Naive Set Theory

Paul R. Halmos

University of Michigan

PREFACE BY THE EDITOR

As the book title says, this is the famous set theory book *Naive Set Theory* by Paul Richard Halmos, first published in 1960 by D. Van Nostrand Company, INC., part of a series called *The University Series in Undergraduate Mathematics*.

What the title doesn't say is that this version is an independent re-edition. The original work is currently public domain in Hathi Trust Digital Library — the reader probably found (or could find) the original digitized book on Google by just searching for its title. This version was written in LaTeX and first released on July 14, 2023, available for free to download on my Github repository. After the initial release, some people have already made contributions in fixing typos and further improving the re-edition. I extend here my gratitude to these people for helping me and keeping the spirit of the re-edition alive.

Even though the book was freely available online, there are three reasons for this project. First, the book in its digitized state is perfectly readable, but it doesn't allow searching words with `Ctrl-F` (windows), `Command-F` (mac), `Ctrl-F` (linux) and doesn't have an interactable summary with it. The second is to update the book by correcting the errors in the original version, following the published errata - as noticed, and updated in this edition, by Michał Zdunek. The third reason is purely personal, I have a passion and gratitude for this book and, while I want to learn OCR, I decided to re-edit it as a homage.

Some notes on this re-edition are necessary. The book page format is B5 paper with font size 12pt. The margins of the book should be perfectly suitable for printing. The mainly differences with the original editions are the cover and the chapters title page designs. The mathematical symbol which denotes *set inclusion* in the original is (ϵ), but I opted to use ($\in$) since it's used regularly nowadays for this case. Besides this, I didn't change anything from the text. Therefore, any mistakes — which I hope are non-existent or, at least, few — are solely mine, and if someone finds any please contact me via e-mail.

As mentioned, the original book is public domain and, so, freely available in the internet. Therefore, the resulting re-edition of the book at the end of this project has no lucrative ends by any means. This re-edition cannot be used for any commercial purposes.

I thought about writting a short story about Paul R. Halmos, since it's common for books to do this specially after the author has deceased. However, I couldn't do a better job than someone just searching on Google and/or Wikipedia. So, for now, I will just say that this book has a special place in my heart. It was one of the first works that introduced and helped me push througth writting proofs. And at the end, I fell in love not only with it, but with mathematics overall. I hope that anybody that found this version can have the same outcome as I did. Now read it, absorb it and forget it.

Matheus Girola Macedo Barbosa - 01/07/2024

PREFACE BY THE AUTHOR

Every mathematician agrees that every mathematician must know some set theory; the disagreement begins in trying to decide how much is some. This book contains my answer to that question. The purpose of the book is to tell the beginning student of advanced mathematics the basic set-theoretic facts of life, and to do so with the minimum of philosophical discourse and logical formalism. The point of view throughout is that prospective mathematician anxious to study groups, or integrals, or manifolds. From this point of view the concepts and methods of this book are merely some of the standard mathematical tools; the expert specialist will find nothing new here.

Scholarly bibliographical credits and references are out of place in a purely expository book such as this one. The student who gets interested in set theory for its own sake should know, bowever, that there is much more to the subject than there is in this book. One of the most beautiful sources of set-theoretic wisdom is still Hausdorff's *Set theory*. A recent and highly readable addition to the literature, with an extensive and up-to-date bibliography, is *Axiomatic set theory* by Suppes.

In set theory "naive" and "axiomatic" are contrasting words. The present treatment might best be described as axiomatic set theory from the naive point of view. It is axiomatic in that some axioms for set theory are stated and used as the basis of all subsequent proofs. It is naive in that the language and notation are those of ordinary informal (but formalizable) mathematics. A more important way in which the naive point view predominates is that set theory is regarded as a body of facts, of which the axioms are a brief and convenient summary; in the orthodox axiomatic view the logical relations among various axioms are the central objects of study. Analogously, a study of geometry might be regarded purely naive if it proceeded on the paper-folding kind of intuition alone; the other extreme, the purely axiomatic one, is the one in which axioms for the various non-Euclidean geometries are studied with the same amount of

attention as Euclid's. The analogue of the point of view of this book is the study of just one sane set of axioms with the intention of describing Euclidean geometry only.

Instead of *Naive set theory* a more honest title for the book would have been *An outline of the elements of naive set theory.* "Elements" would warn the reader that not everything is here; "outline" would warn him that even what is here needs filling in. The style is usually informal to the point of conversational. There are very few displayed theorems; most of the facts are just stated and followed by a sketch of a proof, very much as they might be in a general descriptive lecture. There are only a few exercises, officially so labelled, but, in fact, most of the book is nothing but a long chain of exercises with hints. The reader should continually ask himself whether he knows how to jump from one hint to the next, and, accordingly, he should not be discouraged if he finds that his reading rate is considerably slower than normal.

This is not to say that the contents of this book are unusually difficult or profound. What is true is that the concepts are very general and very abstract, and that, therefore, they may take some getting used to. It is a mathematical truism, however, that the more generally a theorem applies, the less deep it is. The student's task in learning set theory is to steep himself in unfamiliar but essentially shallow generalities till they become so familiar that they can be used with almost no conscious effort. In other words, general set theory is pretty trivial stuff really, but, if you want to be a mathematician, you need some, and here it is; read it, absorb it, and forget it.

P. R. H.

CONTENTS

CHAPTER 1

THE AXIOM OF EXTENSION

A pack of wolves, a bunch of grapes, or a flock of pigeons are all examples of sets of things. The mathematical concept of a set can be used as the foundation for all known mathematics. The purpose of this little book is to develop the basic properties of sets. Incidentally, to avoid terminological monotony, we shall sometimes say *collection* instead of *set*. The word "class" is also used in this context, but there is a slight danger in doing so. The reason is that in some approaches to set theory "class" has a special technical meaning. We shall have occasion to refer to this again a little later.

One thing that the development will not include is a definition of sets. The situation is analogous to the familiar axiomatic approach to elementary geometry. That approach does not offer a definition of points and lines; instead it describes what it is that one can do with those objects. The semi-axiomatic point of view adopted here assumes that the reader has the ordinary, human, intuitive (and frequently erroneous) understanding of what sets are; the purpose of the exposition is to delineate some of the many things that one can correctly do with them.

Sets, as they are usually conceived, have *elements* or *members*. An element of a set may be a wolf, a grape, or a pigeon. It is important to know that a set itself may also be an element of some other set. Mathematics is full of examples of sets of sets. A line, for instance; is a set of points; the set of all lines in the plane is a natural example of a set of sets (of points). What may be surprising is not so much that sets may occur as elements, but that for mathematical purposes no other elements need ever be considered. In this book, in particular, we shall study set, and sets of sets, and similar towers of sometimes frightening height and complexity — and nothing else. By way of examples we might occasionally

speak of sets of cabbages, and kings, and the like, but such usage is always to be construed as an illuminating parable only, and not as a part of the theory that is being developed.

The principal concept of set theory, the one that in completely axiomatic studies is the principal primitive (undefined) concept, is that of *belonging*. If x belongs to A (x is an element of A, x is *contained* in A), we shall write

$$x \in A.$$

This version of the Greek letter epsilon is so often used to denote belonging that its use to denote anything else is almost prohibited. Most authors relegate $\in$ to its set-theoretic use forever and use ε when they need the fifth letter of the Greek alphabet.

Perhaps a brief digression on alphabetic etiquette in set theory might be helpful. There is no compelling reason for using small and capital letters as in the preceding paragraph; we might have written, and often will write, things like $x \in y$ and $A \in B$. Whenever possible, however, we shall informally indicate the status of a set in a particular hierarchy under consideration by means of the convention that letters at the beginning of the alphabet denote elements, and letters at the end denote sets containing them; similarly letters of a relatively simple kind denote elements, and letters of the larger and gaudier fonts denote sets containing them. Examples: $x \in A$, $A \in X$, $X \in \mathcal{C}$.

A possible relation between sets, more elementary than belonging, is *equality*. The equality of two sets A and B is universally denoted by the familiar symbol

$$A = B;$$

the fact that A and B are not equal is expressed by writing

$$A \neq B.$$

The most basic property of belonging is its relation to equality, which can be formulated as follows.

Axiom 1.1 (Axiom of extension)**.** *Two sets are equal if and only if they have the same elements.*

With greater pretentiousness and less clarity: a set is determined by its extension.

It is valuable to understand that the axiom of extension is not just a logically necessary property of equality but a non-trivial statement about belonging. One way to come to understand the point is to consider a partially analogous situation in which the analogue of the axiom of extension does not hold. Suppose, for

instance, that we consider human beings instead of sets, and that, if x and A are human beings, we write $x \in A$ whenever x is an ancestor of A. (The ancestors of a human being are his parents, his parents' parents, their parents, etc., etc.) The analogue of the axiom of extension would say here that if two human beings are equal, then they have the same ancestors (this is the "only if" part, and it is true), and also that if two human being the same ancestors, then they are equal (this is the "if" part, and it is false).

If A and B are sets and if every element of A is an element of B, we say that A is a *subset* of B, or B *includes* A, and we write

$$A \subset B$$

or

$$A \supset B.$$

The wording of the definition implies that each set must be considered to be included in itself ($A \subset A$); this fact is described by saying that set inclusion is *reflexive*. (Note that; in the same sense of the word, equality also is reflexive.) If A and B are sets such that $A \subset B$ and $A \neq B$, the word *proper* is used (proper subset, proper inclusion). If A, B, and C are sets such that $A \subset B$ and $B \subset C$, then $A \subset C$; this fact is described by saying that set inclusion is *transitive*. (This property is also shared by equality.)

If A and B are sets such that $A \subset B$ and $B \subset A$, then A and B have the same elements and therefore, by the axiom of extension, $A = B$. This fact is described by saying that set inclusion is *antisymmetric*. (In this respect set inclusion behaves differently from equality. Equality is *symmetric*, in the sense that if $A = B$, then necessarily $B = A$.) The axiom of extension can, in fact, be reformulated in these terms: if A and B are sets, then a necessary and sufficient condition that $A = B$ is that both $A \subset B$ and $B \subset A$. Correspondingly , almost all proofs of equalities between two sets A and B are split into two parts; first show that $A \subset B$, and then show that $B \subset A$.

Observe that belonging ($\in$) and inclusion ($\subset$) are conceptually very different indeed. One important difference has already manifested itself above: inclusion is always reflexive, whereas it is not at all clear that belonging is ever reflexive. That is: $A \subset A$ is always true; is $A \in A$ ever true? It is certainly not true of any reasonable set that anyone has ever seen. Observe, along the same lines, that inclusion is transitive, whereas belonging is not. Everyday examples, involving, for instance, super-organizations whose members are organizations, will readily occur to the interested reader.

CHAPTER 2

THE AXIOM OF SPECIFICATION

All the basic principles of set theory, except only the axiom of extension, are designed to make new sets out of old ones. The first and most important of these basic principles of set manufacture says, roughly speaking, that anything intelligent one can assert about the elements of a set specifies a subset, namely, the subset of those elements about which the assertion is true.

Before formulating this principle in exact terms, we look at a heuristic example. Let A be the set of all men. The sentence "x is married" is true for some of the elements x of A and false for others. The principle we are illustrating is the one that justifies the passage from the given set A to the subset (namely, the set of all married men) specified by the given sentence. To indicate the generation of the subset, it is usually denoted by

$$\{x \in A : x \textit{ is married}\}.$$

Similarly

$$\{x \in A : x \textit{ is not married}\}$$

is the get of all bachelors;

$$\{x \in A : \textit{ the father of } x \textit{ is } Adam\}$$

is the set that contains Seth, Cain and Abel and nothing else; and

$$\{x \in A : x \textit{ is the father of } Abel\}$$

is the set that contains Adam and nothing else. Warning: a box that contains a hat and nothing else is not the same thing as a hat, and, in the same way, the

last set in this list of examples is not to be confused with Adam. The analogy between sets and boxes has many weak points, but sometimes it gives a helpful picture of the facts.

All that is lacking for the precise general formulation that underlies the examples above is a definition of *sentence* . Here is a quick and informal one. There are two basic types of sentences, namely, assertions of belonging,

$$x \in A,$$

and assertions of equality,

$$A = B;$$

all other sentences are obtained from such *atomic* sentences by repeated applications of the usual logical operators, subject only to the minimal courtesies of grammar and unambiguity. To make the definition more explicit (and longer) it is necessary to append to it a list of the "usual logical operators" and the rules of syntax. An adequate (and, in fact, redundant) list of the former contains seven items:

and,
or (in the sense of "either — or — or both"),
not,
if—then—(or implies),
if and only if,
for some (or there exists),
for all.

As for the rules of sentence construction, they can be described as follows. (i) Put "not" before a sentence and enclose the result between parentheses. (The reason for parentheses, here and below, is to guarantee unambiguity. Note, incidentally, that they make all other punctuation marks unnecessary. The complete parenthetical equipment that the definition of sentences calls for is rarely needed. We shall always omit as many parentheses as it seems safe to omit without leading to confusion. In normal mathematical practice, to be followed in this book, several different sizes and shapes of parentheses are used, but that is for visual convenience only.) (ii) Put "and" or "or" or "if and only if" between two sentences and enclose the result between parentheses. (iii) Replace the dashes in "if—then—" by sentences and enclose the result in parentheses. (iv) Replace the dash in "for some—" or in "for all—" by a letter, follow the result by a sentence, and enclose the whole in parentheses. (If the letter used does not occur in the sentence, no harm is done. According to the usual and natural convention "for some y $(x \in A)$" just means "$x \in A$". It is equally harmless if the letter used has already been used with "for some—." Recall that

"for some x $(x \in A)$" means the same as "for some y $(y \in A)$"; it follows that a judicious change of notation will always avert alphabetic collisions.)

We are now ready to formulate the major principle of set theory, often referred to by its German name *Aussonderungsaxiom*.

Axiom 2.1 (Axiom of specification)**.** *To every set A and to every condition $S(x)$ corresponds a set B whose elements are exactly those elements x of A for which $S(x)$ holds.*

A "condition" here is just a sentence. The symbolism is intended to indicate the letter x is *free* in the sentence $S(x)$; that means that x occurs in $S(x)$ at least once without being introduced by one of the phrases "for some x" or "for all x". It is an immediate consequence of the axiom of extension that the axiom of specification determines the set B uniquely. To indicate the way B is obtained from A and from $S(x)$ it is customary to write

$$B = \{x \in A : S(x)\}.$$

To obtain an amusing and instructive application of the axiom of specification, consider, in the role of $S(x)$, the sentence

$$\text{not } (x \in x).$$

It will be convenient, here and throughout, to write "$x \notin A$" instead of "not $(x \in A)$"; in this notation, the role of $S(x)$ is now played by

$$x \notin x.$$

It follows that, whatever the set A may be, if $B = x \in A : x \notin x$, then, for all y,

$$y \in B \text{ if and only if } (y \in A \text{ and } y \notin y). \tag{2.1}$$

Can it be that $B \in A$? We proceed to prove that the answer is no. Indeed, if $B \in A$, then either $B \in B$ also (unlikely, but not obviously impossible), or else $B \notin B$. If $B \in B$, then, by Equation 2.1, the assumption $B \in A$ yields $B \notin B$—a contradiction. If $B \notin B$, then, by Equation 2.1 again, the assumption $B \in A$ yields $B \in B$—a contradiction again. This completes the proof that is impossible, so that we must have $B \notin A$. The most interesting part of this conclusion is that there exists something (namely B) that does not belong to A. The set A in this argument was quite arbitrary. We have proved, in other words, that

nothing contains everything,

or, more spectacularly,

there is no universe.

"Universe" here is used in the sense of "universe of discourse," meaning, in any particular discussion, a set that contains all the objects that enter into that discussion.

In older (pre-axiomatic) approaches to set theory, the existence of universe was taken for granted, and the argument in the preceding paragraph was known as the *Russell's paradox*. The moral is that it is impossible, especially in mathematics, to get something for nothing. To specify a set, it is not enough to pronounce some magic words (which may form a sentence such as "$x \notin x$"); it is necessary also to have at hand a set to whose elements the magic words apply.

CHAPTER 3

UNORDERED PAIRS

For all that has been said so far, we might have been operating in a vacuum. To give the discussion some substance, let us now officially assume that

there exists a set.

Since later on we shall formulate a deeper and more useful existential assumption, this assumption plays a temporary role only. One consequence of this innocuous seeming assumption is that there exists a set without any elements at all. Indeed, if A is a set, apply the axiom of specification to A with the sentence "$x \neq x$" (or, for that matter, with any other universally false sentence). The result is the set $\{x \in A : x \neq x\}$, and that set, clearly, has no elements. The axiom of extension implies that there can be only one set with no elements. The usual symbol for that set is

$$\emptyset;$$

the set is called the *empty set*.

The empty set is a subset of every set, or, in other words, $\emptyset \subset A$ for every A. To establish this, we might argue as follows. It is to be proved that every element in $\emptyset$ belongs to A; since there are no elements in $\emptyset$, the condition is automatically fulfilled. The reasoning is correct but perhaps unsatisfying. Since it is a typical example of a frequent phenomenon, a condition holding in the "vacuous" sense, a word of advice to the inexperienced reader might be in order. To prove that something is true about the empty set, prove that it cannot be false. How, for instance, could it be false that $\emptyset \subset A$? It could be false only if $\emptyset$ had an element that did not belong to A. Since $\emptyset$ has no elements at all, this is absurd. Conclusion: $\emptyset \subset A$ is not false, and therefore $\emptyset \subset A$ for every A.

The set theory developed so far is still a pretty poor thing; for all we know there is only one set and that one is empty. Are there enough sets to ensure that every set is an element of some set? Is it true that for any two sets there is a third one that they both belong to? What about three sets, or four, or any number? We need a new principle of set construction to resolve such questions. The following principle is a good beginning.

Axiom 3.1 (Axiom of pairing)**.** *For any two sets there exists a set that they both belong to.*

Note that this is just the affirmative answer to the second question above.

To reassure worriers, let us hasten to observe that words such as "two," "three," and "four," used above, do not refer to the mathematical concepts bearing those names, which will be defined later; at present such words are merely the ordinary linguistic abbreviations for "something and then something else" repeated an appropriate number of times. Thus, for instance, the axiom of pairing, in unabbreviated form, says that if a and b are sets, then there exists a set A such that $a \in A$ and $b \in A$.

One consequence (in fact an equivalent formulation) of the axiom of pairing is that for any two sets there exists a set that contains both of them and nothing else. Indeed, if a and b are sets, and if A is a set such that $a \in A$ and $b \in A$, then we can apply the axiom of specification to A with the sentence "$x = a$ *or* $x = b$." The result is the set

$$\{x \in A : x = a \text{ or } x = b\},$$

and that set, clearly, contains just a and b. The axiom of extension implies that there can be only one set with this property. The usual symbol for that set is

$$\{a, b\};$$

the set is called the *pair* (or, by way of emphatic comparison with a subsequent concept, the *unordered pair*) formed by a and b.

If, temporarily, we refer to the sentence "$x = a$ *or* $x = b$" as $S(x)$, we may express the axiom of pairing by saying that there exists a set B such that

$$x \in B \text{ if and only if } S(x). \tag{3.1}$$

The axiom of specification, applied to a set A, asserts the existence of a set B such that

$$x \in B \text{ if and only if } (x \in A \text{ and } S(x)). \tag{3.2}$$

The relation between Equation 3.1 and Equation 3.2 typifies something that occurs quite frequently. All the remaining principles of set construction are pseudo-special cases of the axiom of specification in the sense in which Equation 3.1 is a pseudo-special case of Equation 3.2. They all assert the existence of a set specified by a certain condition; if it were known in advance that there exists a set containing all the specified elements, then the existence of a set containing just them would indeed follow as a special case of the axiom of specification.

If a is a set, we may form the unordered pairs $\{a, a\}$. That unordered pair is denoted by

$$\{a\}$$

and is called the *singleton* of a; it is uniquely characterized by the statement that it has a as its only element. Thus, for instance, $\emptyset$ and $\{\emptyset\}$ are very different sets; the former has no elements, whereas the latter has the unique element $\emptyset$. To say that $a \in A$ is equivalent to saying that $\{a\} \subset A$.

The axiom of pairing ensures that every set is an element of some set and that any two sets are simultaneously elements of some one and the same set. (The corresponding questions for three and four and more sets will be answered later.) Another pertinent comment is that from the assumptions we have made so far we can infer the existence of very many sets indeed. For examples consider the sets $\emptyset, \{\emptyset\}, \{\{\emptyset\}\}, \{\{\{\emptyset\}\}\}$, etc.; consider the pairs, such as $\{\emptyset, \{\emptyset\}\}$, formed by any two of them; consider the pairs formed by any two such pairs, or else the mixed pairs formed by any singleton and any pair; proceed so on ad infinitum.

Exercise 3.1. Are all the sets obtained in this way distinct from one another?

Before continuing our study of set theory, we pause for a moment to discuss a notational matter. It seems natural to denote the set B described in Equation 3.1 by $\{x : S(x)\}$; in the special case that was there considered

$$\{x : x = a \text{ or } x = b\} = \{a, b\}.$$

We shall use this symbolism whenever it is convenient and permissible to do so. If, that is, $S(x)$ is a condition on x such that the x's that $S(x)$ specifies constitute a set, then we may denote that set by

$$\{x : S(x)\}.$$

In case A is a set and $S(x)$ is $(x \in A)$, then it is permissible to form $\{x : S(x)\}$; in fact

$$\{x : x \in A\} = A.$$

If A is a set and $S(x)$ is an arbitrary sentence, it is permissible to form $\{x : x \in A \textit{ and } S(x)\}$; this set is the same as $\{x \in A : S(x)\}$. As further examples, we note that

$$\{x : x \neq x\} = \emptyset$$

and

$$\{x : x = a\} = \{a\}.$$

In case $S(x)$ is $(x \notin x)$, or in case $S(x)$ is $(x = x)$, the specified x's do not constitute a set.

Despite the maxim about never getting something for nothing, it seems a little harsh to be told that certain sets are not really sets and even their names must never be mentioned. Some approaches to set theory try to soften the blow by making systematic use of such illegal sets but just not calling them sets; the customary word is "class". A precise explanation of what classes really are and how they are used is irrelevant in the present approach. Roughly speaking, a class may be identified with a condition (sentence), or, rather, with the "extension" of a condition.

CHAPTER 4

UNIONS AND INTERSECTIONS

If A and B are sets, it is sometimes natural to wish to unite their elements into one comprehensive set. One way of describing such a comprehensive set is to require it to contain all the elements that belong to at least one of the two members of the pair $\{A, B\}$. This formulation suggests a sweeping generalization of itself; surely a similar construction should apply to arbitrary collections of sets and not just to pairs of them. What is wanted, in other words, is the following principle of set construction.

Axiom 4.1 (Axiom of unions)**.** *For every collection of sets there exists a set that contains all the elements that belong to at least one set of the given collection.*

Here it is again: for every collection $\mathcal{C}$ there exists a set U such that if $x \in X$ for some X in $\mathcal{C}$, then $x \in U$. (Note that "at least one" is the same as "some.")

The comprehensive set U described above may be too comprehensive; it may contain elements that belong to none of the sets X in the collection $\mathcal{C}$. This is easy to remedy; just apply the axiom of specification to form the set

$$\{x \in U : x \in X \text{ for some } X \text{ in } \mathcal{C}\}.$$

(The condition here is a translation into idiomatic usage of the mathematically more acceptable "*for some* X $(x \in X$ *and* $X \in \mathcal{C})$.") It follows that, for every x, a necessary and sufficient condition that x belong to this set is that x belong to X for some X in $\mathcal{C}$. If we change notation and call the new set U again, then

$$U = \{x : x \in X \text{ for some } X \text{ in } \mathcal{C}\}.$$

This set U is called the *union* of the collection $\mathcal{C}$ of sets; note that the axiom of extension guarantees its uniqueness. The simplest symbol for U that is in use at all is not very popular in mathematical circles; it is

$$\bigcup \mathcal{C}.$$

Most mathematicians prefer something like

$$\bigcup \{X : X \in \mathcal{C}\}$$

or

$$\bigcup_{X \in \mathcal{C}} X.$$

Further alternatives are available in certain important special cases; they will be described in due course.

For the time being we restrict our study of the theory of unions to the simplest facts only. The simplest fact of all is that

$$\bigcup \{X : X \in \emptyset\} = \emptyset,$$

and the next simplest fact is that

$$\bigcup \{X : X \in \{A\}\} = A.$$

In the brutally simple notation mentioned above these facts are expressed by

$$\bigcup \emptyset = \emptyset$$

and

$$\bigcup \{A\} = A.$$

The proofs are immediate from the definitions.

There is a little more substance in the union of pairs of sets (which is what started this whole discussion anyway). In that case special notation is used:

$$\bigcup \{X : X \in \{A, B\}\} = A \cup B.$$

The general definition of unions implies in the special case that $x \in A \cup B$ if and only if x belongs to either A or B or both; it follows that

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

Here are some easily proved facts about the unions of pairs:

$$A \cup \emptyset = A,$$

$$A \cup B = B \cup A \ (commutatitivity),$$

$$A \cup (B \cup C) = (A \cup B) \cup C \ (associativity),$$

$$A \cup A = A \ (idempotence),$$

$$A \subset B \text{ if and only if } A \cup B = B.$$

Every student of mathematics should prove these things for himself at least once in his life. The proofs are based on the corresponding elementary properties of the logical operator *or*.

An equally simple but quite suggestive fact is that

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

What this suggests is the way to generalize pairs. Specifically, we write

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

The equation defines its left side. The right side should by rights have at least one pair of parentheses in it, but, in view of the associative law, their omission can lead to no misunderstanding. Since it is easy to prove that

$$\{a, b, c\} = \{x : x = a \text{ or } x = b \text{ or } x = c\},$$

we know now that for every three sets there exists a set that contains them and nothing else; it is natural to call that uniquely determined set the (*unordered*) *triple* formed by them. The extension of the notation and terminology thus introduced to more terms (*quadruples*, etc.) is obvious.

The formation of unions has many points of similarity with another set-theoretic operation. If A and B are sets, the *intersection* of A and B is the set

$$A \cap B$$

defined by

$$A \cap B = \{x \in A : x \in B\}.$$

The definition is symmetric in A and B even if it looks otherwise; we have

$$A \cap B = \{x \in B : x \in A\},$$

and, in fact, since $x \in A \cap B$ if and only if x belongs to both A and B, it follows that

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

The basic facts about intersections, as well as their proofs, are similar to the basic facts about unions:

$$A \cap \emptyset = \emptyset,$$

$$A \cap B = B \cap A,$$

$$A \cap (B \cap C) = (A \cap B) \cap C,$$

$$A \cap A = A,$$

$$A \subset B \text{ if and only if } A \cap B = A.$$

Pairs of sets with an empty intersection occur frequently enough to justify the use of a special word: if $A \cap B = \emptyset$, the sets A and B are called *disjoint*. The same word is sometimes applied to a collection of sets to indicate that any two distinct sets of the collection are disjoint; alternatively we may speak in such a situation of a *pairwise disjoint* collection.

Two useful facts about unions and intersections involve both the operations at the same time:

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

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

These identities are called the *distributive laws*. By way of a sample of a set-theoretic proof, we prove the second one. If x belongs to the left side, then x belongs either to A or to both B and C; if x is in A, then x is in both $A \cup B$ and $A \cup C$, and if x is in both B and C, then, again, x is in both $A \cup B$ and $A \cup C$; it follows that, in any case, x belongs to the right side. This proves that the right side includes the left. To prove the reverse inclusion, just observe that if x belongs to both $A \cup B$ and $A \cup C$, then x belongs either to A or to both B and C.

The formation of the intersection of two sets A and B, or, we might as well say, the formation of the intersection of a pair $\{A, B\}$ of sets, is a special case of a much more general operation. (This is another respect in which the theory of intersections imitates that of unions.) The existence of the general operation of intersection depends on the fact that for each non-empty collection of sets there

exists a set that contains exactly those elements that belong to every set of the given collection. In other words: for each collection $\mathcal{C}$, other than $\emptyset$, there exists a set V such that $x \in V$ if and only if $x \in X$ for every X in $\mathcal{C}$. To prove this assertion, let A be any particular set in $\mathcal{C}$ (this step is justified by the fact that $\mathcal{C} \neq \emptyset$) and write

$$V = \{x \in A : x \in X \text{ for every } X \text{ in } \mathcal{C}\}.$$

(The condition means "*for all* X (*if* $X \in \mathcal{C}$, *then* $x \in X$).") The dependence of V on the arbitrary choice of A is illusory; in fact

$$V = \{x : x \in X \text{ for every } X \text{ in } \mathcal{C}\}.$$

The set V is called the *intersection* of the collection $\mathcal{C}$ of sets; the axiom of extension guarantees its uniqueness. The customary notation is similar the one for unions: instead of the unobjectionable but unpopular

$$\bigcap \mathcal{C},$$

the set V is usually denoted by

$$\bigcap \{X : X \in \mathcal{C}\}$$

or

$$\bigcap_{X \in \mathcal{C}} X.$$

Exercise 4.1. A necessary and sufficient condition that $(A \cap B) \cup C = A \cap (B \cup C)$ is that $C \subset A$. Observe that the condition has nothing to do with the set B.

CHAPTER 5

COMPLEMENTS AND POWERS

If A and B are sets, the *difference* between A and B, more often known as the *relative complement* of B in A, is the set $A - B$ defined by

$$A - B = \{x \in A : x \notin B\}.$$

Note that in this definition it is not necessary to assume that $B \subset A$. In order to record the basic facts about complementation as simply as possible, we assume nevertheless (in this section only) that all the sets to be mentioned are subsets of one and the same set E and that all complements (unless otherwise specified) are formed relative to that E. In such situations (and they are quite common) it is easier to remember the underlying set E than to keep writing it down, and this makes it possible to simplify the notation. An often used symbol for the temporarily absolute (as opposed to relative) complement of A is A'. In terms of this symbol the basic facts about complementation can be stated as follows:

$$(A')' = A,$$

$$\emptyset' = E,\ E' = \emptyset,$$

$$A \cap A' = \emptyset,\ A \cup A' = E,$$

$$A \subset B \text{ *if and only if* } B' \subset A'.$$

The most important statements about complements are the so-called *De Morgan laws*:

$$(A \cup B)' = A' \cap B',\ (A \cap B)' = A' \cup B'.$$

(We shall see presently that the De Morgan laws hold for the unions and intersections of larger collections of sets than just pairs.) These facts about complementation imply that the theorems of set usually come in pairs. If in an inclusion equation involving unions, intersections, and complements of subsets of E we replace each set by its complement, interchange unions and intersections, and reverse all inclusions, the result is another theorem. This fact is sometimes referred to as the *principle of duality* for sets.

Here are some easy exercises on complementation.

$$A - B = A \cap B'.$$

$$A \subset B \text{ if and only if } A - B = \emptyset.$$

$$A - (A - B) = A \cap B.$$

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

$$A \cap B \subset (A \cap C) \cup (B \cap C').$$

$$(A \cup C) \cap (B \cup C') \subset A \cup B.$$

If A and B are sets, the *symmetric difference* (or *Boolean sum*) of A and B is the set $A + B$ defined by

$$A + B = (A - B) \cup (B - A).$$

This operation is commutative ($A+B = B+A$) and associative ($A+(B+C) = (A + B) + C$), and is such that $A + \emptyset = A$ and $A + A = \emptyset$.

This may be the right time to straighten out a trivial but occasionally puzzling part of the theory of intersections. Recall, to begin with, that intersections were defined for non-empty collections only. The reason is that the same approach to the empty collection does not define a set. Which x's are specified by the sentence

$$x \in X \text{ for every } X \text{ in } \emptyset?$$

As usual for questions about $\emptyset$ the answer is easier to see for the corresponding negative question. Which x's do *not* satisfy the stated condition? If it is not true that $x \in X$ for every X in $\emptyset$, then there must exist an X in $\emptyset$ such that $x \notin X$; since, however, there do not exist any X's in $\emptyset$ at all, this is absurd. Conclusion: no x fails to satisfy the stated condition, or, equivalently, every x does satisfy it. In other words, the x's that the condition specifies exhaust the (nonexistent) universe. There is no profound problem here; it is merely a nuisance to be forced always to be making qualifications and exceptions just because some set somewhere along some construction might turn out to be empty. There is nothing to be done about this; it is just a fact of life.

If we restrict our attention to subsets of a particular set E, as we have temporarily agreed to do, then the unpleasantness described in the preceding paragraph appears to go away. The point is that in that case we can define the intersection of a collection $\mathcal{C}$ (of subsets of E) to be the set

$$\{x \in E : x \in X \text{ for every } X \in \mathcal{C}\}.$$

This is nothing revolutionary; for each non-empty collection, the new definition agrees with the old one. The difference is in the way the old and the new definitions treat the empty collection; according to the new definition $\bigcap_{X\in\emptyset} X$ is equal to E. (For which elements x of E can it be false that $x \in X$ for every X in $\emptyset$?) The difference is just a matter of language. A little reflection reveals that the "new" definition offered for intersection of a collection $\mathcal{C}$ of subsets of E is really the same as the old definition of the intersection of the collection $\mathcal{C} \cup \{E\}$, and the latter is never empty.

We have been considering the subsets of a set E; do those subsets themselves constitute a set? The following principle guarantees that the answer is yes.

Axiom 5.1 (Axiom of powers)**.** *For each set there exists a collection of sets that contains among its elements all the subsets of the given set.*

In other words, if E is a set, then there exists a set (collection) $\mathcal{P}$ such that if $X \subset E$, then $X \in \mathcal{P}$.

The set $\mathcal{P}$ described above may be larger than wanted; it may contain elements other than subsets of E. This is easy to remedy; just apply the axiom of specification to form the set $\{X \in \mathcal{P} : X \subset E\}$. (Recall that "$X \subset E$" says the same thing as " *for all* x (*if* $x \in X$ *then* $x \in E$).") Since, for every X, a necessary and sufficient condition that X belong to this set is that X be a subset of E, it follows that if we change notation and call this set $\mathcal{P}$ again, then

$$\mathcal{P} = \{X : X \subset E\}.$$

The set $\mathcal{P}$ is called the *power set* of E; the axiom of extension guarantees its uniqueness. The dependence of $\mathcal{P}$ on E is denoted by writing $\mathcal{P}(E)$ instead of just $\mathcal{P}$.

Because the set $\mathcal{P}(E)$ is very big in comparison with E, it is not easy to give examples. If $E = \emptyset$, the situation is clear enough; the set $\mathcal{P}(\emptyset)$ is the singleton $\{\emptyset\}$. The power sets of singletons and pairs are also easily describable; we have

$$\mathcal{P}(\{a\}) = \{\emptyset, \{a\}\}$$

and

$$\mathcal{P}(\{a, b\}) = \{\emptyset, \{a\}, \{b\}, \{a, b\}\}.$$

The power set of a triple has eight elements. The reader can probably guess (and is hereby challenged to prove) the generalization that includes all these statements: the power set of a finite set with, say, n elements has 2^n elements. (Of course concepts like "finite" and "2^n" have no official standing for us yet; this should not prevent them from being unofficially understood.) The occurrence of n as an exponent (the n-th power of 2) has something to do with the reason why power set bears its name.

If $\mathcal{C}$ is a collection of subsets of a set E (that is, $\mathcal{C}$ is a subcollection of $\mathcal{P}(E)$), then write

$$\mathcal{D} = \{X \in \mathcal{P}(E) : X' \in \mathcal{C}\}.$$

(To be certain that the condition used in the definition of $\mathcal{D}$ is a sentence in the precise technical sense, it must be rewritten in something like the form

for some Y [$Y \in \mathcal{C}$ *and for all* x ($x \in X$ *if and only if* ($x \in E$ *and* $x \notin Y$))].

Similar comments often apply when we wish to use defined abbreviations instead of logical and set-theoretic primitives only. The translation rarely requires any ingenuity and we shall usually omit it.) It is customary to denote the union and the intersection of the collection $\mathcal{D}$ by the symbols

$$\bigcup_{X \in \mathcal{C}} X' \text{ and } \bigcap_{X \in \mathcal{C}} X'.$$

In this notation the general forms of the De Morgan laws become

$$\left(\bigcup_{X \in \mathcal{C}} X \right)' = \bigcap_{X \in \mathcal{C}} X'.$$

and

$$\left(\bigcap_{X \in \mathcal{C}} X \right)' = \bigcup_{X \in \mathcal{C}} X'.$$

The proofs of these equations are immediate consequences of the appropriate definitions.

Exercise 5.1. Prove that $\mathcal{P}(E) \cap \mathcal{P}(F) = \mathcal{P}(E \cap F)$ and $\mathcal{P}(E) \cup \mathcal{P}(F) \subset \mathcal{P}(E \cup F)$. These assertions can be generalized to

$$\bigcap_{X \in \mathcal{C}} \mathcal{P}(X) = \mathcal{P}\left(\bigcap_{X \in \mathcal{C}} X\right)$$

and

$$\bigcup_{X \in \mathcal{C}} \mathcal{P}(X) \subset \mathcal{P}\left(\bigcup_{X \in \mathcal{C}} X\right);$$

find a reasonable interpretation of the notation in which these generalizations were here expressed and then prove them. Further elementary facts:

$$\bigcap_{X \in \mathcal{P}(E)} X = \emptyset,$$

and

$$\textit{if } E \subset F, \textit{ then } \mathcal{P}(E) \subset \mathcal{P}(F).$$

A curious question concerns the commutativity of the operators $\mathcal{P}$ and $\bigcup$. Show that E is always equal $\bigcup_{X \in \mathcal{P}(E)} X$ (that is $E = \bigcup \mathcal{P}(E)$), but that the result of applying $\mathcal{P}$ and $\bigcup$ to E in the other order is a set that includes E as a subset, typically a proper subset.

CHAPTER 6

ORDERED PAIRS

What does it mean to arrange the elements of a set A in some order? Suppose, for instance, that the set A is the quadruple $\{a, b, c, d\}$ of distinct elements, and suppose that we want to consider its elements in the order

$$c \ b \ d \ a.$$

Even without a precise definition of what this means, we can do something set-theoretically intelligent with it. We can, namely, consider, for each particular spot in the ordering, the set of all those elements that occur at or before that spot; we obtain in this way the sets

$$\{c\} \ \{c, b\} \ \{c, b, d\} \ \{c, b, d, a\}.$$

We can go on then to consider the set (or collection, if that sounds better)

$$\mathcal{C} = \{\{a, b, c, d\}, \{b, c\}, \{b, c, d\}, \{c\}\}$$

that has exactly those sets for its elements. In order to emphasize that the intuitively based and possibly unclear concept of order has succeeded in producing something solid and simple, namely a plain, unembellished set $\mathcal{C}$, the elements of $\mathcal{C}$, and *their* elements, are presented above in a scrambled manner. (The lexicographically inclined reader might be able to see a method in the manner of scrambling.)

Let us continue to pretend for a while that we do know what order means. Suppose that in a hasty glance at the preceding paragraph all we could catch is the set $\mathcal{C}$; can we use it to recapture the order that gave rise to it? The answer is easily seen to be yes. Examine the elements of $\mathcal{C}$ (they themselves are sets,

of course) to find one that is included in all the others; since $\{c\}$ fills the bill (and nothing else does) we know that c must have been the first element. Look next for the next smallest element of $\mathcal{C}$, i.e., the one that is included in all the ones that remain after $\{c\}$ is removed; since $\{b, c\}$ fills the bill (and nothing else does), we know that b must have been the second element. Proceeding thus (only two more steps are needed) we pass from the set $\mathcal{C}$ to the given ordering of the given set A.

The moral is this: we may not know precisely what it means to order the elements of a set A, but with each order we can associate a set $\mathcal{C}$ of subsets of A in such a way that the given order can be uniquely recaptured from $\mathcal{C}$. (Here is a non-trivial exercise: find an intrinsic characterization of those sets of subsets of A that correspond to some order in A. Since "order" has no official meaning for us yet, the whole problem is officially meaningless. Nothing that follows depends on the solution, but the reader would learn something valuable by trying to find it.) The passage from an order in A to the set $\mathcal{C}$, and back, was illustrated above for a quadruple; for a pair everything becomes at least twice as simple. If $A = \{a, b\}$ and if, in the desired order, a comes first, then $\mathcal{C} = \{\{a\}, \{a, b\}\}$; if, however, b comes first, then $\mathcal{C} = \{\{b\}, \{a, b\}\}$.

The *ordered pair* of a and b, with *first coordinate* a and *second coordinate* b, is the set (a, b) defined by

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

However convincing the motivation of this definition may be, we must still prove that the result has the main property that an ordered pair must have to deserve its name. We must show that if (a, b) and (x, y) are ordered pairs and if $(a, b) = (x, y)$, then $a = x$ and $b = y$. To prove this, we note first that if a and b happen to be equal, then the ordered pair (a, b) is the same as the singleton $\{\{a\}\}$. If, conversely, (a, b) is a singleton, then $\{a\} = \{a, b\}$, so that $b \in \{a\}$, and therefore $a = b$. Suppose now that $(a, b) = (x, y)$. If $a = b$, then both (a, b) and (x, y) are singletons, so that $x = y$; since $\{x\} \in (a, b)$ and $\{a\} \in (x, y)$, it follows that a, b, x, and y are all equal. If $a \neq b$, then both (a, b) and (x, y) contain exactly one singleton, namely $\{a\}$ and $\{x\}$ respectively, so that $a = x$. Since in this case it is also true that both (a, b) and (x, y) contain exactly one unordered pair that is not a singleton, namely $\{a, b\}$ and $\{x, y\}$ respectively, it follows that $\{a, b\} = \{x, y\}$, and therefore, in particular, $b \in \{x, y\}$. Since b cannot be x (for then we should have $a = x$ and $b = x$, and, therefore, $a = b$), we must have $b = y$, and the proof is complete.

If A and B are sets, does there exist a set that contains all the ordered pairs (a, b) with a in A and b in B? It is quite easy to see that the answer is yes. Indeed, if $a \in A$ and $b \in B$, then $\{a\} \subset A$ and $\{b\} \subset B$, and therefore $\{a, b\} \subset A \cup B$.

Since also $\{a\} \subset A \cup B$, it follows that both $\{a\}$ and $\{a, b\}$ are elements of $\mathcal{P}(A \cup B)$. This implies that $\{\{a\}, \{a, b\}\}$ is a subset of $\mathcal{P}(A \cup B)$, and hence that it is an element of $\mathcal{P}(\mathcal{P}(A \cup B))$; in other words $(a, b) \in \mathcal{P}(\mathcal{P}(A \cup B))$ whenever $a \in A$ and $b \in B$. Once this is known, it is a routine matter to apply the axiom of specification and the axiom of extension to produce the unique set $A \times B$ that consists exactly of the ordered pairs (a, b) with a in A and b in B. This set is called the *Cartesian product* of A and B; it is characterized by the fact that

$$A \times B = \{x : x = (a, b) \textit{ for some } a \textit{ in } A \textit{ and for some } b \textit{ in } B\}.$$

The Cartesian product of two sets is a set of ordered pairs (that is, a set each of whose elements is an ordered pair), and the same is true of every subset of a Cartesian product. It is of technical importance to know that we can go in the converse direction also: every set of ordered pairs is a subset of the Cartesian product of two sets. In other words: if R is a set such that every element of R is an ordered pair, then there exist two sets A and B such that $R \subset A \times B$. The proof is elementary. Suppose indeed that $x \in R$, so that $x = \{\{a\}, \{a, b\}\}$ for some a and for some b. The problem is to dig out a and b from under the braces. Since the elements of R are sets, we can form the union of the sets in R; since x is one of the sets in R, the elements of x belong to that union. Since $\{a, b\}$ is one of the elements of x, we may write, in what has been called the brutal notation above, $\{a, b\} \in \bigcup R$. One set of braces has disappeared; let us do the same thing again to make the other set go away. Form the union of the sets in $\bigcup R$. Since $\{a, b\}$ is one of those sets, it follows that the elements of $\{a, b\}$ belong to that union, and hence both a and b belong to $\bigcup \bigcup R$. This fulfills the promise made above; to exhibit R as a subset of some $A \times B$, we may take both A and B to be $\bigcup \bigcup R$. It is often desirable to take A and B as small as possible. To do so, just apply the axiom of specification to produce the sets

$$A = \{a : \textit{ for some } b\ ((a, b) \in R)\}$$

and

$$B = \{b : \textit{ for some } a\ ((a, b) \in R)\}.$$

These sets are called the *projections* of R onto the first and second coordinates respectively.

However important set theory may be now, when it began some scholars considered it a disease from which, it was to be hoped, mathematics would soon recover. For this reason many set-theoretic considerations were called pathological, and the word lives on in mathematical usage; it often refers to

something the speaker does not like. The explicit definition of an ordered pair ($(a, b) = \{\{a\}, \{a, b\}\}$) is frequently relegated to pathological set theory. For the benefit of those who think that in this case the name is deserved, we note that the definition has served its purpose by now and will never be used again. We need to know that ordered pairs are determined by and uniquely determine their first and second coordinates, that Cartesian products can be formed, and that every set of ordered pairs is a subset of some Cartesian product; which particular approach is used to achieve these ends is immaterial.

It is easy to locate the source of the mistrust and suspicion that many mathematicians feel toward the explicit definition of ordered pair given above. The trouble is not that there is anything wrong or anything missing; the relevant properties of the concept we defined are all correct (that is, in accord with the demands of intuition) and all the correct properties are present. The trouble is that the concept has some irrelevant properties that are accidental and distracting. The theorem that $(a, b) = (x, y)$ if and only if $a = x$ and $b = y$ is the sort of thing we expect to learn about ordered pairs. The fact that $\{a, b\} \in (a, b)$, on the other hand, seems accidental; it is a freak property of the definition rather than an intrinsic property of the concept.

The charge of artificiality is true; but it is not too high a price to pay for conceptual economy. The concept of an ordered pair could have been introduced as an additional primitive, axiomatically endowed with just the right properties, no more and no less. In some theories this is done. The mathematician's choice is between having to remember a few more axioms and having to forget a few accidental facts; the choice is pretty clearly a matter of taste. Similar choices occur frequently in mathematics; in this book, for instance, we shall encounter them again in connection with the definitions of numbers of various kinds.

Exercise 6.1. If A, B, X, and Y are sets, then

$$(i)\ (A \cup B) \times X = (A \times X) \cup (B \times X),$$
$$(ii)\ (A \cap B) \times (X \cap Y) = (A \times X) \cap (B \times Y),$$
$$(iii)\ (A - B) \times X = (A \times X) - (B \times X).$$

If either $A = \emptyset$ or $B = \emptyset$, then $A \times B = \emptyset$, and conversely. If $A \subset X$ and $B \subset Y$, then $A \times B \subset X \times Y$, and (provided $A \times B \neq \emptyset$) conversely.

CHAPTER 7

RELATIONS

Using ordered pairs, we can formulate the mathematical theory of relations in set-theoretic language. By a relation we mean here something like marriage (between men and women) belonging (between elements and sets). More explicitly, what we shall call a relation is sometimes called a *binary* relation. An example of a ternary relation is parenthood for people (Adam and Eve are the parents of Cain). In this book we shall have no occasion to treat the theory of relations that are ternary, quaternary, or worse.

Looking at any specific relation, such as marriage for instance, we might be tempted to consider certain ordered pairs (x, y), namely just those for which x is man, y is a woman, and x is married to y. We have not yet seen the definition of the general concept of a relation, but it seems plausible that, just as in this marriage example, every relation should uniquely determine the set of all those ordered pairs for which the first coordinate does stand in that relation to the second. If we know the relation, we know the set, and, better yet, if we know the set, we know the relation. If, for instance, we were presented with the set of ordered pairs of people that corresponds to marriage, then, even if we forgot the definition of marriage, we could always tell when a man x is married to a woman y and when not; we would just have to see whether the ordered pair (x, y) does or does not belong to the set.

We may not know what a relation is, but we do know what a set is, and the preceding considerations establish a close connection between relations and sets. The precise set-theoretic treatment of relations takes advantage of that heuristic connection; the simplest to do is to define a relation to be the corresponding set. This is what we do; we hereby define a *relation* as a set of ordered pairs. Explicitly: a set R is a relation if each element of R is an ordered pair; this

means, of course, that if $z \in R$, then there exist x and y so that $z = (x, y)$. If R is relation, it is sometimes convenient to express the fact that $(x, y) \in R$ by writing

$$xRy$$

and saying, as in everyday language, that x stands in the relation R to y.

The least exciting relation is the empty one. (To prove that $\emptyset$ is a set of ordered pairs, look for an element of $\emptyset$ that is not an ordered pair.) Another dull example is the Cartesian product of any two sets X and Y. Here is a slightly more interesting example: let X be any set, and let R be the set of all those pairs (x, y) in $X \times X$ for which $x = y$. The relation R is just the relation of equality between elements of X; if x and y are in X, then xRy means the same as $x = y$. One more example will suffice for now: let X be any set, and let R be the set of all those pairs (x, A) in $X \times \mathcal{P}(X)$ for which $x \in A$. This relation R is just the relation of belonging between elements of X and subsets of X; if $x \in X$ and $A \in \mathcal{P}(X)$, then xRA means the same as $x \in A$.

In the preceding section we saw that associated with every set R of ordered pairs there are two sets called the projections of R onto the first and second coordinates. In the theory of relations these sets are known as the *domain* and the *range* of R (abbreviated dom R and ran R); we recall that they are defined by

$$\text{dom } R = \{x : \textit{for some } y\ (xRy)\}$$

and

$$\text{ran } R = \{y : \textit{ for some } x\ (xRy)\}.$$

If R is the relation of marriage, so that xRy means that x is a man, y is a woman, and x and y are married to one another; then dom R is the set of married men and ran R is the set of married womem. Both the domain and the range of $\emptyset$ are equal to $\emptyset$. If $R = X \times Y$, then dom $R = X$ and ran $R = Y$. If R is equality in X, then dom R = ran $R = X$. If R is belonging, between X and $\mathcal{P}(X)$, then dom $R = X$ and ran $R = \mathcal{P}(X) - \{\emptyset\}$.

If R is a relation included in a Cartesian product $X \times Y$ (so that dom $R \subset X$ and ran $R \subset Y$), it is sometimes convenient to say that R is a relation *from* X *to* Y; instead of a relation from X to X we may speak of a relation *in* X. A relation R in X is *reflexive* if xRx for every x in X; it is *symmetric* if xRy implies that yRx; and it is *transitive* if xRy and yRz imply that xRz. (Exercise: for each of these three possible properties, find a relation that does not have that property but does have the other two.) A relation in a set is an *equivalence relation* if

it is reflexive, symmetric, and transitive. The smallest equivalence relation in a set X is the relation of equality in X; the largest equivalence relation in X is $X \times X$.

There is an intimate connection between equivalence relations in a set X and certain collections (called partitions) of subsets of X. A *partition* of X is a disjoint collection $\mathcal{C}$ of non-empty subsets of X whose union is X. If R is an equivalence relation in X, and if x is in X, the *equivalence class* of x with respect to R is the set of all those elements y in X for which xRy. (The weight of tradition makes the use of the word "class" at this point unavoidable.) Examples: if R is equality in X, then each equivalence class is a singleton; if $R = X \times X$, then the set X itself is the only equivalence class. There is no standard notation for the equivalence class of x with respect to R; we shall usually denote it by x/R, and we shall write X/R for the set of all equivalence classes. (Pronounce X/R as "X modulo R," or, in abbreviated form, "X mod R." Exercise: show that X/R is indeed a set by exhibiting a condition that specifies exactly the subset X/R of the power set $\mathcal{P}(X)$.) Now forget R for a moment and begin anew with a partition $\mathcal{C}$ of X. A relation, which we shall call $X/\mathcal{C}$, is defined in X by writing

$$x \; X/\mathcal{C} \; y$$

just in case x and y belong to the same set of the collection $\mathcal{C}$. We shall call $X/\mathcal{C}$ the relation *induced* by the partition $\mathcal{C}$.

In the preceding paragraph we saw how to associate a set of subsets of X with every equivalence relation in X and how to associate a relation in X with every partition of X. The connection between equivalence relations and partitions can be described by saying that the passage from $\mathcal{C}$ to $X/\mathcal{C}$ is exactly the reverse of the passage from R to X/R. More explicitly: if R is an equivalence relation in X, then the set of equivalence classes is a partition of X that induces the relation R, and if $\mathcal{C}$ is a partition of X, then the induced relation is an equivalence relation whose set of equivalence classes is exactly $\mathcal{C}$.

For the proof, let us start with an equivalence relation R. Since each x belongs to some equivalence class (for instance $x \in x/R$), it is clear that the union of the equivalence classes is all X. If $z \in x/R \cap y/R$, then xRz and zRy, and therefore xRy. This implies that if two equivalence classes have an element in common, then they are identical, or, in other words, that two distinct equivalence classes are always disjoint. The set of equivalence classes is therefore a partition. To say that two elements belong to the same set (equivalence class) of this partition means, by definition, that they stand in the relation R to one another. This proves the first half of our assertion.

The second half is easier. Start with a partition $\mathcal{C}$ and consider the induced

relation. Since every element of X belongs to some set of $\mathcal{C}$, reflexivity just says that x and x are in the same set of $\mathcal{C}$. Symmetry says that if x and y are in the same set of $\mathcal{C}$, then y and x are in the same set of $\mathcal{C}$, and this is obviously true. Transitivity says that if x and y are in the same set of $\mathcal{C}$ and if y and z are in the same set of $\mathcal{C}$, then x and z are in the same set of $\mathcal{C}$, and this too is obvious. The equivalence class of each x in X is just the set of $\mathcal{C}$ to which x belongs. This completes the proof of everything that was promised.

CHAPTER 8

FUNCTIONS

If X and Y are sets, a *function* from (or *on*) X *to* (or *into*) Y is a relation f such that dom $f = X$ and such that for each x in X there is a unique element y in Y with $(x, y) \in f$. The uniqueness condition can be formulated explicitly as follows: if $(x, y) \in f$ and $(x, z) \in f$, then $y = z$. For each x in X, the unique y in Y such that $(x, y) \in f$ is denoted by $f(x)$. For functions this notation and its minor variants supersede the others used for more general relations; from now on, if f is a function, we shall write $f(x) = y$ instead of $(x, y) \in f$ or xfy. The element y is called the *value* that the function f *assumes* (or *takes on*) at the *argument* x; equivalently we may say that f *sends* or *maps* or *transforms* x into y. The words *map* or *mapping, transformation, correspondence,* and *operator* are among some of the many that are sometimes used as synonyms for *function.* The symbol

$$f : X \to Y$$

is sometimes used as an abbreviation for "f is a function from X to Y." The set of all functions X to Y is a subset of the power set $\mathcal{P}(X \times Y)$; it will be denoted by Y^X.

The connotations of activity suggested by the synonyms listed above make some scholars dissatisfied with the definition according to which function does not *do* anything but merely *is.* This dissatisfaction is reflected in a different use of the vocabulary: *function* is reserved for the undefined object that is somehow active, and the set of ordered pairs that we have called the function is then called the *graph* of the function. It is easy to find examples of functions in the precise set-theoretic sense of the word in both mathematics and everyday life; all we have to look for is information, not necessarily numerical, in tabulated

form. One example is a city directory; the arguments of the function are, in this case, the inhabitants of the city, and the values are their addresses.

For relations in general, and hence for functions in particular, we have defined the concepts of domain and range. The domain of a function f from X into Y is, by definition, equal to X, but its range need not be equal to Y; the range consists of those elements y of Y for which there exists an x in X such that $f(x) = y$. If the range of f is equal to Y, we say that f maps X *onto* Y. If A is a subset of X, we may want to consider the set of all those elements y of Y for which there exists an x in the subset A such that $f(x) = y$. This subset of Y is called the *image* of A under f and is frequently denoted by $f(A)$. The notation is bad but not catastrophic. What is bad about it is that if A happens to be both an element of X and a subset of X (an unlikely situation, but far from an impossible one), then the symbol $f(A)$ is ambiguous. Does it mean the value of f at A or does it mean the set of values of f at the elements of A? Following normal mathematical custom, we shall use the bad notation, relying on context, and, on the rare occasions when it is necessary, adding verbal stipulations, to avoid confusion. Note that the image of X itself is the range of f; the "onto" character of f can be expressed by writing $f(X) = Y$.

If X is a subset of a set Y, the function f defined by $f(x) = x$ for each x in X is called the *inclusion map* (or *embedding*, or the *injection*) of X into Y. The phrase "the function f defined by …" is a very common one in such contexts. It is intended to imply, of course, that there does indeed exist a unique function satisfying the stated condition. In the special case at hand this is obvious enough; we are being invited to consider the set of all those ordered pairs (x, y) in $X \times Y$ for which $x = y$. Similar considerations apply in every case, and, following normal mathematical practice, we shall usually describe a function by describing its value y at each argument x. Such a description is sometimes longer and more cumbersome than a direct description of the set (of ordered pairs) involved, but, nevertheless, most mathematicians regard the argument-value description as more perspicuous than any other.

The inclusion map of X into X is called the *identity map* on X. (In the language of relations, the identity map on X is the same as the relation of equality in X.) If, as before, $X \subset Y$, then there is a connection between the inclusion map of X into Y and the identity map on Y; that connection is a special case of a general procedure for making small functions out of large ones. If f is a function from Y to Z, say, and if X is a subset of Y, then there is a natural way of constructing a function g from X to Z; define $g(x)$ to be equal to $f(x)$ for each x in X. The function g is called the *restriction* of f to X, and f is called an *extension* of g to Y; it is customary to write $g = f \mid X$. The definition of restriction can be expressed by writing $(f \mid X)(x) = f(x)$ for each x in X; observe also that $\operatorname{ran}(f \mid X) = f(X)$. The inclusion map of a subset of

Y is the restriction to that subset of the identity map on Y.

Here is a simple but useful example of a function. Consider any two sets X and Y, and define a function f from $X \times Y$ onto X by writing $f(x, y) = x$. (The purist will have noted that we should have written $f((x, y))$ instead of $f(x, y)$, but nobody ever does.) The function f is called the *projection* from $X \times Y$ onto X; if, similarly, $g(x, y) = y$, then g is the projection from $X \times Y$ onto Y. The terminology here is at variance with an earlier one, but not badly. If $R = X \times Y$, then what was earlier called the projection of R onto the first coordinate is, in the present language, the range of the projection f.

A more complicated and correspondingly more valuable example of a function can be obtained as follows. Suppose R is an equivalence relation in X, and let f be the function from X onto X/R defined by $f(x) = x/R$. The function f is sometimes called the *canonical map* from X to X/R.

If f is an arbitrary function, from X onto Y, then there is a natural way of defining an equivalence relation R in X; write aRb (where a and b are in X) in case $f(a) = f(b)$. For each element y of Y, let $g(y)$ be the set of all those elements x in X for which $f(x) = y$. The definition of R implies that $g(y)$ is, for each y, an equivalence class of the relation R; in other words, g is a function from Y onto the set X/R of all equivalence classes of R. The function g has the following special property: if u and v are distinct elements of Y, then $g(u)$ and $g(v)$ are distinct elements of X/R. A function that always maps distinct elements onto distinct elements is called *one-to-one* (usually a *one-to-one correspondence*). Among the examples above the inclusion maps are one-to-one, but, except in some trivial special cases, the projections are not. (Exercise: what special cases?)

To introduce the next aspect of the elementary theory of functions we must digress for a moment and anticipate a tiny fragment of our ultimate definition of natural numbers. We shall not find it necessary to define all the natural numbers now; all we need is the first three of them. Since this is not the appropriate occasion for lengthy heuristic preliminaries, we shall proceed directly to the definition, even at the risk of temporarily shocking or worrying some readers. Here it is: we define 0, 1, and 2 by writing

$$0 = \emptyset,\ 1 = \{\emptyset\},\ \text{and}\ 2 = \{\emptyset, \{\emptyset\}\}.$$

In other words, 0 is empty, 1 is the singleton $\{0\}$, and 2 is the pair $\{0, 1\}$. Observe that there is some method in this apparent madness; the number of elements in the sets 0, 1, or 2 (in the ordinary everyday sense of the word) is, respectively, zero, one, or two.

If A is a subset of a set X, the *characteristic function* of A is the function χ from X to 2 such that $\chi(x) = 1$ or 0 according as $x \in A$ or $x \in X - A$. The

dependence of the characteristic function of A on the set A may be indicated by writing χ_A instead of χ. The function that assigns to each subset A of X (that is, to each element of $\mathcal{P}(X)$) the characteristic function of A (that is an element of 2^X) is a one-to-one correspondence between $\mathcal{P}(X)$ and 2^X. (Parenthetically: instead of the phrase "the function that assigns to each A in $\mathcal{P}(X)$ the element χ_A in 2^X" it is customary to use the abbreviation "the function $A \to \chi_A$." In this language, the projection from $X \times Y$ onto X, for instance, may be called the function $(x, y) \to x$, and the canonical map from a set X with a relation R onto X/R may be called the function $x \to x/R$.)

Exercise 8.1. (*i*) $Y^\emptyset$ has exactly one element, namely $\emptyset$, wheter Y is empty or not, and (*ii*) if X is not empty, then $\emptyset^X$ is empty.

CHAPTER 9

FAMILIES

There are occasions when the range of a function is deemed to be more important than the function itself. When that is the case, both the terminology and the notation undergo radical alterations. Suppose, for instance, that x is a function from a set I to a set X. (The very choice of letters indicates that something strange is afoot.) An element of the domain I is called an *index*, I is called the *index set*, the range of the function is called an *indexed set*, the function itself is called a *family*, and the value of the function x at an index i, called a *term* of the family, is denoted by x_i. (This terminology is not absolutely established, but it is one of the standard choices among related slight variants; in the sequel it and it alone will be used.) An unacceptable but generally accepted way of communicating the notation and indicating the emphasis is to speak of a family $\{x_i\}$ in X, or of a family $\{x_i\}$ of whatever the elements of X may be; when necessary, the index set I is indicated by some such parenthetical expression as $(i \in I)$. Thus, for instance, the phrase "a family $\{A_i\}$ of subsets of X" is usually understood to refer to a function A, from some set I of indices, into $\mathcal{P}(X)$.

If $\{A_i\}$ is a family of subsets of X, the union of the range of the family is called the union of the family $\{A_i\}$, or the union of the sets A_i; the standard notation for it is

$$\bigcup_{i \in I} A_i \quad \text{or} \quad \bigcup_i A_i,$$

according as it is or is not important to emphasize the index set I. It follows immediately from the definition of unions that $x \in \bigcup_i A_i$ if and only if x belongs to A_i for at least one i. If $I = 2$, so that the range of the family $\{A_i\}$ is the unordered pair $\{A_0, A_i\}$, then $\bigcup_i A_i = A_0 \cup A_1$. Observe that there is no loss of

generality in considering families of sets instead of arbitrary collections of sets; every collection of sets is the range of some family. If, indeed, $\mathcal{C}$ is a collection of sets, let $\mathcal{C}$ itself play the role of the index set, and consider the identity mapping on $\mathcal{C}$ in the role the family.

The algebraic laws satisfied by the operation of union for pairs can be generalized to arbitrary unions. Suppose, for instance, that $\{I_j\}$ is a family of sets with domain J, say; write $K = \bigcup_j I_j$, and let $\{A_k\}$ be a family of sets with domain K. It is then not difficult to prove that

$$\bigcup_{k\in K} A_k = \bigcup_{j\in J} \left(\bigcup_{i\in I_j} A_i \right);$$

this is the generalized version of the associative law for unions. Exercise: formulate and prove a generalized version of the commutative law.

An empty union makes sense (and is empty), but an empty intersection does not make sense. Except for this triviality, the terminology and notation for intersections parallels that for unions in every respect. Thus, for instance, if $\{A_i\}$ is a non-empty family of sets, the intersection of the range of the family is called the intersection of the family $\{A_i\}$, or the intersection of the sets A_i; the standard notation for it is

$$\bigcap_{i\in I} A_i \text{ or } \bigcap_i A_i,$$

according as it is or is not important to emphasize the index set I. (By a "non-empty family" we mean a family whose domain I is not empty.) It follows immediately from the definition of intersections that if $I \neq \emptyset$, then a necessary and sufficient condition that x belong $\bigcap_i A_i$ is that x belong to A_i for all i.

The generalized commutative and associative laws for intersections can be formulated and proved the same way as for unions, or, alternatively, De Morgan's laws can be used to derive them from the facts for unions. This is almost obvious, and, therefore, it is not of much interest. The interesting algebraic identities are the ones that involve both unions and intersections. Thus, for instance, if $\{A_i\}$ is a family of subsets of X and $B \subset X$, then

$$B \cap \bigcup_i A_i = \bigcup_i (B \cap A_i)$$

and

$$B \cup \bigcap_i A_i = \bigcap_i (B \cup A_i);$$

these equations are a mild generalization of the distributive laws.

Exercise 9.1. If both $\{A_i\}$ and $\{B_i\}$ are families of sets, then

$$\left(\bigcup_i A_i\right) \cap \left(\bigcup_j B_j\right) = \bigcup_{i,j}(A_i \cap B_j)$$

and

$$\left(\bigcap_i A_i\right) \cup \left(\bigcap_j B_j\right) = \bigcap_{i,j}(A_i \cup B_j).$$

Explanation of notation: a symbol such as $\bigcup_{i,j}$ is an abbreviation for $\bigcup_{(i,j)\in I\times J}$.

The notation of families is the one normally used in generalizing the concept of Cartesian product. The Cartesian product of two sets X and Y was defined as the set of all ordered pairs (x, y) with x in X and y in Y. There is a natural one-to-one correspondence between this set and a certain set of families. Consider, indeed, any particular unordered pair $\{a, b\}$, with $a \neq b$, and consider the set Z of all families z, indexed by $\{a, b\}$, such that $z_a \in X$ and $z_b \in Y$. If the function f from Z to $X \times Y$ is defined by $f(z) = (z_a, z_b)$, then f is the promised one-to-one correspondence. The difference between Z and $X \times Y$ is merely matter of notation. The generalization of Cartesian products generalizes Z rather than $X \times Y$ itself. (As a consequence there is a little terminological friction in the passage from the special case to the general. There is no help for it; that is how mathematical language is in fact used nowadays.) The generalization is now straightforward. If $\{X_i\}$ is a family of sets $(i \in I)$, the *Cartesian product* of the family is, by definition, the set of all families $\{x_i\}$ with $x_i \in X_i$ for each i in I. There are several symbols for the Cartesian product in more or less current usage; in this book we shall denote it by

$$\bigtimes_{i\in I} X_i \quad \text{or} \quad \bigtimes_i X_i.$$

It is clear that if every X_i is equal to one and the same set X, then $\bigtimes_i X_i = X^I$. If I is a pair $\{a, b\}$, with $a \neq b$, then it is customary to identify $\bigtimes_{i\in I} X_i$ with the Cartesian product $X_a \times X_b$ as defined earlier, and if I is a singleton $\{a\}$, then, similarly, we identify $\bigtimes_{i\in I} X_i$ with X_a itself. *Ordered triples, ordered quadruples*, etc., may be defined as families whose index sets are unordered triples, quadruples, etc.

Suppose that $\{X_i\}$ is a family of sets $(i \in I)$ and let X be its Cartesian product. If J is a subset of I, then to each element of X there corresponds in a natural way an element of the partial Cartesian product $\bigtimes_{i\in J} X_i$. To define the

correspondence, recall that each element x of X is itself a family $\{x_i\}$, that is, in the last analysis, a function on I; the corresponding element, say y, of $\bigtimes_{i\in J} X_i$ is obtained by simply restricting that function to J. Explicitly, we write $y_i = x_i$ whenever $i \in J$. The correspondence $x \to y$ is called the projection from X onto $\bigtimes_{i\in J} X_i$; we shall temporarily denote it by f_J. If, in particular, J is a singleton, say $J = \{j\}$, then we shall write f_j (instead of $f_{\{j\}}$) for f_J. The word "projection" has a multiple use; if $x \in X$, the value of f_j at x, that is x_j, is also called the projection of x onto X_j, or, alternatively, the *j-coordinate* of x. A function on a Cartesian product such as X is called a function of *several variables*, and, in particular, a function on a Cartesian product $X_a \times X_b$ is called a function of two variables.

Exercise 9.2. Prove that $(\bigcup_i A_i) \times (\bigcup_j B_j) = \bigcup_{i,j}(A_i \times B_j)$, and that the same equation holds for intersections (provided that the domains of the families involved are not empty). Prove also (with appropriate provisos about empty families) that $\bigcap_i X_i \subset X_j \subset \bigcup_i X_i$ for each index j and that intersection and union can in fact be characterized as the extreme solutions of these inclusions. This means that if $X_j \subset Y$ for each index j, then $\bigcup_i X_i \subset Y$, and that $\bigcup_i X_i$ is the only set satisfying this minimality condition; the formulation for intersections is similar.

CHAPTER 10

INVERSES AND COMPOSITES

Associated with every function f, from X to Y, say, there is a function from $\mathcal{P}(X)$ to $\mathcal{P}(Y)$, namely the function (frequently called f also) that assigns to each subset A of X the image subset $f(A)$ of Y. The algebraic behavior of the mapping $A \to f(A)$ leaves something to be desired. It is true that if $\{A_i\}$ is a family of subsets of X, then $f(\bigcup_i A_i) = \bigcup_i f(A_i)$ (proof?), but the corresponding equation for intersections is false in general (example?), and the connection between images and complements is equally unsatisfactory.

A correspondence between the elements of X and the elements of Y does always induce a well-behaved correspondence between the subsets of X and the subsets of Y, not forward, by the formation of images, but backward, by the formation of inverse images. Given a function f from X to Y, let f^{-1}, the *inverse* of f, be the function from $\mathcal{P}(Y)$ to $\mathcal{P}(X)$ such that if $B \subset Y$, then

$$f^{-1}(B) = \{x \in X : f(x) \in B\}.$$

In words: $f^{-1}(B)$ consists of exactly those elements of X that f maps into B; the set $f^{-1}(B)$ is called the *inverse image* of B under f. A necessary and sufficient condition that f map X onto Y is that the inverse image under f of each non-empty subset of Y be a non-empty subset of X. (Proof?) A necessary and sufficient condition that f be one-to-one is that the inverse image under f of each singleton in the range of f be a singleton in X.

If the last condition is satisfied, then the symbol f^{-1} is frequently assigned a second interpretation, namely as the function whose domain is the range of f, and whose value for each y in the range of f is the unique x in X for which $f(x) = y$. In other words, for one-to-one functions f we may write $f^{-1}(y) = x$ if and only if $f(x) = y$. This use of the notation is mildly inconsistent with our

first interpretation of f^{-1}, but the double meaning is not likely to lead to any confusion.

The connection between images and inverse images is worth a moment's consideration.

If $B \subset Y$, then

$$f(f^{-1}(B)) \subset B.$$

Proof. If $y \in f(f^{-1}(B))$, then $y = f(x)$ for some x in $f^{-1}(B)$; this means that $y = f(x)$ and $f(x) \in B$, and therefore $y \in B$. □

If f maps X onto Y, then

$$f(f^{-1}(B)) = B.$$

Proof. If $y \in B$, then $y = f(x)$ for some x in X, and therefore for some x in $f^{-1}(B)$; this means that $y \in f(f^{-1}(B))$. □

If $A \subset X$, then

$$A \subset f^{-1}(f(A))$$

Proof. If $x \in A$, then $f(x) \in f(A)$; this means that $x \in f^{-1}(f(A))$. □

If f is one-to-one, then

$$A = f^{-1}(f(A))$$

Proof. If $x \in f^{-1}(f(A))$, then $f(x) \in f(A)$, and therefore $f(x) = f(u)$ for some u in A; this implies that $x = u$ and hence that $x \in A$. □

The algebraic behavior of f^{-1} is unexceptionable. If $\{B_i\}$ is a family of subsets of Y, then

$$f^{-1}\left(\bigcup_i B_i\right) = \bigcup_i f^{-1}(B_i)$$

and

$$f^{-1}\left(\bigcap_i B_i\right) = \bigcap_i f^{-1}(B_i)$$

The proofs are straightforward. If, for instance, $x \in f^{-1}(\bigcap_i B_i)$, then $f(x) \in B_i$ for all i, so that $x \in f^{-1}(B_i)$ for all i, and therefore $x \in \bigcap_i f^{-1}(B_i)$; all the

steps in this argument are reversible. The formation of inverse images commutes with complementation also; i.e.,

$$f^{-1}(Y - B) = X - f^{-1}(B)$$

for each subset B of Y. Indeed: if $x \in f^{-1}(Y - B)$, then $f(x) \in Y - B$, so that $x \notin f^{-1}(B)$, and therefore $x \in X - f^{-1}(B)$; the steps are reversible. (Observe that the last equation is indeed a kind of commutative law: it says that complementation followed by inversion is the same as inversion followed by complementation.)

The discussion of inverses shows that what a function does can in a certain sense be undone; the next thing we shall see is that what two functions do can sometimes be done in one step. If, to be explicit, f is a function from X to Y and g is a function from Y to Z, then every element in the range of f belongs to the domain of g, and, consequently, $g(f(x))$ makes sense for each x in X. The function h from X to Z, defined by $h(x) = g(f(x))$ is called the *composite* of the functions f and g; it is denoted by $g \circ f$ or, more simply, by gf. (Since we shall not have occasion to consider any other kind of multiplication for functions, in this book we shall use the latter, simpler notation only.)

Observe that the order of events is important in the theory of functional composition. In order that gf be defined, the range of f must be included in the domain of g, and this can happen without it necessarily happening in the other direction at the same time. Even if both fg and gf are defined, which happens if, for instance, f maps X into Y and g maps Y into X, the functions fg and gf need not be the same; in other words, functional composition is not necessarily commutative.

Functional composition may not be commutative, but it is always associative. If f maps X into Y, if g maps Y into Z, and if h maps Z into U, then we can form the composite of h with gf and the composite of hg with f; it is a simple exercise to show that the result is the same in either case.

The connection between inversion and composition is important; something like it crops up all over mathematics. If f maps X into Y and g maps Y into Z, then f^{-1} maps $\mathcal{P}(Y)$ into $\mathcal{P}(X)$ and g^{-1} maps $\mathcal{P}(Z)$ into $\mathcal{P}(Y)$. In this situation, the composites that are formable are gf and $f^{-1}g^{-1}$; the assertion is that the latter is the inverse of the former. Proof: if $x \in (gf)^{-1}(C)$, where $x \in X$ and $C \subset Z$, then $g(f(x)) \in C$, so that $f(x) \in g^{-1}(C)$, and therefore $x \in f^{-1}(g^{-1}(C))$; the steps of the argument are reversible.

Inversion and composition for functions are special cases of similar operations for relations. Thus, in particular, associated with every relation R from X to Y there is the *inverse* (or *converse*) relation R^{-1} from Y to X; by definition $yR^{-1}x$ means that xRy. Example: if R is the relation of belonging, from X to $\mathcal{P}(X)$,

then R^{-1} is the relation of containing, from $\mathcal{P}(X)$ to X. It is an immediate consequence of the definitions involved that dom $R^{-1} = \text{ran } R$ and ran $R^{-1} = \text{dom } R$. If the relation R is a function, then the equivalent assertions xRy and $yR^{-1}x$ can be written in the equivalent forms $R(x) = y$ and $x \in R^{-1}(\{y\})$.

Because of difficulties with commutativity, the generalization of functional composition has to be handled with care. The composite of the relations R and S is defined in case R is a relation from X to Y and S is a relation from Y to Z. The composite relation T, from X to Z, is denoted by $S \circ R$, or, simply, by SR; it is defined so that xTz if and only if there exists an element y in Y such that xRy and ySz. For an instructive example, let R mean "son" and let S mean "brother" in the set of human males, say. In other words, xRy means that x is son of y, and ySz means that y is a brother of z. In this case the composite relation SR means "nephew." (Query: what do R^{-1}, S^{-1}, RS, and $R^{-1}S^{-1}$ mean?) If both R and S are functions, then xRy and ySz can be rewritten as $R(x) = y$ and $S(y) = z$ respectively. It follows that $S(R(x)) = z$ if and only if xTz, so that functional composition is indeed a special case of what is sometimes called the *relative product*.

The algebraic properties of inversion and composition are the same for relations as for functions. Thus, in particular, composition is commutative by accident only, but it is always associative, and it is always connected with inversion via the equation $(SR)^{-1} = R^{-1}S^{-1}$. (Proofs?)

The algebra of relations provides some amusing formulas. Suppose that, temporarily, we consider relations in one set X only, and, in particular, let the relation of equality in X (which is the same as the identity mapping on X). The relation I acts as a multiplicative unit; this means that $IR = RI = R$ for every relation R in X. Query: is there a connection among I, RR^{-1}, and $R^{-1}R$? The three defining properties of an equivalence relation can be formulated in algebraic terms as follows: reflexivity means $I \subset R$, symmetry means $R \subset R^{-1}$, and transitivity means $RR \subset R$.

Exercise 10.1. (Assume in each case that f is a function from X to Y) (*i*) If g is a function from Y to X such that gf is the identity on X, then f is one-to-one and g maps Y onto X. (*ii*) A necessary and sufficient condition that $f(A \cap B) = f(A) \cap f(B)$ for all subsets A and B of X is that f be one-to-one. (*iii*) A necessary and sufficient condition that $f(X - A) \subset Y - f(A)$ for all subsets A of X is that f be one-to-one. (*iv*) A necessary and sufficient condition that $Y - f(A) \subset f(X - A)$ for all subsets A of X is that f map X onto Y.

CHAPTER 11

NUMBERS

How much is two? How, more generally, are we to define numbers? To prepare for the answer, let us consider a set X and let us form the collection P of all unordered pairs $\{a, b\}$, with a in X, b in X, and $a \neq b$. It seems clear that all the sets in the collection P have a property in common, namely the property of consisting of two elements. It is tempting to try to define "twoness" as the common property of all the sets in the collection P, but the temptation must be resisted; such a definition is, after all, mathematical nonsense. What is a "property"? How do we know that there is only one property in common to all the sets in P?

After some cogitation we might hit upon a way of saving the idea behind the proposed definition without using vague expressions such as "the common property". It is ubiquitous mathematical practice to identify a property with a set, namely with the set of all objects that possess the property; why not do it here? Why not, in other words, define "two" as the set P? Something like this is done at times, but it is not completely satisfying. The trouble is that our present modified proposal depends on P, and hence ultimately on X. At best the proposal defines twoness for subsets of X; it gives no hint as to when we may attribute twoness to a set that is not included in X.

There are two ways out. One way is to abandon the restriction to a particular set and to consider instead all possible unordered pairs $\{a, b\}$ with $a \neq b$. These unordered pairs do not constitute a set; in order to base the definition of "two" on them, the entire theory under consideration would have to be extended to include the "unsets" (classes) of another theory. This can be done, but it will not be done here; we shall follow a different route.

How would a mathematician define a meter? The procedure analogous to the

one sketched above would involve the following two steps. First, select an object that is one of the intended models of the concept being defined—an object, in other words, such that on intuitive or practical grounds it deserves to be called one meter if anything does. Second, form the set of all objects in the universe that are of the same length as the selected one (note that this does not depend on knowing what a meter is), and define a meter as the set so formed.

How in fact is a meter defined? The example was chosen so that the answer to this question should suggest an approach to the definition of numbers. The point is that in the customary definition of a meter the second step is omitted. By a more or less arbitrary convention an object is selected and its length is called a meter. If the definition is accused of circularity (what does "length" mean?), it can easily be converted into an unexceptionable demonstrative definition; there is after all nothing to stop us from defining a meter as equal to the selected object. If this demonstrative approach is adopted, it is just as easy to explain as before when "one-meter-ness" shall be attributed to some other object, namely, just in case the new object has the same length as the selected standard. We comment again that to determine whether two objects have the same length depends on a simple act of comparison only, and does not depend on having a precise definition of length.

Motivated by the considerations described above, we have earlier defined 2 as some particular set with (intuitively speaking) exactly two elements. How was that standard set selected? How should other such standard sets for other numbers be selected? There is no compelling mathematical reason for preferring one answer to this question to another; the whole thing is largely a matter of taste. The selection should presumably be guided by considerations of simplicity and economy. To motivate the particular selection that is usually made, suppose that a number, say 7, has already been defined as a set (with seven elements). How in this case, should we define 8? Where, in other words, can we find a set consisting of exactly eight elements? We can find seven elements in the set 7; what shall we use as an eighth to adjoin to them? A reasonable answer to the last question is the number (set) 7 itself; the proposal is to define 8 to be the set consisting of the seven elements of 7, together with 7. Note that according to this proposal each number will be equal to the set of its own predecessors.

The preceding paragraph motivates a set-theoretic construction that makes sense for every set, but that is of interest in the construction of numbers only. For every set x we define the *successor* x^+ of x to be the set obtained by adjoining x to the elements of x; in other words,

$$x^+ = x \cup \{x\}.$$

(The successor of x is frequently denoted by x'.)

We are now ready to define the natural numbers. In defining 0 to be a set with zero elements, we have no choice; we must write (as we did)

$$0 = \emptyset$$

If every natural number is to be equal to the set of its predecessors, we have no choice in defining 1, or 2, or 3 either; we must write

$$\begin{aligned} 1 &= 0^+ (= \{0\}), \\ 2 &= 1^+ (= \{0, 1\}), \\ 3 &= 2^+ (= \{0, 1, 2\}), \end{aligned}$$

etc. The "etc." means that we hereby adopt the usual notation, and, in what follows, we shall feel free to use numerals such as "4" or "956" without any further explanation or apology.

From what has been said so far it does not follow that the construction of successors can be carried out ad infinitum within one and the same set. What we need is a new set-theoretic principle.

Axiom 11.1 (Axiom of infinity)**.** *There exists a set containing* 0 *and containing the successor of each of its elements.*

The reason for the name of the axiom should be clear. We have not yet given a precise definition of infinity, but it seems reasonable that sets such as the ones that the axiom of infinity describes deserve to be called infinite.

We shall say, temporarily, that a set A is a *successor set* if $0 \in A$ and if $x^+ \in A$ whenever $x \in A$. In this language the axiom of infinity simply says that there exists a successor set A. Since the intersection of every (non-empty) family of successor sets is a successor set itself (proof?), the intersection of all the successor sets included in A is a successor set ω. The set ω is a subset of every successor set. If, indeed, B is an arbitrary successor set, then so is $A \cap B$. Since $A \cap B \subset A$, the set $A \cap B$ is one of the sets that entered into the definition of ω; it follows that $\omega \subset A \cap B$, and, consequently, that $\omega \subset B$. The minimality property so established uniquely characterizes ω; the axiom of extension guarantees that there can be only one successor set that is included in every other successor set. A *natural number* is, by definition, an element of the minimal successor set ω. This definition of natural numbers is the rigorous counterpart of the intuitive description according to which they consist of $0, 1, 2, 3$, "and so on." Incidentally, the symbol we are using for the set of all natural numbers (ω) has a plurality of the votes of the writers on the subject, but nothing like a clear majority. In this book that symbol will be used systematically and exclusively in the sense defined above.

The slight feeling of discomfort that the reader may experience in connection with the definition of natural numbers is quite common and in most cases temporary. The trouble is that here, as once before (in the definition of ordered pairs), the object defined has some irrelevant structure, which seems to get in the way (but is in fact harmless). We want to be told that the successor of 7 is 8, but to be told that 7 is a subset of 8 or that 7 is an element of 8 is disturbing. We shall make use of this superstructure of natural numbers just long enough to derive their most important natural properties; after that the superstructure may safely be forgotten.

A family $\{x_i\}$ whose index set is either a natural number or else the set of all natural numbers is called a *sequence* (*finite* or *infinite*, respectively). If $\{A_i\}$ is a sequence of sets, where the index set is the natural number n^+, then the union of the sequence is denoted by

$$\bigcup_{i=0}^{n} A_i \text{ or } A_0 \cup \cdots \cup A_n.$$

If the index set is ω, the notation is

$$\bigcup_{i=0}^{\infty} A_i \text{ or } A_0 \cup A_1 \cup A_2 \cup \cdots.$$

Intersections and Cartesian products of sequences are denoted similarly by

$$\bigcap_{i=0}^{n} A_i, \quad A_0 \cup \cdots \cup A_n,$$
$$\bigtimes_{i=0}^{n} A_i, \quad A_0 \times \cdots \times A_n,$$

and

$$\bigcap_{i=0}^{\infty} A_i, \quad A_0 \cup A_1 \cup A_2 \cup \cdots,$$
$$\bigtimes_{i=0}^{\infty} A_i, \quad A_0 \times A_1 \times A_2 \times \cdots.$$

The word "sequence" is used in a few different ways in the mathematical literature, but the differences among them are more notational than conceptual. The most common alternative starts at 1 instead of 0; in other words, it refers to a family whose index set is $\omega - \{0\}$ instead of ω.

CHAPTER 12

THE PEANO AXIOMS

We enter now into a minor digression. The purpose of the digression is to make fleeting contact with the arithmetic theory of natural numbers. From the set-theoretic point of view this is a pleasant luxury.

The most important thing we know about the set ω of all natural numbers is that it is the unique successor set that is a subset of every successor set. To say that ω is a successor set means that

$$0 \in \omega \tag{12.1}$$

(where, of course, $0 = \emptyset$), and that

$$\textit{if } n \in \omega, \textit{ then } n^+ \in \omega \tag{12.2}$$

(where $n^+ = n \cup \{n\}$). The minimality property of ω can be expressed by saying that if a subset S of ω is a successor set, then $S = \omega$. Alternatively, and in more primitive terms,

$$\textit{if } S \subset \omega, \textit{ if } 0 \in S, \textit{ and if } n^+ \in S \textit{ whenever } n \in S, \textit{ then } S = \omega. \tag{12.3}$$

Property (Equation 12.3) is known as the **principle of mathematical induction.** We shall now add to this list of properties of ω two others:

$$n^+ \neq 0 \textit{ for all } n \in \omega, \tag{12.4}$$

and

$$\textit{if } n \textit{ and } m \textit{ are in } \omega, \textit{ and if } n^+ = m^+, \textit{ then } n = m. \tag{12.5}$$

The proof of Equation 12.4 is trivial; since n^+ contains n, and since 0 is empty, it is clear that n^+ is different from 0. The proof of Equation 12.5 is not trivial; it depends on a couple of auxiliary propositions. The first one asserts that something that ought not to happen indeed does not happen. Even if the considerations that the proof involves seem to be pathological and foreign to the arithmetic spirit that we expect to see in the theory of natural numbers, the end justifies the means. The second proposition refers to behavior that is quite similar to the one just excluded. This time, however, the apparently artificial considerations end in an affirmative result: something mildly surprising always does happen. The statements are as follows: (i) *no natural number is a subset of any of its elements*, and (ii) *every element of a natural number is a subset of it*. Sometimes a set with the property that it includes ($\subset$) everything that it contains ($\in$) is called a *transitive* set. More precisely, to say that E is transitive means that if $x \in y$ and $y \in E$, then $x \in E$. (Recall the slightly different use of the word that we encountered in the theory of relations.) In this language, (ii) says that every natural number is transitive.

The proof of (i) is a typical application of the principle of mathematical induction. Let S be the set of all those natural numbers n that are not included in any of their elements. (Explicitly: $n \in S$ if and only if $n \in \omega$ and n is not a subset of x for any x in n.) Since 0 is not a subset of any of its elements, it follows that $0 \in S$. Suppose now that $n \in S$. Since n is a subset of n, we may infer that n is not an element of n, and hence that n^+ is not a subset of n. What can n^+ be a subset of? If $n^+ \subset x$, then $n \subset x$, and therefore (since $n \in S$) $x \notin n$. It follows that n^+ cannot be a subset of n, and n^+ cannot be a subset of any element of n. This means that n^+ cannot be a subset of any element of n^+, and hence that $n^+ \in S$. The desired conclusion (i) is now a consequence of Equation 12.3.

The proof of (ii) is also inductive. This time let S be the set of all transitive natural numbers. (Explicitly: $n \in S$ if and only if $n \in \omega$ and x is a subset of n for every x in n.) The requirement that $0 \in S$ is vacuously satisfied. Suppose now that $n \in S$. If $x \in n^+$, then either $x \in n$ or $x = n$. In the first case $x \subset n$ (since $n \in S$) and therefore $x \subset n^+$; in the second case $x \subset n^+$ for even more trivial reasons. It follows that every element of n^+ is a subset of n^+, or, in other words, that $n^+ \in S$. The desired conclusion (ii) is a consequence of Equation 12.3.

We are now ready to prove Equation 12.5. Suppose indeed that n and m are natural numbers and that $n^+ = m^+$. Since $n \in n^+$ it follows that $n \in m^+$, and hence that either $n \in m$ or $n = m$. Similarly, either $m \in n$ or $m = n$. If $n \neq m$, then we must have $n \in m$ and $m \in n$. Since, by (ii), n is transitive, it follows that $n \in n$. Since, however, $n \subset n$, this contradicts (i), and the proof is complete.

The assertions Equation 12.1 — Equation 12.5 are known as the Peano axioms; they used to be considered as the fountainhead of all mathematical knowledge. From them (together with the set-theoretic principles we have already met) it is possible to define integers, rational numbers, real numbers, and complex numbers, and to derive their usual arithmetic and analytic properties. Such a program is not within the scope of this book; the interested reader should have no difficulty in locating and studying it elsewhere.

Induction is often used not only to prove things but also to define things. Suppose, to be specific, that f is a function from a set X into the same set X, and suppose that a is an element of X. It seems natural to try to define an infinite sequence $\{u(n)\}$ of elements of X (that is, a function u from ω to X) in some such way as this: write $u(0) = a$, $u(1) = f(u(0))$, $u(2) = f(u(1))$, and so on. If the would-be definer were pressed to explain the "and so on," he might lean on induction. What it all means, he might say, is that we define $u(0)$ as a, and then, inductively, we define $u(n^+)$ as $f(u(n))$ for every n. This may sound plausible, but, as justification for an existential assertion, it is insufficient. The principle of mathematical induction does indeed prove, easily, that there can be at most one function satisfying all the stated conditions, but it does not establish the existence of such a function. What is needed is the following result.

Theorem 12.1 (Recursion theorem)**.** *If a is an element of a set X, and if f is a function from X into X, then there exists a function u from ω into X such that $u(0) = a$ and such that $u(n^+) = f(u(n))$ for all n in ω.*

Proof. Recall that a function from ω to X is a certain kind of subset of $\omega \times X$; we shall construct u explicitly as a set of ordered pairs. Consider, for this purpose, the collection $\mathcal{C}$ of all those subsets A of $\omega \times X$ for which $(0, a) \in A$ and for which $(n^+, f(x)) \in A$ whenever $(n, x) \in A$. Since $\omega \times X$ has these properties, the collection $\mathcal{C}$ is not empty. We may, therefore, form the intersection of all the sets of the collection $\mathcal{C}$. Since it is easy to see that u itself belongs to $\mathcal{C}$, it remains only to prove that u is a function. We are to prove, in other words, that for each natural number n there exists at most one element x of X such that $(n, x) \in u$. (Explicitly: if both (n, x) and (n, y) belong to u, then $x = y$.) The proof is inductive. Let S be the set of all those natural numbers n for which it is indeed true that $(n, x) \in u$ for at most one x. We shall prove that $0 \in S$ and that if $n \in S$, then $n^+ \in S$.

Does 0 belong to S? If not, then $(0, b) \in u$ for some b distinct from a. Consider, in this case, the set $u - \{(0, b)\}$. Observe that this diminished set still contains $(0, a)$ (since $a \neq b$), and that if the diminished set contains (n, x), then it contains $(n^+, f(x))$ also. The reason for the second assertion is that since $n^+ \neq 0$, the discarded element is not equal to $(n^+, f(x))$. In other words,

$u - \{(0, b)\} \in \mathcal{C}$. This contradicts the fact that u is the smallest set in $\mathcal{C}$, and we may conclude that $0 \in S$.

Suppose now that $n \in S$; this means that there exists a unique element x in X such that $(n, x) \in u$. Since $(n, x) \in u$, it follows that $(n^+, f(x)) \in u$. If n^+ does not belong to S, then $(n^+, y) \in u$ for some y different from $f(x)$. Consider, in this case, the set $u - \{(n^+, y)\}$. Observe that this diminished set contains $(0, a)$ (since $n^+ \neq 0$), and that if the diminished set contains (m, t), say, then it contains $(m^+, f(t))$ also. Indeed, if $m = n$, then t must be x, and the reason the diminished set contains $(n^+, f(x))$ is that $f(x) \neq y$; if, on the other hand, $m \neq n$, then the reason the diminished set contains $(m^+, f(t))$ is that $m^+ \neq n^+$. In other words, $u - \{(n^+, y)\} \in \mathcal{C}$. This again contradicts the fact that u is the smallest set in $\mathcal{C}$, and we may conclude that $n^+ \in S$. □

The proof of the recursion theorem is complete. An application of the recursion theorem is called *definition by induction.*

Exercise 12.1. Prove that if n is a natural number, then $n \neq n^+$; if $n \neq 0$, then $n = m^+$ for some natural number m. Prove that ω is transitive. Prove that if E is a non-empty subset of some natural number, then there exists an element k in E such that $k \in m$ whenever m is an element of E distinct from k.

INDEX