.2.1, it would follow that $\kappa \approx \lambda$. But κ is initial and $\lambda < \kappa$, so that is impossible. Therefore $\lambda < \kappa$ cannot occur. By ordinal trichotomy, $\kappa \leq \lambda$.

For part (ii), first assume $\kappa < \lambda$. Then part (i) gives $\kappa \preceq \lambda$. If also $\lambda \preceq \kappa$, then part (i) would imply $\lambda \leq \kappa$, contradicting $\kappa < \lambda$. So $\lambda \not\preceq \kappa$.

Conversely, suppose $\kappa \preceq \lambda$ and $\lambda \not\preceq \kappa$. By part (i), $\kappa \leq \lambda$. If $\kappa = \lambda$, then certainly $\lambda \preceq \kappa$, contrary to assumption. Hence $\kappa \neq \lambda$, so $\kappa < \lambda$. □

Corollary 13.3.7 (Mutual injections between cardinals force equality)**.** *Assume the axiom of choice, and let κ and λ be cardinals. If $\kappa \preceq \lambda$ and $\lambda \preceq \kappa$, then $\kappa = \lambda$.*

Proof. By part (i) of Theorem 13.3.6, the assumptions imply both $\kappa \leq \lambda$ and $\lambda \leq \kappa$. Therefore $\kappa = \lambda$. □

Remark 13.3.8 (Why initial ordinals are convenient). Theorem 13.3.6 shows the practical advantage of representing cardinals by initial ordinals. The abstract statement "there exists an injection from one size into another" turns into the concrete ordinal comparison $\kappa \leq \lambda$. So the ordinal hierarchy from Chapters 11 and 12 becomes a hierarchy of sizes as soon as we restrict attention to the initial ordinals.

Remark 13.3.9 (Equivalence classes versus representatives). If one works in a more foundationally formal style, a cardinal may be introduced as an equivalence class of sets under equinumerosity. The initial-ordinal approach used here avoids class language and is better suited to our gradual development. The price is that it uses the axiom of choice to guarantee that every set has a well-ordered representative. Later, in Chapter 15, we will revisit the foundational background more explicitly.

13.4 Aleph Numbers and Familiar Infinite Cardinals

Once cardinals are represented by initial ordinals, the infinite sizes we have already met acquire standard names. The first infinite cardinal is simply the cardinality of $\mathbb{N}$, hence also of $\mathbb{Z}$, $\mathbb{Q}$, and any other countably infinite set. The first uncountable cardinal is the cardinality of the first uncountable ordinal ω_1. These are written $\aleph_0$ and $\aleph_1$.

The notation is more than symbolism. It emphasizes that the infinite cardinals form an ordered hierarchy, just as the ordinals do. What is subtle is that many different ordinals can determine the same cardinal. Thus the aleph notation helps us keep separate two questions that are easy to confuse:

> *What is the order type of this well-order, and what is merely its size?*

The first infinite cardinal

Definition 13.4.1 ($\aleph_0$)**.** The cardinal ω is called *aleph-zero* and is denoted by

$$\aleph_0.$$

Because ω is initial by Example 13.3.2, this definition is legitimate. It gives a distinguished name to the size of all countably infinite sets.

Theorem 13.4.2 (Countability in cardinal language)**.** *Assume the axiom of choice, and let A be a set.*

(i) The set A is finite if and only if $\mathrm{Card}(A) = n$ *for some* $n \in \mathbb{N}_0$.

(ii) The set A is countably infinite if and only if $\mathrm{Card}(A) = \aleph_0$.

(iii) The set A is countable if and only if $\mathrm{Card}(A) \leq \aleph_0$.

Proof. For part (i), if A is finite, then by definition there is some $n \in \mathbb{N}_0$ with $A \approx n$. Since n is initial, Theorem 13.3.3 gives $\mathrm{Card}(A) = n$. Conversely, if $\mathrm{Card}(A) = n$ for some finite ordinal n, then $A \approx n$, so A is finite.

For part (ii), if A is countably infinite, then by Definition 8.2.1, $A \approx \mathbb{N}$. Since $\mathbb{N} \approx \omega$, we have $A \approx \omega = \aleph_0$, and uniqueness in Theorem 13.3.3 yields $\mathrm{Card}(A) = \aleph_0$.

Conversely, if $\mathrm{Card}(A) = \aleph_0$, then $A \approx \aleph_0 = \omega$. Because $\omega \approx \mathbb{N}$, the set A is countably infinite.

For part (iii), first suppose A is countable. If A is finite, part (i) shows that $\mathrm{Card}(A) = n$ for some finite ordinal $n < \omega$, so $\mathrm{Card}(A) < \aleph_0$. If A is countably infinite, part (ii) gives $\mathrm{Card}(A) = \aleph_0$. Thus in either case $\mathrm{Card}(A) \leq \aleph_0$.

Conversely, suppose $\mathrm{Card}(A) \leq \aleph_0$. If $\mathrm{Card}(A) = \aleph_0$, then A is countably infinite by part (ii). If $\mathrm{Card}(A) < \aleph_0 = \omega$, then $\mathrm{Card}(A)$ is a finite ordinal, so part (i) shows that A is finite. Therefore A is countable. □

Example 13.4.3 (Familiar countably infinite sets)**.** Theorems 8.3.1, 8.3.3, and 8.3.5 show that

$$\mathrm{Card}(\mathbb{N}) = \mathrm{Card}(\mathbb{Z}) = \mathrm{Card}(\mathbb{Q}) = \aleph_0.$$

So although these sets look very different — one is ordered like the positive integers, one extends in both directions, and one is densely ordered — they all determine the same infinite cardinal.

The first uncountable cardinal

Chapter 12 introduced the first uncountable ordinal ω_1. That ordinal was defined as the least ordinal that is not countable. Since every smaller ordinal is countable, ω_1 turns out to be initial automatically.

Theorem 13.4.4 (ω_1 is the least uncountable cardinal)**.** *The ordinal ω_1 is an initial ordinal. Consequently* $\mathrm{Card}(\omega_1) = \omega_1$*. This cardinal is denoted by*

$$\aleph_1.$$

Moreover, $\aleph_1$ is the least uncountable cardinal.

Proof. Let $\beta < \omega_1$. By Theorem 12.5.8, the ordinal β is countable. But ω_1 itself is uncountable by definition. So β cannot be equinumerous with ω_1. This shows that no smaller ordinal is equinumerous with ω_1, and therefore ω_1 is initial.

Since ω_1 is initial, its cardinality is itself: $\mathrm{Card}(\omega_1) = \omega_1$. We call this cardinal $\aleph_1$.

Finally, let κ be any uncountable cardinal. Because κ is uncountable, it cannot satisfy $\kappa \leq \omega$; otherwise Theorem 13.4.2 would imply that κ is countable. By ordinal trichotomy, we must therefore have $\omega < \kappa$. Since ω_1 is the least uncountable ordinal, this means $\omega_1 \leq \kappa$. So $\aleph_1 = \omega_1$ is the least uncountable cardinal. □

Example 13.4.5 (The real numbers sit at least at $\aleph_1$)**.** Chapter 9 showed that $\mathbb{R}$ is uncountable. By Theorem 13.4.2, its cardinality is therefore strictly larger than $\aleph_0$. In fact, Theorem 13.4.4 shows that

$$\aleph_1 \leq \mathrm{Card}(\mathbb{R}).$$

So the continuum is at least as large as the first uncountable cardinal. The next chapter will reinterpret this more sharply in terms of the power-set operation, leading to the notation $2^{\aleph_0}$.

Remark 13.4.6 (Many ordinals collapse to one cardinal). The ordinals

$$\omega, \qquad \omega + 1, \qquad \omega + 2, \qquad \omega \cdot 2, \qquad \omega^2, \qquad \omega^\omega$$

are all countable by Chapter 12. Therefore Theorem 13.4.2 gives

$$\mathrm{Card}(\omega) = \mathrm{Card}(\omega + 1) = \mathrm{Card}(\omega + 2) = \mathrm{Card}(\omega \cdot 2) = \mathrm{Card}(\omega^2) = \mathrm{Card}(\omega^\omega) = \aleph_0.$$

This is a vivid reminder that ordinals and cardinals answer different questions. Ordinals distinguish all of these order types; cardinality forgets the ordering and sees only one countably infinite size.

Remark 13.4.7 (The aleph hierarchy). After $\aleph_0$ and $\aleph_1$, one writes $\aleph_2$ for the least cardinal larger than $\aleph_1$, $\aleph_3$ for the least cardinal larger than $\aleph_2$, and so on. More generally there is an $\aleph_\alpha$ for every ordinal α. The next section gives an important glimpse of why the phrase "the next cardinal" is not empty notation: Hartogs' theorem proves that after any set there is a strictly larger well-ordered size.

13.5 Hartogs' Theorem as a Glimpse of Deeper Structure

At this point one could easily believe that the hierarchy of cardinals is being built only because we already assumed choice and already decided to represent sizes by ordinals. Hartogs'

theorem shows that something transfinite is happening much more deeply. It does not start by well-ordering the set A itself. Instead, it looks at all the well-ordered pieces that can be found inside A, records their order types, and then forms an ordinal larger than all of them.

The result is astonishingly strong. For every set A, there is an ordinal that cannot inject into A. So no matter how large A is, there is always some well-ordered size beyond it. In the presence of choice, this becomes the existence of the next cardinal. Without choice, it still remains a genuine theorem and is one of the first signs that the transfinite hierarchy cannot be exhausted by any single set.

A larger ordinal built from the well-orders inside a set

Theorem 13.5.1 (Hartogs' theorem)**.** *For every set A, there exists an ordinal $h(A)$ such that no injection $h(A) \to A$ exists. Moreover, $h(A)$ may be chosen to be the least ordinal with this property.*

Proof. Consider the set of all well-orders carried by subsets of A:

$$\mathcal{W}_A = \left\{ \langle B, R\rangle \in \mathcal{P}(A) \times \mathcal{P}(A \times A) \;\middle|\; B \subseteq A \text{ and } R \text{ well-orders } B \right\}.$$

For each $\langle B, R\rangle \in \mathcal{W}_A$, let $\mathrm{otp}(B, R)$ be the unique ordinal order-isomorphic to the well-ordered set (B, R). Define

$$H(A) = \{\mathrm{otp}(B, R) \mid \langle B, R\rangle \in \mathcal{W}_A\}.$$

Thus $H(A)$ is the set of all ordinals that occur as order types of well-ordered subsets of A.

Now define

$$h(A) = \bigcup_{\alpha \in H(A)} S(\alpha).$$

Because each $S(\alpha)$ is an ordinal, $h(A)$ is an ordinal. By construction,

$$\alpha < h(A) \qquad \text{for every } \alpha \in H(A).$$

We claim first that no injection $h(A) \to A$ exists.

Suppose, toward a contradiction, that $f : h(A) \to A$ were injective. Then $f[h(A)]$ is a subset of A, and we may transport the well-order $\in$ on $h(A)$ across f to obtain a well-order on $f[h(A)]$: define

$$x \triangleleft y \qquad \text{if and only if} \qquad f^{-1}(x) \in f^{-1}(y).$$

Then $(f[h(A)], \triangleleft)$ is well-ordered and order-isomorphic to $h(A)$. Therefore $h(A)$ is an element of $H(A)$, because it is the order type of a well-ordered subset of A.

But by the construction of $h(A)$, every element of $H(A)$ is strictly smaller than $h(A)$. So $h(A) < h(A)$, which is impossible. Therefore no injection $h(A) \to A$ exists.

It remains to show leastness. Let $\beta < h(A)$. Because $h(A) = \bigcup_{\alpha \in H(A)} S(\alpha)$, there exists some $\alpha \in H(A)$ such that $\beta \in S(\alpha)$, that is, $\beta \leq \alpha$.

Since $\alpha \in H(A)$, some well-ordered subset of A has order type α. In particular, there is an injection $\alpha \to A$. Because $\beta \leq \alpha$, the inclusion $\beta \hookrightarrow \alpha$ is injective. Composing these injections gives an injection $\beta \to A$.

So every ordinal smaller than $h(A)$ does inject into A, while $h(A)$ itself does not. Hence $h(A)$ is the least ordinal not injecting into A. □

Example 13.5.2 (Finite sets and the natural numbers)**.** If n is a finite ordinal, then the least ordinal not injecting into n is $n + 1$. So

$$h(n) = n + 1.$$

For the natural numbers, Hartogs' theorem gives

$$h(\mathbb{N}) = \omega_1.$$

Indeed, every $\beta < \omega_1$ is countable by Theorem 12.5.8, hence injects into $\mathbb{N}$ by Theorem 8.2.6. On the other hand, ω_1 itself cannot inject into $\mathbb{N}$, because that would make ω_1 countable, contradicting Theorem 12.5.8 again.

Hartogs' theorem becomes especially transparent when we apply it to a cardinal rather than to an arbitrary set.

Corollary 13.5.3 (Hartogs gives the next cardinal)**.** *Assume the axiom of choice, and let κ be a cardinal. Then $h(\kappa)$ is an initial ordinal with*

$$\kappa < h(\kappa).$$

Moreover, $h(\kappa)$ is the least cardinal strictly larger than κ.

Proof. Because κ injects into itself, the leastness statement in Theorem 13.5.1 implies that $\kappa < h(\kappa)$.

We next show that $h(\kappa)$ is initial. Suppose that some ordinal $\lambda < h(\kappa)$ satisfies $\lambda \approx h(\kappa)$. Since $\lambda < h(\kappa)$, Theorem 13.5.1 gives an injection $\lambda \to \kappa$. Composing with a bijection $h(\kappa) \to \lambda$, we would obtain an injection $h(\kappa) \to \kappa$, contradicting the defining property of $h(\kappa)$. Thus no smaller ordinal is equinumerous with $h(\kappa)$, so $h(\kappa)$ is initial.

Now let μ be any cardinal with $\kappa < \mu$. If $\mu < h(\kappa)$, then again by Hartogs' theorem there would be an injection $\mu \to \kappa$. Since $\kappa < \mu$, the inclusion $\kappa \hookrightarrow \mu$ is injective as well. The Cantor–Bernstein Theorem 13.2.1 would then imply $\kappa \approx \mu$, contradicting the fact that both are cardinals and $\kappa < \mu$. Therefore $h(\kappa) \leq \mu$.

So every cardinal strictly larger than κ lies at or above $h(\kappa)$. Since $h(\kappa)$ itself is a cardinal and is larger than κ, it is the least cardinal strictly larger than κ. □

Remark 13.5.4 (Successor cardinals). One often writes κ^+ for the least cardinal strictly larger than κ. Corollary 13.5.3 says that, under choice,

$$\kappa^+ = h(\kappa).$$

In particular,

$$\aleph_1 = \aleph_0^+, \qquad \aleph_2 = \aleph_1^+, \qquad \aleph_3 = \aleph_2^+, \quad \text{and so on.}$$

So Hartogs' theorem is the mechanism that keeps the aleph hierarchy moving upward.

Remark 13.5.5 (What Hartogs' theorem does *not* say). Hartogs' theorem does not by itself provide a well-order of the original set A. It only constructs an ordinal that cannot be injected into A. That distinction matters. The theorem is true even in settings where one cannot prove that every set is well-orderable. Choice enters only when we pass from the existence of a larger ordinal to the claim that A itself has a cardinal representative as an initial ordinal.

Remark 13.5.6 (A foundational caution). The proof of Hartogs' theorem packages together all well-orders on subsets of A and then forms the set of their order types. In the fully axiomatic treatment of Chapter 15, that construction will be justified using the usual ZF axioms, especially Power Set, Separation, and Replacement. At the present intuitive stage, the proof should be understood as a guided preview of what the formal theory later makes precise.

Remark 13.5.7 (Further reading). Hartogs' theorem is a standard turning point in modern set theory. Good discussions may be found in Levy [12], Jech [10], Kunen [11], and Potter [14]. It is also one of the theorems that shows most clearly how ordinal methods can produce information about size even before one has a fully global theory of cardinals.

Looking ahead

This chapter reorganized the book's earlier theory of size. We began with direct comparison of sets by injections and surjections, proved via Cantor–Bernstein that mutual injectability is already equality of size, and then used the well-ordering theorem to represent cardinalities by initial ordinals. The familiar infinite sizes $\aleph_0$ and $\aleph_1$ emerged from ω and ω_1, while Hartogs' theorem showed that the transfinite hierarchy always continues upward.

The next chapter turns from *which* sizes exist to *how* sizes combine. We will define cardinal addition, multiplication, and exponentiation by disjoint unions, Cartesian products, and function sets. There the contrast between finite and infinite arithmetic will become one of the central themes, and the continuum will reappear as $2^{\aleph_0}$.

Chapter 14

Cardinal Arithmetic

In Chapter 12 we defined ordinal addition, ordinal multiplication, and ordinal exponentiation. There the point was not to measure size, but to record the shape of a well-order. The expressions

$$1 + \omega = \omega \qquad \text{and} \qquad \omega + 1 > \omega$$

were memorable precisely because ordinal arithmetic is sensitive to where a new block is attached.

In the present chapter the same symbols return with a different meaning. A *cardinal* remembers only size. So when we add two cardinals, we are asking for the size of a disjoint union; when we multiply them, we are asking for the size of a Cartesian product; and when we exponentiate them, we are asking for the size of a set of functions. The operations are now designed to forget order and retain only multiplicity.

That change of viewpoint produces some of the most striking phenomena in all of elementary set theory. Two countably infinite sets put together still make only a countably infinite set. A countable square remains countable. More generally, under the axiom of choice, the sum and product of two infinite cardinals collapse to the larger one. By contrast, exponentiation behaves very differently: the power-set operation always creates a strictly larger cardinal, so $2^\kappa > \kappa$ for every cardinal κ.

The chapter also sharpens a theme that has been visible since Chapter 9. The continuum is not merely the real line as a geometric object; it is a specific cardinal, namely $2^{\aleph_0}$. Once that fact is understood, one of the most natural questions in set theory becomes unavoidable: is $2^{\aleph_0}$ equal to $\aleph_1$, or is there a cardinal in between? That is the continuum hypothesis, and its fate explains why the next chapter must finally turn from intuition to axioms.

> *Ordinal arithmetic records ordered shape. Cardinal arithmetic records how many objects there are, regardless of order.*

14.1 Sum, Product, and Exponentiation of Cardinals

The guiding idea of cardinal arithmetic is simple: whenever ordinary finite arithmetic counts something, we ask for the corresponding set construction in general. Putting two piles of objects together suggests disjoint union. Choosing one object from one set and one from another suggests Cartesian product. Making a list of choices indexed by a set suggests a function set. Cardinal operations are just the cardinalities of those constructions.

From operations on sets to operations on cardinals

There is one small technical point to settle at the beginning. If two sets happen to overlap, then their ordinary union does not faithfully represent "taking one copy of each." For cardinal addition we therefore use a *tagged disjoint union*, in which the two copies are explicitly marked.

Definition 14.1.1 (Tagged disjoint union)**.** For sets A and B, the *tagged disjoint union* of A and B is the set

$$A \sqcup B = \big(A \times \{0\}\big) \cup \big(B \times \{1\}\big).$$

Even when A and B are disjoint already, the tags do no harm. If A and B overlap, the tags are essential: the elements $\langle x, 0\rangle$ and $\langle x, 1\rangle$ are different, so an object coming from A is never confused with the same-looking object coming from B.

Example 14.1.2 (Why the tags matter)**.** Let $A = B = \{1, 2, 3\}$. Then the ordinary union is just

$$A \cup B = \{1, 2, 3\},$$

which has only three elements. But the tagged disjoint union has six members:

$$A \sqcup B = \{\langle 1, 0\rangle, \langle 2, 0\rangle, \langle 3, 0\rangle, \langle 1, 1\rangle, \langle 2, 1\rangle, \langle 3, 1\rangle\}.$$

So cardinal addition must be based on $A \sqcup B$, not on $A \cup B$.

The next proposition says that the three constructions relevant to cardinal arithmetic depend only on size, not on the particular sets used to represent that size.

Proposition 14.1.3 (The basic constructions are well defined up to bijection)**.** *Let A, A', B, B' be sets with $A \approx A'$ and $B \approx B'$. Then:*

(i) $A \sqcup B \approx A' \sqcup B'$;

(ii) $A \times B \approx A' \times B'$;

(iii) the function sets B^A and $(B')^{A'}$ are in bijection.

Proof. Choose bijections $f : A \to A'$ and $g : B \to B'$.

For part (i), define

$$h : A \sqcup B \to A' \sqcup B'$$

by

$$h(\langle a, 0\rangle) = \langle f(a), 0\rangle, \qquad h(\langle b, 1\rangle) = \langle g(b), 1\rangle.$$

Because the tags are preserved, this map is well defined. It is clearly injective and surjective, hence bijective.

For part (ii), define

$$k : A \times B \to A' \times B', \qquad k(a, b) = \langle f(a), g(b)\rangle.$$

Again, because f and g are bijections, so is k.

For part (iii), let $\mathcal{F}$ be the set of functions $u\colon A \to B$, and let $\mathcal{F}'$ be the set of functions $v\colon A' \to B'$. Define

$$\Phi\colon \mathcal{F} \to \mathcal{F}'$$

by

$$\Phi(u) = g \circ u \circ f^{-1}.$$

This is a function $A' \to B'$. Its inverse is given by

$$\Psi(v) = g^{-1} \circ v \circ f.$$

Thus Φ is a bijection. □

We may therefore define cardinal operations by choosing any convenient representative sets. Since Chapter 13 identifies each cardinal with a particular initial ordinal, the most economical choice is to use the cardinals themselves as the representing sets.

Definition 14.1.4 (Cardinal addition, multiplication, and exponentiation)**.** Assume the axiom of choice. Let κ and λ be cardinals. We define:

(i) the *cardinal sum*

$$\kappa + \lambda = \operatorname{Card}\big((\kappa \times \{0\}) \cup (\lambda \times \{1\})\big);$$

(ii) the *cardinal product*

$$\kappa \cdot \lambda = \operatorname{Card}(\kappa \times \lambda);$$

(iii) the *cardinal exponentiation*

$$\kappa^{\lambda} = \operatorname{Card}\big(\{f \mid f\colon \lambda \to \kappa\}\big).$$

Because of Proposition 14.1.3, these definitions agree with the more informal rule "take any sets of sizes κ and λ, perform the corresponding set construction, and then take the resulting cardinality."

Example 14.1.5 (The smallest cardinals)**.** Using the empty set and singleton sets, we immediately obtain

$$0 + \kappa = \kappa, \qquad 1 \cdot \kappa = \kappa, \qquad \kappa^0 = 1.$$

The last identity reflects a familiar fact from Chapter 3: there is exactly one function from the empty set to any set.

Also,

$$0^0 = 1,$$

because there is exactly one function $\varnothing \to \varnothing$, namely the empty function.

Algebraic laws inherited from set constructions

Once the definitions are in place, the formal laws of cardinal arithmetic come from explicit bijections between the corresponding set constructions. This is one of the clearest advantages of defining the operations at the set level first.

Proposition 14.1.6 (The basic algebraic laws for sum and product)**.** *Assume the axiom of choice. For all cardinals κ, λ, μ:*

(i) $\kappa + \lambda = \lambda + \kappa$;

(ii) $(\kappa + \lambda) + \mu = \kappa + (\lambda + \mu)$;

(iii) $\kappa \cdot \lambda = \lambda \cdot \kappa$;

(iv) $(\kappa \cdot \lambda) \cdot \mu = \kappa \cdot (\lambda \cdot \mu)$;

(v) $\kappa \cdot (\lambda + \mu) = \kappa \cdot \lambda + \kappa \cdot \mu$.

Proof. For part (i), define a bijection

$$\big((\kappa \times \{0\}) \cup (\lambda \times \{1\})\big) \to \big((\lambda \times \{0\}) \cup (\kappa \times \{1\})\big)$$

by swapping the tags:

$$\langle \alpha, 0\rangle \mapsto \langle \alpha, 1\rangle, \qquad \langle \beta, 1\rangle \mapsto \langle \beta, 0\rangle.$$

So $\kappa + \lambda = \lambda + \kappa$.

For part (ii), both sides are cardinalities of three tagged copies, namely

$$(\kappa \times \{0\}) \cup (\lambda \times \{1\}) \cup (\mu \times \{2\}).$$

More explicitly, the map that simply forgets the parentheses and records which of the three original tags occurred gives a bijection between the two representatives.

For part (iii), the map

$$\kappa \times \lambda \to \lambda \times \kappa, \qquad (\alpha, \beta) \mapsto (\beta, \alpha)$$

is a bijection.

For part (iv), the map

$$((\alpha, \beta), \gamma) \mapsto (\alpha, (\beta, \gamma))$$

is a bijection from $(\kappa \times \lambda) \times \mu$ to $\kappa \times (\lambda \times \mu)$.

For part (v), write $\lambda + \mu$ as the tagged disjoint union $(\lambda \times \{0\}) \cup (\mu \times \{1\})$. Then

$$\kappa \times \big((\lambda \times \{0\}) \cup (\mu \times \{1\})\big)$$

is in bijection with

$$(\kappa \times \lambda \times \{0\}) \cup (\kappa \times \mu \times \{1\}),$$

by sending

$$(\alpha, (\beta, 0)) \mapsto ((\alpha, \beta), 0), \qquad (\alpha, (\gamma, 1)) \mapsto ((\alpha, \gamma), 1).$$

The right-hand side is a tagged disjoint union of a copy of $\kappa \times \lambda$ and a copy of $\kappa \times \mu$, so its cardinality is $\kappa \cdot \lambda + \kappa \cdot \mu$. □

Proposition 14.1.7 (The basic laws of cardinal exponentiation)**.** *Assume the axiom of choice. Let κ, λ, μ be cardinals. Then:*

(i) $\kappa^1 = \kappa$, $1^\lambda = 1$, *and if* $\lambda > 0$ *then* $0^\lambda = 0$;

(ii) $\kappa^{\lambda+\mu} = \kappa^\lambda \cdot \kappa^\mu$;

(iii) $(\kappa^\lambda)^\mu = \kappa^{\lambda\cdot\mu}$;

(iv) $(\kappa\cdot\lambda)^\mu = \kappa^\mu\cdot\lambda^\mu$.

Proof. For part (i), a function $1 \to \kappa$ is determined by the image of its unique input, so the set of such functions is in bijection with κ. A function $\lambda \to 1$ must be constant, so there is exactly one such function. If $\lambda > 0$, then there is no function from a nonempty set to the empty set, so $0^\lambda = 0$.

For part (ii), let

$$D = (\lambda \times \{0\}) \cup (\mu \times \{1\}).$$

A function $f : D \to \kappa$ is exactly the same thing as a pair consisting of a function $f_0 \colon \lambda \to \kappa$ and a function $f_1 \colon \mu \to \kappa$, obtained by restriction:

$$f_0(\alpha) = f(\alpha, 0), \qquad f_1(\beta) = f(\beta, 1).$$

Conversely, any such pair determines a unique function on D. So the set of functions $D \to \kappa$ is in bijection with $\kappa^\lambda \times \kappa^\mu$.

For part (iii), a function

$$F \colon \mu \to \{g \mid g \colon \lambda \to \kappa\}$$

may be viewed as a function of two variables,

$$\widetilde{F} \colon \mu \times \lambda \to \kappa, \qquad \widetilde{F}(\beta, \alpha) = F(\beta)(\alpha).$$

This is the usual currying correspondence, and it is bijective. Since $\mu \times \lambda$ and $\lambda \times \mu$ have the same cardinality by part (iii) of Proposition 14.1.6, we obtain

$$(\kappa^\lambda)^\mu = \kappa^{\lambda\cdot\mu}.$$

For part (iv), a function $h \colon \mu \to \kappa \times \lambda$ corresponds exactly to a pair of functions obtained by projection:

$$h_1 = \pi_1 \circ h \colon \mu \to \kappa, \qquad h_2 = \pi_2 \circ h \colon \mu \to \lambda.$$

Conversely, any pair (u, v) with $u \colon \mu \to \kappa$ and $v \colon \mu \to \lambda$ determines a unique map

$$h(\gamma) = (u(\gamma), v(\gamma)).$$

So the function set $(\kappa \times \lambda)^\mu$ is in bijection with $\kappa^\mu \times \lambda^\mu$, and taking cardinalities gives the result. □

Remark 14.1.8 (Cardinal arithmetic is not ordinal arithmetic). The notation is intentionally parallel, but the operations are different. For ordinals, Chapter 12 defined $\alpha + \beta$, $\alpha \cdot \beta$, and α^β as order constructions. For cardinals, the same symbols now describe sizes of set-theoretic constructions. Thus

$$\omega + 1 > \omega$$

in ordinal arithmetic, while

$$\aleph_0 + 1 = \aleph_0$$

in cardinal arithmetic, as we will soon prove.

14.2 Finite versus Infinite Arithmetic

The first test for cardinal arithmetic is whether it reproduces ordinary counting when the sets involved are finite. It does. The second test is whether it reveals genuinely new behavior when infinite sets are present. It does that too, and the contrast is one of the most useful lessons of the chapter.

Finite cardinals really do behave like ordinary numbers

Proposition 14.2.1 (Agreement with finite arithmetic)**.** *Let $m, n \in \mathbb{N}_0$ be finite cardinals. Then:*

(i) *$m + n$ in cardinal arithmetic agrees with the sum on $\mathbb{N}_0$ defined in Chapter 6;*

(ii) *$m \cdot n$ in cardinal arithmetic agrees with the product on $\mathbb{N}_0$;*

(iii) *m^n in cardinal arithmetic agrees with the ordinary finite exponentiation on $\mathbb{N}_0$.*

Proof. Take finite sets A and B with $\mathrm{Card}(A) = m$ and $\mathrm{Card}(B) = n$.

By Theorem 7.4.1, the tagged disjoint union $A \sqcup B$ has $m+n$ elements. Therefore the cardinal sum of m and n is the same number $m + n$.

By Theorem 7.4.3, the Cartesian product $A \times B$ has mn elements. So the cardinal product agrees with ordinary multiplication.

Finally, Theorem 7.4.5 says that when A has n elements and B has m elements, the set of functions $A \to B$ has m^n elements. Hence cardinal exponentiation agrees with ordinary finite exponentiation. □

So cardinal arithmetic is not a new arithmetic invented from nowhere. It is an extension of the one we already know, but an extension in which infinite sets make the rules look very different.

Countably infinite self-similarity

The countably infinite world already shows the collapse of finite intuition. The simplest examples come from splitting the natural numbers into two infinite pieces and from arranging the integer lattice in a list.

Theorem 14.2.2 ($\aleph_0 + \aleph_0 = \aleph_0$)**.** *In cardinal arithmetic,*

$$\aleph_0 + \aleph_0 = \aleph_0.$$

Proof. The cardinal $\aleph_0$ is the cardinality of $\mathbb{N}$. So the sum $\aleph_0 + \aleph_0$ is the cardinality of $\mathbb{N} \sqcup \mathbb{N} = (\mathbb{N} \times \{0\}) \cup (\mathbb{N} \times \{1\})$.

Define

$$f : \mathbb{N} \sqcup \mathbb{N} \to \mathbb{N}$$

by

$$f(n, 0) = 2n - 1, \qquad f(n, 1) = 2n.$$

This maps the first copy of $\mathbb{N}$ onto the odd positive integers and the second copy onto the even positive integers. It is therefore a bijection. Hence

$$\mathrm{Card}(\mathbb{N} \sqcup \mathbb{N}) = \mathrm{Card}(\mathbb{N}) = \aleph_0,$$

which is exactly the desired identity. □

Theorem 14.2.3 ($\aleph_0 \cdot \aleph_0 = \aleph_0$)**.** *In cardinal arithmetic,*

$$\aleph_0 \cdot \aleph_0 = \aleph_0.$$

Proof. By definition, $\aleph_0 \cdot \aleph_0$ is the cardinality of $\mathbb{N} \times \mathbb{N}$. But Theorem 8.3.3 shows that $\mathbb{N} \times \mathbb{N}$ is countably infinite. Therefore

$$\mathrm{Card}(\mathbb{N} \times \mathbb{N}) = \aleph_0.$$

So $\aleph_0 \cdot \aleph_0 = \aleph_0$. □

Corollary 14.2.4 (Finite powers of a countable set)**.** *For every positive finite cardinal n,*

$$\aleph_0^n = \aleph_0.$$

Also $\aleph_0^0 = 1$.

Proof. The statement $\aleph_0^0 = 1$ was noted in Example 14.1.5. We prove the positive case by induction on n.

For $n = 1$, Proposition 14.1.7(i) gives $\aleph_0^1 = \aleph_0$.

Assume $\aleph_0^n = \aleph_0$. Then by Proposition 14.1.7(ii) with $\mu = 1$,

$$\aleph_0^{n+1} = \aleph_0^n \cdot \aleph_0^1.$$

Using the inductive hypothesis, Proposition 14.1.7(i), and Theorem 14.2.3, we obtain

$$\aleph_0^{n+1} = \aleph_0 \cdot \aleph_0 = \aleph_0.$$

This completes the induction. □

Example 14.2.5 (A first comparison with ordinal arithmetic)**.** The same visual operation can mean two very different things depending on whether one remembers order or remembers only size.

Expression	*As an ordinal statement*	*As a cardinal statement*
add one point on the left	$1 + \omega = \omega$	$1 + \aleph_0 = \aleph_0$
add one point on the right	$\omega + 1 > \omega$	$\aleph_0 + 1 = \aleph_0$
two finite blocks repeated	$2 \cdot \omega = \omega$	$2 \cdot \aleph_0 = \aleph_0$
two ω-blocks in a row	$\omega \cdot 2 > \omega$	$\aleph_0 \cdot 2 = \aleph_0$

The ordinal column distinguishes where new blocks are placed; the cardinal column counts only how many points occur altogether.

Remark 14.2.6 (Why the countable identities are surprising). For finite sets, combining two nonempty sets always makes something strictly larger. The identities

$$\aleph_0 + \aleph_0 = \aleph_0, \qquad \aleph_0 \cdot \aleph_0 = \aleph_0$$

therefore look paradoxical only if we keep finite intuition too long. What they really say is that a countably infinite set has room to be rearranged into many countably infinite pieces without changing its size.

14.3 Infinite Cardinal Arithmetic with Choice

The countable examples are not isolated curiosities. Under the axiom of choice, the same collapse phenomenon holds for every infinite cardinal: sums and products stop creating genuinely new infinite sizes. The hard part is to prove that an infinite cardinal can absorb a second copy of itself.

Monotonicity and comparison principles

Before proving the main theorem, it is useful to record the fact that all three operations respect injections.

Proposition 14.3.1 (Monotonicity of cardinal operations)**.** *Assume the axiom of choice. Let $\kappa, \kappa', \lambda, \lambda'$ be cardinals.*

(i) If $\kappa \le \kappa'$ and $\lambda \le \lambda'$, then
$$\kappa + \lambda \le \kappa' + \lambda'.$$

(ii) If $\kappa \le \kappa'$ and $\lambda \le \lambda'$, then
$$\kappa \cdot \lambda \le \kappa' \cdot \lambda'.$$

(iii) If $\kappa \le \kappa'$, then
$$\kappa^\lambda \le (\kappa')^\lambda.$$

(iv) If $0 < \kappa$ and $\lambda \le \lambda'$, then
$$\kappa^\lambda \le \kappa^{\lambda'}.$$

Proof. For part (i), let $i\colon \kappa \to \kappa'$ and $j\colon \lambda \to \lambda'$ be injections. Define
$$F\colon (\kappa \times \{0\}) \cup (\lambda \times \{1\}) \to (\kappa' \times \{0\}) \cup (\lambda' \times \{1\})$$
by
$$F(\alpha, 0) = (i(\alpha), 0), \qquad F(\beta, 1) = (j(\beta), 1).$$
This is injective, so the corresponding cardinal inequality follows.

For part (ii), define
$$G\colon \kappa \times \lambda \to \kappa' \times \lambda', \qquad G(\alpha, \beta) = (i(\alpha), j(\beta)).$$
This is injective.

For part (iii), every function $f\colon \lambda \to \kappa$ yields a function $i \circ f\colon \lambda \to \kappa'$. The assignment $f \mapsto i \circ f$ is injective.

For part (iv), because $0 < \kappa$ and κ is a cardinal, we have $0 \in \kappa$. Since $\lambda \le \lambda'$, we may view λ as an initial segment of λ'. Define
$$E\colon \{f \mid f\colon \lambda \to \kappa\} \to \{g \mid g\colon \lambda' \to \kappa\}$$

by extension with the constant value 0:

$$E(f)(\xi) = \begin{cases} f(\xi), & \text{if } \xi < \lambda, \\ 0, & \text{if } \lambda \le \xi < \lambda'. \end{cases}$$

If $E(f) = E(h)$, then restricting to λ gives $f = h$. So E is injective. □

The square of an infinite cardinal

Theorem 14.3.2 (The square of an infinite cardinal)**.** *Assume the axiom of choice. If κ is an infinite cardinal, then*

$$\kappa \cdot \kappa = \kappa.$$

Proof. We already know the theorem for $\kappa = \aleph_0$ by Theorem 14.2.3. We now prove the statement for all infinite cardinals by transfinite induction on the ordinal κ, using the fact that cardinals are initial ordinals.

Fix an infinite cardinal κ, and assume as inductive hypothesis that for every infinite cardinal $\mu < \kappa$, one has $\mu \cdot \mu = \mu$.

We first claim that for every ordinal $\xi < \kappa$,

$$\operatorname{Card}(\xi \times \xi) < \kappa.$$

Let $\mu = \operatorname{Card}(\xi)$. Since κ is initial and $\xi < \kappa$, we have $\mu < \kappa$. If μ is finite, then $\mu \cdot \mu$ is finite and therefore still below the infinite cardinal κ. If μ is infinite, then the inductive hypothesis gives $\mu \cdot \mu = \mu < \kappa$. Because $\xi \approx \mu$, Proposition 14.1.3 shows that $\xi \times \xi$ is equinumerous with $\mu \times \mu$, so $\operatorname{Card}(\xi \times \xi) = \mu \cdot \mu < \kappa$. This proves the claim.

Now define a well-order $\lhd$ on $\kappa \times \kappa$ by ordering pairs first by the maximum of their coordinates and then lexicographically inside each layer. More precisely,

$$(\alpha, \beta) \lhd (\gamma, \delta)$$

if either

(i) $\max\{\alpha, \beta\} < \max\{\gamma, \delta\}$, or

(ii) $\max\{\alpha, \beta\} = \max\{\gamma, \delta\}$ and $\alpha < \gamma$, or

(iii) the maxima and first coordinates agree, and $\beta < \delta$.

This is a well-order because the set of layers κ is well-ordered and each layer is lexicographically well-ordered.

By the discussion in Chapter 11, every well-ordered set is order-isomorphic to an ordinal. Let θ be the order type of $(\kappa \times \kappa, \lhd)$, and let $e : \theta \to \kappa \times \kappa$ be the order isomorphism.

We define an injection $f : \kappa \times \kappa \to \kappa$ by transfinite recursion along θ. Suppose $\xi < \theta$ and that $f(e(\eta))$ has already been defined for all $\eta < \xi$. Write

$$e(\xi) = (\alpha, \beta), \qquad m = \max\{\alpha, \beta\}.$$

Every $\lhd$-predecessor of (α, β) lies in

$$(m \times m) \cup \big(\{m\} \times (m+1)\big) \cup \big((m+1) \times \{m\}\big).$$

By the claim, $m \times m$ has cardinality $< \kappa$. The other two sets each have cardinality at most $\mathrm{Card}(m+1) < \kappa$, because $m + 1 < \kappa$ and κ is initial. Therefore the whole predecessor set has cardinality $< \kappa$.

The already used values of f therefore form a subset of κ of cardinality $< \kappa$. Such a subset cannot be all of κ, so there exists some ordinal below κ not yet used. Let $f(e(\xi))$ be the least such ordinal.

Because at each stage we choose a value not used earlier, the resulting function f is injective.

We now have an injection $f : \kappa \times \kappa \to \kappa$. The map

$$\kappa \to \kappa \times \kappa, \qquad \alpha \mapsto (\alpha, 0)$$

is also injective. Hence the Cantor–Bernstein Theorem (Theorem 13.2.1) implies

$$\kappa \times \kappa \approx \kappa.$$

Therefore $\kappa \cdot \kappa = \kappa$. □

The max law for infinite sums and products

Corollary 14.3.3 (Absorbing a smaller cardinal)**.** *Assume the axiom of choice. Let κ be an infinite cardinal, and let λ be a cardinal with $1 \leq \lambda \leq \kappa$. Then*

$$\kappa + \lambda = \kappa \qquad \textit{and} \qquad \kappa \cdot \lambda = \kappa.$$

Proof. Because $\lambda \leq \kappa$, there is an injection $i : \lambda \to \kappa$.

For the sum, define

$$F : (\kappa \times \{0\}) \cup (\lambda \times \{1\}) \to \kappa \times \kappa$$

by

$$F(\alpha, 0) = (\alpha, 0), \qquad F(\beta, 1) = (i(\beta), 1).$$

This is injective, so

$$\kappa + \lambda \leq \kappa \cdot \kappa = \kappa$$

by Theorem 14.3.2. On the other hand, the first summand injects into the disjoint union, so $\kappa \leq \kappa + \lambda$. Therefore $\kappa + \lambda = \kappa$.

For the product, the map

$$G : \kappa \times \lambda \to \kappa \times \kappa, \qquad G(\alpha, \beta) = (\alpha, i(\beta))$$

is injective. Hence

$$\kappa \cdot \lambda \leq \kappa \cdot \kappa = \kappa.$$

Because λ is a nonzero cardinal, $0 \in \lambda$. So the map

$$\kappa \to \kappa \times \lambda, \qquad \alpha \mapsto (\alpha, 0)$$

is injective, and thus $\kappa \leq \kappa \cdot \lambda$. Therefore $\kappa \cdot \lambda = \kappa$. □

Theorem 14.3.4 (Infinite sums and products are the maximum)**.** *Assume the axiom of choice. Let*

κ and λ be cardinals, and suppose that at least one of them is infinite and neither is 0. Then

$$\kappa + \lambda = \kappa \cdot \lambda = \max\{\kappa, \lambda\}.$$

Proof. Because cardinals are ordinals, they are linearly ordered. Without loss of generality, assume $\lambda \leq \kappa$. Since at least one of the cardinals is infinite and $\lambda \leq \kappa$, the cardinal κ is infinite. Corollary 14.3.3 then shows that

$$\kappa + \lambda = \kappa \qquad \text{and} \qquad \kappa \cdot \lambda = \kappa.$$

But $\kappa = \max\{\kappa, \lambda\}$. □

Corollary 14.3.5 (Finite powers of an infinite cardinal)**.** *Assume the axiom of choice. Let κ be an infinite cardinal. Then for every positive finite cardinal n,*

$$\kappa^n = \kappa.$$

Also $\kappa^0 = 1$.

Proof. The statement $\kappa^0 = 1$ comes from Example 14.1.5. We prove the positive case by induction on n.

For $n = 1$, Proposition 14.1.7(i) gives $\kappa^1 = \kappa$.

Assume $\kappa^n = \kappa$. Then Proposition 14.1.7(ii) yields

$$\kappa^{n+1} = \kappa^n \cdot \kappa.$$

Using the inductive hypothesis and Theorem 14.3.4, we obtain

$$\kappa^{n+1} = \kappa \cdot \kappa = \kappa.$$

This completes the induction. □

Example 14.3.6 (Finite sequences over an infinite alphabet)**.** Let κ be an infinite cardinal, and let $\kappa^{<\omega}$ denote the set of all finite sequences of elements of κ. Then

$$\mathrm{Card}(\kappa^{<\omega}) = \kappa.$$

Indeed, for each fixed length n, the set of sequences of length n has cardinality κ^n, which is κ for $n \geq 1$ by Corollary 14.3.5, and is 1 for $n = 0$. So $\kappa^{<\omega}$ is a countable union of sets of size at most κ, whence

$$\mathrm{Card}(\kappa^{<\omega}) \leq \kappa \cdot \aleph_0 = \kappa$$

by Theorem 14.3.4. The one-term sequences already form a subset of size κ, so equality holds.

Theorem 14.3.7 (Cantor's theorem in cardinal form)**.** *For every cardinal κ,*

$$\kappa < 2^\kappa.$$

Proof. By definition, 2^κ is the cardinality of the set of all functions $\kappa \to 2$. By Proposition 4.4.10 from Chapter 4, this function set is in bijection with the power set $\mathcal{P}(\kappa)$. Therefore

$$2^\kappa = \mathrm{Card}(\mathcal{P}(\kappa)).$$

Cantor's theorem for power sets (Theorem 9.3.1) shows that $\kappa < \mathcal{P}(\kappa)$. Taking cardinalities, we get

$$\kappa < 2^{\kappa}.$$

□

Remark 14.3.8 (Addition and multiplication stabilize; exponentiation does not). Theorem 14.3.4 says that for infinite cardinals, addition and multiplication do not create new sizes. By contrast, Theorem 14.3.7 shows that the operation $\kappa \mapsto 2^{\kappa}$ always jumps upward. So the power-set operation is fundamentally different from the other two.

14.4 The Continuum and $2^{\aleph_0}$

We now return to the real line. Chapter 9 showed that the continuum is larger than $\aleph_0$. Chapter 13 refined that by observing that $\aleph_1 \leq \mathrm{Card}(\mathbb{R})$. Cardinal exponentiation now lets us say exactly what the size of the continuum is.

The continuum as a power set

Definition 14.4.1 (The cardinal of the continuum)**.** Assume the axiom of choice. The cardinal

$$\mathfrak{c} = \mathrm{Card}(\mathbb{R})$$

is called the *cardinal of the continuum*.

Theorem 14.4.2 (The continuum is $2^{\aleph_0}$)**.** *Assume the axiom of choice. Then*

$$\mathfrak{c} = 2^{\aleph_0}.$$

Proof. We show first that $2^{\aleph_0} \leq \mathfrak{c}$. Since $\aleph_0 = \mathrm{Card}(\mathbb{N})$, the cardinal $2^{\aleph_0}$ is $\mathrm{Card}(\mathcal{P}(\mathbb{N}))$. Proposition 9.4.5 from Chapter 9 gives a bijection from $\mathcal{P}(\mathbb{N})$ onto a subset of $(0, 1)$. Because $(0, 1) \subseteq \mathbb{R}$, this yields an injection $\mathcal{P}(\mathbb{N}) \to \mathbb{R}$. Hence

$$2^{\aleph_0} \leq \mathfrak{c}.$$

For the reverse inequality, it is enough by Theorem 9.4.8 to inject $(0, 1)$ into $\mathcal{P}(\mathbb{N})$. Let

$$D = \mathbb{Q} \cap (0, 1).$$

By Theorem 8.3.5, the set D is countably infinite, so $\mathrm{Card}(D) = \aleph_0$ and therefore $\mathrm{Card}(\mathcal{P}(D)) = 2^{\aleph_0}$.

Define

$$\Phi \colon (0, 1) \to \mathcal{P}(D)$$

by

$$\Phi(x) = \{q \in D \mid q < x\}.$$

We claim that Φ is injective. Suppose $x < y$. By the density remark from Chapter 9 (Remark 9.2.5), there exists a rational number q with $x < q < y$. Then $q \notin \Phi(x)$ but $q \in \Phi(y)$, so $\Phi(x) \neq \Phi(y)$. Thus Φ is injective.

Therefore $(0,1) \preceq \mathcal{P}(D)$, so

$$\mathfrak{c} = \mathrm{Card}((0,1)) \leq 2^{\aleph_0}.$$

Together with the first inequality and Corollary 13.3.7, this proves $\mathfrak{c} = 2^{\aleph_0}$. □

Remark 14.4.3 (One cardinal, many disguises). Theorem 14.4.2 says that the following sets all have the same cardinality:

$$\mathbb{R}, \qquad (0,1), \qquad \mathcal{P}(\mathbb{N}), \qquad \{0,1\}^{\mathbb{N}}.$$

They look utterly different, but from the viewpoint of cardinality they all represent the same infinite size.

Binary sequences, Baire space, and Euclidean space

The preceding theorem identifies the continuum with the set of all binary sequences. It is equally natural to ask about sequences of arbitrary natural numbers.

Theorem 14.4.4 (The set of countable sequences of natural numbers has cardinality of the continuum)**.** *Assume the axiom of choice. Then*

$$\aleph_0^{\aleph_0} = 2^{\aleph_0} = \mathfrak{c}.$$

In other words, the set $\mathbb{N}^{\mathbb{N}}$ of all sequences of natural numbers has the same cardinality as the real line.

Proof. Because $2 \leq \aleph_0$, monotonicity in the base (Proposition 14.3.1(iii)) gives

$$2^{\aleph_0} \leq \aleph_0^{\aleph_0}.$$

For the reverse inequality, send each function $f : \mathbb{N} \to \mathbb{N}$ to its graph

$$G_f = \{(n, f(n)) \mid n \in \mathbb{N}\} \subseteq \mathbb{N} \times \mathbb{N}.$$

If $f \neq g$, then there exists some n with $f(n) \neq g(n)$, so $G_f \neq G_g$. Thus the map $f \mapsto G_f$ is injective from $\mathbb{N}^{\mathbb{N}}$ into $\mathcal{P}(\mathbb{N} \times \mathbb{N})$.

By Theorem 8.3.3, the set $\mathbb{N} \times \mathbb{N}$ is countably infinite, so

$$\mathrm{Card}(\mathcal{P}(\mathbb{N} \times \mathbb{N})) = 2^{\aleph_0}.$$

Hence

$$\aleph_0^{\aleph_0} \leq 2^{\aleph_0}.$$

Combining this with the opposite inequality and using Corollary 13.3.7, we obtain

$$\aleph_0^{\aleph_0} = 2^{\aleph_0}.$$

The final equality with $\mathfrak{c}$ follows from Theorem 14.4.2. □

Corollary 14.4.5 (Finite-dimensional Euclidean space has the cardinality of the continuum)**.** *Assume the axiom of choice. For every positive finite cardinal n,*

$$\mathrm{Card}(\mathbb{R}^n) = \mathfrak{c}.$$

Proof. By Definition 14.4.1, we have $\mathrm{Card}(\mathbb{R}) = \mathfrak{c}$. Therefore

$$\mathrm{Card}(\mathbb{R}^n) = \mathfrak{c}^n.$$

Since $\mathfrak{c}$ is an infinite cardinal, Corollary 14.3.5 gives $\mathfrak{c}^n = \mathfrak{c}$ for every positive finite n. □

Example 14.4.6 (Countable sequences of real numbers)**.** The set $\mathbb{R}^{\mathbb{N}}$ of all countable sequences of real numbers also has size $\mathfrak{c}$. Indeed,

$$\mathrm{Card}(\mathbb{R}^{\mathbb{N}}) = \mathfrak{c}^{\aleph_0} = (2^{\aleph_0})^{\aleph_0} = 2^{\aleph_0 \cdot \aleph_0} = 2^{\aleph_0} = \mathfrak{c},$$

where we used Proposition 14.1.7(iii), Theorem 14.2.3, and Theorem 14.4.2.

So even the set of all real-valued sequences is no larger, in cardinal terms, than the real line itself.

Remark 14.4.7 (The next power-set jump). Cantor's theorem applied to $\mathbb{R}$ says that

$$\mathfrak{c} < 2^{\mathfrak{c}} = \mathrm{Card}(\mathcal{P}(\mathbb{R})).$$

So the set of all subsets of the real line is strictly larger than the real line. The continuum is therefore not the end of the transfinite story, but only one rung in a much longer hierarchy.

14.5 A Glimpse of the Continuum Hypothesis and Independence

Once we know that the continuum has size $2^{\aleph_0}$, the most obvious remaining question is how that cardinal sits inside the aleph hierarchy. We know from Chapter 13 that $\aleph_1 \leq \mathfrak{c}$. We also know from Cantor's theorem that $\mathfrak{c} > \aleph_0$. The missing information is whether there is a new cardinal strictly between $\aleph_0$ and the continuum.

The first natural question after Cantor's theorem

Definition 14.5.1 (Continuum hypothesis)**.** Assume the axiom of choice. The *continuum hypothesis* (CH) is the statement

$$2^{\aleph_0} = \aleph_1.$$

Equivalently, CH says that there is no cardinal strictly between $\aleph_0$ and the continuum.

The continuum hypothesis is perhaps the most natural question one can ask after learning that $\mathbb{R}$ is uncountable. Countable sets have size $\aleph_0$. The first uncountable cardinal is $\aleph_1$. The continuum is some uncountable cardinal $2^{\aleph_0}$. Must that next step already be $\aleph_1$?

Proposition 14.5.2 (A useful reformulation of CH)**.** *Assume the axiom of choice. The continuum hypothesis is equivalent to the statement:*

Every uncountable subset of $\mathbb{R}$ has cardinality $\mathfrak{c}$.

Proof. Assume first that CH holds, so $\mathfrak{c} = 2^{\aleph_0} = \aleph_1$. Let $A \subseteq \mathbb{R}$ be uncountable. Then A is not countable, so Theorem 13.4.2 implies $\aleph_0 < \mathrm{Card}(A)$. Because $A \subseteq \mathbb{R}$, we have $\mathrm{Card}(A) \leq \mathfrak{c} = \aleph_1$. There is no cardinal strictly between $\aleph_0$ and $\aleph_1$, so necessarily $\mathrm{Card}(A) = \aleph_1 = \mathfrak{c}$.

Conversely, assume that every uncountable subset of $\mathbb{R}$ has cardinality $\mathfrak{c}$. By Theorem 13.4.4, the cardinal $\aleph_1$ is the least uncountable cardinal. Since $\aleph_1 \leq \mathfrak{c}$, Proposition 13.1.5 from Chapter 13 shows that there exists a subset $A \subseteq \mathbb{R}$ with $\mathrm{Card}(A) = \aleph_1$. The set A is uncountable, so by hypothesis it must satisfy $\mathrm{Card}(A) = \mathfrak{c}$. Therefore $\aleph_1 = \mathfrak{c} = 2^{\aleph_0}$, which is CH. □

Remark 14.5.3 (Why CH feels plausible from both directions). CH can feel plausible because $\aleph_1$ is the first uncountable cardinal, and the real line is the first uncountable set many students meet in practice. But it can also feel implausible, because the power set $\mathcal{P}(\mathbb{N})$ looks much richer than merely "the next step after countable." Both intuitions are reasonable, and one of the surprises of modern set theory is that the usual axioms do not settle the matter.

Independence and why axioms matter

Definition 14.5.4 (Independence from an axiomatic system)**.** A statement is said to be *independent* of an axiomatic system if, assuming the system itself is consistent, the statement can neither be proved nor disproved from those axioms.

Remark 14.5.5 (CH is independent of ZFC). The continuum hypothesis is independent of the usual axioms of set theory, ZFC. More precisely, assuming ZFC is consistent, Kurt G"odel showed that CH cannot be disproved from ZFC, and Paul Cohen later showed that CH cannot be proved from ZFC [25, 26, 27]. Thus ZFC leaves the exact size of the continuum undecided.

Remark 14.5.6 (Independence is not meaninglessness). To say that CH is independent of ZFC is not to say that CH is vague or meaningless. The statement itself is perfectly precise. What fails is not the clarity of the question, but the power of the chosen axioms to answer it. That is one of the deepest reasons set theory eventually had to become axiomatic.

Remark 14.5.7 (A broader pattern). The continuum hypothesis is only the first instance of a broader issue. Cantor's theorem always tells us that $\kappa < 2^\kappa$, but it does not tell us *which* cardinal lies at 2^κ. Questions about power-set sizes therefore belong among the central structural questions of modern set theory [10, 11, 14].

Looking ahead

This chapter defined cardinal addition, multiplication, and exponentiation by the underlying set constructions of disjoint union, Cartesian product, and function set. We saw that finite cardinals obey ordinary arithmetic, while infinite cardinals behave in a sharply different way: under the axiom of choice, sums and products collapse to the larger infinite cardinal, but exponentiation can jump strictly upward. The continuum emerged as the specific cardinal $2^{\aleph_0}$, and the continuum hypothesis appeared as the first natural question about where that cardinal sits in the aleph hierarchy.

The next chapter finally returns to a warning that has hovered over the whole book from the beginning. We have worked intuitively with sets, subsets, functions, power sets, ordinals, cardinals, and products, but what exactly licenses those constructions? Chapter 15 will step back and survey the axioms of ZF and ZFC, explaining both why naive set theory needs restraint and how the formal axioms support the transfinite mathematics developed in this book.

Part V

Foundations, Axioms, and Outlook

Chapter 15

From Intuition to Axioms: ZF and ZFC

At the end of Chapter 1 we made a promise. We said that the early parts of the book would treat sets informally, not because foundations are unimportant, but because foundations become more meaningful after one has already learned what sets, functions, relations, ordinals, and cardinals are supposed to do. By now we have used sets to define the natural numbers, to organize families and products, to compare infinite sizes, to state the axiom of choice, and to carry out transfinite recursion. We have reached the right moment to ask a foundational question in a serious way:

> *What restrictions must we impose on set formation so that the subject remains both powerful and contradiction-free?*

This chapter is not a full course in logic, and it is not meant to replace a systematic text on axiomatic set theory. Its aim is more specific. We want to understand why the naive idea of "a set is any collection describable in words" eventually fails, how the main axioms of ZF repair that failure, why ZFC adds the axiom of choice as a further commitment, and where the language of classes enters.

There is also a pedagogical reason for postponing this discussion until now. If one begins with a formal list of axioms before the reader has a feel for unions, power sets, transfinite stages, or countability, the axioms can look arbitrary. After fourteen chapters of experience, however, they look much less mysterious. Extensionality says that sets are determined by their members, which is exactly how we have reasoned since Chapter 2. The axiom of union legitimizes the operation $\bigcup A$ that we used in Chapter 4 and again in the theory of ordinals. The axiom of infinity guarantees a home for the natural numbers from Chapter 6. Replacement supports the long transfinite constructions of Chapter 12. In short, the axioms are not an alien overlay on the earlier material; they are a disciplined account of what we have been doing.

A second theme will run alongside the list of axioms. Set theory is not merely a catalogue of permissible constructions. It is also a picture of the mathematical universe. The modern picture is cumulative: sets are built in stages from earlier sets, rather than all appearing at once. That picture explains why some collections are too large to be sets at all. We met one important example in Theorem 11.5.6: the ordinals do not form a set. By the end of the chapter we will see how the contrast between *sets* and *proper classes* fits naturally into the axiomatic story.

A final warning is worth stating clearly. The chapter title contains "ZF" and "ZFC," but we will not formalize everything down to the last symbol. Instead we will work at the level appropriate to this book: the axioms will be stated carefully in mathematical English, often with a symbolic companion line, and their roles will be explained through the constructions

you already know. Readers who later study logic in more detail will learn how these statements are coded in first-order language; for now, our goal is conceptual clarity.

15.1 Why Naive Set Theory Needs Restraint

At first glance, the naive picture of a set is irresistible. We say "the set of primes less than 100," "the set of solutions of an equation," or "the set of points in an interval," and ordinary mathematical life proceeds smoothly. Most of the time, this informal language causes no trouble at all. Indeed, without it, the opening chapters of the book would have been unnecessarily heavy.

The problem is not that the naive viewpoint is useless. The problem is that it is too generous if taken literally. If every property were allowed to determine a set, then one could attempt to form collections so large, or so self-referential, that contradiction appears. The basic issue is therefore not whether we may speak informally, but which kinds of set formation are legitimate.

Collections described in words

Definition 15.1.1 (Naive comprehension)**.** The *naive comprehension principle* (or *unrestricted comprehension*) is the informal rule that for every property $P(x)$, there exists a set

$$\{x \mid P(x)\}$$

whose members are exactly the objects satisfying P.

The appeal of this principle is obvious. It matches ordinary language: describe the objects you want, then gather them into a set. Many perfectly harmless examples fit this model.

Example 15.1.2 (Descriptions that look completely safe)**.** Each of the following descriptions feels natural and causes no immediate alarm.

(i) $\{n \in \mathbb{N}_0 \mid n \text{ is even}\}$, the even natural numbers;

(ii) $\{x \in \mathbb{R} \mid 0 \le x \le 1\}$, the closed unit interval;

(iii) $\{p \in \mathbb{N}_0 \mid p \text{ is prime and } p < 100\}$, a finite set of primes;

(iv) $\{f : A \to B \mid f \text{ is bijective}\}$, the bijections from A to B.

The first three are very concrete, and the fourth is simply the familiar practice of selecting certain functions from a previously given function set.

The subtle point is that Example 15.1.2 contains two different kinds of set formation. In the first three examples we quietly started with a known ambient set and then carved out a subset. In the fourth we again worked inside an ambient set, namely the set of all functions from A to B. That is already a much more disciplined rule than unrestricted comprehension.

Remark 15.1.3 (The hidden safety feature). Throughout the early chapters, most displayed set-builder expressions had the form

$$\{x \in A \mid P(x)\}.$$

The phrase "$x \in A$" matters. It means that we are not creating a set out of absolutely everything satisfying P; we are selecting certain elements from an already existing set A. Later

in this chapter, the axiom schema of Separation will formalize exactly this kind of bounded construction.

Why restriction is not merely pedantic

A beginner may wonder whether this distinction is just a matter of style. Why should it matter whether we write $\{x \mid P(x)\}$ or $\{x \in A \mid P(x)\}$? The answer is that the first expression tries to create a set from the whole universe at once, while the second only cuts a piece out of previously available material.

That difference becomes especially vivid if we compare complements. In Chapter 2 we used complements relative to a chosen ambient set. If $A \subseteq U$, then the complement of A in U is

$$U \setminus A = \{x \in U \mid x \notin A\}.$$

This is a bounded construction. By contrast, an *absolute* complement would require a universal set of all objects under discussion. As we shall soon see, that global viewpoint is incompatible with the usual axioms of set theory.

Definition 15.1.4 (Axiomatic set theory)**.** *Axiomatic set theory* is the program of specifying, by explicit axioms, which sets exist and which methods of forming new sets from old ones are allowed.

The point of axiomatic set theory is not to abandon intuition, but to discipline it. One still thinks of sets as collections, but one refuses to accept every verbal description as automatically legitimate. Modern set theory is built on this disciplined compromise; standard treatments include Suppes [9], Enderton [3], and Jech [10].

Remark 15.1.5 (What restraint is supposed to accomplish). A good axiom system for set theory must do two things at once. First, it must be strong enough to support ordinary mathematics: numbers, functions, products, quotients, ordinals, cardinals, and so on. Second, it must be restrictive enough to block the paradoxes that arise from unlimited set formation. The tension between those two goals is what makes the subject foundationally interesting.

15.2 Russell's Paradox and Related Warnings

The best-known warning sign is Russell's paradox. We already met it in Chapter 1, where it served as an honest warning label for our early informal development. Now we revisit it in a more foundational setting. The point is not merely to repeat a clever trick, but to see exactly which assumption fails and what kind of repair is needed.

The contradiction from unrestricted comprehension

Theorem 15.2.1 (Russell's paradox revisited)**.** *If unrestricted comprehension holds, then a contradiction follows.*

Proof. Assume that every property $P(x)$ determines a set $\{x \mid P(x)\}$. Consider the property

$$P(x) :\Longleftrightarrow\ x \notin x.$$

By the assumed principle, there exists a set

$$R = \{x \mid x \notin x\}.$$

Now ask whether $R \in R$.

If $R \in R$, then by the defining condition of R we must have $R \notin R$. If $R \notin R$, then again by the defining condition, $R \in R$. In either case we obtain a contradiction. Therefore unrestricted comprehension cannot be correct. □

Remark 15.2.2 (The real lesson of the paradox). Russell's paradox does *not* show that set theory is impossible. It shows that *unrestricted* set formation is impossible. The problem lies not in the idea of a set, but in the claim that every property determines a set.

Remark 15.2.3 (A self-reference warning). The expression $x \notin x$ looks deceptively simple, but it mixes an object with a membership statement about itself. Axiomatic set theory does not ban all self-reference in ordinary language, but it does refuse to turn every such description into a set.

No universal set

Russell's paradox can be repackaged in a way that is often even more instructive for beginners. Suppose we did not accept unrestricted comprehension, but we still hoped there might be a single set containing all sets. Then bounded subset formation inside that universal set would already recreate the paradox.

Proposition 15.2.4 (There is no universal set in ZF-style set theory)**.** *Assume that for every set A and every property $P(x)$, the subset $\{x \in A \mid P(x)\}$ exists. Then there is no set U such that every set belongs to U.*

Proof. Suppose, toward a contradiction, that such a set U exists. Apply bounded subset formation inside U to the property $x \notin x$. Then the set

$$R = \{x \in U \mid x \notin x\}$$

exists. Since U contains every set, the set R itself belongs to U. Consequently,

$$R \in R \iff R \notin R,$$

which is impossible. Therefore no universal set can exist. □

Example 15.2.5 (Why complements must be relative)**.** If there were a universal set U, then for every set A the absolute complement

$$U \setminus A$$

would be available. Proposition 15.2.4 shows why our earlier complement notation was always relative to a background set. In ZF set theory there is no global ambient set of all objects.

This is one reason the language of modern set theory may feel slightly different from elementary Venn-diagram intuition. In elementary logic texts, a "universe" is often fixed for convenience. In foundational set theory, by contrast, the totality of all sets is too large to be a set.

Other warning signs: ordinals and size

Russell's paradox is not the only reason to reject the idea that every interesting collection is a set. By the time we reached ordinals, we encountered another large collection that cannot itself be a set.

Remark 15.2.6 (The ordinals are too large to form a set). Theorem 11.5.6 showed that there is no set of all ordinals. This is sometimes called the *Burali–Forti phenomenon*. It is not exactly the same argument as Russell's paradox, but it carries the same moral: some collections are simply too large to be sets.

Remark 15.2.7 (Historical note). Russell's paradox became widely known in the early twentieth century and forced mathematicians to rethink the foundations of set theory; see Russell's book [21]. Zermelo's 1908 axiomatization [23] was one of the decisive responses. Later refinements by Fraenkel, Skolem, and others led to the modern system now called ZF; see [13].

15.3 The Core Axioms: Extensionality, Empty Set, Pairing, Union, Power Set, Infinity

Once we accept that not every collection is automatically a set, we must say positively which basic constructions are allowed. The usual axioms of ZF do exactly that. They do not list every set one by one. Instead, they provide a small family of principles from which the familiar set operations of mathematics can be developed.

A useful way to read these axioms is not as arbitrary declarations, but as answers to questions that had already arisen in the earlier chapters. Why were we allowed to identify two sets by their members? Why does the empty set exist? Why may we form $\{a, b\}$, $\bigcup A$, $\mathcal{P}(A)$, or an inductive set from which the natural numbers are built? The next list addresses exactly those issues.

Axiom	*What it guarantees*	*Seen earlier in*
Extensionality	sets are determined entirely by their elements	Chapters 2–5
Empty set	there is a set with no elements	Chapter 2
Pairing	from a and b we can form $\{a, b\}$	Chapters 2–3
Union	from a set of sets we can collect all their members	Chapters 4, 11, 12
Power set	from A we can form the set of all subsets of A	Chapters 2, 9, 14
Infinity	there exists an inductive set	Chapter 6

Extensionality

Definition 15.3.1 (Axiom of Extensionality)**.** One axiom of ZF says that if two sets have exactly the same elements, then they are equal. Symbolically,

$$\forall x\, \forall y\, \Big[(\forall z\, (z \in x \iff z \in y)) \to x = y\Big].$$

This axiom is so natural that beginners often fail to notice that it is an axiom at all. It expresses the basic viewpoint that a set has no hidden internal structure beyond membership.

Example 15.3.2 (Why double inclusion proves equality)**.** In Chapter 2, when we proved $A = B$ by showing $A \subseteq B$ and $B \subseteq A$, we were using extensionality. Indeed, the two inclusions together mean that an object lies in A if and only if it lies in B, and extensionality then gives $A = B$.

Remark 15.3.3 (Names do not matter). The set $\{1, 2, 3\}$ can also be described as $\{n \in \mathbb{N}_0 \mid 1 \leq n \leq 3\}$. Extensionality says that once the members coincide, the two descriptions determine the same set. This is why mathematics can move freely between roster notation and set-builder notation.

The empty set

Definition 15.3.4 (Axiom of the Empty Set)**.** One axiom of ZF asserts that there exists a set with no elements. We denote it by $\varnothing$. Symbolically,

$$\exists x \, \forall y \, (y \notin x).$$

Example 15.3.5 (The role of $\varnothing$)**.** The empty set appeared constantly from the beginning: it is the neutral object for union, it is a subset of every set, and in Chapter 6 it became the von Neumann number 0. The empty set axiom guarantees that this basic object exists.

Remark 15.3.6 (Different presentations). Some axiom systems do not list the empty set separately because it can be obtained from other axioms. There is no harm, however, in naming it explicitly; pedagogically it is often the clearest presentation.

Pairing

Definition 15.3.7 (Axiom of Pairing)**.** For any sets a and b, there exists a set whose elements are exactly a and b. We write this set as $\{a, b\}$. In symbols,

$$\forall a \, \forall b \, \exists p \, \forall x \, (x \in p \iff (x = a \vee x = b)).$$

Example 15.3.8 (Singletons and unordered pairs)**.** If we apply Pairing with $a = b$, we obtain the singleton set $\{a\}$. If $a \neq b$, we obtain the two-element set $\{a, b\}$. Thus a single axiom supports both singleton and pair formation.

Remark 15.3.9 (Why Pairing matters later). Once singleton and pair formation are available, one can code ordered pairs as sets, for example by Kuratowski's definition

$$\langle a, b \rangle = \big\{\{a\}, \{a, b\}\big\}.$$

That construction from Chapter 3 depends ultimately on the possibility of forming singleton and pair sets.

Union

Definition 15.3.10 (Axiom of Union)**.** For every set A, there exists a set whose members are exactly the elements of the members of A. This set is denoted by $\bigcup A$. Symbolically,

$$\forall A\, \exists U\, \forall x\, \big(x \in U \iff \exists B\, (B \in A \wedge x \in B)\big).$$

Example 15.3.11 (General unions)**.** In Chapter 4 we defined general unions such as

$$\bigcup_{i \in I} A_i.$$

If we package the family as the set $\{A_i \mid i \in I\}$, then the axiom of union guarantees that the total collection of all elements belonging to some A_i is again a set.

Remark 15.3.12 (Union in transfinite work). The union axiom reappeared in Chapters 11 and 12, where unions of sets of ordinals produced suprema and limit stages. This is a good illustration of the same axiom serving both elementary and transfinite purposes.

Power set

Definition 15.3.13 (Axiom of Power Set)**.** For every set A, there exists a set $\mathcal{P}(A)$ whose elements are exactly the subsets of A. Symbolically,

$$\forall A\, \exists P\, \forall x\, \big(x \in P \iff x \subseteq A\big).$$

Example 15.3.14 (A familiar source of new sets)**.** The power set axiom underlies the notation $\mathcal{P}(A)$ from Chapter 2. It also lies behind Cantor's theorem in Chapter 9 and the cardinal exponentiation $2^{\aleph_0}$ in Chapter 14. Once A is a set, the collection of all subsets of A is not merely a conceivable idea; the axiom says it exists as a set.

Remark 15.3.15 (Power sets are strong). Among the basic axioms, the power set axiom is one of the strongest in terms of growth. The jump from A to $\mathcal{P}(A)$ was already the key to Cantor's theorem that $\mathcal{P}(A)$ is strictly larger than A. Axiomatic set theory does not suppress this phenomenon; it legitimizes it.

Infinity

Definition 15.3.16 (Axiom of Infinity)**.** One axiom of ZF asserts that there exists an *inductive set*, that is, a set I such that

(i) $\varnothing \in I$, and

(ii) whenever $x \in I$, the successor set $x \cup \{x\}$ also lies in I.

Example 15.3.17 (Why Infinity is needed)**.** Without the axiom of infinity, the earlier construction of the natural numbers in Chapter 6 would have no guaranteed home. Infinity says that at least one set contains

$$\varnothing, \quad \{\varnothing\}, \quad \{\varnothing, \{\varnothing\}\}, \quad \dots$$

and all their later successors.

Remark 15.3.18 (Infinity gives a large enough ambient set). The axiom of infinity does not directly say "the natural numbers exist." It says that some inductive set exists. To isolate the least inductive part of such a set — the von Neumann natural numbers $\mathbb{N}_0$ — one also uses Separation, which appears in the next section.

Remark 15.3.19 (A first summary). The axioms of extensionality, empty set, pairing, union, power set, and infinity already recover much of the concrete set-building that we used intuitively in the first half of the book. But they do not yet explain how to carve definable subsets out of a set, how to collect the values of a definable function on a set, or why membership loops are excluded. Those tasks belong to Separation, Replacement, and Foundation.

15.4 Separation, Replacement, and Foundation

The previous axioms tell us that certain canonical constructions produce sets. But much of actual mathematics proceeds by a more flexible rule: start with an ambient set, then select the members satisfying a further condition; or start with an indexed set of inputs, then collect the corresponding outputs of a definite rule. These are the roles of Separation and Replacement.

Foundation has a different flavor. It does not create sets. Instead, it places a structural restriction on the membership relation itself. Informally, it says that sets are built from earlier sets rather than circling back into themselves.

Separation: subsets cut from an existing set

Definition 15.4.1 (Axiom schema of Separation)**.** For every set A and every property $P(x, p_1, \dots, p_n)$ involving parameters $p_1, \dots, p_n$, there exists a set

$$B = \{x \in A \mid P(x, p_1, \dots, p_n)\}$$

whose elements are exactly those members of A satisfying the property. Symbolically,

$$\forall A\, \exists B\, \forall x \left(x \in B \iff (x \in A \wedge P(x, p_1, \dots, p_n))\right).$$

The word "schema" matters. We do not have one single sentence called Separation; rather, we have one axiom for each admissible formula P. That sounds technical, but the underlying idea is simple: any definable subcollection of a given set is again a set.

Example 15.4.2 (Subsets we formed throughout the book)**.** Separation legitimizes many familiar constructions:

(i) $\{n \in \mathbb{N}_0 \mid n \text{ is even}\}$, the even naturals;

(ii) $A \cap B = \{x \in A \mid x \in B\}$;

(iii) $A \setminus B = \{x \in A \mid x \notin B\}$;

(iv) the preimage $f^{-1}(C) = \{x \in \mathrm{dom}(f) \mid f(x) \in C\}$.

Each of these begins with an already known set and then selects the members satisfying an additional condition.

Remark 15.4.3 (Why Separation blocks Russell's paradox). Separation does *not* allow the unrestricted set $\{x \mid x \notin x\}$. It only allows

$$\{x \in A \mid x \notin x\}$$

for a previously given set A. This boundedness is exactly what prevents the Russell construction from ranging over everything at once.

Remark 15.4.4 (A schema, not a single axiom). The same comment applies later to Replacement. In a careful first-order presentation, ZF contains infinitely many instances of Separation and Replacement, one for each formula of the language of set theory; see Enderton [3] or Jech [10]. For an introductory chapter, however, it is better to remember the guiding idea: definable subsets of a set are sets.

Replacement: images of sets under definite rules

Definition 15.4.5 (Axiom schema of Replacement)**.** Suppose that a property $P(x, y, p_1, \ldots, p_n)$ assigns a unique output y to each input x in a set A. Then the collection of all such outputs is a set. Symbolically, if

$$\forall x \in A \, \exists! y \, P(x, y, p_1, \ldots, p_n),$$

then there exists a set B such that

$$y \in B \iff \exists x \in A \, P(x, y, p_1, \ldots, p_n).$$

If Separation says "you may cut down an existing set," then Replacement says "you may push an existing set through a definite rule." In everyday mathematics, that is the difference between taking a subset and taking an image.

Proposition 15.4.6 (Images of sets are sets)**.** *Assume Replacement. If f is a function and $A \subseteq \mathrm{dom}(f)$ is a set, then the image*

$$f[A] = \{f(a) \mid a \in A\}$$

is a set.

Proof. Consider the property $P(x, y)$ given by $y = f(x)$. Because f is a function, for each $x \in A$ there exists a unique y satisfying $P(x, y)$, namely $f(x)$. By Replacement, the collection of all such outputs is a set. That collection is exactly $f[A]$. □

Example 15.4.7 (Why images matter)**.** In Chapter 3, the range of a function was defined as the image of its domain. Proposition 15.4.6 shows that this ordinary idea has an axiomatic basis: once the domain is a set, the total collection of output values is also a set.

Remark 15.4.8 (Replacement in transfinite constructions). Replacement becomes especially important in Chapters 11 and 12. When we define objects stage by stage along an ordinal, we repeatedly need the earlier values to form a set so that unions, suprema, or further images may be taken. Replacement is one of the axioms that makes those long constructions legitimate rather than merely suggestive.

Foundation: no membership loops

Definition 15.4.9 (Axiom of Foundation)**.** Every nonempty set A has an $\in$-minimal element. That is, if $A \neq \emptyset$, then there exists $a \in A$ such that

$$a \cap A = \emptyset.$$

Equivalently, every nonempty set contains an element having no members in common with the set itself.

The intuitive picture is that the membership relation points downward rather than in circles. Sets are assembled from earlier material; they are not allowed to contain themselves or to form finite membership loops.

Proposition 15.4.10 (Foundation rules out self-membership and finite cycles)**.** *Assume Foundation.*

(i) There is no set x such that $x \in x$.

(ii) There is no finite cycle

$$x_0 \in x_1 \in \cdots \in x_{n-1} \in x_0.$$

Proof. For part (i), suppose that $x \in x$. Then the singleton $\{x\}$ is a nonempty set. By Foundation it has an element a with $a \cap \{x\} = \emptyset$. The only possible choice is $a = x$. But since $x \in x$, we have $x \in x \cap \{x\}$, contradicting $x \cap \{x\} = \emptyset$.

For part (ii), suppose that

$$x_0 \in x_1 \in \cdots \in x_{n-1} \in x_0.$$

Let

$$A = \{x_0, x_1, \ldots, x_{n-1}\}.$$

This set is nonempty. By Foundation there exists $x_k \in A$ with $x_k \cap A = \emptyset$. But the element immediately preceding x_k in the cycle belongs both to x_k and to A, contradicting $x_k \cap A = \emptyset$. □

Corollary 15.4.11 (No infinite descending membership chain)**.** *Assume Foundation and Replacement. Then there is no sequence*

$$x_0 \ni x_1 \ni x_2 \ni x_3 \ni \cdots$$

of sets descending forever by membership.

Proof. Assume such a sequence exists. By Replacement, the image

$$A = \{x_n \mid n \in \mathbb{N}_0\}$$

of $\mathbb{N}_0$ under the map $n \mapsto x_n$ is a set. The set A is nonempty, so by Foundation it contains an element x_k with $x_k \cap A = \varnothing$. But $x_{k+1} \in x_k$ and also $x_{k+1} \in A$, so $x_k \cap A$ is not empty. This contradiction shows that no such descending sequence exists. □

Remark 15.4.12 (Foundation and the cumulative picture). Foundation encourages the view that sets are arranged in cumulative stages. Very roughly, one starts from $\varnothing$, then forms sets of objects already available, then sets of those, and so on. In this picture, membership always points from a later stage to an earlier one. That intuition is developed more systematically in the cumulative hierarchy discussed below.

Remark 15.4.13 (The cumulative hierarchy as a picture of the universe). A standard conceptual picture of set theory is the *cumulative hierarchy*, informally given by

$$V_0 = \varnothing, \qquad V_{\alpha+1} = \mathcal{P}(V_\alpha), \qquad V_\lambda = \bigcup_{\beta<\lambda} V_\beta$$

for limit ordinals λ. The idea is that stage $\alpha + 1$ contains all sets built from what was already available at stage α, while a limit stage gathers together everything from earlier stages. A full development of this picture requires transfinite methods and careful formalization, but even at an intuitive level it explains why modern set theory thinks of sets as *built up* rather than given all at once; see [9, 10].

15.5 Choice, Classes, and the Picture of ZF versus ZFC

We have now met the main axioms that control ordinary set formation and transfinite construction. One major principle remains to be placed in this landscape: the axiom of choice from Chapter 10. That axiom is often separated from the others not because it is less important, but because mathematics can proceed in two parallel styles: with it or without it.

A second issue also comes to the foreground. Once we accept that some collections are too large to be sets, we need language for talking about those collections. This is where classes enter. They let us speak about "all ordinals" or "all sets" without pretending that these totalities are themselves sets.

ZF and ZFC

Definition 15.5.1 (ZF and ZFC)**.** The system *ZF* (Zermelo–Fraenkel set theory) consists of the usual axioms of extensionality, empty set, pairing, union, power set, infinity, separation, replacement, and foundation.

The system *ZFC* is ZF together with the *axiom of choice*.

System	*Description*
ZF	the standard core axioms without the axiom of choice
ZFC	ZF plus the axiom of choice

Remark 15.5.2 (Why Choice is separated from the others). Most mathematicians work comfortably in ZFC, and many classical theorems are stated there by default. But Chapter 10 showed

that Choice has a distinct flavor: it often asserts the existence of global selectors without providing an explicit rule. Because of that, it is customary to state clearly when a theorem uses Choice.

Example 15.5.3 (A theorem that belongs to ZFC rather than bare ZF)**.** The well-ordering theorem from Chapter 10 — every set can be well-ordered — is equivalent to the axiom of choice. So it is not regarded as a theorem of ZF alone. It becomes available exactly when we pass from ZF to ZFC.

Remark 15.5.4 (What earlier chapters used Choice). Chapters 1 through 9 were largely developed without serious dependence on Choice. By contrast, Chapters 10 through 14 used Choice in essential ways when we well-ordered arbitrary sets, identified cardinals with initial ordinals, and studied the arithmetic of arbitrary infinite cardinals. This is one reason the distinction between ZF and ZFC is mathematically meaningful, not merely terminological.

Classes and proper classes

Definition 15.5.5 (Class and proper class)**.** A *class* is an informal or metamathematical way of speaking about a collection of sets specified by a condition. A class is called a *proper class* if it is not a set.

The word "class" lets us express large totalities without forcing them into the category of sets. This is precisely what we need for the collections that earlier chapters proved to be too large.

Example 15.5.6 (Standard proper classes)**.** Three important examples are:

(i) Ord, the class of all ordinals, which is not a set by Theorem 11.5.6;

(ii) Card, the class of all cardinals, which is likewise too large to be a set;

(iii) V, the total collection of all sets, which cannot be a set by Proposition 15.2.4.

Remark 15.5.7 (Classes are not gigantic sets). A proper class is not a very large set sitting somewhere beyond the ordinary ones. In the usual ZF or ZFC framework, classes are a convenient language used from the outside to describe collections of sets. They are not additional objects that may themselves be members of sets. If one wants a formal theory in which classes appear as official objects, one studies systems such as NBG or MK. That lies beyond the scope of this book.

Remark 15.5.8 (Why classes are useful anyway). Although classes are not sets, the language is extremely convenient. It is natural to say "every ordinal belongs to Ord" or "the cumulative hierarchy fills the universe V." This usage lets us keep ordinary mathematical sentences readable while still respecting the fact that some totalities are too large to be sets.

The cumulative universe and foundational perspective

The distinction between sets and proper classes fits neatly with the cumulative picture mentioned in Remark 15.4.13. Each stage V_α is meant to be a set-sized portion of the universe. The whole collection of all stages, however, is not itself a set. In this way the hierarchy builds the universe locally while remaining globally open-ended.

Remark 15.5.9 (How the axioms fit the cumulative picture). The cumulative hierarchy provides an intuitive role for many axioms:

(i) Pairing and Union say that once some objects are available, small ways of grouping them remain available.

(ii) Power Set says that once a stage is available, all of its subsets are available at the next stage.

(iii) Infinity says that the hierarchy does not stop after finitely many steps.

(iv) Separation says that definable subcollections of an existing stage remain within the universe.

(v) Replacement says that definable images of a set-sized stage remain set-sized.

(vi) Foundation says that sets are built from earlier material rather than by circular membership.

This is not a proof of the axioms, but it is an excellent mental model for why they hang together.

Remark 15.5.10 (Philosophical caution). Different philosophers and set theorists interpret the set-theoretic universe in different ways. Some regard the cumulative hierarchy as a picture of a single intended universe of sets; others treat it more as a structural framework. For a beginner, the most important point is much more modest: the axioms are not random. They work together to express a coherent idea of how sets are built and compared; see Potter [14] for a thoughtful discussion.

A brief word about independence

By Chapter 14 we had already met the continuum hypothesis and heard that some natural questions may be undecidable from the usual axioms. The same kind of issue also surrounds the axiom of choice itself.

Remark 15.5.11 (Independence in one sentence). Very roughly speaking, Gödel showed that if ZF is consistent, then ZF cannot disprove Choice or the continuum hypothesis in certain strong forms [25]; Cohen later showed that if ZF is consistent, then ZF cannot prove Choice or the continuum hypothesis either [26, 27]. Thus moving from ZF to ZFC is a genuine additional assumption, not a mere change of notation.

We will return to this foundational horizon in the final chapter. For now, the important practical lesson is simply this: when mathematicians say they are working in ZFC, they are naming the background axiom system within which the whole cumulative story of sets is being developed.

Remark 15.5.12 (A final summary of the chapter). The route from naive to axiomatic set theory can now be described in a single sentence. We begin with intuitive collections, discover through Russell's paradox and related phenomena that unrestricted comprehension fails, replace it by carefully limited axioms for forming sets from sets, and then decide whether or not to add the axiom of choice. The outcome is the standard framework of modern set theory: ZF, or ZFC if Choice is included.

Looking ahead

This chapter deliberately changed the point of view. Instead of asking what new set-theoretic constructions we can perform, we asked why the whole enterprise does not collapse into contradiction. The answer was not a single trick but an entire discipline of set formation: the axioms of ZF, with Choice added when one works in ZFC. We also saw that the language of proper classes provides a natural home for large totalities such as all ordinals or all sets.

The final chapter steps back from technical development and looks across the whole book. We will retrace the path from finite counting to the transfinite, reflect on the role of set theory as a language for later mathematics, and glance at directions that lie beyond this introductory text: descriptive set theory, large cardinals, forcing, and other deeper foundational questions.

Chapter 16

Looking Back and Further Directions

This chapter has a different task from the chapters that came before it. We are not introducing a new central construction on the scale of functions, ordinals, or cardinals. Instead we are stepping back and asking what the whole journey has amounted to. We began with ordinary mathematical language, elementary set operations, and the idea that a function is a rule or a graph. We then moved through relations and orders, built the natural numbers inside set theory, discovered that infinite sets can sometimes be listed and sometimes cannot, used the axiom of choice to connect choice with well-ordering, and finally climbed through ordinals, cardinals, cardinal arithmetic, and the axioms of ZF and ZFC.

Seen from a distance, the story has two great arcs. One arc concerns *size*: finite sets, countable sets, uncountable sets, power sets, cardinals, and ever larger infinities. The other arc concerns *stage* or *order type*: the natural numbers, well-orders, ordinals, transfinite induction, and transfinite recursion. The first arc asks, "How many?" The second asks, "In what order can a process unfold?" One of the central lessons of set theory is that these are not the same question. The sets ω and $\omega + 1$, for example, have the same cardinality but different ordinal structure.

There is also a third theme that may be even more important in the long run. Set theory is not only a subject in its own right; it is a *language* in which large portions of mathematics can be expressed. A sequence is a function on $\mathbb{N}$. A matrix is a function on a finite Cartesian product. A graph is a set with an adjacency relation. A topology is a family of subsets. A group is a set with an operation satisfying laws. Even when later courses do not mention set theory explicitly, their basic objects are quietly built from sets, functions, and relations.

Finally, the last two chapters have shown that the subject does not end with mastering the standard constructions. Once one reaches axiomatic set theory, new questions appear. Which statements can be proved from ZF or ZFC, and which are independent of those axioms? What stronger axioms might one adopt to clarify the infinite further? How does set theory interact with topology, algebra, analysis, logic, and category theory? This final chapter is meant to open those doors without rushing through them. Its aim is not technical mastery, but orientation.

16.1 From Counting to the Transfinite

The natural way to look back over this book is to notice how far the notion of "counting" had to be stretched before it became adequate for modern mathematics. At the beginning, counting seemed almost trivial. A finite set has n elements if it can be put in bijection with $\{0, 1, \ldots, n-1\}$. But Chapter 7 showed that this already contains a first abstraction: we compare size by bijection, not by the names or arrangement of the elements. The set $\{a, b, c\}$ and the set $\{17, 23, 101\}$ have the same size because there is a bijection between them.

Once that abstraction is accepted, the finite world begins to point beyond itself. In Chapter 8 we allowed bijections with $\mathbb{N}$ and obtained the notion of a *countably infinite set*. That move is easy to state and hard to absorb the first time one sees it. The even numbers are no fewer in number than all natural numbers. The integers are listable. The rational numbers are listable. Counting in set theory therefore means "listing in principle," not "ending after finitely many steps."

Then Chapter 9 broke the illusion that all infinities might still be the same size. Cantor's diagonal argument produced a set that escapes every purported list, and the power-set construction $X \mapsto \mathcal{P}(X)$ provided a general mechanism for making a larger set from any given one. The infinite stopped being a single vague idea and became stratified. One then needed a language for comparing these different sizes, and that is what Chapters 13 and 14 supplied.

In parallel, Chapters 11 and 12 showed that infinity is not only about size. A well-ordered process can continue past every finite stage. After all finite stages comes ω; after ω comes $\omega + 1$; after many further stages come $\omega \cdot 2$, ω^2, and so on. Ordinals make precise the idea of proceeding stage by stage through the transfinite, while cardinals make precise the idea of comparing sheer sizes. The two ideas often cooperate, but they answer different questions.

Stage of the story	*Central question*	*Main chapters*
Finite counting	When do two finite sets have the same size?	6–7
Countability	What does it mean to list an infinite set?	8
Uncountability	Which sets escape every list?	9
Well-order and transfinite process	How can a construction continue beyond the finite?	10–12
Cardinals and their arithmetic	How do we compare and combine infinite sizes?	13–14
Axioms	Which set-forming principles justify the whole story?	15

The summary table is useful, but it can also make the subject look more linear than it really is. In practice the themes constantly feed one another. The axiom of choice from Chapter 10 connects arbitrary sets with well-orders. Hartogs' theorem from Chapter 13 shows that every set has some larger well-order type available beyond it. The cumulative picture from Chapter 15 explains why taking power sets and unions over stages are both foundationally natural. Looking back, one sees that the book was not climbing a single ladder but moving through a network of related ideas.

Two endless growth principles

One excellent way to summarize the subject is to notice that set theory contains at least two systematic ways of moving beyond any previously given stage: one by *making a larger collection of subsets*, and the other by *moving to a larger well-order type*.

Proposition 16.1.1 (There is no final stage of set-theoretic growth)**.** *Let X be any set.*

(i) The power set $\mathcal{P}(X)$ has strictly larger cardinality than X.

(ii) There is an ordinal that does not embed into X as a well-ordered subset.

Consequently there is no largest cardinal, and there is no set of all ordinals.

Proof. Part (i) is exactly Cantor's theorem from Chapter 9: there is no surjection from X onto $\mathcal{P}(X)$, so $|X| < |\mathcal{P}(X)|$. Part (ii) is the content of Hartogs' theorem from Chapter 13: from any set X one can construct an ordinal $h(X)$ that is not order-isomorphic to any well-ordered subset of X.

These two statements imply the conclusions immediately. If there were a largest cardinal, applying part (i) to a set of that size would produce a strictly larger one. If there were a set of all ordinals, then applying part (ii) to that set would produce an ordinal not already in it, contradicting Theorem 11.5.6 from Chapter 11. □

Remark 16.1.2 (Why Proposition 16.1.1 matters). Many introductory subjects culminate in a fixed classification or a final formula. Set theory teaches a different kind of lesson. Its basic operations and its basic comparison principles already guarantee that there is no last infinite size and no final transfinite stage. In that sense the subject is inexhaustible from the start.

Order and size are not the same thing

One of the most important conceptual corrections made by set theory is that order and size must be separated. A beginner naturally tends to conflate them. If one list has an extra item tacked on at the end, surely it must be larger. That intuition is correct in the finite world but fails in the infinite world.

Example 16.1.3 (Same size, different order type)**.** The ordinals ω and $\omega + 1$ do not have the same order structure. In $\omega + 1$ there is a greatest element, namely ω, whereas ω has no greatest element. Therefore they cannot be order-isomorphic.

Nevertheless they have the same cardinality. An explicit bijection $f : \omega + 1 \to \omega$ is given by

$$f(\omega) = 0, \qquad f(n) = n + 1 \text{ for every } n \in \omega.$$

This map is one-to-one and onto. Thus ω and $\omega + 1$ have the same size even though they have different order types. More generally, Chapter 12 showed that adding finitely many points to a countably infinite set does not change its cardinality.

Remark 16.1.4 (Two questions, two answers). Cardinals answer the question *how many up to bijection?* Ordinals answer the question *what is the order type of a well-ordered process?* In the presence of the axiom of choice, every set can be well-ordered, and then cardinals can be represented by *initial ordinals*. Even so, the conceptual difference remains essential.

What the book really achieved

A reader finishing this book should not leave with the impression that set theory is merely a collection of tricks about infinite sets. A more accurate summary is this:

> *Set theory gives a disciplined way to build mathematical objects, compare them by functions and relations, and continue those comparisons and constructions beyond the finite.*

That sentence contains nearly everything. “Build” refers to pairing, union, products, power sets, recursion, and axioms. “Compare” refers to inclusion, equivalence relations, orders, injections, surjections, and bijections. “Beyond the finite” refers to countability, uncountability, ordinals, cardinals, and transfinite induction.

By the time one arrives at Chapter 15, the earlier technical material starts to look less like a string of separate topics and more like a single coherent style of thought. The cumulative hierarchy does not merely justify the previous chapters after the fact; it also explains why those chapters fit together so naturally. Union, power set, separation, replacement, and infinity are exactly the kinds of principles needed to support the constructions we carried out.

16.2 Set Theory as a Language for the Rest of Mathematics

A student sometimes asks, after a first course in set theory, whether the subject will later disappear. In one sense the answer is yes. A course in algebra or analysis usually does not spend much time proving again that ordered pairs are sets or that functions are special sets of pairs. But in a deeper sense the answer is no. Set theory remains present as a kind of background grammar. The main objects of later mathematics are usually sets together with extra structure, and the main constructions are usually expressed in terms of subsets, products, relations, functions, and quotients.

Definition 16.2.1 (Mathematical structure)**.** In the broad sense used throughout modern mathematics, a *mathematical structure* is a set, or a collection built from sets, together with specified functions, relations, subsets, or operations satisfying certain laws.

Definition 16.2.1 is intentionally broad. It is not a formal model-theoretic definition. Its purpose is to capture the everyday working viewpoint of mathematics. One starts with an underlying collection of objects and then records how those objects can be compared, combined, measured, or transformed.

Example 16.2.2 (Several familiar structures viewed set-theoretically)**.** The following examples all fit Definition 16.2.1.

(i) A *group* is a set G together with a function $G \times G \to G$, usually written $\langle x, y \rangle \mapsto xy$, satisfying associativity, identity, and inverse laws.

(ii) A *graph* may be encoded by a vertex set V together with an edge relation $E \subseteq V \times V$.

(iii) A *metric space* is a set X together with a distance function $d \colon X \times X \to \mathbb{R}$ satisfying certain axioms.

(iv) A *topological space* is a set X together with a family $\tau \subseteq \mathcal{P}(X)$ of subsets, called the open sets, satisfying closure properties under unions and finite intersections.

(v) A *sequence* of elements of a set A is simply a function $a \colon \mathbb{N} \to A$, or sometimes $a \colon \mathbb{N}_0 \to A$, depending on where indexing begins.

Remark 16.2.3 (The same raw materials keep returning). Example 16.2.2 uses only the kinds of objects we studied in the first five chapters: sets, subsets, Cartesian products, functions, and relations. That is why set theory is so widely useful as a foundational language. It does not

usually supply the special theorems of later subjects, but it does provide a common format in which their objects can be described.

Object	*Underlying set or index set*	*Additional data*
Sequence	$\mathbb{N}$ or $\mathbb{N}_0$	a function into a target set
Matrix	a finite product $I \times J$	a function $I \times J \to \mathbb{R}$
Graph	a vertex set V	an edge relation on V
Order	a set X	a relation $\leqslant$ on X
Topology	a set X	a family $\tau \subseteq \mathcal{P}(X)$
Algebraic structure	a set X	operations and laws on X

Functions organize dependence

Perhaps the most underestimated concept in the whole book is the function. Once one learns to think of functions as first-class objects, large parts of mathematics suddenly become easier to organize. A sequence is a function on $\mathbb{N}$. A parametrized curve is a function from an interval into the plane. A matrix can be seen as a function on a finite grid. A coordinate system is a choice of bijection between an abstract object and a more concrete model. Even the general product $\prod_{i \in I} A_i$ from Chapter 4 is a set of functions.

Example 16.2.4 (A matrix as a function)**.** Suppose one has an $m \times n$ matrix with real entries. If $I = \{1, \dots, m\}$ and $J = \{1, \dots, n\}$, then the matrix is nothing more than a function

$$a : I \times J \to \mathbb{R},$$

where $a(i, j)$ is the entry in row i, column j. The familiar array notation is convenient, but the set-theoretic content is exactly a function on a finite Cartesian product.

Example 16.2.4 is pedagogically important because it shows that set-theoretic language is not reserved for abstract or exotic objects. It reaches back and reorganizes material the reader already knows.

Relations organize sameness and comparison

Relations are just as pervasive. Equality itself is an equivalence relation. Congruence modulo n is an equivalence relation. A partial order is a relation. Adjacency in a graph is a relation. Divisibility on $\mathbb{N}$ is a relation. Once this is noticed, it becomes much easier to see why Chapter 5 was not a detour but a central piece of mathematical language.

Example 16.2.5 (Quotients appear whenever sameness is weakened)**.** Whenever mathematics decides that two different descriptions should count as "the same" for some purpose, an equivalence relation is usually not far away. Fractions identify $\frac{1}{2}$ with $\frac{2}{4}$. Congruence identifies integers that differ by a multiple of n. Geometric symmetry identifies points or figures up to rotation or reflection. In each case one passes from raw objects to *equivalence classes*.

The same pattern recurs throughout mathematics. Quotients in algebra, quotient spaces in topology, and identification of states in geometry or physics all begin from the relation

viewpoint: one decides which objects are to count as indistinguishable, then forms classes or orbits.

Subsets and power sets organize possibility

Subsets seem elementary when first introduced, but they become one of the most flexible tools in all later mathematics. A topology is a selected family of subsets. A measurable structure, studied later in analysis, is again a selected family of subsets. The algebra of events in probability theory is a family of subsets. In geometry and topology one constantly organizes a space by specifying which subsets are open, closed, compact, connected, or otherwise distinguished.

Remark 16.2.6 (Why the power-set operation keeps reappearing). The power set $\mathcal{P}(X)$ is a space of possibilities. Each element of $\mathcal{P}(X)$ is a possible subcollection of X. Once one begins to specify which subcollections have a certain property, one is working inside a power set. That is one reason the operation $X \mapsto \mathcal{P}(X)$ is both mathematically rich and foundationally important.

Set theory as grammar rather than plot

It is useful to distinguish two ways a subject can support later mathematics. Sometimes a subject provides the *plot*: its main results continue to be the focus. At other times it provides the *grammar*: the basic forms in which later ideas are expressed.

Set theory is often the grammar of later mathematics even when it is not the plot.

That is why this book spent so much time on definitions and basic constructions that may have looked humble when they first appeared. Ordered pairs, Cartesian products, functions, general unions, and relations are not merely preliminary formalities. They are part of the common working language of modern mathematics.

16.3 Independence, Large Cardinals, and Further Foundational Questions

Chapter 15 ended with a remarkable change of mood. Up to that point many of our questions had the form "what follows from the concepts and constructions we already know?" But once axioms enter the story, another question becomes unavoidable: "which statements follow from these axioms, and which do not?" The subject becomes not only the study of sets, but also the study of the limits of our formal framework.

Definition 16.3.1 (Independence)**.** Let T be an axiom system and let φ be a statement in the same language. We say that φ is *independent* of T if, assuming T is consistent, neither φ nor $\neg\varphi$ can be proved from T.

Definition 16.3.1 expresses an idea that every student of foundations eventually meets: the axioms may leave some natural questions undecided. This does not mean that the question is badly posed. It means that the current axioms are not strong enough to settle it.

Remark 16.3.2 (Why consistency matters). The phrase "assuming T is consistent" is essential. If an axiom system were inconsistent, then every statement would be provable from it, and independence would become meaningless. Independence is therefore a relative notion: it measures what cannot be decided *within a coherent theory*.

The first independence results one meets

The standard examples for this book are the axiom of choice and the continuum hypothesis. In Chapter 10 we saw that choice has many equivalent forms and many striking consequences. In Chapter 14 we saw that the continuum hypothesis asks whether $2^{\aleph_0}$ is the very next cardinal after $\aleph_0$.

Remark 16.3.3 (Choice and the continuum hypothesis). Gödel showed that if ZF is consistent, then so is ZF+AC+GCH, by working in the constructible universe L [25]. Cohen later introduced forcing and showed that if ZF is consistent, then so is $ZF + \neg AC$, and if ZFC is consistent, then so is $ZFC + \neg CH$ [26, 27]. Thus AC is independent of ZF, and CH is independent of ZFC.

Remark 16.3.4 (What independence teaches philosophically). Independence results force us to distinguish two questions that are easy to blur together. The first is *what is true in the intended universe of sets?* The second is *what can be proved from a given axiom system?* In elementary mathematics those questions often feel as though they coincide. In set theory they can come apart dramatically. Potter [14] offers a thoughtful discussion of this foundational tension.

Forcing as a method of building new universes

The technical machinery of forcing lies far beyond the scope of this book, but the basic idea can still be described meaningfully.

Definition 16.3.5 (Forcing, informal)**.** *Forcing* is a method for passing from one model of set theory to a larger model by adjoining a suitably generic object so as to control which statements become true in the extension.

This informal definition suppresses almost all of the details, but it captures the guiding picture. One begins with a universe of sets that already satisfies the axioms one cares about. One then enlarges that universe in a carefully organized way. The new universe contains old sets plus additional objects, and the extension can be designed so that particular statements hold or fail there. Cohen's proof of the independence of CH used precisely this kind of construction [26, 27]; modern expositions include Kunen [11] and Jech [10].

Remark 16.3.6 (Why forcing was revolutionary). Before forcing, one could hope that the standard axioms might eventually settle every natural question about sets. Forcing showed that the set-theoretic universe is, in a precise technical sense, much more flexible than that hope suggests. It made independence a central theme rather than an occasional curiosity.

Large cardinals and the higher infinite

Once one accepts that ZFC does not settle every natural question, a new kind of problem emerges. Should one add new axioms? If so, which ones, and why should they be believed? One influential family of answers comes from *large cardinal axioms*.

Definition 16.3.7 (Large cardinal axiom, informal)**.** A *large cardinal axiom* is an axiom asserting the existence of a cardinal with very strong structural properties, far beyond what ZFC alone proves.

The word "large" here does not mean merely "bigger than $\aleph_0$." The subject already contains endlessly many cardinals. Instead it means that certain cardinals have exceptionally strong reflection, compactness, or closure properties. Their existence often has consequences far below their own level, especially for the structure of definable sets of reals and for consistency questions.

Example 16.3.8 (Names one encounters after a first course)**.** Among the first large-cardinal notions that a student may hear about are *inaccessible cardinals*, *measurable cardinals*, and *supercompact cardinals*. Their formal definitions require tools beyond this book, but their role is roughly this: they express very strong forms of set existence and often organize the landscape of consistency strength.

Remark 16.3.9 (Why very large infinities affect ordinary mathematics). One of the surprises of modern set theory is that axioms about extremely large cardinals can influence the behavior of sets of real numbers, regularity properties, and classification questions that look much more concrete. Thus the "higher infinite" is not merely remote decoration; it can shape the structure of mathematics closer to the ground. A classic guide to this theme is Kanamori's *The Higher Infinite* [15].

Further foundational questions

Independence and large cardinals do not exhaust the foundational landscape. They point toward a broader collection of questions.

(i) Which statements are *absolute*, in the sense that they remain true when one passes between related models of set theory?

(ii) Which new axioms best capture the intuitive universe of sets, and which are best viewed as useful hypotheses for particular areas?

(iii) How should one compare competing foundational viewpoints: ZFC, stronger large-cardinal frameworks, constructibility-based approaches, or alternative foundations such as category-theoretic ones?

(iv) When a statement is independent, should one search for new axioms to settle it, or accept a pluralism of possible set-theoretic universes?

None of these questions has a one-line answer. They belong to the living frontier of foundations. What matters most for a first reading is not solving them, but learning to recognize them as legitimate and deep mathematical questions.

16.4 Further Directions in Set Theory

Set theory branches in many directions after the material of this book. The branch names can sound intimidating when first encountered, but each one grows out of themes we have already studied. Descriptive set theory expands the study of subsets of $\mathbb{R}$ and other Polish spaces; infinite combinatorics expands our work with countable sets, orders, and partitions; model-theoretic viewpoints expand the idea that axioms may have different universes; category-theoretic viewpoints reconsider the whole role of sets and functions as foundational primitives.

Descriptive set theory

Definition 16.4.1 (Descriptive set theory, informal)**.** *Descriptive set theory* studies subsets of spaces such as $\mathbb{R}$ by analyzing how definable they are and how complicated their construction is.

A first hint of descriptive set theory already appeared in our discussion of the real line. Intervals are simple subsets of $\mathbb{R}$. Countable unions of intervals are still fairly concrete. Taking complements and countable intersections leads to wider and wider classes of sets. The resulting hierarchy begins with the *Borel sets*, which are generated from open intervals by repeatedly taking countable unions, countable intersections, and complements.

What makes descriptive set theory so attractive is that it keeps one foot in concrete analysis and another in deep logic. It asks questions such as these: when must a definable set of reals be measurable or have the property of Baire? When can a definable equivalence relation be classified by countable data? How do games, determinacy, and large cardinals interact with regularity properties of sets of reals? A reader who enjoyed the real line and diagonalization in Chapter 9 will often find descriptive set theory a natural next destination.

Combinatorial set theory

Definition 16.4.2 (Combinatorial set theory, informal)**.** *Combinatorial set theory* studies infinite sets by asking which patterns, colorings, partitions, trees, or order-theoretic configurations must or may occur.

If Chapters 7 through 14 made you ask "what remains true when finite combinatorics is pushed into the infinite?", then you are already thinking like a combinatorial set theorist. Ramsey-type statements, partition relations, stationary and club sets, trees, and singular-cardinal phenomena all belong to this region of the subject. The intuition is that large infinite structures still contain unavoidable patterns, but the precise forms of those patterns can be subtle and surprising.

Remark 16.4.3 (A familiar seed of infinite combinatorics). Even the pigeonhole principle from Chapter 7 can be seen as a finite ancestor of deeper infinite partition principles. One of the recurring themes of combinatorial set theory is that statements which are trivial for finite sets can become profound when the sets are infinite and especially when large cardinals enter the picture.

Model theory and universes of sets

Definition 16.4.4 (Model of set theory, informal)**.** A *model* of a theory such as ZFC is a universe of objects in which the axioms of that theory hold.

The language of models lets us state independence results precisely. Rather than saying vaguely that a statement is "undecidable," one shows that there is one model of the axioms in

which the statement holds and another in which it fails. Forcing and inner-model theory both belong to this wider model-theoretic perspective.

Remark 16.4.5 (Why models matter to a beginner). A first course in mathematics often encourages the feeling that every well-posed statement has a fixed truth value waiting to be discovered. Model-theoretic thinking refines that expectation. It teaches us to ask not only whether a statement is true, but also in *which axiomatic universe* it is true and what assumptions are needed to make it so.

Category-theoretic viewpoints

Definition 16.4.6 (Category-theoretic viewpoint, informal)**.** A *category-theoretic viewpoint* shifts the focus from elements and membership toward objects, morphisms, and the way constructions are characterized by universal mapping properties.

From this perspective, the category **Set** of sets and functions is itself a mathematical object to be studied. Products, coproducts, equalizers, quotients, exponentials, and other constructions are then described not primarily by what their elements look like, but by how maps into and out of them behave.

This does not replace ordinary set theory so much as reinterpret it. Some alternative foundations begin with the category of sets rather than with the membership relation $\in$. The *Elementary Theory of the Category of Sets (ETCS)* is a famous example. More advanced approaches use *toposes*, which support internal forms of logic and set-like reasoning. These ideas lie well beyond our scope, but they are worth knowing about because they show that foundations can be organized around mappings and structure, not only around membership.

Remark 16.4.7 (Why set theory still matters here). Even when one adopts a category-theoretic viewpoint, sets and functions remain the first great example. In that sense the opening chapters of this book still matter. One cannot appreciate why products, quotients, or universal properties are powerful unless one already understands them well in the concrete setting of sets.

A subject with many borders

Set theory has unusually porous borders. It touches topology through the study of spaces of reals and of compactness phenomena. It touches analysis through measure, category, and regularity questions. It touches algebra through infinite combinatorics and through the role of choice in bases, products, and well-orderings. It touches logic through models, forcing, recursion, and definability. It touches philosophy through questions about mathematical existence and the nature of the infinite.

A student should therefore not think of "continuing in set theory" as a single narrow road. Some readers move toward pure logic and foundations. Others move toward topology or analysis and carry set-theoretic methods with them. Others become interested in category theory and alternative foundations. The best next step depends less on chapter titles than on which questions in the book felt most alive.

16.5 Suggested Further Reading

A good reading path after this book depends strongly on what kind of continuation you want. Some readers want a second, more systematic first course; some want axioms and formalism; some want independence and forcing; some want philosophy or history. The references already listed in the bibliography can be grouped accordingly.

For consolidating the basics

For a reader who wants another pass through elementary set theory from a slightly different angle, Halmos's *Naive Set Theory* [2] remains a classic. It is short, elegant, and surprisingly rich for its length. Devlin's *The Joy of Sets* [5] is more expansive and especially good for readers who enjoy seeing basic set theory placed in a broader modern context. Pinter's *A Book of Set Theory* [6] is also a friendly next choice, with a direct and accessible style.

These books make excellent companions for Chapters 2 through 9. They are especially useful if you want a firmer command of the core constructions before moving toward formal axioms.

For a more systematic axiomatic treatment

If Chapter 15 made you want a careful treatment of ZF and ZFC, Enderton's *Elements of Set Theory* [3] is one of the most natural next books. It is clear, direct, and widely used for a first serious course. Hrbacek and Jech's *Introduction to Set Theory* [4] provides another very good bridge from the introductory level to more advanced ideas. Moschovakis's *Notes on Set Theory* [7] is especially valuable if you liked ordinals, recursion, and the conceptual side of the subject.

Suppes's *Axiomatic Set Theory* [9] is an older classic whose style repays careful reading. Levy's *Basic Set Theory* [12] and Fraenkel, Bar-Hillel, and Levy's *Foundations of Set Theory* [13] take the reader further into the axiomatic landscape.

For the axiom of choice and its equivalents

Readers who were especially struck by Chapter 10 may enjoy a focused study of choice itself. Jech's *The Axiom of Choice* [16] is a standard specialized reference, while Herrlich's *Axiom of Choice* [17] presents the subject in a more modern lecture-note style. These books are where the many equivalent formulations of choice, and their surprising uses and nonuses, truly come into view.

For independence, forcing, and advanced set theory

Jech's *Set Theory* [10] and Kunen's *Set Theory* [11] are among the standard advanced references. Both lead the reader into forcing, independence, and the deeper fine structure of the subject. Jech is encyclopedic and broad; Kunen is often used as a graduate-level route into forcing and related techniques. Either one becomes especially meaningful after Chapters 14 and 15, where independence first entered our story.

For the world of large cardinals, Kanamori's *The Higher Infinite* [15] is a landmark. It is not a beginner's book, but it is the standard historical and conceptual guide to how the higher infinite developed and why large-cardinal ideas became central.

For philosophy and foundational reflection

If the last two chapters raised philosophical questions for you, Potter's *Set Theory and Its Philosophy* [14] is highly recommended. It discusses the meaning of set-theoretic language, the status of axioms, the significance of independence, and different ways of thinking about the universe of sets. It is particularly valuable for readers who want to understand not only the mathematics, but also why that mathematics has inspired so much foundational debate.

For historical perspective

Some readers learn best by tracing how the subject actually emerged. Cantor's papers [18, 19] show the birth of the modern theory of infinite sets. Dedekind's essay on the natural numbers [20] remains profound and readable. Zermelo's papers [22, 23] show the early axiomatic turn and the first systematic use of well-ordering and choice. Von Neumann's 1923 paper [24] is a classic source for the ordinal viewpoint used in this book. Gödel's monograph [25] and Cohen's work [26, 27] mark the decisive independence era.

Historical reading should not be treated as decorative. It is often one of the best ways to see which questions were truly difficult and why the modern organization of the subject looks the way it does.

Remark 16.5.1 (How to choose your next book). If you loved Chapters 7 through 9, choose Halmos, Devlin, or Pinter next. If you loved Chapters 11 through 15, choose Enderton, Hrbacek–Jech, or Moschovakis. If the brief discussion of independence caught your imagination, move toward Kunen or Jech. If the foundational questions themselves felt most compelling, read Potter. If you want to see just how high the infinite can rise, keep Kanamori in view as a long-term destination.

Looking ahead

There is no Chapter 17, but there is certainly a next step. The point of an introductory text is not to close a subject but to make later study possible and meaningful. If this book has done its job, words such as *countable*, *power set*, *ordinal*, *cardinal*, *choice*, and *independence* no longer sound like isolated technical vocabulary. They now belong to a connected picture of how mathematics builds, compares, and extends collections.

You may next meet sets in algebra, topology, analysis, logic, category theory, or another course entirely. When that happens, one of the most useful things you can remember from this book is that set theory is not only about collections of objects. It is about the patterns by which mathematics organizes those objects: functions, relations, quotients, products, stages, and sizes. Those patterns will return again and again.

The transfinite, too, should not be remembered only as a catalogue of strange symbols. It is a disciplined extension of ordinary mathematical thinking. We count. We compare. We order. We build by stages. Set theory shows that none of those activities stops where the finite world ends. That is why the subject remains both foundational and endlessly open.

Bibliography

[1] D. J. Velleman, *How to Prove It: A Structured Approach*, 3rd ed., Cambridge University Press, Cambridge, 2019.

[2] P. R. Halmos, *Naive Set Theory*, Undergraduate Texts in Mathematics, Springer, New York, 1974. Originally published by D. Van Nostrand, Princeton, NJ, 1960.

[3] H. B. Enderton, *Elements of Set Theory*, Academic Press, New York, 1977.

[4] K. Hrbacek and T. Jech, *Introduction to Set Theory*, 3rd ed., Monographs and Textbooks in Pure and Applied Mathematics, Vol. 220, Marcel Dekker, New York, 1999.

[5] K. Devlin, *The Joy of Sets: Fundamentals of Contemporary Set Theory*, 2nd ed., Undergraduate Texts in Mathematics, Springer, New York, 1993.

[6] C. C. Pinter, *A Book of Set Theory*, Dover Books on Mathematics, Dover, Mineola, NY, 2014.

[7] Y. N. Moschovakis, *Notes on Set Theory*, 2nd ed., Undergraduate Texts in Mathematics, Springer, New York, 2006.

[8] B. A. Davey and H. A. Priestley, *Introduction to Lattices and Order*, 2nd ed., Cambridge Mathematical Textbooks, Cambridge University Press, Cambridge, 2002.

[9] P. Suppes, *Axiomatic Set Theory*, Dover, New York, 1972. Reprint of the 1960 edition.

[10] T. Jech, *Set Theory*, The Third Millennium Edition, revised and expanded, Springer Monographs in Mathematics, Springer, Berlin, 2003.

[11] K. Kunen, *Set Theory*, Studies in Logic: Mathematical Logic and Foundations, Vol. 34, College Publications, London, 2011.

[12] A. Levy, *Basic Set Theory*, Perspectives in Mathematical Logic, Springer, Berlin, 1979.

[13] A. A. Fraenkel, Y. Bar-Hillel, and A. Levy, *Foundations of Set Theory*, 2nd ed., Studies in Logic and the Foundations of Mathematics, Vol. 67, North-Holland, Amsterdam, 1973.

[14] M. Potter, *Set Theory and Its Philosophy: A Critical Introduction*, Oxford University Press, Oxford, 2004.

[15] A. Kanamori, *The Higher Infinite: Large Cardinals in Set Theory from Their Beginnings*, 2nd ed., Springer Monographs in Mathematics, Springer, Berlin, 2003.

[16] T. Jech, *The Axiom of Choice*, Studies in Logic and the Foundations of Mathematics, Vol. 75, North-Holland, Amsterdam, 1973.

[17] H. Herrlich, *Axiom of Choice*, Lecture Notes in Mathematics, Vol. 1876, Springer, Berlin, 2006.

[18] G. Cantor, "Ueber eine Eigenschaft des Inbegriffs aller reellen algebraischen Zahlen," *Journal für die reine und angewandte Mathematik* **77** (1874), 258–262.

[19] G. Cantor, "Ueber eine elementare Frage der Mannigfaltigkeitslehre," *Jahresbericht der Deutschen Mathematiker-Vereinigung* **1** (1890/91), 75–78.

[20] R. Dedekind, *Was sind und was sollen die Zahlen?*, Vieweg, Braunschweig, 1888. English translation in *Essays on the Theory of Numbers*, Dover, New York, 1963.

[21] B. Russell, *The Principles of Mathematics*, Cambridge University Press, Cambridge, 1903.

[22] E. Zermelo, "Beweis, dass jede Menge wohlgeordnet werden kann (Aus einem an Herrn Hilbert gerichteten Briefe)," *Mathematische Annalen* **59** (1904), 514–516.

[23] E. Zermelo, "Untersuchungen über die Grundlagen der Mengenlehre. I," *Mathematische Annalen* **65** (1908), 261–281.

[24] J. von Neumann, "Zur Einführung der transfiniten Zahlen," *Acta Scientiarum Mathematicarum (Szeged)* **1** (1923), 199–208.

[25] K. Gödel, *The Consistency of the Axiom of Choice and of the Generalized Continuum-Hypothesis with the Axioms of Set Theory*, Annals of Mathematics Studies, No. 3, Princeton University Press, Princeton, NJ, 1940.

[26] P. J. Cohen, "The independence of the continuum hypothesis," *Proceedings of the National Academy of Sciences of the United States of America* **50** (1963), no. 6, 1143–1148.

[27] P. J. Cohen, *Set Theory and the Continuum Hypothesis*, W. A. Benjamin, New York, 1966.

Index

n-element set, 101

absoluteness, 248
addition
    on $\mathbb{N}_0$, 92
$\aleph_0$, 204
algebraic real number, 136
antisymmetric relation, 66
axiom, 14
axiom of choice, 61, 144, 237
axiomatic set theory, 229
axiomatic system, 14

biconditional, 6
bijective function, 42
binary sequence, 133
Borel set, 249
Burali–Forti phenomenon, 231

Cantor's theorem, 131
cardinal addition, 211
cardinal exponentiation, 211
cardinal multiplication, 211
cardinal number, 195, 202, 209
cardinality, 203
    finite set, 103
Cartesian product, 33, 35
    general, 56
category of sets, 250
category theory, 250
chain, 149
characteristic function, 58
choice function, 59
class, 238
codomain, 38
combinatorial set theory, 249
comparable elements, 71
comparison of sets, 196
complement, 23
composition of functions, 41
conjunction, 6
constructible universe, 247
continuum, 136
    cardinal of the, 220
continuum hypothesis, 222
contrapositive, 9
converse, 6
countable choice, 144
countable ordinal, 188
countable set, 118
countably infinite set, 118, 242
counterexample, 8, 12
cover relation, 75
cumulative hierarchy, 237

Dedekind-infinite set, 116
descriptive set theory, 249
diagonal argument, 127
diagonalization, 129
difference, 23
direct proof, 9
disjoint sets, 23
disjoint union, 54
    tagged, 54
disjunction, 6
domain, 38
    of a relation, 65
domain of discourse, 5

element, 17
empty set, 20
enumeration, 125
equinumerous sets, 100
equivalence class, 68, 245
equivalence relation, 67
ETCS, 250

existential quantifier, 7
exponentiation
    on $\mathbb{N}_0$, 97
extensionality, 19

factorial, 111
family of sets, 47
finite ordinal, 168
finite set, 101
first infinite ordinal, 168
first uncountable ordinal, 189
forcing, 247
function, 33, 36

general Cartesian product, 56
general intersection, 49
general union, 49
grammar of mathematics, 246
graph, 244
graph of a function, 37
greatest element, 72
group, 244

Hasse diagram, 75

identity function, 41
image, 38
    of a subset, 39
implication, 6
inaccessible cardinal, 248
incomparable elements, 71
independence, 246
independence from an axiomatic system, 223
index set, 48
indexed family, 48
inductive set, 84, 233
infinite set, 116
informal set theory, 14
initial ordinal, 202, 243
initial segment, 145
injection
    as comparison of size, 196
injective function, 42
intersection, 22
inverse function, 43
inverse image, 39
inverse relation, 65

language of mathematics, 241
large cardinal axiom, 247
least element, 72
lexicographic order, 74
limit ordinal, 169
linear order, *see* total order

mathematical structure, 244
maximal element, 72
measurable cardinal, 248
metric space, 244
minimal element, 72
model of set theory, 249
multiplication
    on $\mathbb{N}_0$, 94

naive comprehension, 228
natural number
    von Neumann natural number, 84
negation, 6

ω, 168
    as cardinal, 204
ω_1, 189
order isomorphism, 158
order type, 159, 241
order-preserving map, 158
ordered pair, 33, 34
ordinal, 164
ordinal addition, 181
ordinal exponentiation, 186
ordinal multiplication, 184

pair set, 20
pairwise disjoint family, 54
partial order, 63, 71
partition, 55
permutation, 111
    of a set, 111
poset, 71
power set, 25
predicate, 4
preimage, 39
principal initial segment, 160
product order, 72
proof, 9
proof by cases, 9
proof by contradiction, 9

proper class, 227, 238
proper subset, 21
proposition, 4

quantifier, 7
quotient set, 69

range
    of a relation, 65
recursive specification, 89
reflexive relation, 66
relation, 63
    from one set to another, 64
    on a set, 64
relative complement, 23
restriction of a function, 46
right inverse, 153
roster notation, 18

section, 153
sequence, 244
set, 3, 17, 227
set of representatives, 152
set-builder notation, 18
singleton, 20
size, 241
size of a set
    finite set, 103
stage, 241
strictly smaller set, 196
subset, 21
successor, 83
successor ordinal, 169
supercompact cardinal, 248
supremum, 171
surjective function, 42
symmetric relation, 66

tagged disjoint union, 210
topological space, 244
topos, 250
total order, 63, 73
transcendental real number, 136
transfinite recursion
    approximation, 178
transitive relation, 66
transitive set, 83
transversal, 152

uncountable set, 128
union, 22
universal quantifier, 7
upper bound, 149
upper bound, 170

vacuous truth, 8
von Neumann natural number, 81

well-order, 76
well-ordered set, 76
witness, 8

ZF, 237
ZFC, 237
Zorn's lemma, 149