CHAPTER 2

Axiomatic Set Theory

2.1 Introduction

Why should students of mathematics want to know something about axiomatic set theory? Here is one answer: set theory provides a natural and efficient framework in which all of contemporary mathematics can be unified. We live in a time when the specialization within mathematics (and many other disciplines, of course) is mind-boggling, and even though that specialization is a tribute to the great success of mathematics and mathematicians, it can be alienating. Not too many centuries ago, the few professional mathematicians generally understood just about all of the mathematics of their day, and many of them also did important work in one or more branches of science and engineering. Now it is almost impossible to be this "broad." Mathematics has major branches such as real analysis, algebraic geometry, number theory, and topology, but it is quite difficult to master even one of these branches. Most mathematicians work in much more narrowly defined areas and, if they are motivated, may be able to understand most of the research in one of these major branches. Again, this observation is not intended as any sort of criticism! I believe the current situation is the almost inevitable consequence of the enormous progress that has been made in almost all areas of mathematics in the last hundred years or so.

Personally, I find it somewhat comforting to know that, formally at least, all the different branches and fragments of mathematics can be nicely packaged together. It may also be valuable to explain this to nonmathematicians who see our discipline as a complex hodgepodge of barely related subjects. Of course, one must not overstate this position. It would be absurd to claim that all of mathematics *is* set theory. The objects of study and the creative activity of most mathematicians are not about sets. Indeed, the great variety of concepts and methods in mathematics makes it all the more remarkable that they can all be embedded in a single, apparently simple theory.

We briefly discussed the elegant simplicity of ZFC set theory in Chapter 1, but the point bears repeating. In the intended interpretation of this first-order theory, the only objects are sets—the same basic "clumps" of things that one learns about in middle school or even elementary school. The only atomic formulas allowed, besides equations, are statements saying that one set is an element of another. With a couple of exceptions, the axioms of ZFC make completely plausible assertions about sets and are not difficult to understand. It seems almost magical that the deepest results of modern mathematics, in all of these varied branches of the subject, can be formally stated and proved in this "sparse" theory.

It is also possible to develop axiomatic set theory with additional variables for objects that are not sets. Such objects are called **urelements**, **individuals**, or **atoms**. For example, we might want variables for natural numbers and perhaps even for real numbers, with appropriate axioms, included as a basic part of our set theory. The obvious way to do this is with a many-sorted first-order theory. There is no harm in this approach, but it doesn't really gain anything either (once the strangeness of developing number systems within pure set theory wears off), so we will not discuss it further in our development of set theory.

Another reason for mathematicians to know something about set theory is that it induces us to think about the meaning of what we do. Specialists in most branches of mathematics do not need to think very often about foundational questions. Number theorists, analysts,

and even algebraists have little incentive to spend much time wondering whether the objects they study are "real," or what their theorems "really mean." But when one starts thinking about set theory, and especially the independence results that tell us, for example, how unlikely it is that we will ever know whether the continuum hypothesis is "true," it becomes natural to ask such questions about the more abstract objects of mathematics. And mental exercises of this sort, while they may be unsettling, are also a valuable philosophical endeavor.

A third reason to be familiar with set theory is that its history is so interesting and so intertwined with developments in other parts of mathematics. In order to highlight this, much of this chapter and Chapter 6 are arranged historically, outlining three major phases in the development of the subject while also presenting the main concepts and results of set theory.

For a more thorough introduction to set theory at an elementary level, see [Gol], [Vau], [Roi], or [Sup]. A more advanced treatment can be found in [Jech78], [JW], or [TZ].

2.2 "Naive" set theory

The first phase in the development of set theory, which extended from the 1870s until about 1900, was marked by the attempts of Dedekind, Georg Cantor, and others to gain acceptance of the use of infinite sets in mathematics. Through his early work on trigonometric series, Cantor came to realize that the efforts of the time to establish a rigorous theory of the real numbers, primarily by Dedekind and Weierstrass, were essentially based on infinite sets. From today's perspective, it seems surprising just how much resistance this new subject sparked. But Carl Friedrich Gauss, certainly the most influential mathematician of the first half of the nineteenth century, shared the ancient Greek "horror of the infinite." Thus he believed that infinite collections should be considered only as incomplete, potential objects, never as completed ones that can be "grasped as a whole." Many mathematicians of the latter

Georg Ferdinand Cantor (1845–1918) is generally considered to be the main founder of set theory. Cantor's father wanted him to study engineering, but Georg was more interested in philosophy, theology and mathematics. Eventually, Cantor concentrated on mathematics and received his doctorate from the University of Berlin in 1867. In 1874, Cantor published one of the first papers that seriously considered infinite sets as actual objects, and he devoted the rest of his career to this subject.

Cantor's work encountered a degree of resistance that, in retrospect, seems quite unfair and regrettable. Kronecker in particular was often vicious in his criticisms of other mathematicians. His attacks on the free use of the infinite angered Weierstrass and Dedekind, but had a more profound effect on Cantor. Kronecker used his influence to block Cantor's applications for positions at Germany's most prestigious universities; thus Cantor spent his entire 44-year career at the relatively minor Halle University. Cantor became exhausted and discouraged by the resistance to his work, and began having bouts of severe depression and mental illness in 1884. Cantor did very little new research during the last thirty years of his life, and even though his work finally received proper recognition after the turn of the century, he died in a mental institution in Halle.

part of the century, notably Leopold Kronecker, shared Gauss's **finitist** philosophy and refused to accept Cantor's radical ideas.

Set theory during this period was based on two very simple axioms. One, called the **comprehension** axiom, says that any collection of objects that can be clearly specified can be considered to be a set. The other axiom, **extensionality**, asserts that two sets are equal if and

only if they have the same elements. This theory was meant to be combined with the rest of mathematics, not to replace it.

Cantor not only viewed infinite sets as actual objects; he defined operations on them and developed an elaborate theory of their sizes, called **cardinal arithmetic**. This program was especially offensive to Kronecker and to many mathematicians of the next generation such as Henri Poincaré. When Russell's amazingly short "paradox" (1902) showed that set theory based on the full comprehension axiom is inconsistent, Poincaré was particularly pleased, stating, "Later mathematicians will regard set theory as a disease from which we have recovered." Hilbert, who supported the new subject, countered by saying that "no one will evict us from the paradise that Cantor has built for us."

Russell's paradox, which Ernst Zermelo actually discovered independently a bit before Russell, begins by letting A be the set of all sets that are not members of themselves. In symbols, $A = \{B \mid B \notin B\}$. By definition of this set, $A \in A$ if and only if $A \notin A$, and so we have a contradiction. (The word paradox, which usually means an *apparent* contradiction, understates the situation here.) While the Burali–Forti paradox (described in the next section) was discovered about five years earlier, it was based on more sophisticated concepts and was not viewed as a major threat to the subject. But Russell's paradox is so simple that it put an end to set theory as it was practiced at the time, which is now called *naive* set theory. Cantor had an inkling of this development, but Frege, who was also a pioneer of the subject, was crushed by Russell's discovery and did no serious mathematics thereafter. (To be fair, there were other important factors in Frege's depression and retirement, such as the death of his wife.)

Russell's paradox was later popularized as the **barber paradox**: in a certain town the barber, who is a man, shaves exactly those men who don't shave themselves. Who shaves the barber? This question has no consistent answer.

Two of the most important problems in modern set theory arose from Cantor's study of cardinality, so we will devote the rest of this section to this topic. Here are the fundamental definitions, which are among Cantor's major contributions. While parts of naive set theory

had to be discarded, the definitions and theorems in the rest of this section are for the most part not in this category and are an essential part of contemporary set theory. More of the basics of cardinal arithmetic are outlined in Appendix C.

Definitions. For any sets A and B:

(a) $A \preceq B$ means there is a one-to-one function from A to B.

(b) $A \sim B$ means there is a **bijection** or **one-to-one correspondence** between A and B, that is, a one-to-one function from A *onto* B.

(c) $A \prec B$ means $A \preceq B$ but not $A \sim B$.

There are many ways of reading $A \sim B$: A and B are **equinumerous**, or **equipollent**, or A and B have the same **cardinality**, or the same **size**. It is clear that this defines an equivalence relation on all sets. Frege defined a **cardinal** as an equivalence class of $\sim$. For example, the cardinal 3 would be the class of all sets with three members. This definition is intuitively appealing, but it is not permissible in ZFC. (We'll say more about this in the next section.) Two other definitions of this term that are permissible in ZFC are given later in this chapter, and in Appendix C.

Cantor was very interested in the ordering on sets based on cardinality. The relation $\preceq$ is a preordering on sets, but it is more natural to think of $\preceq$ and $\prec$ as orderings on cardinals. One of the first nontrivial accomplishments of set theory was to show that this makes sense:

Theorem 2.1 (Cantor–Schröder–Bernstein (CSB) Theorem). *If $A \preceq B$ and $B \preceq A$, then $A \sim B$.*

Proof. Assume that $A \preceq B$ and $B \preceq A$. So there are one-to-one functions $f : A \to B$ and $g : B \to A$. Let $C = Rng(f)$ and $D = Rng(g)$. So $g^{-1} : D \to B$ is a bijection. We will define a bijection $h : A \to B$ such that, for every x in A, $h(x)$ is either $f(x)$ or $g^{-1}(x)$. We always let $h(x) = f(x)$ unless "forced" to do otherwise.

Suppose y is any element of $B - C$, and let $x_1 = g(y)$. If we let $h(x_1) = f(x_1)$, then y will not be in the range of h, because y is not in the range of f, and x_1 is the only element of A such that

$g^{-1}(x_1) = y$. Therefore, we must let $h(x_1) = g^{-1}(x_1)$. But then, by the same reasoning, h must be defined to be g^{-1} on the following elements of A: $x_2 = g(f(x_1))$, $x_3 = g(f(x_2))$, $x_4 = g(f(x_3))$, etc. So, for each element of $B - C$, we get an entire infinite sequence of elements of A on which h must be defined to be g^{-1}. (See Figure 2.1.) On all other elements of A, we let h be f. It is routine to show that this h is indeed a bijection between A and B. ■

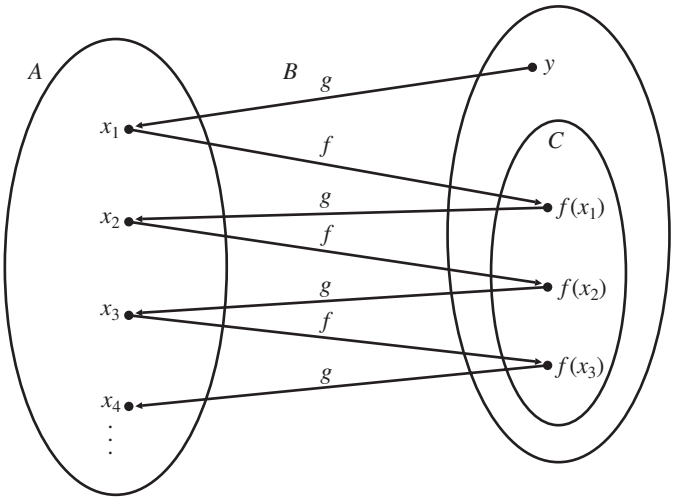

Figure 2.1. Construction of the sequence (x_n) in the proof of the CSB theorem

Example 1. Let's use the proof of the CSB theorem to define a bijection between the intervals $(-1, 1)$ and $[-1, 1]$. We have simple one-to-one functions $f : (-1, 1) \to [-1, 1]$ and $g : [-1, 1] \to (-1, 1)$ defined by $f(x) = x$ and $g(x) = x/2$. In the notation of the above proof, we have $B - C = \{-1, 1\}$. So we must set $h(1/2) = 1$, $h(1/4) = 1/2$, $h(1/8) = 1/4$, etc. Similarly, $h(-1/2) = -1$, $h(-1/4) = -1/2$, $h(-1/8) = -1/4$, etc. For all other elements of $(-1, 1)$, we let $h(x) = x$. It follows that $(-1, 1) \sim [-1, 1]$.

Note that this function h is not continuous. It is not hard to show that there cannot be a continuous bijection between an open interval and a closed interval.

Exercise 1. Complete the details of the proof of the CSB theorem.

The bulk of the proof of the CSB theorem was provided by Cantor's student Felix Bernstein at the age of nineteen. Without this theorem, there would be another natural equivalence relation based on size of a set, defined by ($A \preceq B$ and $B \preceq A$). That the two equivalence relations actually coincide is much more appealing. Another way of stating the CSB theorem is that the relation $\prec$ is strongly antisymmetric: $A \prec B$ and $B \prec A$ cannot hold simultaneously.

Other questions regarding cardinality were more elusive. Cantor naturally hoped to prove that the ordering on sets is total: $\forall A$, $B(A \preceq B \vee B \preceq A)$ or, equivalently, $\forall A, B(A \prec B \vee B \prec A \vee A \sim B)$. He was able to do this, but only by assuming the **well-ordering principle**, that every set can be well ordered. For some time Cantor claimed to have proved this principle from more elementary assumptions, but later he realized that he could not do so.

Here is another important achievement of Cantor's study of cardinality:

Definition. For any set A, its **power set**, denoted $\mathcal{P}(A)$, is the set of all subsets of A.

Theorem 2.2 (Cantor's Theorem). *For any set A, $A \prec \mathcal{P}(A)$.*

Proof. The function $f : A \to \mathcal{P}(A)$ defined by $f(u) = \{u\}$ is clearly one-to-one, so $A \preceq \mathcal{P}(A)$. Now we must show that $A \not\sim \mathcal{P}(A)$. Assume that g is any one-to-one function from A to $\mathcal{P}(A)$. Now let $B = \{u \in A \mid u \notin g(u)\}$. The set B is in $\mathcal{P}(A)$ but cannot be in the range of g, because if we assume that $g(u) = B$, we find that $u \in B$ if and only if $u \notin B$, a contradiction. Thus g is not a bijection, so we conclude that $A \not\sim \mathcal{P}(A)$. ∎

This proof was the first example of a **diagonalization argument**, which has since become a powerful tool. Note the similarity to Rus-

sell's paradox, except that here we don't reach a contradiction. We just show that a certain function can't exist. We will encounter several more diagonalization arguments in this book, mostly in Chapters 3 and 4.

Cantor's theorem implies that there is no largest cardinal, and more specifically that there are **uncountable** sets, sets greater in size than $\mathbb{N}$. Cantor also proved that $\mathbb{R} \sim \mathcal{P}(\mathbb{N})$, so in particular the reals are uncountable. Now, if A is a finite set, say with n elements, then $\mathcal{P}(A)$ has 2^n elements. Unless $n = 0$ or 1, this means that there are sets that are strictly between A and $\mathcal{P}(A)$ in cardinality. But if A is infinite, no such "intermediate" sets present themselves. Cantor conjectured, but could not prove, that there are no sets that are between $\mathbb{N}$ and $\mathcal{P}(\mathbb{N})$ in cardinality. This conjecture is called the **continuum hypothesis** (CH). The more general conjecture obtained by replacing $\mathbb{N}$ with an arbitrary infinite set is called the **generalized continuum hypothesis** (GCH).

Cantor's theorem also provides an alternative proof of the inconsistency of naive set theory, almost as short as Russell's paradox: in naive set theory, we can define the set of all sets A. But then $\mathcal{P}(A)$ must be larger than A in size, which is absurd because $\mathcal{P}(A)$ is clearly a set of sets, and is therefore a subset of A.

Here is a direct proof of the uncountability of $\mathbb{R}$, by a modification of the proof of Cantor's theorem that makes it more clear where the term "diagonalization argument" comes from. In the following proof, we assume for notational simplicity that $0 \notin \mathbb{N}$. Only a slight modification is required if $0 \in \mathbb{N}$.

Proposition 2.3. $\mathbb{N} \prec \mathbb{R}$.

Proof. First of all, $\mathbb{N} \preceq \mathbb{R}$ because $\mathbb{N} \subseteq \mathbb{R}$. To complete the proof, we must show that there is no bijection between $\mathbb{N}$ and $\mathbb{R}$. We will prove a bit more, namely that if $f : \mathbb{N} \to \mathbb{R}$, then the range of f cannot contain the entire interval $[0, 1]$.

So let $f : \mathbb{N} \to \mathbb{R}$. We will construct a real number c between 0 and 1 that is not in the range of f. For each $n \in \mathbb{N}$, let the digit in the nth decimal place of c be obtained by increasing or decreasing the digit in the nth decimal place of $f(n)$ by 5. (This is one of many satisfactory procedures for constructing c.) For instance, if $f(1) = 17.374\ldots$ and

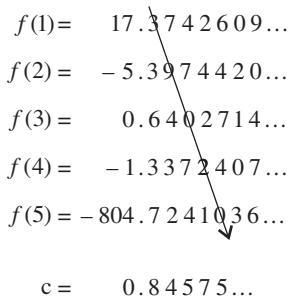

Figure 2.2. Diagonalization argument used to define the number c

$f(2) = -5.397\ldots$, then c begins $0.84\ldots$. (See Figure 2.2.) Since c differs in at least one decimal place from each $f(n)$, c is not in the range of f. ■

This proof assumes the ability to represent real numbers as decimals and glosses over the fact that some reals have two different decimal forms. For instance, $0.999\ldots = 1$ and $7.47999\ldots = 7.48$. This is the only type of ambiguity in decimal representation: real numbers with terminating decimal expansions are the only ones with more than one decimal form.

Exercise 2. Show that the interval [0, 1] is uncountable.

Further discussion of cardinality will be given in Sections 2.5 and 3.2, as well as Appendix C. In Chapter 6 we will return to the two conjectures that eluded Cantor, the well-ordering principle and the continuum hypothesis, and see the prominent role they played in the development of set theory.

2.3 Zermelo–Fraenkel set theory

The second phase in the history of set theory began with efforts to free set theory, and hopefully all of mathematics, from contradictions such as Russell's paradox. Obviously, the thought that a branch of mathe-

matics, especially such a simple-looking one, could turn out to be inconsistent was quite disturbing.

In the early years of the twentieth century, three movements emerged whose goals included ridding mathematics of contradictions: logicism, formalism, and intuitionism. We briefly mentioned logicism and formalism in Chapter 1. Intuitionism, founded by Brouwer, continued and expanded the tradition of Gauss and Kronecker by insisting that mathematical activity should be confined to "constructive" operations. We will discuss intuitionism further in Chapter 8. While all of these movements made important contributions to mathematics, none of them accomplished the main goal.

Attempts to fix set theory were much more successful. Zermelo was the first (in 1908) to create a set of axioms for set theory that replaced the unrestricted comprehension axiom with a more cautious list of principles for the existence of sets. His ideas were refined by Abraham Fraenkel, Thoralf Skolem, John von Neumann, and others in the 1920s, creating the theory ZF that has withstood eighty years of extensive use and scrutiny. Even though Gödel's incompleteness theorem creates a substantial obstacle to proving that ZF is consistent, almost all mathematicians are confident that it is.

We now list the axioms of ZF set theory. For the most part, the axioms are written completely formally, except that we use the standard abbreviation $x \subseteq y$ for $\forall u(u \in x \rightarrow u \in y)$, as well as the restricted quantifier notation introduced in Section 1.3. Also, starting with axiom 5, we will use "terms" to shorten the axioms.

Proper axioms of ZF set theory

1. **Extensionality:** $\forall x, y[x = y \leftrightarrow \forall u(u \in x \leftrightarrow u \in y)]$. (Two sets are equal if and only if they have the same elements.)
2. **Pairing:** $\forall x, y\, \exists z\, \forall u(u \in z \leftrightarrow u = x \vee u = y)$. (For any x and y, the set $\{x, y\}$ exists.)
3. **Union:** $\forall x\, \exists y\, \forall u(u \in y \leftrightarrow \exists w \in x(u \in w))$. (For any x, the union of all the sets in x exists. This union is denoted $\bigcup x$.)

4. **Empty Set:** $\exists x\, \forall y \sim (y \in x)$. (The empty set $\emptyset$ exists.)
5. **Infinity:** $\exists x[\emptyset \in x \wedge \forall y \in x((y \cup \{y\}) \in x)]$. (There exists an infinite set.) We will explain this axiom more fully in the next section. Note that the way we have written this axiom is not within the first-order language of set theory because it includes terms (in the sense of Section 1.4) like $\emptyset$ and $y \cup \{y\}$. This situation is discussed in the two examples following this list of axioms.
6. **Power Set:** $\forall x\, \exists y\, \forall u(u \in y \leftrightarrow u \subseteq x)$. (For any set x, its power set $\mathcal{P}(x)$ exists.)
7. **Replacement:**

$$[\forall x \in a\, \exists! y \mathrm{P}(x, y)] \rightarrow [\exists b\, \forall y(y \in b \leftrightarrow \exists x \in a \mathrm{P}(x, y))].$$

   (If the formula $\mathrm{P}(x, y)$, which cannot contain b as a free variable, defines a function on the domain a, then there is a set b that is the range of this function.) Replacement is an axiom schema since there are infinitely many choices for the formula P.
8. **Regularity or Foundation:** $\forall x[x \neq \emptyset \rightarrow \exists y \in x(x \cap y = \emptyset)]$. (A nonempty set must contain an element that is disjoint from it.) We will thoroughly discuss the significance of this axiom at the end of this section.

ZFC set theory is obtained from ZF by adding one more axiom:

9. **Axiom of Choice (AC):**

$$[\forall u \in x(u \neq \emptyset) \wedge \forall u, v \in x(u \neq v \rightarrow u \cap v = \emptyset)]$$
$$\rightarrow \exists y\, \forall u \in x\, \exists! w \in u(w \in y).$$

   (If x is a set of nonempty, pairwise disjoint sets, then there is a set y that consists of exactly one member of every set in x. Such a y is called a **choice set** for x.)

Example 2. The empty set axiom asserts the existence of a set with no members. By extensionality, this set is unique. Therefore, it is permissible and reasonable to introduce the term $\emptyset$ to denote this set. So the first-order Skolem form of the empty set axiom would be $\forall y(y \notin \emptyset)$.

When we write something like $\emptyset \in x$, as in the axiom of infinity, this is an abbreviation for what would be a much longer formula in the sparse language of ZF, namely: $\exists z[z \in x \wedge \forall y(y \notin z)]$. It would be extremely cumbersome to carry out the development of set theory without introducing terms for sets. The next example will continue the discussion of terms.

There are many other natural statements that are equivalent (in ZF) to the axiom of choice, such as the well-ordering principle and the totality of the ordering on cardinals, mentioned in the previous section. Another concise version of AC is that the Cartesian product of any family of nonempty sets is nonempty. The axiom of choice will be discussed further in Chapter 6.

We will not give a systematic development of basic set theory from the axioms of ZF or ZFC. To see how this is done in detail, refer to [Sup] or [Jech78]. We will just mention a few of the most useful basic results and then move on to more specific topics.

Set-builder notation $\{x \mid \mathrm{P}(x)\}$ (read "the set of all x such that $\mathrm{P}(x)$") is very convenient and commonly used throughout mathematics. Of course, the intended meaning is that if $y = \{x \mid \mathrm{P}(x)\}$, then, for every x, $x \in y$ if and only if $\mathrm{P}(x)$. Russell's paradox makes it clear that we cannot expect this notation to define a set in all cases. The more cautious viewpoint of modern set theory is that we should at least be able to assert the existence of any set of the form $\{x \in a \mid \mathrm{P}(x)\}$, "the set of all x in a, such that $\mathrm{P}(x)$." The idea is that since what we are asserting to exist in this way is a subset of some set a that already exists, we can't end up with a set that is "too big," such as the set of all sets. The principle that such subsets always exist is called the axiom (schema) of **separation**.

So, in ZF or ZFC, one cannot define the set of all sets, the set of all rings, etc. Informally, it's convenient and harmless to refer to such collections as **classes**. If a class is known to be too large to be a set, as these are, then it is called a **proper class**. Specifically, we can talk about the class of all sets x such that $\mathrm{P}(x)$, for any formula $\mathrm{P}(x)$. But there are no variables for such classes, and a proper class can never be a member of a set.

John von Neumann (1903–1957) was born to a well-to-do and intellectual family in Budapest. He was a true child prodigy who could divide eight-digit numbers in his head at age six, and he learned calculus at age eight. He would also show off his photographic memory by reading a page of a telephone book and then repeating all the names, addresses and phone numbers by heart. Von Neumann obtained a university degree in chemical engineering in 1925, and then got his PhD in mathematics just one year later. Like Albert Einstein and Gödel, he left Europe in the 1930s to become a permanent member of the Institute for Advanced Study in Princeton.

Von Neumann made important contributions to many fields within mathematics and science, from the very abstract to the very practical. In the 1920s, he worked with Hilbert and Paul Bernays on the formalist program and the foundations of set theory. When Gödel publicly announced his incompleteness theorem, von Neumann was the first member of the audience to grasp the significance of Gödel's accomplishment. During this period, von Neumann also studied the mathematical foundations of quantum mechanics, and in 1932 published a very successful textbook in that field. His next major achievement was in game theory; he collaborated with the economist Oskar Morgenstern to publish, in 1944, *The Theory of Games and Economic Behavior*, which is the primary reference for modern game theory. During World War II he made substantial contributions to the American atom bomb project, and much of his work in the forties and fifties was on military projects.

From a practical standpoint, von Neumann's most important achievement was his pioneering work on the development of the

(*continued*)

John von Neumann *continued*

computer. Toward the end of World War II, the army built a machine called ENIAC, considered to be the first digital computer. Not only was it huge—over 100 feet long—but it was also extremely awkward and complicated to instruct it what to do. Von Neumann realized that it would be more efficient to give the computer instructions using a "stored program" that one could create outside the computer and then insert into the machine. In other words, he essentially invented the notion of a computer program. Von Neumann built a computer at the Institute (a project that was very controversial at one of the world's "purest" ivory towers), and his work helped IBM develop the machines that launched the computer age. He also pioneered related fields such as cellular automata theory and the theory of self-reproducing machines.

Given his huge output of important work, von Neumann spent a surprising amount of time having fun. His personality was outgoing and friendly, and he loved money, fine wine, women, noise, dirty limericks, and fast cars—several of which he wrecked. He gave large parties, often more than one per week, which were legendary in the Princeton area. In short, von Neumann was a rare combination of genius and "party animal."

In another important version of axiomatic set theory, created by von Neumann, Bernays, and Gödel and therefore called VBG, the variables represent classes, which may be sets or proper classes. But only a set can be a member of a class. So, for instance, Frege's definition of a cardinal becomes acceptable in VBG: the cardinal of any set x is a legitimate object. Unless x is empty, it's a proper class. VBG proves exactly the same theorems about *sets* as ZFC does and, unlike ZF and ZFC, it is finitely axiomatizable. In spite of these desirable features of VBG, most contemporary treatments of set theory use ZFC exclusively, and we will also.

Proposition 2.4. *The full separation schema is derivable in ZF set theory. In other words, for any formula* $\mathrm{Q}(x)$ *in which* b *is not a free variable, the following formula is provable in ZF:*

$$\forall a\, \exists b\, \forall x[x \in b \leftrightarrow (x \in a \wedge \mathrm{Q}(x))].$$

Proof. We give a rather informal proof that can easily be formalized in ZF; technically, the formal proof consists of an infinite set of proofs, one for each Q.

Let a be given. We consider two cases. If there are no members of a for which $\mathrm{Q}(x)$ holds, then let $b = \emptyset$, and we are done.

If there are members of a for which $\mathrm{Q}(x)$ holds, let c be one of these. Now define the formula $\mathrm{P}(x, y)$ to be

$$(\mathrm{Q}(x) \wedge y = x) \vee (\sim \mathrm{Q}(x) \wedge y = c).$$

Then apply the replacement axiom to this P and a. The set b that must exist by replacement is easily shown to be $\{x \in a \mid \mathrm{Q}(x)\}$. ∎

Zermelo's original version of set theory did not include the regularity axiom and had separation instead of replacement. By the previous proposition, Zermelo's theory is a subtheory of ZF, and it turns out to be a proper subtheory, but it is powerful enough to prove the great majority of important mathematical results outside of foundations.

With the exception of extensionality, all of the axioms of ZFC assert the existence of sets with certain properties. The existence of other familiar sets can easily be derived in ZF.

Example 3. The set that is asserted to exist by the pairing axiom is denoted $\{x, y\}$. From this, we can write $\{x\}$ for $\{x, x\}$. The ordinary union of two sets x and y, denoted $x \cup y$, is $\bigcup(\{x, y\})$, whose existence follows from pairing and union. Then $\{x, y, z\} = \{x, y\} \cup \{z\}$, and from this we can define $x \cup y \cup z$, etc. No special axiom is needed for intersections, since their existence follows from separation: $x \cap y = \{z \in x \mid z \in y\}$. Similarly, the set $x - y$, defined as $\{z \in x \mid z \notin y\}$, exists by separation. By the extensionality axiom, all of these terms denote sets that are unique, for any given values of the

variables appearing in them. Therefore, it is natural to think of these terms as defining Skolem functions.

The notation $x - y$ may be read "x minus y," but this set is more correctly called the **relative complement of** y **in** x. There is also the related concept of the **symmetric difference** of any two sets x and y, denoted $x \Delta y$, and defined to be $(x \cup y) - (x \cap y)$ (or, equivalently, $(x - y) \cup (y - x)$. It is worth noting that, in ZF or ZFC, all complements are relative. That is, if x is a set, then $\{z : z \notin x\}$ cannot be a set; it is always a proper class.

Occasionally, one must resort to an artificial definition in order to "embed" some mathematical notion smoothly into ZFC. One such definition is Kazimierz Kuratowski's definition of the **ordered pair** of any two objects: $(x, y) = \{\{x\}, \{x, y\}\}$. The set on the right side of this equation has no conceptual connection with ordered pairs. It is used simply because it allows us to prove, in ZF, the two essential properties of ordered pairs: that the ordered pair of any two sets exists, and that $(x, y) = (u, v)$ if and only if $x = u$ and $y = v$.

Exercise 3. Prove these two properties of ordered pairs, in ZF. Don't try to make your proof too formal, but make sure your steps follow from the axioms of ZF.

Once ordered pairs are available, one can prove (in ZF) the existence of various other important sets, such as the Cartesian product $A \times B$ and the set B^A of all functions from A to B, for any sets A and B.

Exercise 4. Outline a proof (in ZF) that the Cartesian product of any two sets exists. You will need to use the replacement axiom twice and the union axiom once.

The regularity axiom

To conclude this section, we will examine the regularity axiom in some detail. What does it say? What is it about? Superficially, it asserts the existence of a certain type of set, just as all the other axioms of ZFC

except extensionality do. But it really has a different flavor from the other axioms, in that the set asserted to exist is an element of a given set x. So, in an important sense, it doesn't assert the existence of any *new* sets.

The regularity axiom may be viewed as the result of the following line of thinking: naive set theory suffered from paradoxes, and paradoxes in logic and mathematics are almost always traceable to some sort of circular reasoning or definition. In set theory, one is constantly defining sets by specifying their members, and a prudent rule of thumb to avoid circular definitions is to require that all the members of a set must already be defined or "constructed" before we can define that set. This would imply, among other things, that a set cannot be a member of itself. For instance, note that the set of all sets, which we have shown cannot exist because it leads to paradoxes, would violate this principle.

With this in mind, let's consider what regularity says and some of its consequences. One immediate consequence is that no set is a member of itself. For if $y \in y$, then letting $x = \{y\}$ violates this axiom. Another consequence is that we cannot have $y \in z$ and $z \in y$ simultaneously, for then $x = \{y, z\}$ would violate regularity. Generalizing this, the regularity axiom guarantees that there cannot be any finite *cycles* in the relation $\in$, and this is clearly one desirable result if we are trying to eliminate circularity in the construction of sets.

Here is an even more significant consequence of regularity: imagine an infinite sequence of sets $x_0, x_1, x_2, \ldots$ such that $x_{n+1} \in x_n$ for every n. Then the set $\{x_0, x_1, x_2, \ldots .\}$ violates regularity. In other words, no such sequence can exist; we say that regularity prevents **infinite descending sequences** under $\in$. So if we start with any set x_0 and try to generate such a sequence, we inevitably find that $x_n = \emptyset$ for some n. (Remember, every object is a set.) This result corresponds to the notion that if sets may not be defined in a circular way, then they must be defined "from scratch," in "stages." "From scratch" can only mean from the empty set. And, in order for the idea of stages to make sense, it should not be possible to have an infinite sequence of earlier and earlier stages. More mathematically, what regularity says

is precisely that $\in$ is a **well-founded** relation—hence the alternative name foundation. (It is important not to overstate this message. We are not saying that every set must be definable from $\emptyset$ in a finite number of stages. As we will see, there can be infinite *increasing* sequences of sets under $\in$.)

The well-foundedness of $\in$ has important ramifications in set theory. The fact that there are no infinite descending sequences in $\mathbb{N}$ is essentially equivalent to the principle of mathematical induction. In fact, Fermat's **method of infinite descent**, considered the first clear statement of induction, was based on the postulate that every decreasing sequence of natural numbers must terminate. We will soon see that the well-foundedness of $\in$ is useful for the development of the theory of ordinals as well as for embedding the theory of $\mathbb{N}$ in set theory.

2.4 Ordinals

In this section we outline an essential and fascinating part of set theory that is not well known to most mathematicians outside of foundations. In less theoretical treatments, an ordinal is usually defined to be an equivalence class of well-orderings. Here is the more rigorous definition that can be formalized in ZF:

Definitions. A set is **transitive** if every member of it is a subset of it (that is, every member of a member of it is a member of it). An **ordinal** is a transitive set, all of whose members are also transitive.

Example 4. The sets $\emptyset$, $\{\emptyset\}$, and $\{\emptyset, \{\emptyset\}\}$ are ordinals. But the set $\{\{\emptyset\}\}$ is not even transitive, because its only member is not a subset. Why such strange examples? Remember that in pure set theory, all sets must be built up from $\emptyset$.

We now present some basic facts about transitive sets and ordinals. We will usually omit the words "The following is provable in ZF" from the beginning of such results.

Proposition 2.5.

(a) *The power set of a transitive set is transitive.*

(b) *The union and intersection of a collection of transitive sets are transitive.*

Proof.

(a) Assume y is transitive. To show that $\mathcal{P}(y)$ is also transitive, consider $x \in \mathcal{P}(y)$. That means $x \subseteq y$. So if $u \in x$, then $u \in y$. Since y is transitive, this implies $u \subseteq y$, so $u \in \mathcal{P}(y)$. So we have $x \subseteq \mathcal{P}(y)$, as desired. ∎

Exercise 5. Prove part (b) of this proposition.

Notation. Lower case Greek letters are used to denote ordinals. So a statement of the form $\forall\alpha \mathrm{P}(\alpha)$ means that P holds for all ordinals.

Proposition 2.6.

(a) *Every member of an ordinal is an ordinal.*

(b) *$\emptyset$ is an ordinal.*

(c) *For any ordinal α, $\alpha \cup \{\alpha\}$ is also an ordinal.*

Proof.

(a) Assume α is an ordinal, and $x \in \alpha$. Then x is transitive by definition of ordinals. To show x is an ordinal, we must also show that every member of x is transitive. But if $u \in x$, then u is a member of a member of α, and thus a member of α since α is transitive. So u must be transitive because α is an ordinal. ∎

Exercise 6. Prove parts (b) and (c) of this proposition.

The ordinal $\alpha \cup \{\alpha\}$ referred to in (c) of this proposition is called the **successor** of α, denoted $S(\alpha)$. If β is of the form $S(\alpha)$, then we say β is a **successor ordinal**, written $\mathrm{Suc}(\beta)$. If λ is neither empty nor a successor, then we say λ is a **limit ordinal**, written $\mathrm{Lim}(\lambda)$. We also write 0 for the ordinal $\emptyset$.

Exercise 7. Prove that the successor operation is one-to-one, not just on ordinals but on arbitrary sets, that is: $S(x) = S(y) \to x = y$.

Lemma 2.7 (Trichotomy). *Any two ordinals are comparable under $\in$, that is,*

$$\forall \alpha, \beta (\alpha \in \beta \vee \alpha = \beta \vee \beta \in \alpha).$$

Proof. Let's abbreviate what we want to prove as $\forall \alpha, \beta\ \mathrm{C}(\alpha, \beta)$. Assuming it's false, choose α_0 such that $\exists \beta \sim \mathrm{C}(\alpha_0, \beta)$. Then let $A = \{\alpha \in S(\alpha_0) \mid \exists \beta \sim \mathrm{C}(\alpha, \beta)\}$. A is a set by separation, and $A \neq 0$ because $\alpha_0 \in A$. So by regularity, there is an α_1 that is an $\in$-minimal member of A. So every member of α_1 is comparable with every ordinal.

Since α_1 is incomparable to some β, we can choose β_0 that is incomparable to α_1. Just as in the previous paragraph, we can then get an $\in$-minimal β_1 such that $\sim \mathrm{C}(\alpha_1, \beta_1)$. So every member of β_1 is comparable with α_1.

We claim that $\beta_1 \subset \alpha_1$. Assume $\gamma \in \beta_1$. Then γ is an ordinal by Proposition 2.6(a). By definition of β_1, $\mathrm{C}(\gamma, \alpha_1)$. But either $\gamma = \alpha_1$ or $\alpha_1 \in \gamma$ contradicts the fact that $\sim \mathrm{C}(\beta_1, \alpha_1)$. Thus $\gamma \in \alpha_1$. So we have shown that $\beta_1 \subseteq \alpha_1$. Since $\sim \mathrm{C}(\alpha_1, \beta_1)$, we know that $\beta_1 \neq \alpha_1$, so $\beta_1 \subset \alpha_1$.

Now let $\gamma \in (\alpha_1 - \beta_1)$. By definition of α_1, every member of it is comparable to everything. In particular, $\mathrm{C}(\gamma, \beta_1)$. Since $\gamma \notin \beta_1$, we must have $\gamma = \beta_1$ or $\beta_1 \in \gamma$. But each of these possibilities contradicts $\sim \mathrm{C}(\alpha_1, \beta_1)$, so our original assumption must be false. ∎

We have included this rather technical proof because it illustrates the power of regularity in a very typical way. Note that regularity is used twice: to define α_1, and then to define β_1 from α_1.

Notation. When α and β are ordinals, we write $\alpha < \beta$ to mean $\alpha \in \beta$.

Exercise 8. Prove:

(a) $\alpha < \beta \leftrightarrow \alpha \subset \beta$.

(b) $\alpha \leq \beta \leftrightarrow \alpha \subseteq \beta$.

(c) $S(\alpha)$ really is the successor of α. That is, $\alpha < S(\alpha)$, but $\sim \exists\beta(\alpha < \beta < S(\alpha))$.

Theorem 2.8.

(a) *The class of all ordinals is well ordered by* $<$.

(b) *If* $\exists\alpha \mathrm{P}(\alpha)$*, then there is a least* α *such that* $\mathrm{P}(\alpha)$.

Proof.

(a) What we mean by this rather informal statement is that the defining properties of an (irreflexive) well-ordering, with the variables ranging over ordinals, can be proved about $<$, in ZF. So we need to show that $<$ is a partial ordering (irreflexive and transitive), well-founded (every nonempty set of ordinals has a minimal element under $<$), and total for ordinals.

By regularity, we know that $\alpha < \alpha$ is always false, so $<$ is irreflexive. The well-foundedness of $<$ on ordinals also follows immediately from the regularity axiom. If $\alpha < \beta$ and $\beta < \gamma$, then $\alpha < \gamma$ because γ is a transitive set. This shows $<$ is transitive on ordinals. Finally, to establish that $<$ is a total ordering on ordinals rather than just a partial ordering, we need trichotomy, which was proved in Lemma 2.7.

(b) Assume $\exists\alpha \mathrm{P}(\alpha)$. We can't form the set $\{\alpha \mid \mathrm{P}(\alpha)\}$, but we can proceed as in the proof of Lemma 2.7: choose a particular β_0 such that $\mathrm{P}(\beta_0)$, and form the set $\{\alpha \leq \beta_0 \mid \mathrm{P}(\alpha)\}$. This set has a least element, by (a). ■

Corollary 2.9. *Each ordinal is well ordered by* $<$.

Proof. Every initial segment of a well-ordering is a well-ordering. ■

Exercise 9. Prove the following near-converse of this corollary: if x is transitive and is totally ordered by $\in$, then x is an ordinal.

Proposition 2.10. *If* x *is any set of ordinals, then* $\bigcup x$ *is an ordinal, which is also the least upper bound of* x.

Exercise 10. Prove this proposition.

This proposition is quite useful. It tells us that every set of ordinals is bounded above in the class of all ordinals, and in fact has a least upper bound, which is simply its own union.

Corollary 2.11. *There is no set that contains all ordinals.*

Proof. Given any set x, let y be the set of all ordinals in x. Then let $\alpha = S(\bigcup y)$. By the previous proposition, α is an ordinal that contains every ordinal in x. Since $\alpha \notin \alpha$ by regularity, we have $\alpha \notin x$. So x does not contain all ordinals. ∎

The paradox that results from assuming the existence of the set of all ordinals and then arguing as above is called the **Burali–Forti Paradox**. Note that this corollary provides yet another proof, the third we have seen, that there is no set of all sets.

So the ordinals form a very large collection, a proper class (denoted *Ord*), but they are naturally well ordered by the simplest possible binary relation, $\in$. We will see that the ordinals are perfectly suited to represent the "stages" in the construction of sets mentioned earlier. Also, it is not hard to show that every well-ordering of a set is isomorphic to a unique ordinal. So there is a natural bijection between ordinals and equivalence classes of well-orderings (which are the ordinals in the intuitive sense).

We have not shown that there are any limit ordinals; it's time to fix that.

Theorem 2.12. *There exists a limit ordinal.*

Proof. Let x be a set satisfying the axiom of infinity, and define y to be the set of all ordinals in x. We have $0 \in y$ and $\forall\alpha(\alpha \in y \rightarrow S(\alpha) \in y)$. Now let $\beta = \bigcup y = \text{LUB}\,(y)$, as in Proposition 2.10. Since $0 \in \beta$, we know that $\beta \neq 0$. And whenever $\alpha < \beta$, we also have $S(\alpha) < \beta$. Therefore, $\beta \neq S(\alpha)$. In other words, β cannot be a successor. Thus β is a limit ordinal. ∎

It follows from this result that there is a least limit ordinal, which is called ω ("omega"). The members of ω are called **finite** ordinals or

natural numbers. In other words, to a set theorist, $\omega = \mathbb{N}$. Of course, ω and all larger ordinals are called **infinite** ordinals.

We are already writing 0 to mean $\emptyset$. Similarly, 1 means $S(0)$ or $\{0\}$, 2 means $S(1)$ or $\{0, 1\}$, etc. (So the three ordinals mentioned in Example 4 are in fact 0, 1, and 2.) Since each ordinal is the set of all smaller ordinals, each natural number is the set of all smaller natural numbers.

Since the ordinals are well ordered, a principle of proof by induction should hold for them. Here are the two main versions of it:

Theorem 2.13 (Transfinite Induction). *For any formula* $\mathrm{P}(\alpha)$*:*

(a) $\forall\alpha[(\forall\beta < \alpha)\mathrm{P}(\beta) \rightarrow \mathrm{P}(\alpha)] \rightarrow \forall\alpha\mathrm{P}(\alpha)$.

(b) $[\mathrm{P}(0) \wedge \forall\alpha(\mathrm{P}(\alpha) \rightarrow \mathrm{P}(S(\alpha)) \wedge \forall\lambda((\mathrm{Lim}(\lambda) \wedge (\forall\alpha < \lambda)\mathrm{P}(\alpha)) \rightarrow \mathrm{P}(\lambda))] \rightarrow \forall\alpha\mathrm{P}(\alpha)$.

Proof. (a) is just the contrapositive of Theorem 2.8, and is the same principle that we often use for $\mathbb{N}$: if there's no least counterexample to a certain statement, then there's no counterexample at all. Part (b) just takes (a) and breaks it down according to the three types of ordinals (0, successors, and limits). ■

Just as with $\mathbb{N}$, a principle of proof by induction always gives rise to a procedure for defining functions by induction. With ordinals, we may use Theorem 2.13(a) and give a single condition defining $f(\alpha)$ in terms of f's values on the entire domain α. Or we may use Theorem 2.13(b) and give a three-part definition. We may inductively define a function whose domain is some ordinal, or we may (informally, or in VBG) define a proper class function whose domain is all ordinals.

Here are two important binary operations defined on all ordered pairs of ordinals, with a three-part inductive definition on the second variable:

Definition. Ordinal **addition** is defined by induction as follows:

(a) $\alpha + 0 = \alpha$.

(b) $\alpha + S(\beta) = S(\alpha + \beta)$.

(c) For $\mathrm{Lim}(\lambda)$, $\alpha + \lambda = \bigcup_{\beta<\lambda}(\alpha + \beta)$.

Setting $\beta = 0$ in clause (b) implies that $\alpha + 1 = S(\alpha)$ for every α, so we will no longer use the special notation $S(\alpha)$.

Definition. Ordinal **multiplication** is defined by induction as follows:

(a) $\alpha \cdot 0 = 0$.

(b) $\alpha \cdot (\beta + 1) = (\alpha \cdot \beta) + \alpha$.

(c) For $\mathrm{Lim}(\lambda)$, $\alpha \cdot \lambda = \bigcup_{\beta<\lambda}(\alpha \cdot \beta)$.

It is natural to think of α as fixed in these definitions. Parts (a) and (b) of both definitions are the standard inductive definitions for $+$ and $\cdot$ on $\mathbb{N}$, starting from the successor operation. Part (c) just extends this by taking the least upper bound of all previous values when a limit ordinal is reached.

The existence of a set like $\bigcup_{\beta<\lambda}(\alpha+\beta)$ follows from the replacement and union axioms.

What does it mean to define, in ZF, functions like these whose domain is the proper class of all ordinals? Certainly, the function we define is not a set. Technically, when we write such a definition, we are asserting (in ZF) that for any ordinals α and γ, there is a unique function with domain γ that satisfies all the clauses of the definition, for all β in γ (and that one fixed α).

Ordinal arithmetic is an interesting topic in its own right; for one thing, neither $+$ nor $\cdot$ is commutative. We will not go into it further, except to mention that $\alpha + \beta$ is the ordinal whose order type looks like "a copy of α followed by a copy of β," and $\alpha \cdot \beta$ is the ordinal whose order type looks like "β copies of α."

Exercise 11.

(a) Prove that $1 + \omega = \omega$, while $\omega + 1 > \omega$. Thus ordinal addition is not commutative.

(b) Prove that $2 \cdot \omega = \omega$, while $\omega \cdot 2 = \omega + \omega > \omega$. Thus ordinal multiplication is not commutative.

Exercise 12.

(a) Define **ordinal exponentiation**. The definition of α^β should be by transfinite induction on β. The clauses for $\beta = 0$ and $\mathrm{Suc}(\beta)$

should be correct for natural numbers, and the clause for $\mathrm{Lim}(\beta)$ should be the same as for addition and multiplication.

(b) Prove that $\alpha^\beta \cdot \alpha^\gamma = \alpha^{\beta+\gamma}$. (Difficult!)

(c) Prove that $(\alpha^\beta)^\gamma = \alpha^{\beta\cdot\gamma}$. (Difficult!)

It can be shown in ZF that ω, together with the operations $+$, $\cdot$, and S (restricted to ω, of course) and the ordinal 0, satisfies all the axioms of Peano arithmetic. In other words, all the usual mathematics of $\mathbb{N}$ can be carried out in ZF. From there, we can give the usual constructions of $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$, and all the usual mathematics of these number systems can also be carried out in ZF or ZFC. So the ability to carry out all of standard mathematics within axiomatic set theory depends crucially on the theory of ordinals.

2.5 Cardinals and the cumulative hierarchy

In axiomatic set theory, ordinals are also useful for defining some important notions relating to cardinality:

Definitions. A set x is called **finite** if $x \prec \omega$; **denumerable** if $x \sim \omega$; **countable** if $x \preceq \omega$; **infinite** if $\omega \preceq x$; and **uncountable** if $\omega \prec x$.

Note that we have not defined "infinite" and "uncountable" simply as the negations of "finite" and "countable." The next few results will clarify these terms. We will indicate those results whose proofs require the axiom of choice. Make sure not to confuse the relations $<$ and $\leq$ between ordinals with the relations $\prec$ and $\preceq$ between arbitrary sets.

Lemma 2.14. *For any $n \in \omega$, $n \prec n+1$.*

Proof. Since $n \subseteq n+1$, we have $n \preceq n+1$. We also need $n \not\sim n+1$, which we will prove by ordinary induction on n.

Since $0 = \emptyset$ and $1 = \{\emptyset\}$, the only function from 0 to 1 is the function $\emptyset$. This function is not onto 1, so $0 \not\sim 1$.

Now assume $n \not\sim n+1$. We want to show that $n+1 \not\sim n+2$. Assume on the contrary that f is a bijection between $n+1$ and $n+2$.

If $f(n) = n + 1$, then $f - \{(n, n + 1)\}$ is clearly a bijection between n and $n + 1$, contradicting the induction hypothesis. If $f(n) \neq n + 1$, let $k = f(n)$, and $m = f^{-1}(n + 1)$. Now define g to be the relation $[f \cup \{(m, k)\}] - \{(n, k), (m, n+1)\}$. It is easy to see that g is a bijection between n and $n + 1$, again contradicting the induction hypothesis. ■

Corollary 2.15. *If $m < n < \omega$, then $m \prec n \prec \omega$.*

Lemma 2.16. *Let $x \subseteq \omega$. Then:*

(a) *x is finite if and only if it is bounded above.*

(b) *x is denumerable if and only if it is unbounded above.*

Exercise 13. Prove the previous corollary and lemma.

Proposition 2.17. *For any set x,*

(a) *x is finite if and only if $x \sim n$ for some $n \in \omega$.*

(b) *x is infinite if and only if it has a proper subset of the same cardinality as x itself. (Dedekind used this as the definition of infinite sets.)*

(c) *If x is infinite, then it is not finite.*

(d) *(AC) x is infinite if and only if it is not finite.*

(e) *If x is uncountable, then it is not countable.*

(f) *(AC) x is uncountable if and only if it is not countable.*

Proof.

(a) This follows easily from the previous corollary and lemma, and is left as an exercise (see below).

(b) Assume x is infinite. So there is a one-to-one function f from ω to x. Clearly, ω itself is "Dedekind infinite." For example, the function $g(n) = 2n$ is a bijection between ω and the set of even numbers in ω. Now define $h : x \to x$ by

$$h(u) = \begin{cases} u, & \text{if } u \notin Rng(f), \\ f(g(f^{-1}(u))), & \text{otherwise.} \end{cases}$$

It is easy to show that h is a bijection between x and a proper subset of x. (See the next exercise.)

For the converse, assume $f : x \to y$, where f is one-to-one and $y \subset x$. Let c be any element of $x - y$. Define $g : \omega \to x$ inductively by $g(0) = c$, and $g(n + 1) = f(g(n))$. It is easy to show that g is one-to-one. Hence $\omega \preceq x$, making x infinite. (See the next exercise.)

Parts (c) and (e) are trivial. Parts (d) and (f) follow immediately, assuming the fact (which we will not prove) that AC implies that all sets are comparable by cardinality. ■

Exercise 14. Prove part (a) of this proposition, and complete the proof of part (b).

The phenomenon described in part (b) of this proposition was first described by Galileo in the early 1600s, and for over two centuries thereafter it was viewed as paradoxical and an argument against the use of infinite sets. Nowadays, this strange property is viewed as a fact of mathematical life.

Example 5. Since $\mathbb{N}$ (that is, ω) is infinite, it must have proper subsets of the same cardinality. In fact, Lemma 2.16(b) tells us that the set of even natural numbers, the set of primes, the set of perfect squares, etc., are just as big as all of $\mathbb{N}$.

Exercise 15. Appendix C shows that $\mathbb{N} \times \mathbb{N} \sim \mathbb{N}$. (See the discussion following Proposition C.1.) Use this fact and the CSB theorem to prove that $\mathbb{Q} \sim \mathbb{N}$. Many people find this especially surprising because, on a number line, there are an infinite number of rationals between each pair of whole numbers.

Exercise 16. This amusing scenario was concocted by Hilbert to illustrate the surprises that are inherent in the study of infinite sets. You are the desk clerk at Hilbert Hotel, which has a denumerable number of single rooms (numbered 1, 2, 3, etc.), and is currently full.

(a) Suddenly a man comes in, desperately wanting a room. At first you tell him that he can't have one because the hotel is full, but then you realize you can give him a room, provided that you are willing to move people around (but not force people to share rooms). How do you do that?

(b) Later, an even bigger problem occurs. There is another Hilbert Hotel across the street, and it burns down. Suddenly a denumerable set of customers arrives, all wanting rooms in your hotel. How can that be done?

(c) Now comes the true disaster. Across town, there is an infinite sequence of Hilbert Hotels, all full, and they all burn down. All the customers from all those hotels appear at your desk, wanting rooms. How can you accommodate them?

Example 6. The function e^x from $\mathbb{R}$ to $\mathbb{R}$ is one-to-one but not onto; it is a bijection between $\mathbb{R}$ and its proper subset $\mathbb{R}^+$, and so $\mathbb{R} \sim \mathbb{R}^+$. The function $x^3 - x$ from $\mathbb{R}$ to $\mathbb{R}$ is onto but not one-to-one. Such functions cannot exist from a finite set to itself.

Exercise 17. Give examples (or prove the nonexistence) of functions from $\mathbb{N}$ to $\mathbb{N}$ that are:

(a) one-to-one and onto
(b) neither one-to-one nor onto
(c) one-to-one but not onto
(d) onto but not one-to-one.

Exercise 18.

(a) Find a bijection between $\mathbb{R}$ and a bounded open interval. There is at least one such function that you have been familiar with since high school.

(b) Using part (a), the CSB theorem, and other familiar functions, show that all intervals of the forms (a, b), $[a, b]$, $(a, b]$, $[a, b)$, (a, ∞), $[a, \infty)$, $(-\infty, b)$, and $(-\infty, b]$ have the same cardinality as $\mathbb{R}$, provided that $a < b$.

Proposition C.1(c) of Appendix C also tells us that $\mathbb{R} \times \mathbb{R} \sim \mathbb{R}$, and therefore $\mathbb{R}^k \sim \mathbb{R}$ for every positive integer k. This special case

does not require AC, and in fact the idea behind the required bijection is straightforward: since a real number is basically a decimal, two reals can be "coded" into one simply by alternating digits. That is, if we want to define $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, we might let $f(2.5, 1/3) = 20.530303\ldots$, since $2.5 = 2.5000\ldots$ and $1/3 = 0.3333\ldots$. However, there are some sticky points here, such as the treatment of negative numbers and the ambiguity of decimal representation mentioned in Section 2.2. If we allow ourselves to use the fact that $\mathbb{R} \sim 2^{\mathbb{N}}$, then these difficulties disappear, since we can code two infinite sequences of bits into one sequence in the same way.

If we combine the content of the previous paragraph and the previous exercise, we reach a conclusion that seems geometrically absurd: a line segment one millimeter long has the same "number of points" (cardinality) as all of three-dimensional space. Results like this one and the fact that $\mathbb{Q} \sim \mathbb{N}$ contributed to the resistance that Cantor faced in the early years of set theory. Perhaps even more bizarre is the existence of a **space-filling curve**, devised by Peano: a *continuous* function from the unit interval $[0, 1]$ *onto* the unit square $[0, 1] \times [0, 1]$. Such a function cannot be one-to-one, but even so, most people find it difficult to imagine how such a curve could exist.

Exercise 19. Prove that $\mathbb{N}^{\mathbb{N}} \sim 2^{\mathbb{N}}$, and therefore $k^{\mathbb{N}} \sim 2^{\mathbb{N}}$ for every natural number $k > 1$. The main thing you need to show is how to "code" an infinite sequence of natural numbers into an infinite sequence of 0's and 1's.

Here are a few other basic facts involving cardinality, which are restated as parts (f), (h), (i) and (j) of Proposition C.1 in Appendix C. These facts are somewhat abstract in that they deal with functions on functions, but otherwise they are straightforward.

Proposition 2.18.

(a) *For any set A, $\mathcal{P}(A) \sim 2^A$. Here, 2^A has its literal meaning: the set of all functions from A to the ordinal $\{0, 1\}$.*

(b) *If B and C are disjoint, then $A^B \times A^C \sim A^{B \cup C}$.*

(c) *For any sets A, B, and C,* $(A^B)^C \sim A^{B\times C}$.

(d) *For any sets A, B, and C,* $A^C \times B^C \sim (A \times B)^C$.

Proof.

(a) To define a bijection F from $\mathcal{P}(A)$ to 2^A, let $F(B)$ be the characteristic function of B (with domain A). It is then routine to show that F is one-to-one and onto 2^A. (See the exercise below).

(b) Here, we need to define a mapping F that takes as input an ordered pair (g, h), where $g \in A^B$ and $h \in A^C$, and outputs a function from $B \cup C$ to A. What is the obvious way to do that if B and C are disjoint?

(c) This is the most notationally confusing part of this proposition. We want to define a bijection F between these sets. Now, an element of $(A^B)^C$ is by definition a function g such that, for any $y \in C$, $g(y)$ is a function from B to A. So let $F(g)$ be the function with domain $B \times C$ such that $[F(g)](x, y) = [g(y)](x)$, for every $x \in C$ and $y \in C$. Again, it is routine to show that F produces the required bijection.

(d) See the exercise. ∎

Exercise 20. Complete the proof of this proposition. For parts (b) and (d), this requires defining the appropriate mapping. For all four parts, show that the mapping that's been defined really does yield the desired bijection.

Von Neumann cardinals

Here is another use of ordinals. Recall that Frege's definition of a cardinal as an equivalence class of sets under $\sim$ is unsatisfactory in ZFC. Von Neumann gave this more rigorous definition:

Definition. An ordinal α is called a **cardinal** if, for every $\beta < \alpha$, we have $\beta \prec \alpha$. Such an ordinal is also called an **initial** ordinal.

Under this definition, every finite ordinal is also a cardinal, and ω is the first infinite cardinal. Obviously, there are no other countable cardinals. In ZFC, we can show that every set "has" a unique von Neumann cardinal(ity):

Theorem 2.19 (AC). *For every set x there's a unique von Neumann cardinal α such that $x \sim \alpha$.*

Proof. Let x be given. By AC, x can be well ordered. But then, given a well-ordering on x, there's a unique ordinal β such that this well-ordering is isomorphic to the ordering of β under $\in$. (This fact was mentioned without proof in the previous section.) It follows that $x \sim \beta$. Therefore, by Theorem 2.8(b) there's a smallest ordinal α such that $x \sim \alpha$. So this α is the unique initial ordinal of the same size as x. ■

This theorem, in combination with transfinite induction, is a powerful tool for proving things in "ordinary mathematics." This method can be used to prove many results that are usually proved by Zorn's lemma or the Hausdorff maximal principle (defined in Appendix B). Here is a typical example, with two proofs; we will encounter more examples in Sections 6.4 and 6.5.

Theorem 2.20 (AC). *Every vector space has a basis.*

Proof. (By transfinite induction) Given a vector space V over some field, let α be the von Neumann cardinal of V, and then let f be a bijection between α and V. By transfinite induction, we define a subset B of V: for each $\beta < \alpha$, include $f(\beta)$ in B if and only if $f(\beta)$ is not a linear combination of the vectors that have already been included in B for $\gamma < \beta$. It is very easy to show that B is a basis for V. ■

Proof. (By Zorn's lemma) Given a vector space V over some field, let A be the collection of all linearly independent subsets of V. Partially order A by the subset relation, so $S_1 \leq S_2$ means $S_1 \subseteq S_2$. Since the union of any chain of linearly independent sets of vectors is still linearly independent, that union is the least upper bound of the chain. Therefore, every chain in this partial ordering has an upper bound. By

Zorn's lemma, there is a maximal element B, which is easily shown to span V and so must be a basis. ∎

Exercise 21. Complete both of these proofs by showing that B is a basis for V: every vector in V is a linear combination of vectors in B, and no nontrivial linear combination of vectors in B is the zero vector.

Is there a rigorous definition of cardinality for all sets that "works" without the axiom of choice? Yes, there is, but we have to wait until the end of the section to be able to present it.

You might wonder whether the existence of any uncountable ordinals can be proved without AC, since there is no obvious way to define a well-ordering on any familiar uncountable set such as $\mathbb{R}$. But a nice result (of ZF) known as **Hartogs's theorem** states that $\forall x\, \exists \alpha (\alpha \not\preceq x)$. This implies that for any ordinal there's another one of larger cardinality. It follows (using replacement) that there's a one-to-one correspondence between the infinite von Neumann cardinals and all ordinals. The first uncountable von Neumann cardinal is denoted ω_1, the next one is ω_2, etc.; and if $\mathrm{Lim}(\lambda)$, ω_λ is just $\bigcup_{\alpha<\lambda} \omega_\alpha$. Under this definition, there's an ω_α for every α. (Technically, $\omega_0 = \omega$, but this subscript is usually dropped.)

Outside of foundations, Cantor's notation $\aleph_\alpha$ is more common than ω_α. In ZFC, these notations may be used interchangeably. It's often clearer to use alephs when doing cardinal arithmetic, since cardinal arithmetic is different from ordinal arithmetic. For example, CH is usually written $2^{\aleph_0} = \aleph_1$. It can't be written $2^\omega = \omega_1$, since $2^\omega = \omega$ as ordinals.

The cumulative hierarchy

Here is possibly the most important transfinite inductive definition in axiomatic set theory:

Definition.

(a) $V_0 = \emptyset$.

(b) For every α, $V_{\alpha+1} = \mathcal{P}(V_\alpha)$.

(c) If Lim(λ), then $V_\lambda = \bigcup_{\alpha<\lambda} V_\alpha$.

Intuitively, we think of V_α as the set of all sets that are formed *before* stage α.

Proposition 2.21.

(a) *V_α is transitive, for every α.*

(b) *If $\alpha < \beta$, then $V_\alpha \subset V_\beta$.*

Proof.

(a) By transfinite induction: certainly, $\emptyset$ is transitive. If V_α is transitive, then so is $V_{\alpha+1}$, by Proposition 2.5(a). Finally, if Lim(λ) and V_α is transitive for all $\alpha < \lambda$, then $V(\lambda)$ is a union of transitive sets, which must be transitive by Proposition 2.5(b). ■

Exercise 22. Prove part (b) of this proposition.

Lemma 2.22. *For any set, there is a transitive set that contains it (as a subset).*

Proof. Given x, let $y_0 = x$ and, inductively, $y_{n+1} = y_n \cup (\bigcup y_n)$, that is, all the members of y_n together with all of their members. Then let $z = \bigcup_{n<\omega} y_n$. Clearly, $x \subseteq z$, and it is easy to show that z is transitive. ■

Clearly, the set z defined in this proof is the smallest transitive set containing x. This set is called the **transitive closure** of x, denoted $TC(x)$. By the way, the word "contains" is often ambiguous in set theory, as in the statement of this lemma. If we wanted the smallest transitive set containing x as a *member*, we would simply use $TC(\{x\})$.

Here is the most important property of the sets V_α:

Theorem 2.23. *Every set is in some V_α.*

Proof. Assuming that x is in no V_α, let $y = TC(\{x\})$. Then $\{u \in y \mid u$ is in no $V_\alpha\}$ is nonempty, and we may choose an $\in$-minimal element v

of this set, by regularity. Since y is transitive, every element of v is in some V_α. For $w \in v$, let $g(w)$ be the least α such that $w \in V_\alpha$. Then we can form $\{g(w) \mid w \in v\}$ by replacement, and let β be the LUB (union) of this set of ordinals. So $v \subseteq V_\beta$. But then $v \in V_{\beta+1}$, a contradiction. ■

Definition. For every x, the least α such that $x \in V_{\alpha+1}$ is called the **rank** of x.

This definition is set up so that the rank of any ordinal is itself.

Set theorists use a picture to describe the content of this theorem. Think of the class of all ordinals *Ord* as a *very* long, vertical "spine" starting with $\emptyset$ and proceeding upward. For each α, think of the collection of sets of rank α (that is, $V_{\alpha+1} - V_\alpha$) as a horizontal layer at α. Remember, α represents a level or stage of construction. Since the V_α's get bigger as α increases, the width increases as you go up. Thus the entire picture looks like a letter V (see Figure 2.3), and the class of all sets is denoted V. This categorization of sets is called the **cumulative hierarchy**. As we will see in Chapter 6, several variations of this idea have been extremely fruitful in modern set theory.

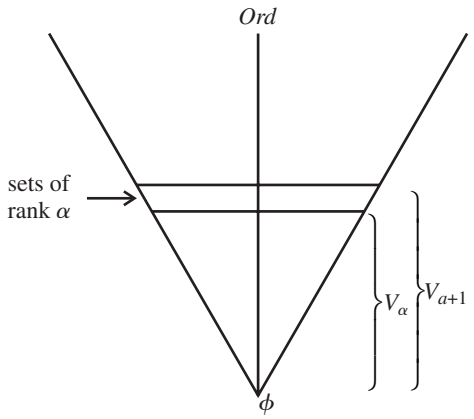

Figure 2.3. The cumulative hierarchy of sets

A set is called **hereditarily finite** if its transitive closure is finite. This means that the entire "membership tree" of the set is finite. For example, $\{\omega\}$ is obviously finite but it is not hereditarily finite. The following characterization of these sets is sometimes useful:

Proposition 2.24. *The set of all hereditarily finite sets exists and is precisely V_ω.*

Proof. We first show that every set in V_ω is hereditarily finite. By Proposition 2.21(a), each V_α is transitive. Also, it is well known (and easy to prove) that the power set of a finite set is finite. Therefore, V_n is finite for every $n \in \omega$. If $x \in V_\omega$, then $x \in V_n$ for some $n \in \omega$, and therefore $TC(x)$ is a subset of the finite set V_n. Therefore, $TC(x)$ is finite.

For the other direction, assume that x is a hereditarily finite set that is not in V_ω. An argument similar to the proof of Theorem 2.23 leads to a contradiction—see the following exercise. ∎

Exercise 23. Complete the second part of this proof. You may use the fact that a finite set of finite ordinals has a finite supremum.

Here, as promised, is the standard way to define cardinality rigorously in ZF. This clever adaptation of Frege's definition is due to Dana Scott:

Definition. The **cardinal** of any set x is the set of all sets *of least rank* that are the same size as x.

In other words, given x, let α be the least ordinal such that $\exists y \in V_{\alpha+1}(x \sim y)$. Then the cardinal of x is $\{y \in V_{\alpha+1} \mid x \sim y\}$.

Under this definition, a set is usually not a member of its own cardinal. But, trivially, x and y have the same cardinal(ity) under this definition if and only if $x \sim y$ in the sense of Section 2.2.

We have not finished our study of set theory. Having covered the basic concepts of the subject, we are almost ready to discuss the brilliant work of Gödel, Paul Cohen, and others that led to the enormous advances in set theory in the second half of the twentieth century. We will take up this discussion in Chapter 6, after we cover some more prerequisite topics.

APPENDIX C

Cardinal Arithmetic

This appendix is related to material in at least three sections of the text: 2.2, 2.5, and 3.2. Chapter 2 explains that the word "cardinal" can be defined in three different ways. Let's review these meanings briefly.

Definition. For any set x, its **cardinal** or **cardinality**, denoted $Card(x)$, is either:

(a) the class of all sets y such that $x \sim y$ (Frege cardinals; not a rigorous definition in ZF or ZFC),

(b) the set of all sets y of least rank such that $x \sim y$ (Scott's adaptation of Frege cardinals; rigorous in ZF or ZFC), or

(c) the least ordinal α such that $x \sim \alpha$ (von Neumann cardinals; rigorous in ZFC but not defined for all sets in ZF).

The material in this appendix is written in accordance with definitions (a) and (b), under which a cardinal is a collection of sets of the same size. It is not hard to rewrite this material to fit definition (c).

The letters κ, μ, and ν, possibly with subscripts, will denote cardinals.

Section 2.2 gives the definitions of the basic relations $x \preceq y$ and $x \prec y$ on *sets*. The relations $\kappa \leq \mu$ and $\kappa < \mu$ on cardinals are defined from these. For example, $\kappa < \mu$ means that $x \prec y$, where $x \in \kappa$, $y \in \mu$. It is very easy to show that this definition is well defined, meaning that it does not depend on the choice of x and y. All of our subsequent

definitions involving cardinals are also well defined. Similarly, words such as "finite" and "uncountable" can be applied to cardinals without ambiguity.

Definitions (Cardinal Arithmetic). Let $Card(A_i) = \kappa_i$ $(i = 1, 2)$. Then:

(a) $\kappa_1 + \kappa_2 = Card[(A_1 \times \{1\}) \cup (A_2 \times \{2\})]$.

(b) $\kappa_1 \cdot \kappa_2 = Card(A_1 \times A_2)$.

(c) $\kappa_1^{\kappa_2} = Card(A_1^{A_2})$.

The set on the right-hand side of part (a) above is called the **formal disjoint union** of A_1 and A_2, denoted $A_1 \coprod A_2$. Clearly, we can't use $A_1 \cup A_2$ there, unless we already know that A_1 and A_2 are disjoint. The set on the right-hand side of (c) is, as usual, the set of all functions from A_2 to A_1.

We now list many of the basic properties of cardinal arithmetic, noting which ones require AC:

Proposition C.1.

(a) *Cardinal addition and multiplication are associative and commutative, and satisfy the distributive law.*

(b) *On finite cardinals, these three operations coincide with the usual operations of arithmetic (and therefore with ordinal arithmetic as well).*

(c) *(AC) If κ or ν is infinite, then $\kappa + \nu = Max(\kappa, \nu)$. If, in addition, neither κ nor ν is zero, then $\kappa \cdot \nu = Max(\kappa, \nu)$. ("Max" stands for "maximum.")*

(d) *(AC) If κ is infinite, then the union of κ sets of cardinality κ has cardinality κ.*

(e) *(AC) If A is infinite, then the set of all finite sequences of members of A, denoted $A^{<\omega}$, has the same cardinality as A.*

(f) *For any set x, $\mathcal{P}(x) \sim 2^x$. In other words, if $Card(x) = \kappa$, then $Card(\mathcal{P}(x)) = 2^\kappa$. (Here, 2 is the ordinal $\{0, 1\}$.)*

(g) **(Cantor's Theorem, restated using (f)).** *For every* κ, $\kappa < 2^\kappa$.

(h) $\kappa^\mu \cdot \kappa^\nu = \kappa^{\mu+\nu}$.

(i) $(\kappa^\mu)^\nu = \kappa^{\mu\cdot\nu}$.

(j) $(\kappa \cdot \mu)^\nu = \kappa^\nu \cdot \mu^\nu$.

Rather than prove any parts of this proposition here, we just illustrate a couple of useful special cases. Suppose we want to establish part (c) for $\kappa = \nu = \aleph_0$, the cardinality of $\mathbb{N}$. This amounts to showing that $\mathbb{N} \times \{1, 2\} \sim \mathbb{N}$ and $\mathbb{N} \times \mathbb{N} \sim \mathbb{N}$. A simple bijection f from $\mathbb{N} \times \{1, 2\}$ to $\mathbb{N}$ is given by

$$f(n, 1) = 2n + 1, \quad \text{and} \quad f(n, 2) = 2n.$$

(Recall that we are assuming that $0 \in \mathbb{N}$.) A simple bijection B_2 from $\mathbb{N} \times \mathbb{N}$ to $\mathbb{N}$ is given by $B_2(m, n) = 2^m(2n+1)-1$. Note that the axiom of choice is not needed to define these bijections.

By iterating the function B_2, we can define a bijection B_k between $\mathbb{N}^k$ and $\mathbb{N}$ for each positive integer k. Specifically, let

$$B_k(a_1, a_2, \ldots, a_k) = B_2[a_1, B_{k-1}(a_2, a_3, \ldots, a_k)],$$

for any $k > 2$. We also define B_1 to be the identity on $\mathbb{N}$.

Similarly, (assuming AC), it follows that $\mu^k = \mu$ whenever μ is infinite and k is a nonzero finite cardinal. By the way, for the purposes of cardinal arithmetic, it doesn't matter whether we define the set A^k by iterating the operation $\times$ or as the set of all functions from k to A.

Part (e) of this proposition can also be proved without AC when $A = \mathbb{N}$. We can explicitly define a bijection B between $\mathbb{N}^{<\omega}$ and $\mathbb{N}$ by letting $B(\emptyset) = 0$, and

$$B(a_1, a_2, \ldots, a_k) = 2^{a_1} \cdot 3^{a_2} \cdot \cdots \cdot {p_k}^{a_k+1} - 1,$$

where p_k denotes the kth prime.

Note that parts (f), (h), (i) and (j) of Proposition C.1 are a direct restatement of Proposition 2.18 in Section 2.5, stated in terms of cardinals rather than individual sets. It should also be clear that parts (h),

(i), and (j) are completely analogous to the main laws of exponents in elementary algebra.

Infinitary cardinal operations

It is also fruitful to define **infinitary** arithmetical operations on cardinals. For the rest of this appendix, we always assume AC. Although several of the definitions we will give have versions that make sense without the axiom of choice, they become much more complex without it.

Definition. Let $\{\kappa_i \mid i \in I\}$ be an indexed family of cardinals. Choose $A_i \in \kappa_i$ for each $i \in I$. Then:

(a) $\sum_{i\in I} \kappa_i = Card[\bigcup_{i\in I}(A_i \times \{i\})]$.
(a) $\prod_{i\in I} \kappa_i = Card(\prod_{i\in I} A_i)$.

The set on the right-hand side of (b) is the usual **Cartesian product** of the indexed family $\{A_i \mid i \in I\}$, that is, the set of all functions f with domain I such that $f(i) \in A_i$ for every i.

Notation. The least cardinal that is greater than κ is denoted κ^+.

In the presence of AC, the existence of κ^+ follows directly from Hartogs's theorem. The following notation is defined in Section 2.5, in a slightly different way:

Definition. The cardinal $\aleph_\alpha$ is defined by induction on α as follows:

$\aleph_0 = Card(\mathbb{N})$.
$\aleph_{\alpha+1} = (\aleph_\alpha)^+$.
For limit ordinals λ, $\aleph_\lambda = \sum_{\alpha\in\lambda} \aleph_\alpha$.

When the von Neumann definition of cardinals is being used, $\aleph_\alpha$ is often written ω_α. The following notation is less common in mainstream mathematics but is often useful:

Definition. The cardinal $\beth_\alpha$ is defined by induction on α as follows ($\beth$ is the Hebrew letter beth):

$\beth_0 = Card(\mathbb{N})$.
$\beth_{\alpha+1} = 2^{\beth_\alpha}$.
For limit ordinals λ, $\beth_\lambda = \sum_{\alpha\in\lambda} \beth_\alpha$.

The notation $\beth_\alpha$ creates a concise way of stating CH: $\beth_1 = \aleph_1$. In ZFC (but not in ZF), GCH becomes $\forall\alpha(\beth_\alpha = \aleph_\alpha)$. So this notation is most useful when GCH is not being assumed. For example, it is clear that $Card(\mathcal{P}(\mathbb{R}))$ is $\beth_2$, but without GCH we cannot say where $\beth_2$, or even $\beth_1$, fits in the hierarchy of $\aleph$'s. Cantor's theorem implies trivially that $\beth_\alpha \geq \aleph_\alpha$ for every α, but not much else is obvious.