Contents

Preface

This book is intended as a text for a one- or two-semester introduction to topology, at the senior or first-year graduate level.

The subject of topology is of interest in its own right, and it also serves to lay the foundations for future study in analysis, in geometry, and in algebraic topology. There is no universal agreement among mathematicians as to what a first course in topology should include; there are many topics that are appropriate to such a course, and not all are equally relevant to these differing purposes. In the choice of material to be treated, I have tried to strike a balance among the various points of view.

Prerequisites. There are no formal subject matter prerequisites for studying most of this book. I do not even assume the reader knows much set theory. Having said that, I must hasten to add that unless the reader has studied a bit of analysis or "rigorous calculus," much of the motivation for the concepts introduced in the first part of the book will be missing. Things will go more smoothly if he or she already has had some experience with continuous functions, open and closed sets, metric spaces, and the like, although none of these is actually assumed. In Part II, we do assume familiarity with the elements of group theory.

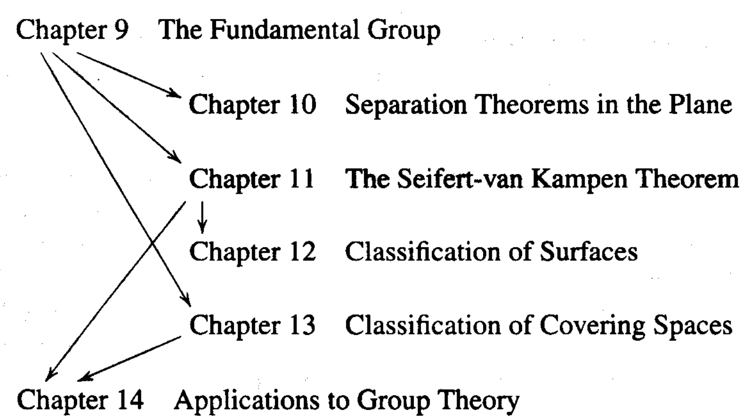
Most students in a topology course have, in my experience, some knowledge of the foundations of mathematics. But the amount varies a great deal from one student to another. Therefore, I begin with a fairly thorough chapter on set theory and logic. It starts at an elementary level and works up to a level that might be described as "semi-sophisticated." It treats those topics (and only those) that will be needed later in the book. Most students will already be familiar with the material of the first few sections, but many of them will find their *expertise* disappearing somewhere about the middle of the chapter. How much time and effort the instructor will need to spend on this chapter will thus depend largely on the mathematical sophistication and experience of the students. Ability to do the exercises fairly readily (and correctly!) should serve as a reasonable criterion for determining whether the student's mastery of set theory is sufficient for the student to begin the study of topology.

Many students (and instructors!) would prefer to skip the foundational material of Chapter 1 and jump right in to the study of topology. One ignores the foundations, however, only at the risk of later confusion and error. What one *can* do is to treat initially only those sections that are needed at once, postponing the remainder until they are needed. The first seven sections (through countability) are needed throughout the book; I usually assign some of them as reading and lecture on the rest. Sections 9 and 10, on the axiom of choice and well-ordering, are not needed until the discussion of compactness in Chapter 3. Section 11, on the maximum principle, can be postponed even longer; it is needed only for the Tychonoff theorem (Chapter 5) and the theorem on the fundamental group of a linear graph (Chapter 14).

How the book is organized. This book can be used for a number of different courses. I have attempted to organize it as flexibly as possible, so as to enable the instructor to follow his or her own preferences in the matter.

Part I, consisting of the first eight chapters, is devoted to the subject commonly called general topology. The first four chapters deal with the body of material that, in my opinion, should be included in any introductory topology course worthy of the name. This may be considered the "irreducible core" of the subject, treating as it does set theory, topological spaces, connectedness, compactness (through compactness of finite products), and the countability and separation axioms (through the Urysohn metrization theorem). The remaining four chapters of Part I explore additional topics; they are essentially independent of one another, depending on only the core material of Chapters 1–4. The instructor may take them up in any order he or she chooses.

Part II constitutes an introduction to the subject of Algebraic Topology. It depends on only the core material of Chapters 1–4. This part of the book treats with some thoroughness the notions of fundamental group and covering space, along with their many and varied applications. Some of the chapters of Part II are independent of one another; the dependence among them is expressed in the following diagram:



Certain sections of the book are marked with an asterisk; these sections may be omitted or postponed with no loss of continuity. Certain theorems are marked similarly. Any dependence of later material on these asterisked sections or theorems is indicated at the time, and again when the results are needed. Some of the exercises also depend on earlier asterisked material, but in such cases the dependence is obvious.

Sets of supplementary exercises appear at the ends of several of the chapters. They provide an opportunity for exploration of topics that diverge somewhat from the main thrust of the book; an ambitious student might use one as a basis for an independent paper or research project. Most are fairly self-contained, but the one on topological groups has as a sequel a number of additional exercises on the topic that appear in later sections of the book.

Possible course outlines. Most instructors who use this text for a course in general topology will wish to cover Chapters 1–4, along with the Tychonoff theorem in Chapter 5. Many will cover additional topics as well. Possibilities include the following: the Stone-Čech compactification (§38), metrization theorems (Chapter 6), the Peano curve (§44), Ascoli's theorem (§45 and/or §47), and dimension theory (§50). I have, in different semesters, followed each of these options.

For a one-semester course in algebraic topology, one can expect to cover most of Part II.

It is also possible to treat both aspects of topology in a single semester, although with some corresponding loss of depth. One feasible outline for such a course would consist of Chapters 1–3, followed by Chapter 9; the latter does not depend on the material of Chapter 4. (The non-asterisked sections of Chapters 10 and 13 also are independent of Chapter 4.)

Comments on this edition. The reader who is familiar with the first edition of this book will find no substantial changes in the part of the book dealing with general topology. I have confined myself largely to "fine-tuning" the text material and the exercises. However, the final chapter of the first edition, which dealt with algebraic topology, has been substantially expanded and rewritten. It has become Part II of this book. In the years since the first edition appeared, it has become increasingly common to offer topology as a two-term course, the first devoted to general topology and the second to algebraic topology. By expanding the treatment of the latter subject, I have intended to make this revision serve the needs of such a course.

Acknowledgments. Most of the topologists with whom I have studied, or whose books I have read, have contributed in one way or another to this book; I mention only Edwin Moise, Raymond Wilder, Gail Young, and Raoul Bott, but there are many others. For their helpful comments concerning this book, my thanks to Ken Brown, Russ McMillan, Robert Mosher, and John Hemperly, and to my colleagues George Whitehead and Kenneth Hoffman.

The treatment of algebraic topology has been substantially influenced by the excellent book by William Massey [M], to whom I express appreciation. Finally, thanks are

due Adam Lewenberg of MacroTeX for his extraordinary skill and patience in setting text and juggling figures.

But most of all, to my students go my most heartfelt thanks. From them I learned at least as much as they did from me; without them this book would be very different.

J.R.M.

A Note to the Reader

Two matters require comment—the exercises and the examples.

Working problems is a crucial part of learning mathematics. No one can learn topology merely by poring over the definitions, theorems, and examples that are worked out in the text. One must work part of it out for oneself. To provide that opportunity is the purpose of the exercises.

They vary in difficulty, with the easier ones usually given first. Some are routine verifications designed to test whether you have understood the definitions or examples of the preceding section. Others are less routine. You may, for instance, be asked to generalize a theorem of the text. Although the result obtained may be interesting in its own right, the main purpose of such an exercise is to encourage you to work carefully through the proof in question, mastering its ideas thoroughly—more thoroughly (I hope!) than mere memorization would demand.

Some exercises are phrased in an "open-ended" fashion. Students often find this practice frustrating. When faced with an exercise that asks, "Is every regular Lindelöf space normal?" they respond in exasperation, "I don't know what I'm supposed to do! Am I suppose to prove it or find a counterexample or what?" But mathematics (outside textbooks) is usually like this. More often than not, all a mathematician has to work with is a conjecture or question, and he or she doesn't know what the correct answer is. You should have some experience with this situation.

A few exercises that are more difficult than the rest are marked with asterisks. But none are so difficult but that the best student in my class can usually solve them.

Another important part of mastering any mathematical subject is acquiring a repertoire of useful examples. One should, of course, come to know those major examples from whose study the theory itself derives, and to which the important applications are made. But one should also have a few counterexamples at hand with which to test plausible conjectures.

Now it is all too easy in studying topology to spend too much time dealing with "weird counterexamples." Constructing them requires ingenuity and is often great fun. But they are not really what topology is about. Fortunately, one does not need too many such counterexamples for a first course; there is a fairly short list that will suffice for most purposes. Let me give it here:

$\mathbb{R}^J$ the product of the real line with itself, in the product, uniform, and box topologies.

$\mathbb{R}_\ell$ the real line in the topology having the intervals $[a, b)$ as a basis.

S_Ω the minimal uncountable well-ordered set.

I_o^2 the closed unit square in the dictionary order topology.

These are the examples you should master and remember; they will be exploited again and again.

Part I
GENERAL TOPOLOGY

Chapter 1

Set Theory and Logic

We adopt, as most mathematicians do, the naive point of view regarding set theory. We shall assume that what is meant by a *set* of objects is intuitively clear, and we shall proceed on that basis without analyzing the concept further. Such an analysis properly belongs to the foundations of mathematics and to mathematical logic, and it is not our purpose to initiate the study of those fields.

Logicians have analyzed set theory in great detail, and they have formulated axioms for the subject. Each of their axioms expresses a property of sets that mathematicians commonly accept, and collectively the axioms provide a foundation broad enough and strong enough that the rest of mathematics can be built on them.

It is unfortunately true that careless use of set theory, relying on intuition alone, can lead to contradictions. Indeed, one of the reasons for the axiomatization of set theory was to formulate rules for dealing with sets that would avoid these contradictions. Although we shall not deal with the axioms explicitly, the rules we follow in dealing with sets derive from them. In this book, you will learn how to deal with sets in an "apprentice" fashion, by observing how we handle them and by working with them yourself. At some point of your studies, you may wish to study set theory more carefully and in greater detail; then a course in logic or foundations will be in order.

§1 Fundamental Concepts

Here we introduce the ideas of set theory, and establish the basic terminology and notation. We also discuss some points of elementary logic that, in our experience, are apt to cause confusion.

Basic Notation

Commonly we shall use capital letters A, B, ... to denote sets, and lowercase letters a, b, ... to denote the ***objects*** or ***elements*** belonging to these sets. If an object a belongs to a set A, we express this fact by the notation

$$a \in A.$$

If a does not belong to A, we express this fact by writing

$$a \notin A.$$

The equality symbol $=$ is used throughout this book to mean *logical identity*. Thus, when we write $a = b$, we mean that "a" and "b" are symbols for the same object. This is what one means in arithmetic, for example, when one writes $\frac{2}{4} = \frac{1}{2}$. Similarly, the equation $A = B$ states that "A" and "B" are symbols for the same set; that is, A and B consist of precisely the same objects.

If a and b are different objects, we write $a \neq b$; and if A and B are different sets, we write $A \neq B$. For example, if A is the set of all nonnegative real numbers, and B is the set of all positive real numbers, then $A \neq B$, because the number 0 belongs to A and not to B.

We say that A is a ***subset*** of B if every element of A is also an element of B; and we express this fact by writing

$$A \subset B.$$

Nothing in this definition requires A to be different from B; in fact, if $A = B$, it is true that both $A \subset B$ and $B \subset A$. If $A \subset B$ and A is different from B, we say that A is a ***proper subset*** of B, and we write

$$A \subsetneq B.$$

The relations $\subset$ and $\subsetneq$ are called ***inclusion*** and ***proper inclusion***, respectively. If $A \subset B$, we also write $B \supset A$, which is read "B ***contains*** A."

How does one go about specifying a set? If the set has only a few elements, one can simply list the objects in the set, writing "A is the set consisting of the elements a, b, and c." In symbols, this statement becomes

$$A = \{a, b, c\},$$

where braces are used to enclose the list of elements.

The usual way to specify a set, however, is to take some set A of objects and some *property* that elements of A may or may not possess, and to form the set consisting of all elements of A having that property. For instance, one might take the set of real numbers and form the subset B consisting of all even integers. In symbols, this statement becomes

$$B = \{x \mid x \text{ is an even integer}\}.$$

Here the braces stand for the words "the set of," and the vertical bar stands for the words "such that." The equation is read "B is the set of all x such that x is an even integer."

The Union of Sets and the Meaning of "or"

Given two sets A and B, one can form a set from them that consists of all the elements of A together with all the elements of B. This set is called the ***union*** of A and B and is denoted by $A \cup B$. Formally, we define

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

But we must pause at this point and make sure exactly what we mean by the statement "$x \in A$ or $x \in B$."

In ordinary everyday English, the word "or" is ambiguous. Sometimes the statement "P or Q" means "P or Q, or both" and sometimes it means "P or Q, but not both." Usually one decides from the context which meaning is intended. For example, suppose I spoke to two students as follows:

> "Miss Smith, every student registered for this course has taken either a course in linear algebra or a course in analysis."
>
> "Mr. Jones, either you get a grade of at least 70 on the final exam or you will flunk this course."

In the context, Miss Smith knows perfectly well that I mean "everyone has had linear algebra or analysis, or both," and Mr. Jones knows I mean "either he gets at least 70 or he flunks, but not both." Indeed, Mr. Jones would be exceedingly unhappy if both statements turned out to be true!

In mathematics, one cannot tolerate such ambiguity. One has to pick just one meaning and stick with it, or confusion will reign. Accordingly, mathematicians have agreed that they will use the word "or" in the first sense, so that the statement "P or Q" always means "P or Q, or both." If one means "P or Q, but not both," then one has to include the phrase "but not both" explicitly.

With this understanding, the equation defining $A \cup B$ is unambiguous; it states that $A \cup B$ is the set consisting of all elements x that belong to A or to B or to both.

The Intersection of Sets, the Empty Set, and the Meaning of "If . . . Then"

Given sets A and B, another way one can form a set is to take the common part of A and B. This set is called the ***intersection*** of A and B and is denoted by $A \cap B$. Formally, we define

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

But just as with the definition of $A \cup B$, there is a difficulty. The difficulty is not in the meaning of the word "and"; it is of a different sort. It arises when the sets A and B happen to have no elements in common. What meaning does the symbol $A \cap B$ have in such a case?

To take care of this eventuality, we make a special convention. We introduce a special set that we call the ***empty set***, denoted by $\varnothing$, which we think of as "the set having no elements."

Using this convention, we express the statement that A and B have no elements in common by the equation

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

We also express this fact by saying that A and B are ***disjoint***.

Now some students are bothered by the notion of an "empty set." "How," they say, "can you have a set with nothing in it?" The problem is similar to that which arose many years ago when the number 0 was first introduced.

The empty set is only a convention, and mathematics could very well get along without it. But it is a very convenient convention, for it saves us a good deal of awkwardness in stating theorems and in proving them. Without this convention, for instance, one would have to prove that the two sets A and B do have elements in common before one could use the notation $A \cap B$. Similarly, the notation

$$C = \{x \mid x \in A \text{ and } x \text{ has a certain property}\}$$

could not be used if it happened that no element x of A had the given property. It is much more convenient to agree that $A \cap B$ and C equal the empty set in such cases.

Since the empty set $\varnothing$ is merely a convention, we must make conventions relating it to the concepts already introduced. Because $\varnothing$ is thought of as "the set with no elements," it is clear we should make the convention that for each object x, the relation $x \in \varnothing$ does not hold. Similarly, the definitions of union and intersection show that for every set A we should have the equations

$$A \cup \varnothing = A \quad \text{and} \quad A \cap \varnothing = \varnothing.$$

The inclusion relation is a bit more tricky. Given a set A, should we agree that $\varnothing \subset A$? Once more, we must be careful about the way mathematicians use the English language. The expression $\varnothing \subset A$ is a shorthand way of writing the sentence, "Every element that belongs to the empty set also belongs to the set A." Or to put it more formally, "For every object x, if x belongs to the empty set, then x also belongs to the set A."

Is this statement true or not? Some might say "yes" and others say "no." You will never settle the question by argument, only by agreement. This is a statement of the form "If P, then Q," and in everyday English the meaning of the "if ... then" construction is ambiguous. It always means that if P is true, then Q is true also. Sometimes that is all it means; other times it means something more: that if P is false, Q must be false. Usually one decides from the context which interpretation is correct.

The situation is similar to the ambiguity in the use of the word "or." One can reformulate the examples involving Miss Smith and Mr. Jones to illustrate the ambiguity. Suppose I said the following:

> "Miss Smith, if any student registered for this course has not taken a course in linear algebra, then he has taken a course in analysis."
>
> "Mr. Jones, if you get a grade below 70 on the final, you are going to flunk this course."

In the context, Miss Smith understands that if a student in the course has not had linear algebra, then he has taken analysis, but if he has had linear algebra, he may or may not have taken analysis as well. And Mr. Jones knows that if he gets a grade below 70, he will flunk the course, but if he gets a grade of at least 70, he will pass.

Again, mathematics cannot tolerate ambiguity, so a choice of meanings must be made. Mathematicians have agreed always to use "if ... then" in the first sense, so that a statement of the form "If P, then Q" means that if P is true, Q is true also, but if P is false, Q may be either true or false.

As an example, consider the following statement about real numbers:

$$\text{If } x > 0, \text{ then } x^3 \neq 0.$$

It is a statement of the form, "If P, then Q," where P is the phrase "$x > 0$" (called the ***hypothesis*** of the statement) and Q is the phrase "$x^3 \neq 0$" (called the ***conclusion*** of the statement). This is a true statement, for in every case for which the hypothesis $x > 0$ holds, the conclusion $x^3 \neq 0$ holds as well.

Another true statement about real numbers is the following:

$$\text{If } x^2 < 0, \text{ then } x = 23;$$

in every case for which the hypothesis holds, the conclusion holds as well. Of course, it happens in this example that there are no cases for which the hypothesis holds. A statement of this sort is sometimes said to be ***vacuously true***.

To return now to the empty set and inclusion, we see that the inclusion $\varnothing \subset A$ does hold for every set A. Writing $\varnothing \subset A$ is the same as saying, "If $x \in \varnothing$, then $x \in A$," and this statement is vacuously true.

Contrapositive and Converse

Our discussion of the "if ... then" construction leads us to consider another point of elementary logic that sometimes causes difficulty. It concerns the relation between a *statement*, its *contrapositive*, and its *converse*.

Given a statement of the form "If P, then Q," its ***contrapositive*** is defined to be the statement "If Q is not true, then P is not true." For example, the contrapositive of the statement

$$\textit{If } x > 0, \textit{ then } x^3 \neq 0,$$

is the statement

$$\textit{If } x^3 = 0, \textit{ then it is not true that } x > 0.$$

Note that both the statement and its contrapositive are true. Similarly, the statement

$$\textit{If } x^2 < 0, \textit{ then } x = 23,$$

has as its contrapositive the statement

$$\textit{If } x \neq 23, \textit{ then it is not true that } x^2 < 0.$$

Again, both are true statements about real numbers.

These examples may make you suspect that there is some relation between a statement and its contrapositive. And indeed there is; they are two ways of saying precisely the same thing. Each is true if and only if the other is true; they are *logically equivalent*.

This fact is not hard to demonstrate. Let us introduce some notation first. As a shorthand for the statement "If P, then Q," we write

$$P \Longrightarrow Q,$$

which is read "P implies Q." The contrapositive can then be expressed in the form

$$(\text{not } Q) \Longrightarrow (\text{not } P),$$

where "not Q" stands for the phrase "Q is not true."

Now the only way in which the statement "$P \Rightarrow Q$" can fail to be correct is if the hypothesis P is true and the conclusion Q is false. Otherwise it is correct. Similarly, the only way in which the statement (not Q) $\Rightarrow$ (not P) can fail to be correct is if the hypothesis "not Q" is true and the conclusion "not P" is false. This is the same as saying that Q is false and P is true. And this, in turn, is precisely the situation in which $P \Rightarrow Q$ fails to be correct. Thus, we see that the two statements are either both correct or both incorrect; they are logically equivalent. Therefore, we shall accept a proof of the statement "not $Q \Rightarrow$ not P" as a proof of the statement "$P \Rightarrow Q$."

There is another statement that can be formed from the statement $P \Rightarrow Q$. It is the statement

$$Q \Longrightarrow P,$$

which is called the ***converse*** of $P \Rightarrow Q$. One must be careful to distinguish between a statement's converse and its contrapositive. Whereas a statement and its contrapositive are logically equivalent, the truth of a statement says nothing at all about the truth or falsity of its converse. For example, the true statement

$$\textit{If } x > 0, \textit{ then } x^3 \neq 0,$$

has as its converse the statement

$$\textit{If } x^3 \neq 0, \textit{ then } x > 0,$$

which is false. Similarly, the true statement

$$\textit{If } x^2 < 0, \textit{ then } x = 23,$$

has as its converse the statement

$$\textit{If } x = 23, \textit{ then } x^2 < 0,$$

which is false.

If it should happen that both the statement $P \Rightarrow Q$ and its converse $Q \Rightarrow P$ are true, we express this fact by the notation

$$P \Longleftrightarrow Q,$$

which is read "P holds if and only if Q holds."

Negation

If one wishes to form the contrapositive of the statement $P \Rightarrow Q$, one has to know how to form the statement "not P," which is called the ***negation*** of P. In many cases, this causes no difficulty; but sometimes confusion occurs with statements involving the phrases "for every" and "for at least one." These phrases are called *logical quantifiers*.

To illustrate, suppose that X is a set, A is a subset of X, and P is a statement about the general element of X. Consider the following statement:

$$(*) \qquad \textit{For every } x \in A, \textit{ statement } P \textit{ holds.}$$

How does one form the negation of this statement? Let us translate the problem into the language of sets. Suppose that we let B denote the set of all those elements x of X for which P holds. Then statement $(*)$ is just the statement that A is a subset of B. What is its negation? Obviously, the statement that A is *not* a subset of B; that is, the statement that there exists at least one element of A that does not belong to B. Translating back into ordinary language, this becomes

$$\textit{For at least one } x \in A, \textit{ statement } P \textit{ does not hold.}$$

Therefore, to form the negation of statement $(*)$, one replaces the quantifier "for every" by the quantifier "for at least one," and one replaces statement P by *its* negation.

The process works in reverse just as well; the negation of the statement

For at least one $x \in A$, statement Q holds,

is the statement

For every $x \in A$, statement Q does not hold.

The Difference of Two Sets

We return now to our discussion of sets. There is one other operation on sets that is occasionally useful. It is the ***difference*** of two sets, denoted by $A - B$, and defined as the set consisting of those elements of A that are not in B. Formally,

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

It is sometimes called the ***complement*** of B relative to A, or the complement of B *in* A.

Our three set operations are represented schematically in Figure 1.1.

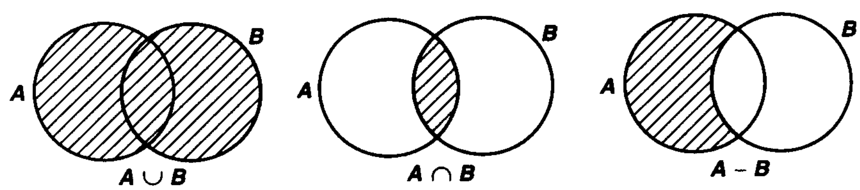


Figure 1.1

Rules of Set Theory

Given several sets, one may form new sets by applying the set-theoretic operations to them. As in algebra, one uses parentheses to indicate in what order the operations are to be performed. For example, $A \cup (B \cap C)$ denotes the union of the two sets A and $B \cap C$, while $(A \cup B) \cap C$ denotes the intersection of the two sets $A \cup B$ and C. The sets thus formed are quite different, as Figure 1.2 shows.

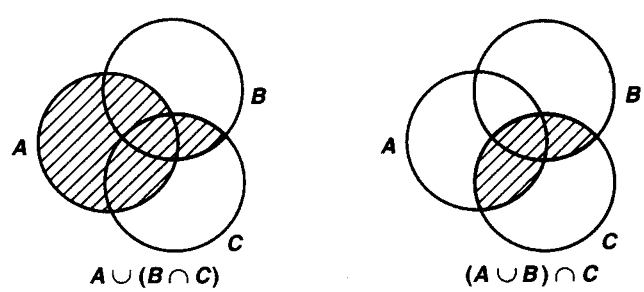


Figure 1.2

Sometimes different combinations of operations lead to the same set; when that happens, one has a rule of set theory. For instance, it is true that for any sets A, B, and C the equation

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

holds. The equation is illustrated in Figure 1.3; the shaded region represents the set in question, as you can check mentally. This equation can be thought of as a "distributive law" for the operations $\cap$ and $\cup$.

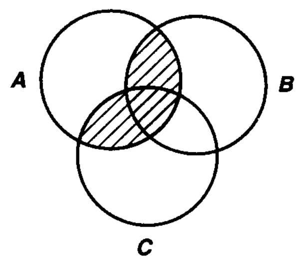


Figure 1.3

Other examples of set-theoretic rules include the second "distributive law,"

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

and *DeMorgan's laws*,

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

We leave it to you to check these rules. One can state other rules of set theory, but these are the most important ones. DeMorgan's laws are easier to remember if you verbalize them as follows:

The complement of the union equals the intersection of the complements.
The complement of the intersection equals the union of the complements.

Collections of Sets

The objects belonging to a set may be of any sort. One can consider the set of all even integers, and the set of all blue-eyed people in Nebraska, and the set of all decks of playing cards in the world. Some of these are of limited mathematical interest, we admit! But the third example illustrates a point we have not yet mentioned: namely, that the objects belonging to a set may *themselves* be sets. For a deck of cards is itself a set, one consisting of pieces of pasteboard with certain standard designs printed on them. The set of all decks of cards in the world is thus a set whose elements are themselves sets (of pieces of pasteboard).

We now have another way to form new sets from old ones. Given a set A, we can consider sets whose elements are subsets of A. In particular, we can consider the set of all subsets of A. This set is sometimes denoted by the symbol $\mathcal{P}(A)$ and is called the ***power set*** of A (for reasons to be explained later).

When we have a set whose elements are sets, we shall often refer to it as a ***collection*** of sets and denote it by a script letter such as $\mathcal{A}$ or $\mathcal{B}$. This device will help us in keeping things straight in arguments where we have to consider objects, and sets of objects, and collections of sets of objects, all at the same time. For example, we might use $\mathcal{A}$ to denote the collection of all decks of cards in the world, letting an ordinary capital letter A denote a deck of cards and a lowercase letter a denote a single playing card.

A certain amount of care with notation is needed at this point. We make a distinction between the object a, which is an *element* of a set A, and the one-element set $\{a\}$, which is a *subset* of A. To illustrate, if A is the set $\{a, b, c\}$, then the statements

$$a \in A, \qquad \{a\} \subset A, \qquad \text{and} \qquad \{a\} \in \mathcal{P}(A)$$

are all correct, but the statements $\{a\} \in A$ and $a \subset A$ are not.

Arbitrary Unions and Intersections

We have already defined what we mean by the union and the intersection of two sets. There is no reason to limit ourselves to just two sets, for we can just as well form the union and intersection of arbitrarily many sets.

Given a collection $\mathcal{A}$ of sets, the ***union*** of the elements of $\mathcal{A}$ is defined by the equation

$$\bigcup_{A \in \mathcal{A}} A = \{x \mid x \in A \text{ for at least one } A \in \mathcal{A}\}.$$

The ***intersection*** of the elements of $\mathcal{A}$ is defined by the equation

$$\bigcap_{A \in \mathcal{A}} A = \{x \mid x \in A \text{ for every } A \in \mathcal{A}\}.$$

There is no problem with these definitions if one of the elements of $\mathcal{A}$ happens to be the empty set. But it is a bit tricky to decide what (if anything) these definitions mean if we allow $\mathcal{A}$ to be the empty collection. Applying the definitions literally, we see that no element x satisfies the defining property for the union of the elements of $\mathcal{A}$. So it is reasonable to say that

$$\bigcup_{A \in \mathcal{A}} A = \varnothing$$

if $\mathcal{A}$ is empty. On the other hand, every x satisfies (vacuously) the defining property for the intersection of the elements of $\mathcal{A}$. The question is, every x in what set? If one has a given large set X that is specified at the outset of the discussion to be one's "universe of discourse," and one considers only subsets of X throughout, it is reasonable to let

$$\bigcap_{A \in \mathcal{A}} A = X$$

when $\mathcal{A}$ is empty. Not all mathematicians follow this convention, however. To avoid difficulty, *we shall not define the intersection when $\mathcal{A}$ is empty.*

Cartesian Products

There is yet another way of forming new sets from old ones; it involves the notion of an "ordered pair" of objects. When you studied analytic geometry, the first thing you did was to convince yourself that after one has chosen an x-axis and a y-axis in the plane, every point in the plane can be made to correspond to a unique ordered pair (x, y) of real numbers. (In a more sophisticated treatment of geometry, the plane is more likely to be *defined* as the set of all ordered pairs of real numbers!)

The notion of ordered pair carries over to general sets. Given sets A and B, we define their cartesian product $A \times B$ to be the set of all ordered pairs (a, b) for which a is an element of A and b is an element of B. Formally,

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

This definition assumes that the concept of "ordered pair" is already given. It can be taken as a primitive concept, as was the notion of "set"; or it can be given a definition in terms of the set operations already introduced. One definition in terms of set operations is expressed by the equation

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

it defines the ordered pair (a, b) as a collection of sets. If $a \neq b$, this definition says that (a, b) is a collection containing two sets, one of which is a one-element set and the other a two-element set. The *first coordinate* of the ordered pair is defined to be the element belonging to both sets, and the *second coordinate* is the element belonging to only one of the sets. If $a = b$, then (a, b) is a collection containing only one set $\{a\}$, since $\{a, b\} = \{a, a\} = \{a\}$ in this case. Its first coordinate and second coordinate both equal the element in this single set.

I think it is fair to say that most mathematicians think of an ordered pair as a primitive concept rather than thinking of it as a collection of sets!

Let us make a comment on notation. It is an unfortunate fact that the notation (a, b) is firmly established in mathematics with two entirely different meanings. One meaning, as an ordered pair of objects, we have just discussed. The other meaning is the one you are familiar with from analysis; if a and b are real numbers, the symbol (a, b) is used to denote the interval consisting of all numbers x such that $a < x < b$. Most of the time, this conflict in notation will cause no difficulty because the meaning will be clear from the context. Whenever a situation occurs where confusion is possible, we shall adopt a different notation for the ordered pair (a, b), denoting it by the symbol

$$a \times b$$

instead.

Exercises

1. Check the distributive laws for $\cup$ and $\cap$ and DeMorgan's laws.

2. Determine which of the following statements are true for all sets A, B, C, and D. If a double implication fails, determine whether one or the other of the possible implications holds. If an equality fails, determine whether the statement becomes true if the "equals" symbol is replaced by one or the other of the inclusion symbols $\subset$ or $\supset$.
   (a) $A \subset B$ and $A \subset C \Leftrightarrow A \subset (B \cup C)$.
   (b) $A \subset B$ or $A \subset C \Leftrightarrow A \subset (B \cup C)$.
   (c) $A \subset B$ and $A \subset C \Leftrightarrow A \subset (B \cap C)$.
   (d) $A \subset B$ or $A \subset C \Leftrightarrow A \subset (B \cap C)$.
   (e) $A - (A - B) = B$.
   (f) $A - (B - A) = A - B$.
   (g) $A \cap (B - C) = (A \cap B) - (A \cap C)$.
   (h) $A \cup (B - C) = (A \cup B) - (A \cup C)$.
   (i) $(A \cap B) \cup (A - B) = A$.
   (j) $A \subset C$ and $B \subset D \Rightarrow (A \times B) \subset (C \times D)$.
   (k) The converse of (j).
   (l) The converse of (j), assuming that A and B are nonempty.
   (m) $(A \times B) \cup (C \times D) = (A \cup C) \times (B \cup D)$.
   (n) $(A \times B) \cap (C \times D) = (A \cap C) \times (B \cap D)$.
   (o) $A \times (B - C) = (A \times B) - (A \times C)$.
   (p) $(A - B) \times (C - D) = (A \times C - B \times C) - A \times D$.
   (q) $(A \times B) - (C \times D) = (A - C) \times (B - D)$.

3. (a) Write the contrapositive and converse of the following statement: "If $x < 0$, then $x^2 - x > 0$," and determine which (if any) of the three statements are true.
   (b) Do the same for the statement "If $x > 0$, then $x^2 - x > 0$."

4. Let A and B be sets of real numbers. Write the negation of each of the following statements:
   (a) For every $a \in A$, it is true that $a^2 \in B$.
   (b) For at least one $a \in A$, it is true that $a^2 \in B$.
   (c) For every $a \in A$, it is true that $a^2 \notin B$.
   (d) For at least one $a \notin A$, it is true that $a^2 \in B$.

5. Let $\mathcal{A}$ be a nonempty collection of sets. Determine the truth of each of the following statements and of their converses:
   (a) $x \in \bigcup_{A \in \mathcal{A}} A \Rightarrow x \in A$ for at least one $A \in \mathcal{A}$.
   (b) $x \in \bigcup_{A \in \mathcal{A}} A \Rightarrow x \in A$ for every $A \in \mathcal{A}$.
   (c) $x \in \bigcap_{A \in \mathcal{A}} A \Rightarrow x \in A$ for at least one $A \in \mathcal{A}$.
   (d) $x \in \bigcap_{A \in \mathcal{A}} A \Rightarrow x \in A$ for every $A \in \mathcal{A}$.

6. Write the contrapositive of each of the statements of Exercise 5.

7. Given sets A, B, and C, express each of the following sets in terms of A, B, and C, using the symbols $\cup$, $\cap$, and $-$.

$$D = \{x \mid x \in A \text{ and } (x \in B \text{ or } x \in C)\},$$
$$E = \{x \mid (x \in A \text{ and } x \in B) \text{ or } x \in C\},$$
$$F = \{x \mid x \in A \text{ and } (x \in B \Rightarrow x \in C)\}.$$

8. If a set A has two elements, show that $\mathcal{P}(A)$ has four elements. How many elements does $\mathcal{P}(A)$ have if A has one element? Three elements? No elements? Why is $\mathcal{P}(A)$ called the power set of A?

9. Formulate and prove DeMorgan's laws for arbitrary unions and intersections.

10. Let $\mathbb{R}$ denote the set of real numbers. For each of the following subsets of $\mathbb{R} \times \mathbb{R}$, determine whether it is equal to the cartesian product of two subsets of $\mathbb{R}$.
(a) $\{(x, y) \mid x \text{ is an integer}\}$.
(b) $\{(x, y) \mid 0 < y \leq 1\}$.
(c) $\{(x, y) \mid y > x\}$.
(d) $\{(x, y) \mid x \text{ is not an integer and } y \text{ is an integer}\}$.
(e) $\{(x, y) \mid x^2 + y^2 < 1\}$.

§2 Functions

The concept of *function* is one you have seen many times already, so it is hardly necessary to remind you how central it is to all mathematics. In this section, we give the precise mathematical definition, and we explore some of the associated concepts.

A function is usually thought of as a *rule* that assigns to each element of a set A, an element of a set B. In calculus, a function is often given by a simple formula such as $f(x) = 3x^2 + 2$ or perhaps by a more complicated formula such as

$$f(x) = \sum_{k=1}^{\infty} x^k.$$

One often does not even mention the sets A and B explicitly, agreeing to take A to be the set of all real numbers for which the rule makes sense and B to be the set of all real numbers.

As one goes further in mathematics, however, one needs to be more precise about what a function is. Mathematicians *think* of functions in the way we just described, but the definition they use is more exact. First, we define the following:

Definition. A ***rule of assignment*** is a subset r of the cartesian product $C \times D$ of two sets, having the property that each element of C appears as the first coordinate of *at most one* ordered pair belonging to r.

Thus, a subset r of $C \times D$ is a rule of assignment if

$$[(c, d) \in r \text{ and } (c, d') \in r] \Longrightarrow [d = d'].$$

We think of r as a way of assigning, to the element c of C, the element d of D for which $(c, d) \in r$.

Given a rule of assignment r, the ***domain*** of r is defined to be the subset of C consisting of all first coordinates of elements of r, and the ***image set*** of r is defined as the subset of D consisting of all second coordinates of elements of r. Formally,

$$\begin{aligned} \text{domain } r &= \{c \mid \text{there exists } d \in D \text{ such that } (c, d) \in r\}, \\ \text{image } r &= \{d \mid \text{there exists } c \in C \text{ such that } (c, d) \in r\}. \end{aligned}$$

Note that given a rule of assignment r, its domain and image are entirely determined.

Now we can say what a function is.

Definition. A ***function*** f is a rule of assignment r, together with a set B that contains the image set of r. The domain A of the rule r is also called the ***domain*** of the function f; the image set of r is also called the ***image set*** of f; and the set B is called the ***range*** of f.[†]

If f is a function having domain A and range B, we express this fact by writing

$$f : A \longrightarrow B,$$

which is read "f is a function from A to B," or "f is a mapping from A into B," or simply "f maps A into B." One sometimes visualizes f as a geometric transformation physically carrying the points of A to points of B.

If $f : A \to B$ and if a is an element of A, we denote by $f(a)$ the unique element of B that the rule determining f assigns to a; it is called the ***value*** of f at a, or sometimes the ***image*** of a under f. Formally, if r is the rule of the function f, then $f(a)$ denotes the unique element of B such that $(a, f(a)) \in r$.

Using this notation, one can go back to defining functions almost as one did before, with no lack of rigor. For instance, one can write (letting $\mathbb{R}$ denote the real numbers)

> "Let f be the function whose rule is $\{(x, x^3 + 1) \mid x \in \mathbb{R}\}$ and whose range is $\mathbb{R}$,"

or one can equally well write

> "Let $f : \mathbb{R} \to \mathbb{R}$ be the function such that $f(x) = x^3 + 1$."

Both sentences specify precisely the same function. But the sentence "Let f be the function $f(x) = x^3 + 1$" is no longer adequate for specifying a function because it specifies neither the domain nor the range of f.

[†]Analysts are apt to use the word "range" to denote what we have called the "image set" of f. They avoid giving the set B a name.

Definition. If $f : A \to B$ and if A_0 is a subset of A, we define the ***restriction*** of f to A_0 to be the function mapping A_0 into B whose rule is

$$\{(a, f(a)) \mid a \in A_0\}.$$

It is denoted by $f|A_0$, which is read "f restricted to A_0."

EXAMPLE 1. Let $\mathbb{R}$ denote the real numbers and let $\bar{\mathbb{R}}_+$ denote the nonnegative reals. Consider the functions

$$\begin{array}{lll} f : \mathbb{R} \longrightarrow \mathbb{R} & \text{defined by} & f(x) = x^2, \\ g : \bar{\mathbb{R}}_+ \longrightarrow \mathbb{R} & \text{defined by} & g(x) = x^2, \\ h : \mathbb{R} \longrightarrow \bar{\mathbb{R}}_+ & \text{defined by} & h(x) = x^2, \\ k : \bar{\mathbb{R}}_+ \longrightarrow \bar{\mathbb{R}}_+ & \text{defined by} & k(x) = x^2. \end{array}$$

The function g is different from the function f because their rules are different subsets of $\mathbb{R} \times \mathbb{R}$; it is the restriction of f to the set $\bar{\mathbb{R}}_+$. The function h is also different from f, even though their rules are the same set, because the range specified for h is different from the range specified for f. The function k is different from all of these. These functions are pictured in Figure 2.1.

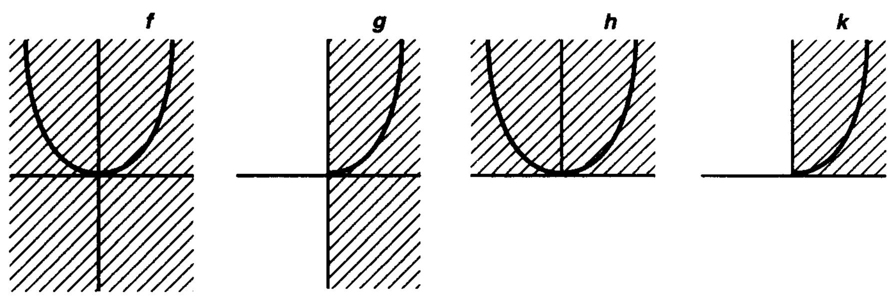


Figure 2.1

Restricting the domain of a function and changing its range are two ways of forming a new function from an old one. Another way is to form the composite of two functions.

Definition. Given functions $f : A \to B$ and $g : B \to C$, we define the ***composite*** $g \circ f$ of f and g as the function $g \circ f : A \to C$ defined by the equation $(g \circ f)(a) = g(f(a))$.

Formally, $g \circ f : A \to C$ is the function whose rule is

$$\{(a, c) \mid \text{For some } b \in B,\ f(a) = b \text{ and } g(b) = c\}.$$

We often picture the composite $g \circ f$ as involving a physical movement of the point a to the point $f(a)$, and then to the point $g(f(a))$, as illustrated in Figure 2.2.

Note that $g \circ f$ is defined only when the range of f *equals* the domain of g.

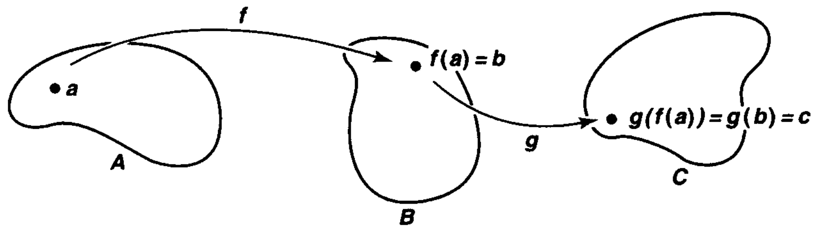


Figure 2.2

EXAMPLE 2. The composite of the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = 3x^2 + 2$ and the function $g : \mathbb{R} \to \mathbb{R}$ given by $g(x) = 5x$ is the function $g \circ f : \mathbb{R} \to \mathbb{R}$ given by

$$(g \circ f)(x) = g(f(x)) = g(3x^2 + 2) = 5(3x^2 + 2).$$

The composite $f \circ g$ can also be formed in this case; it is the quite different function $f \circ g : \mathbb{R} \to \mathbb{R}$ given by

$$(f \circ g)(x) = f(g(x)) = f(5x) = 3(5x)^2 + 2.$$

Definition. A function $f : A \to B$ is said to be ***injective*** (or ***one-to-one***) if for each pair of distinct points of A, their images under f are distinct. It is said to be ***surjective*** (or f is said to map A ***onto*** B) if every element of B is the image of some element of A under the function f. If f is both injective and surjective, it is said to be ***bijective*** (or is called a ***one-to-one correspondence***).

More formally, f is injective if

$$[f(a) = f(a')] \Longrightarrow [a = a'],$$

and f is surjective if

$$[b \in B] \Longrightarrow [b = f(a) \text{ for at least one } a \in A].$$

Injectivity of f depends only on the rule of f; surjectivity depends on the range of f as well. You can check that the composite of two injective functions is injective, and the composite of two surjective functions is surjective; it follows that the composite of two bijective functions is bijective.

If f is bijective, there exists a function from B to A called the ***inverse*** of f. It is denoted by f^{-1} and is defined by letting $f^{-1}(b)$ be that unique element a of A for which $f(a) = b$. Given $b \in B$, the fact that f is surjective implies that there *exists* such an element $a \in A$; the fact that f is injective implies that there is *only one* such element a. It is easy to see that if f is bijective, f^{-1} is also bijective.

EXAMPLE 3. Consider again the functions f, g, h, and k of Figure 2.1. The function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^2$ is neither injective nor surjective. Its restriction g to the nonnegative reals is injective but not surjective. The function $h : \mathbb{R} \to \bar{\mathbb{R}}_+$ obtained from f

by changing the range is surjective but not injective. The function $k : \bar{\mathbb{R}}_+ \to \bar{\mathbb{R}}_+$ obtained from f by restricting the domain *and* changing the range is both injective and surjective, so it has an inverse. Its inverse is, of course, what we usually call the *square-root function*.

A useful criterion for showing that a given function f is bijective is the following, whose proof is left to the exercises:

Lemma 2.1. *Let $f : A \to B$. If there are functions $g : B \to A$ and $h : B \to A$ such that $g(f(a)) = a$ for every a in A and $f(h(b)) = b$ for every b in B, then f is bijective and $g = h = f^{-1}$.*

Definition. Let $f : A \to B$. If A_0 is a subset of A, we denote by $f(A_0)$ the set of all images of points of A_0 under the function f; this set is called the ***image*** of A_0 under f. Formally,

$$f(A_0) = \{b \mid b = f(a) \text{ for at least one } a \in A_0\}.$$

On the other hand, if B_0 is a subset of B, we denote by $f^{-1}(B_0)$ the set of all elements of A whose images under f lie in B_0; it is called the ***preimage*** of B_0 under f (or the "counterimage," or the "inverse image," of B_0). Formally,

$$f^{-1}(B_0) = \{a \mid f(a) \in B_0\}.$$

Of course, there may be no points a of A whose images lie in B_0; in that case, $f^{-1}(B_0)$ is empty.

Note that if $f : A \to B$ is bijective and $B_0 \subset B$, we have two meanings for the notation $f^{-1}(B_0)$. It can be taken to denote the *preimage* of B_0 under the function f or to denote the *image* of B_0 under the function $f^{-1} : B \to A$. These two meanings give precisely the same subset of A, however, so there is, in fact, no ambiguity.

Some care is needed if one is to use the f and f^{-1} notation correctly. The operation f^{-1}, for instance, when applied to subsets of B, behaves very nicely; it preserves inclusions, unions, intersections, and differences of sets. We shall use this fact frequently. But the operation f, when applied to subsets of A, preserves only inclusions and unions. See Exercises 2 and 3.

As another situation where care is needed, we note that it is not in general true that $f^{-1}(f(A_0)) = A_0$ and $f(f^{-1}(B_0)) = B_0$. (See the following example.) The relevant rules, which we leave to you to check, are the following: If $f : A \to B$ and if $A_0 \subset A$ and $B_0 \subset B$, then

$$A_0 \subset f^{-1}(f(A_0)) \quad \text{and} \quad f(f^{-1}(B_0)) \subset B_0.$$

The first inclusion is an equality if f is injective, and the second inclusion is an equality if f is surjective.

EXAMPLE 4. Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = 3x^2 + 2$ (Figure 2.3). Let $[a, b]$ denote the closed interval $a \leq x \leq b$. Then

$$f^{-1}(f([0, 1])) = f^{-1}([2, 5]) = [-1, 1], \quad \text{and}$$
$$f(f^{-1}([0, 5])) = f([-1, 1]) = [2, 5].$$

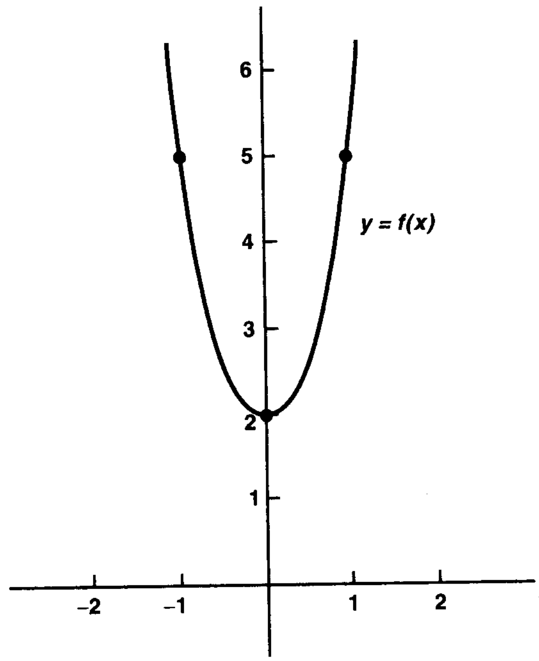


Figure 2.3

Exercises

1. Let $f : A \to B$. Let $A_0 \subset A$ and $B_0 \subset B$.
   (a) Show that $A_0 \subset f^{-1}(f(A_0))$ and that equality holds if f is injective.
   (b) Show that $f(f^{-1}(B_0)) \subset B_0$ and that equality holds if f is surjective.
2. Let $f : A \to B$ and let $A_i \subset A$ and $B_i \subset B$ for $i = 0$ and $i = 1$. Show that f^{-1} preserves inclusions, unions, intersections, and differences of sets:
   (a) $B_0 \subset B_1 \Rightarrow f^{-1}(B_0) \subset f^{-1}(B_1)$.
   (b) $f^{-1}(B_0 \cup B_1) = f^{-1}(B_0) \cup f^{-1}(B_1)$.
   (c) $f^{-1}(B_0 \cap B_1) = f^{-1}(B_0) \cap f^{-1}(B_1)$.
   (d) $f^{-1}(B_0 - B_1) = f^{-1}(B_0) - f^{-1}(B_1)$.
   Show that f preserves inclusions and unions only:
   (e) $A_0 \subset A_1 \Rightarrow f(A_0) \subset f(A_1)$.

(f) $f(A_0 \cup A_1) = f(A_0) \cup f(A_1)$.
(g) $f(A_0 \cap A_1) \subset f(A_0) \cap f(A_1)$; show that equality holds if f is injective.
(h) $f(A_0 - A_1) \supset f(A_0) - f(A_1)$; show that equality holds if f is injective.

3. Show that (b), (c), (f), and (g) of Exercise 2 hold for arbitrary unions and intersections.

4. Let $f : A \to B$ and $g : B \to C$.
(a) If $C_0 \subset C$, show that $(g \circ f)^{-1}(C_0) = f^{-1}(g^{-1}(C_0))$.
(b) If f and g are injective, show that $g \circ f$ is injective.
(c) If $g \circ f$ is injective, what can you say about injectivity of f and g?
(d) If f and g are surjective, show that $g \circ f$ is surjective.
(e) If $g \circ f$ is surjective, what can you say about surjectivity of f and g?
(f) Summarize your answers to (b)–(e) in the form of a theorem.

5. In general, let us denote the ***identity function*** for a set C by i_C. That is, define $i_C : C \to C$ to be the function given by the rule $i_C(x) = x$ for all $x \in C$. Given $f : A \to B$, we say that a function $g : B \to A$ is a ***left inverse*** for f if $g \circ f = i_A$; and we say that $h : B \to A$ is a ***right inverse*** for f if $f \circ h = i_B$.
(a) Show that if f has a left inverse, f is injective; and if f has a right inverse, f is surjective.
(b) Give an example of a function that has a left inverse but no right inverse.
(c) Give an example of a function that has a right inverse but no left inverse.
(d) Can a function have more than one left inverse? More than one right inverse?
(e) Show that if f has both a left inverse g and a right inverse h, then f is bijective and $g = h = f^{-1}$.

6. Let $f : \mathbb{R} \to \mathbb{R}$ be the function $f(x) = x^3 - x$. By restricting the domain and range of f appropriately, obtain from f a bijective function g. Draw the graphs of g and g^{-1}. (There are several possible choices for g.)

§3 Relations

A concept that is, in some ways, more general than that of function is the concept of a *relation*. In this section, we define what mathematicians mean by a relation, and we consider two types of relations that occur with great frequency in mathematics: *equivalence relations* and *order relations*. Order relations will be used throughout the book; equivalence relations will not be used until §22.

Definition. A ***relation*** on a set A is a subset C of the cartesian product $A \times A$.

If C is a relation on A, we use the notation xCy to mean the same thing as $(x, y) \in C$. We read it "x is in the relation C to y."

A rule of assignment r for a function $f : A \to A$ is also a subset of $A \times A$. But it is a subset of a very special kind: namely, one such that each element of A appears as the first coordinate of an element of r exactly once. *Any* subset of $A \times A$ is a relation on A.

EXAMPLE 1. Let P denote the set of all people in the world, and define $D \subset P \times P$ by the equation

$$D = \{(x, y) \mid x \text{ is a descendant of } y\}.$$

Then D is a relation on the set P. The statements "x is in the relation D to y" and "x is a descendant of y" mean precisely the same thing, namely, that $(x, y) \in D$. Two other relations on P are the following:

$$B = \{(x, y) \mid x \text{ has an ancestor who is also an ancestor of } y\},$$
$$S = \{(x, y) \mid \text{the parents of } x \text{ are the parents of } y\}.$$

We can call B the "blood relation" (pun intended), and we can call S the "sibling relation." These three relations have quite different properties. The blood relationship is symmetric, for instance (if x is a blood relative of y, then y is a blood relative of x), whereas the descendant relation is not. We shall consider these relations again shortly.

Equivalence Relations and Partitions

An ***equivalence relation*** on a set A is a relation C on A having the following three properties:

(1) (Reflexivity) xCx for every x in A.

(2) (Symmetry) If xCy, then yCx.

(3) (Transitivity) If xCy and yCz, then xCz.

EXAMPLE 2. Among the relations defined in Example 1, the descendant relation D is neither reflexive nor symmetric, while the blood relation B is not transitive (I am not a blood relation to my wife, although my children are!) The sibling relation S is, however, an equivalence relation, as you may check.

There is no reason one must use a capital letter—or indeed a letter of any sort—to denote a relation, even though it *is* a set. Another symbol will do just as well. One symbol that is frequently used to denote an equivalence relation is the "tilde" symbol $\sim$. Stated in this notation, the properties of an equivalence relation become

(1) $x \sim x$ for every x in A.

(2) If $x \sim y$, then $y \sim x$.

(3) If $x \sim y$ and $y \sim z$, then $x \sim z$.

There are many other symbols that have been devised to stand for particular equivalence relations; we shall meet some of them in the pages of this book.

Given an equivalence relation $\sim$ on a set A and an element x of A, we define a certain subset E of A, called the ***equivalence class*** determined by x, by the equation

$$E = \{y \mid y \sim x\}.$$

Note that the equivalence class E determined by x contains x, since $x \sim x$. Equivalence classes have the following property:

Lemma 3.1. *Two equivalence classes E and E' are either disjoint or equal.*

Proof. Let E be the equivalence class determined by x, and let E' be the equivalence class determined by x'. Suppose that $E \cap E'$ is not empty; let y be a point of $E \cap E'$. See Figure 3.1. We show that $E = E'$.

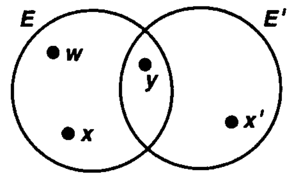


Figure 3.1

By definition, we have $y \sim x$ and $y \sim x'$. Symmetry allows us to conclude that $x \sim y$ and $y \sim x'$; from transitivity it follows that $x \sim x'$. If now w is any point of E, we have $w \sim x$ by definition; it follows from another application of transitivity that $w \sim x'$. We conclude that $E \subset E'$.

The symmetry of the situation allows us to conclude that $E' \subset E$ as well, so that $E = E'$. ■

Given an equivalence relation on a set A, let us denote by $\mathcal{E}$ the collection of all the equivalence classes determined by this relation. The preceding lemma shows that distinct elements of $\mathcal{E}$ are disjoint. Furthermore, the union of the elements of $\mathcal{E}$ equals all of A because every element of A belongs to an equivalence class. The collection $\mathcal{E}$ is a particular example of what is called a partition of A:

Definition. A ***partition*** of a set A is a collection of disjoint nonempty subsets of A whose union is all of A.

Studying equivalence relations on a set A and studying partitions of A are really the same thing. Given any partition $\mathcal{D}$ of A, there is exactly one equivalence relation on A from which it is derived.

The proof is not difficult. To show that the partition $\mathcal{D}$ comes from some equivalence relation, let us define a relation C on A by setting xCy if x and y belong to the same element of $\mathcal{D}$. Symmetry of C is obvious; reflexivity follows from the fact that the union of the elements of $\mathcal{D}$ equals all of A; transitivity follows from the fact that distinct elements of $\mathcal{D}$ are disjoint. It is simple to check that the collection of equivalence classes determined by C is precisely the collection $\mathcal{D}$.

To show there is only one such equivalence relation, suppose that C_1 and C_2 are two equivalence relations on A that give rise to the same collection of equivalence classes $\mathcal{D}$. Given $x \in A$, we show that yC_1x if and only if yC_2x, from which we conclude that $C_1 = C_2$. Let E_1 be the equivalence class determined by x relative to the relation C_1; let E_2 be the equivalence class determined by x relative to the relation C_2. Then E_1 is an element of $\mathcal{D}$, so that it must equal the unique element D of $\mathcal{D}$ that contains x. Similarly, E_2 must equal D. Now by definition, E_1 consists of all y such that yC_1x; and E_2 consists of all y such that yC_2x. Since $E_1 = D = E_2$, our result is proved.

EXAMPLE 3. Define two points in the plane to be equivalent if they lie at the same distance from the origin. Reflexivity, symmetry, and transitivity hold trivially. The collection $\mathcal{E}$ of equivalence classes consists of all circles centered at the origin, along with the set consisting of the origin alone.

EXAMPLE 4. Define two points of the plane to be equivalent if they have the same y-coordinate. The collection of equivalence classes is the collection of all straight lines in the plane parallel to the x-axis.

EXAMPLE 5. Let $\mathcal{L}$ be the collection of all straight lines in the plane parallel to the line $y = -x$. Then $\mathcal{L}$ is a partition of the plane, since each point lies on exactly one such line. The partition $\mathcal{L}$ comes from the equivalence relation on the plane that declares the points (x_0, y_0) and (x_1, y_1) to be equivalent if $x_0 + y_0 = x_1 + y_1$.

EXAMPLE 6. Let $\mathcal{L}'$ be the collection of *all* straight lines in the plane. Then $\mathcal{L}'$ is not a partition of the plane, for distinct elements of $\mathcal{L}'$ are not necessarily disjoint; two lines may intersect without being equal.

Order Relations

A relation C on a set A is called an ***order relation*** (or a ***simple order***, or a ***linear order***) if it has the following properties:

(1) (Comparability) For every x and y in A for which $x \neq y$, either xCy or yCx.

(2) (Nonreflexivity) For no x in A does the relation xCx hold.

(3) (Transitivity) If xCy and yCz, then xCz.

Note that property (1) does not by itself exclude the possibility that for some pair of elements x and y of A, both the relations xCy and yCx hold (since "or" means "one or the other, or both"). But properties (2) and (3) combined do exclude this possibility; for if both xCy and yCx held, transitivity would imply that xCx, contradicting nonreflexivity.

EXAMPLE 7. Consider the relation on the real line consisting of all pairs (x, y) of real numbers such that $x < y$. It is an order relation, called the "usual order relation," on the real line. A less familiar order relation on the real line is the following: Define xCy if $x^2 < y^2$, or if $x^2 = y^2$ and $x < y$. You can check that this is an order relation.

EXAMPLE 8. Consider again the relationships among people given in Example 1. The blood relation B satisfies none of the properties of an order relation, and the sibling relation S satisfies only (3). The descendant relation D does somewhat better, for it satisfies both (2) and (3); however, comparability still fails. Relations that satisfy (2) and (3) occur often enough in mathematics to be given a special name. They are called *strict partial order* relations; we shall consider them later (see §11).

As the tilde, $\sim$, is the generic symbol for an equivalence relation, the "less than" symbol, $<$, is commonly used to denote an order relation. Stated in this notation, the properties of an order relation become

(1) If $x \neq y$, then either $x < y$ or $y < x$.

(2) If $x < y$, then $x \neq y$.

(3) If $x < y$ and $y < z$, then $x < z$.

We shall use the notation $x \leq y$ to stand for the statement "either $x < y$ or $x = y$"; and we shall use the notation $y > x$ to stand for the statement "$x < y$." We write $x < y < z$ to mean "$x < y$ and $y < z$."

Definition. If X is a set and $<$ is an order relation on X, and if $a < b$, we use the notation (a, b) to denote the set

$$\{x \mid a < x < b\};$$

it is called an ***open interval*** in X. If this set is empty, we call a the ***immediate predecessor*** of b, and we call b the ***immediate successor*** of a.

Definition. Suppose that A and B are two sets with order relations $<_A$ and $<_B$ respectively. We say that A and B have the same ***order type*** if there is a bijective correspondence between them that preserves order; that is, if there exists a bijective function $f : A \to B$ such that

$$a_1 <_A a_2 \Longrightarrow f(a_1) <_B f(a_2).$$

EXAMPLE 9. The interval $(-1, 1)$ of real numbers has the same order type as the set $\mathbb{R}$ of real numbers itself, for the function $f : (-1, 1) \to \mathbb{R}$ given by

$$f(x) = \frac{x}{1 - x^2}$$

is an order-preserving bijective correspondence, as you can check. It is pictured in Figure 3.2.

EXAMPLE 10. The subset $A = \{0\} \cup (1, 2)$ of $\mathbb{R}$ has the same order type as the subset

$$[0, 1) = \{x \mid 0 \leq x < 1\}$$

of $\mathbb{R}$. The function $f : A \to [0, 1)$ defined by

$$f(0) = 0,$$
$$f(x) = x - 1 \quad \text{for } x \in (1, 2)$$

is the required order-preserving correspondence.

One interesting way of defining an order relation, which will be useful to us later in dealing with some examples, is the following:

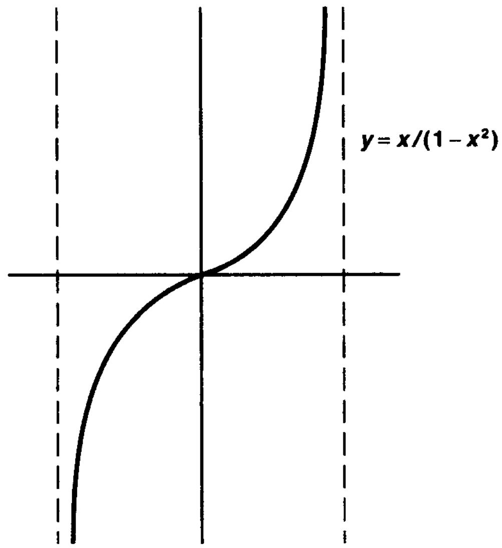


Figure 3.2

Definition. Suppose that A and B are two sets with order relations $<_A$ and $<_B$ respectively. Define an order relation $<$ on $A \times B$ by defining

$$a_1 \times b_1 < a_2 \times b_2$$

if $a_1 <_A a_2$, or if $a_1 = a_2$ and $b_1 <_B b_2$. It is called the ***dictionary order relation*** on $A \times B$.

Checking that this is an order relation involves looking at several separate cases; we leave it to you.

The reason for the choice of terminology is fairly evident. The rule defining $<$ is the same as the rule used to order the words in the dictionary. Given two words, one compares their first letters and orders the words according to the order in which their first letters appear in the alphabet. If the first letters are the same, one compares their second letters and orders accordingly. And so on.

EXAMPLE 11. Consider the dictionary order on the plane $\mathbb{R} \times \mathbb{R}$. In this order, the point p is less than every point lying above it on the vertical line through p, and p is less than every point to the right of this vertical line.

EXAMPLE 12. Consider the set $[0, 1)$ of real numbers and the set $\mathbb{Z}_+$ of positive integers, both in their usual orders; give $\mathbb{Z}_+ \times [0, 1)$ the dictionary order. This set has the same order type as the set of nonnegative reals; the function

$$f(n \times t) = n + t - 1$$

is the required bijective order-preserving correspondence. On the other hand, the set $[0, 1) \times \mathbb{Z}_+$ in the dictionary order has quite a different order type; for example, every element of this ordered set has an immediate successor. These sets are pictured in Figure 3.3.

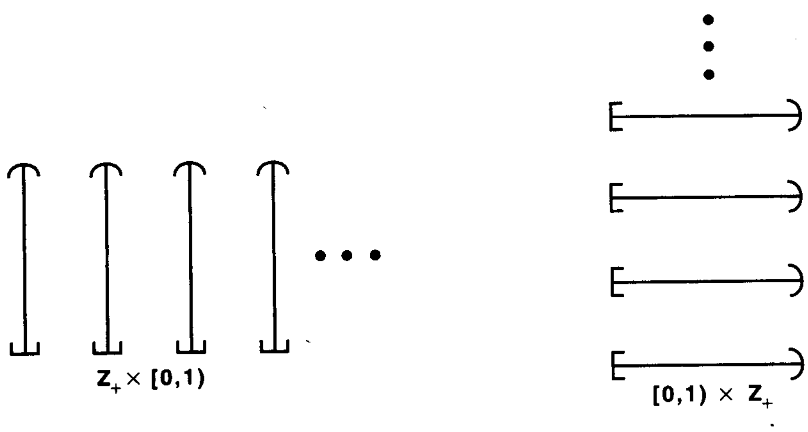


Figure 3.3

One of the properties of the real numbers that you may have seen before is the "least upper bound property." One can define this property for an arbitrary ordered set. First, we need some preliminary definitions.

Suppose that A is a set ordered by the relation $<$. Let A_0 be a subset of A. We say that the element b is the ***largest element*** of A_0 if $b \in A_0$ and if $x \leq b$ for every $x \in A_0$. Similarly, we say that a is the ***smallest element*** of A_0 if $a \in A_0$ and if $a \leq x$ for every $x \in A_0$. It is easy to see that a set has at most one largest element and at most one smallest element.

We say that the subset A_0 of A is ***bounded above*** if there is an element b of A such that $x \leq b$ for every $x \in A_0$; the element b is called an ***upper bound*** for A_0. If the set of all upper bounds for A_0 has a smallest element, that element is called the ***least upper bound***, or the ***supremum***, of A_0. It is denoted by $\sup A_0$; it may or may not belong to A_0. If it does, it is the largest element of A_0.

Similarly, A_0 is ***bounded below*** if there is an element a of A such that $a \leq x$ for every $x \in A_0$; the element a is called a ***lower bound*** for A_0. If the set of all lower bounds for A_0 has a largest element, that element is called the ***greatest lower bound***, or the ***infimum***, of A_0. It is denoted by $\inf A_0$; it may or may not belong to A_0. If it does, it is the smallest element of A_0.

Now we can define the least upper bound property.

Definition. An ordered set A is said to have the ***least upper bound property*** if every nonempty subset A_0 of A that is bounded above has a least upper bound. Analogously, the set A is said to have the ***greatest lower bound property*** if every nonempty subset A_0 of A that is bounded below has a greatest lower bound.

We leave it to the exercises to show that A has the least upper bound property if and only if it has the greatest lower bound property.

EXAMPLE 13. Consider the set $A = (-1, 1)$ of real numbers in the usual order. Assuming the fact that the real numbers have the least upper bound property, it follows that the set A has the least upper bound property. For, given any subset of A having an upper bound *in* A, it follows that its least upper bound (in the real numbers) must be in A. For example, the subset $\{-1/2n \mid n \in \mathbb{Z}_+\}$ of A, though it has no largest element, does have a least upper bound in A, the number 0.

On the other hand, the set $B = (-1, 0) \cup (0, 1)$ does not have the least upper bound property. The subset $\{-1/2n \mid n \in \mathbb{Z}_+\}$ of B is bounded above by any element of $(0, 1)$, but it has no least upper bound in B.

Exercises

Equivalence Relations

1. Define two points (x_0, y_0) and (x_1, y_1) of the plane to be equivalent if $y_0 - x_0^2 = y_1 - x_1^2$. Check that this is an equivalence relation and describe the equivalence classes.

2. Let C be a relation on a set A. If $A_0 \subset A$, define the ***restriction*** of C to A_0 to be the relation $C \cap (A_0 \times A_0)$. Show that the restriction of an equivalence relation is an equivalence relation.

3. Here is a "proof" that every relation C that is both symmetric and transitive is also reflexive: "Since C is symmetric, aCb implies bCa. Since C is transitive, aCb and bCa together imply aCa, as desired." Find the flaw in this argument.

4. Let $f : A \to B$ be a surjective function. Let us define a relation on A by setting $a_0 \sim a_1$ if

$$f(a_0) = f(a_1).$$

(a) Show that this is an equivalence relation.

(b) Let A^* be the set of equivalence classes. Show there is a bijective correspondence of A^* with B.

5. Let S and S' be the following subsets of the plane:

$$S = \{(x, y) \mid y = x + 1 \text{ and } 0 < x < 2\},$$
$$S' = \{(x, y) \mid y - x \text{ is an integer}\}.$$

(a) Show that S' is an equivalence relation on the real line and $S' \supset S$. Describe the equivalence classes of S'.

(b) Show that given any collection of equivalence relations on a set A, their intersection is an equivalence relation on A.

(c) Describe the equivalence relation T on the real line that is the intersection of all equivalence relations on the real line that contain S. Describe the equivalence classes of T.

Order Relations

6. Define a relation on the plane by setting

$$(x_0, y_0) < (x_1, y_1)$$

if either $y_0 - x_0^2 < y_1 - x_1^2$, or $y_0 - x_0^2 = y_1 - x_1^2$ and $x_0 < x_1$. Show that this is an order relation on the plane, and describe it geometrically.

7. Show that the restriction of an order relation is an order relation.

8. Check that the relation defined in Example 7 is an order relation.

9. Check that the dictionary order is an order relation.

10. (a) Show that the map $f : (-1, 1) \to \mathbb{R}$ of Example 9 is order preserving.
(b) Show that the equation $g(y) = 2y/[1 + (1 + 4y^2)^{1/2}]$ defines a function $g : \mathbb{R} \to (-1, 1)$ that is both a left and a right inverse for f.

11. Show that an element in an ordered set has at most one immediate successor and at most one immediate predecessor. Show that a subset of an ordered set has at most one smallest element and at most one largest element.

12. Let $\mathbb{Z}_+$ denote the set of positive integers. Consider the following order relations on $\mathbb{Z}_+ \times \mathbb{Z}_+$:
(i) The dictionary order.
(ii) $(x_0, y_0) < (x_1, y_1)$ if either $x_0 - y_0 < x_1 - y_1$, or $x_0 - y_0 = x_1 - y_1$ and $y_0 < y_1$.
(iii) $(x_0, y_0) < (x_1, y_1)$ if either $x_0 + y_0 < x_1 + y_1$, or $x_0 + y_0 = x_1 + y_1$ and $y_0 < y_1$.
In these order relations, which elements have immediate predecessors? Does the set have a smallest element? Show that all three order types are different.

13. Prove the following:
Theorem. If an ordered set A has the least upper bound property, then it has the greatest lower bound property.

14. If C is a relation on a set A, define a new relation D on A by letting $(b, a) \in D$ if $(a, b) \in C$.
(a) Show that C is symmetric if and only if $C = D$.
(b) Show that if C is an order relation, D is also an order relation.
(c) Prove the converse of the theorem in Exercise 13.

15. Assume that the real line has the least upper bound property.
(a) Show that the sets

$$[0, 1] = \{x \mid 0 \le x \le 1\},$$
$$[0, 1) = \{x \mid 0 \le x < 1\}$$

have the least upper bound property.
(b) Does $[0, 1] \times [0, 1]$ in the dictionary order have the least upper bound property? What about $[0, 1] \times [0, 1)$? What about $[0, 1) \times [0, 1]$?

§4 The Integers and the Real Numbers

Up to now we have been discussing what might be called the *logical foundations* for our study of topology—the elementary concepts of set theory. Now we turn to what we might call the *mathematical foundations* for our study—the integers and the real number system. We have already used them in an informal way in the examples and exercises of the preceding sections. Now we wish to deal with them more formally.

One way of establishing these foundations is to *construct* the real number system, using only the axioms of set theory—to build them with one's bare hands, so to speak. This way of approaching the subject takes a good deal of time and effort and is of greater logical than mathematical interest.

A second way is simply to assume a set of axioms for the real numbers and work from these axioms. In the present section, we shall sketch this approach to the real numbers. Specifically, we shall give a set of axioms for the real numbers and shall indicate how the familiar properties of real numbers and the integers are derived from them. But we shall leave most of the proofs to the exercises. If you have seen all this before, our description should refresh your memory. If not, you may want to work through the exercises in detail in order to make sure of your knowledge of the mathematical foundations.

First we need a definition from set theory.

Definition. A ***binary operation*** on a set A is a function f mapping $A \times A$ into A.

When dealing with a binary operation f on a set A, we usually use a notation different from the standard functional notation introduced in §2. Instead of denoting the value of the function f at the point (a, a') by $f(a, a')$, we usually write the symbol for the function *between* the two coordinates of the point in question, writing the value of the function at (a, a') as afa'. Furthermore (just as was the case with relations), it is more common to use some symbol other than a letter to denote an operation. Symbols often used are the plus symbol $+$, the multiplication symbols $\cdot$ and $\circ$, and the asterisk $*$; however, there are many others.

Assumption

We assume there exists a set $\mathbb{R}$, called the set of ***real numbers***, two binary operations $+$ and $\cdot$ on $\mathbb{R}$, called the addition and multiplication operations, respectively, and an order relation $<$ on $\mathbb{R}$, such that the following properties hold:

Algebraic Properties

(1) $(x + y) + z = x + (y + z)$,
$(x \cdot y) \cdot z = x \cdot (y \cdot z)$ for all x, y, z in $\mathbb{R}$.

(2) $x + y = y + x$,
$x \cdot y = y \cdot x$ for all x, y in $\mathbb{R}$.

(3) There exists a unique element of $\mathbb{R}$ called ***zero***, denoted by 0, such that $x + 0 = x$ for all $x \in \mathbb{R}$.
There exists a unique element of $\mathbb{R}$ called ***one***, different from 0 and denoted by 1, such that $x \cdot 1 = x$ for all $x \in \mathbb{R}$.

(4) For each x in $\mathbb{R}$, there exists a unique y in $\mathbb{R}$ such that $x + y = 0$.
For each x in $\mathbb{R}$ different from 0, there exists a unique y in $\mathbb{R}$ such that $x \cdot y = 1$.

(5) $x \cdot (y + z) = (x \cdot y) + (x \cdot z)$ for all $x, y, z \in \mathbb{R}$.

A Mixed Algebraic and Order Property

(6) If $x > y$, then $x + z > y + z$.
If $x > y$ and $z > 0$, then $x \cdot z > y \cdot z$.

Order Properties

(7) The order relation $<$ has the least upper bound property.

(8) If $x < y$, there exists an element z such that $x < z$ and $z < y$.

From properties (1)–(5) follow the familiar "laws of algebra." Given x, one denotes by $-x$ that number y such that $x + y = 0$; it is called the ***negative*** of x. One defines the ***subtraction operation*** by the formula $z - x = z + (-x)$. Similarly, given $x \neq 0$, one denotes by $1/x$ that number y such that $x \cdot y = 1$; it is called the ***reciprocal*** of x. One defines the ***quotient*** z/x by the formula $z/x = z \cdot (1/x)$. The usual laws of signs, and the rules for adding and multiplying fractions, follow as theorems. These laws of algebra are listed in Exercise 1 at the end of the section. We often denote $x \cdot y$ simply by xy.

When one adjoins property (6) to properties (1)–(5), one can prove the usual "laws of inequalities," such as the following:

$$\text{If } x > y \text{ and } z < 0, \text{ then } x \cdot z < y \cdot z.$$
$$-1 < 0 \text{ and } 0 < 1.$$

The laws of inequalities are listed in Exercise 2.

We define a number x to be ***positive*** if $x > 0$, and to be ***negative*** if $x < 0$. We denote the positive reals by $\mathbb{R}_+$ and the nonnegative reals (for reasons to be explained later) by $\bar{\mathbb{R}}_+$. Properties (1)–(6) are familiar properties in modern algebra. Any set with two binary operations satisfying (1)–(5) is called by algebraists a ***field***; if the field has an order relation satisfying (6), it is called an ***ordered field***.

Properties (7) and (8), on the other hand, are familiar properties in topology. They involve only the order relation; any set with an order relation satisfying (7) and (8) is called by topologists a ***linear continuum***.

Now it happens that when one adjoins to the axioms for an ordered field [properties (1)–(6)] the axioms for a linear continuum [properties (7) and (8)], the resulting list contains some redundancies. Property (8), in particular, can be proved as a consequence of the others; given $x < y$ one can show that $z = (x + y)/(1 + 1)$ satisfies the requirements of (8). Therefore, in the standard treatment of the real numbers, properties (1)–(7) are taken as axioms, and property (8) becomes a theorem. We have included (8) in our list merely to emphasize the fact that it and the least upper bound property are the two crucial properties of the order relation for $\mathbb{R}$. From these two properties many of the topological properties of $\mathbb{R}$ may be derived, as we shall see in Chapter 3.

Now there is nothing in this list as it stands to tell us what an integer is. We now *define* the integers, using only properties (1)–(6).

Definition. A subset A of the real numbers is said to be ***inductive*** if it contains the number 1, and if for every x in A, the number $x+1$ is also in A. Let $\mathcal{A}$ be the collection of all inductive subsets of $\mathbb{R}$. Then the set $\mathbb{Z}_+$ of ***positive integers*** is defined by the equation

$$\mathbb{Z}_+ = \bigcap_{A \in \mathcal{A}} A.$$

Note that the set $\mathbb{R}_+$ of positive real numbers is inductive, for it contains 1 and the statement $x > 0$ implies the statement $x + 1 > 0$. Therefore, $\mathbb{Z}_+ \subset \mathbb{R}_+$, so the elements of $\mathbb{Z}_+$ are indeed positive, as the choice of terminology suggests. Indeed, one sees readily that 1 is the smallest element of $\mathbb{Z}_+$, because the set of all real numbers x for which $x \geq 1$ is inductive.

The basic properties of $\mathbb{Z}_+$, which follow readily from the definition, are the following:

(1) $\mathbb{Z}_+$ is inductive.

(2) (Principle of induction). If A is an inductive set of positive integers, then $A = \mathbb{Z}_+$.

We define the set $\mathbb{Z}$ of ***integers*** to be the set consisting of the positive integers $\mathbb{Z}_+$, the number 0, and the negatives of the elements of $\mathbb{Z}_+$. One proves that the sum, difference, and product of two integers are integers, but the quotient is not necessarily an integer. The set $\mathbb{Q}$ of quotients of integers is called the set of ***rational numbers***.

One proves also that, given the integer n, there is no integer a such that $n < a < n + 1$.

If n is a positive integer, we use the symbol S_n to denote the set of all positive integers less than n; we call it a ***section*** of the positive integers. The set S_1 is empty, and S_{n+1} denotes the set of positive integers between 1 and n, inclusive. We also use the notation

$$\{1, \ldots, n\} = S_{n+1}$$

for the latter set.

Now we prove two properties of the positive integers that may not be quite so familiar, but are quite useful. They may be thought of as alternative versions of the induction principle.

Theorem 4.1 (Well-ordering property). *Every nonempty subset of $\mathbb{Z}_+$ has a smallest element.*

Proof. We first prove that, for each $n \in \mathbb{Z}_+$, the following statement holds: *Every nonempty subset of* $\{1, \dots, n\}$ *has a smallest element.*

Let A be the set of all positive integers n for which this statement holds. Then A contains 1, since if $n = 1$, the only nonempty subset of $\{1, \dots, n\}$ is the set $\{1\}$ itself. Then, supposing A contains n, we show that it contains $n + 1$. So let C be a nonempty subset of the set $\{1, \dots, n + 1\}$. If C consists of the single element $n + 1$, then that element is the smallest element of C. Otherwise, consider the set $C \cap \{1, \dots, n\}$, which is nonempty. Because $n \in A$, this set has a smallest element, which will automatically be the smallest element of C also. Thus A is inductive, so we conclude that $A = \mathbb{Z}_+$; hence the statement is true for all $n \in \mathbb{Z}_+$.

Now we prove the theorem. Suppose that D is a nonempty subset of $\mathbb{Z}_+$. Choose an element n of D. Then the set $A = D \cap \{1, \dots, n\}$ is nonempty, so that A has a smallest element k. The element k is automatically the smallest element of D as well. ∎

Theorem 4.2 (Strong induction principle). *Let A be a set of positive integers. Suppose that for each positive integer n, the statement $S_n \subset A$ implies the statement $n \in A$. Then $A = \mathbb{Z}_+$.*

Proof. If A does not equal all of $\mathbb{Z}_+$, let n be the smallest positive integer that is not in A. Then every positive integer less than n is in A, so that $S_n \subset A$. Our hypothesis implies that $n \in A$, contrary to assumption. ∎

Everything we have done up to now has used only the axioms for an ordered field, properties (1)–(6) of the real numbers. At what point do you need (7), the least upper bound axiom?

For one thing, you need the least upper bound axiom to prove that the set $\mathbb{Z}_+$ of positive integers has no upper bound in $\mathbb{R}$. This is the ***Archimedean ordering property*** of the real line. To prove it, we assume that $\mathbb{Z}_+$ has an upper bound and derive a contradiction. If $\mathbb{Z}_+$ has an upper bound, it has a least upper bound b. There exists $n \in \mathbb{Z}_+$ such that $n > b - 1$; for otherwise, $b - 1$ would be an upper bound for $\mathbb{Z}_+$ smaller than b. Then $n + 1 > b$, contrary to the fact that b is an upper bound for $\mathbb{Z}_+$.

The least upper bound axiom is also used to prove a number of other things about $\mathbb{R}$. It is used for instance to show that $\mathbb{R}$ has the greatest lower bound property. It is also used to prove the existence of a unique positive square root $\sqrt{x}$ for every positive real number. This fact, in turn, can be used to demonstrate the existence of real numbers that are not rational numbers; the number $\sqrt{2}$ is an easy example.

We use the symbol 2 to denote $1 + 1$, the symbol 3 to denote $2 + 1$, and so on through the standard symbols for the positive integers. It is a fact that this procedure assigns to each positive integer a unique symbol, but we never need this fact and shall not prove it.

Proofs of these properties of the integers and real numbers, along with a few other properties we shall need later, are outlined in the exercises that follow.

Exercises

1. Prove the following "laws of algebra" for $\mathbb{R}$, using only axioms (1)–(5):
(a) If $x + y = x$, then $y = 0$.
(b) $0 \cdot x = 0$. [*Hint:* Compute $(x + 0) \cdot x$.]
(c) $-0 = 0$.
(d) $-(-x) = x$.
(e) $x(-y) = -(xy) = (-x)y$.
(f) $(-1)x = -x$.
(g) $x(y - z) = xy - xz$.
(h) $-(x + y) = -x - y$; $-(x - y) = -x + y$.
(i) If $x \neq 0$ and $x \cdot y = x$, then $y = 1$.
(j) $x/x = 1$ if $x \neq 0$.
(k) $x/1 = x$.
(l) $x \neq 0$ and $y \neq 0 \Rightarrow xy \neq 0$.
(m) $(1/y)(1/z) = 1/(yz)$ if $y, z \neq 0$.
(n) $(x/y)(w/z) = (xw)/(yz)$ if $y, z \neq 0$.
(o) $(x/y) + (w/z) = (xz + wy)/(yz)$ if $y, z \neq 0$.
(p) $x \neq 0 \Rightarrow 1/x \neq 0$.
(q) $1/(w/z) = z/w$ if $w, z \neq 0$.
(r) $(x/y)/(w/z) = (xz)/(yw)$ if $y, w, z \neq 0$.
(s) $(ax)/y = a(x/y)$ if $y \neq 0$.
(t) $(-x)/y = x/(-y) = -(x/y)$ if $y \neq 0$.

2. Prove the following "laws of inequalities" for $\mathbb{R}$, using axioms (1)–(6) along with the results of Exercise 1:
(a) $x > y$ and $w > z \Rightarrow x + w > y + z$.
(b) $x > 0$ and $y > 0 \Rightarrow x + y > 0$ and $x \cdot y > 0$.
(c) $x > 0 \Leftrightarrow -x < 0$.
(d) $x > y \Leftrightarrow -x < -y$.
(e) $x > y$ and $z < 0 \Rightarrow xz < yz$.
(f) $x \neq 0 \Rightarrow x^2 > 0$, where $x^2 = x \cdot x$.
(g) $-1 < 0 < 1$.
(h) $xy > 0 \Leftrightarrow x$ and y are both positive or both negative.
(i) $x > 0 \Rightarrow 1/x > 0$.
(j) $x > y > 0 \Rightarrow 1/x < 1/y$.
(k) $x < y \Rightarrow x < (x + y)/2 < y$.

3. (a) Show that if $\mathcal{A}$ is a collection of inductive sets, then the intersection of the elements of $\mathcal{A}$ is an inductive set.
(b) Prove the basic properties (1) and (2) of $\mathbb{Z}_+$.

4. (a) Prove by induction that given $n \in \mathbb{Z}_+$, every nonempty subset of $\{1, \ldots, n\}$ has a largest element.
(b) Explain why you cannot conclude from (a) that every nonempty subset of $\mathbb{Z}_+$ has a largest element.

5. Prove the following properties of $\mathbb{Z}$ and $\mathbb{Z}_+$:
(a) $a, b \in \mathbb{Z}_+ \Rightarrow a + b \in \mathbb{Z}_+$. [*Hint:* Show that given $a \in \mathbb{Z}_+$, the set $X = \{x \mid x \in \mathbb{R} \text{ and } a + x \in \mathbb{Z}_+\}$ is inductive.]
(b) $a, b \in \mathbb{Z}_+ \Rightarrow a \cdot b \in \mathbb{Z}_+$.
(c) Show that $a \in \mathbb{Z}_+ \Rightarrow a - 1 \in \mathbb{Z}_+ \cup \{0\}$. [*Hint:* Let $X = \{x \mid x \in \mathbb{R}$ and $x - 1 \in \mathbb{Z}_+ \cup \{0\}$; show that X is inductive.]
(d) $c, d \in \mathbb{Z} \Rightarrow c + d \in \mathbb{Z}$ and $c - d \in \mathbb{Z}$. [*Hint:* Prove it first for $d = 1$.]
(e) $c, d \in \mathbb{Z} \Rightarrow c \cdot d \in \mathbb{Z}$.

6. Let $a \in \mathbb{R}$. Define inductively

$$a^1 = a,$$
$$a^{n+1} = a^n \cdot a$$

for $n \in \mathbb{Z}_+$. (See §7 for a discussion of the process of inductive definition.) Show that for $n, m \in \mathbb{Z}_+$ and $a, b \in \mathbb{R}$,

$$a^n a^m = a^{n+m},$$
$$(a^n)^m = a^{nm},$$
$$a^m b^m = (ab)^m.$$

These are called the ***laws of exponents***. [*Hint:* For fixed n, prove the formulas by induction on m.]

7. Let $a \in \mathbb{R}$ and $a \neq 0$. Define $a^0 = 1$, and for $n \in \mathbb{Z}_+$, $a^{-n} = 1/a^n$. Show that the laws of exponents hold for $a, b \neq 0$ and $n, m \in \mathbb{Z}$.

8. (a) Show that $\mathbb{R}$ has the greatest lower bound property.
(b) Show that $\inf\{1/n \mid n \in \mathbb{Z}_+\} = 0$.
(c) Show that given a with $0 < a < 1$, $\inf\{a^n \mid n \in \mathbb{Z}_+\} = 0$. [*Hint:* Let $h = (1 - a)/a$, and show that $(1 + h)^n \geq 1 + nh$.]

9. (a) Show that every nonempty subset of $\mathbb{Z}$ that is bounded above has a largest element.
(b) If $x \notin \mathbb{Z}$, show there is exactly one $n \in \mathbb{Z}$ such that $n < x < n + 1$.
(c) If $x - y > 1$, show there is at least one $n \in \mathbb{Z}$ such that $y < n < x$.
(d) If $y < x$, show there is a rational number z such that $y < z < x$.

10. Show that every positive number a has exactly one positive square root, as follows:
(a) Show that if $x > 0$ and $0 \leq h < 1$, then

$$(x + h)^2 \leq x^2 + h(2x + 1),$$
$$(x - h)^2 \geq x^2 - h(2x).$$

(b) Let $x > 0$. Show that if $x^2 < a$, then $(x + h)^2 < a$ for some $h > 0$; and if $x^2 > a$, then $(x - h)^2 > a$ for some $h > 0$.

(c) Given $a > 0$, let B be the set of all real numbers x such that $x^2 < a$. Show that B is bounded above and contains at least one positive number. Let $b = \sup B$; show that $b^2 = a$.
(d) Show that if b and c are positive and $b^2 = c^2$, then $b = c$.

11. Given $m \in \mathbb{Z}$, we say that m is ***even*** if $m/2 \in \mathbb{Z}$, and m is ***odd*** otherwise.
(a) Show that if m is odd, $m = 2n + 1$ for some $n \in \mathbb{Z}$. [*Hint:* Choose n so that $n < m/2 < n + 1$.]
(b) Show that if p and q are odd, so are $p \cdot q$ and p^n, for any $n \in \mathbb{Z}_+$.
(c) Show that if $a > 0$ is rational, then $a = m/n$ for some $m, n \in \mathbb{Z}_+$ where not both m and n are even. [*Hint:* Let n be the smallest element of the set $\{x \mid x \in \mathbb{Z}_+ \text{ and } x \cdot a \in \mathbb{Z}_+\}$.]
(d) *Theorem.* $\sqrt{2}$ *is irrational.*

§5 Cartesian Products

We have already defined what we mean by the cartesian product $A \times B$ of two sets. Now we introduce more general cartesian products.

Definition. Let $\mathcal{A}$ be a nonempty collection of sets. An ***indexing function*** for $\mathcal{A}$ is a surjective function f from some set J, called the ***index set***, to $\mathcal{A}$. The collection $\mathcal{A}$, together with the indexing function f, is called an ***indexed family of sets***. Given $\alpha \in J$, we shall denote the set $f(\alpha)$ by the symbol A_α. And we shall denote the indexed family itself by the symbol

$$\{A_\alpha\}_{\alpha \in J},$$

which is read "the family of all A_α, as α ranges over J." Sometimes we write merely $\{A_\alpha\}$, if it is clear what the index set is.

Note that although an indexing function is required to be surjective, it is not required to be *injective*. It is entirely possible for A_α and A_β to be the same set of $\mathcal{A}$, even though $\alpha \neq \beta$.

One way in which indexing functions are used is to give a new notation for arbitrary unions and intersections of sets. Suppose that $f : J \to \mathcal{A}$ is an indexing function for $\mathcal{A}$; let A_α denote $f(\alpha)$. Then we define

$$\bigcup_{\alpha \in J} A_\alpha = \{x \mid \text{for at least one } \alpha \in J, x \in A_\alpha\},$$

and

$$\bigcap_{\alpha \in J} A_\alpha = \{x \mid \text{for every } \alpha \in J, x \in A_\alpha\}.$$

These are simply new notations for previously defined concepts; one sees at once (using the surjectivity of the index function) that the first equals the union of all the elements of $\mathcal{A}$ and the second equals the intersection of all the elements of $\mathcal{A}$.

Two especially useful index sets are the set $\{1, \dots, n\}$ of positive integers from 1 to n, and the set $\mathbb{Z}_+$ of all positive integers. For these index sets, we introduce some special notation. If a collection of sets is indexed by the set $\{1, \dots, n\}$, we denote the indexed family by the symbol $\{A_1, \dots, A_n\}$, and we denote the union and intersection, respectively, of the members of this family by the symbols

$$A_1 \cup \cdots \cup A_n \quad \text{and} \quad A_1 \cap \cdots \cap A_n.$$

In the case where the index set is the set $\mathbb{Z}_+$, we denote the indexed family by the symbol $\{A_1, A_2, \dots\}$, and the union and intersection by the respective symbols

$$A_1 \cup A_2 \cup \cdots \quad \text{and} \quad A_1 \cap A_2 \cap \cdots .$$

Definition. Let m be a positive integer. Given a set X, we define an ***m-tuple*** of elements of X to be a function

$$\mathbf{x} : \{1, \dots, m\} \to X.$$

If $\mathbf{x}$ is an m-tuple, we often denote the value of $\mathbf{x}$ at i by the symbol x_i rather than $\mathbf{x}(i)$ and call it the ith ***coordinate*** of $\mathbf{x}$. And we often denote the function $\mathbf{x}$ itself by the symbol

$$(x_1, \dots, x_m).$$

Now let $\{A_1, \dots, A_m\}$ be a family of sets indexed with the set $\{1, \dots, m\}$. Let $X = A_1 \cup \cdots \cup A_m$. We define the ***cartesian product*** of this indexed family, denoted by

$$\prod_{i=1}^{m} A_i \quad \text{or} \quad A_1 \times \cdots \times A_m,$$

to be the set of all m-tuples $(x_1, \dots, x_m)$ of elements of X such that $x_i \in A_i$ for each i.

EXAMPLE 1. We now have two definitions for the symbol $A \times B$. One definition is, of course, the one given earlier, under which $A \times B$ denotes the set of all ordered pairs (a, b) such that $a \in A$ and $b \in B$. The second definition, just given, defines $A \times B$ as the set of all functions $\mathbf{x} : \{1, 2\} \to A \cup B$ such that $\mathbf{x}(1) \in A$ and $\mathbf{x}(2) \in B$. There is an obvious bijective correspondence between these two sets, under which the ordered pair (a, b) corresponds to the function $\mathbf{x}$ defined by $\mathbf{x}(1) = a$ and $\mathbf{x}(2) = b$. Since we commonly denote this function $\mathbf{x}$ in "tuple notation" by the symbol (a, b), the notation itself suggests the correspondence. Thus for the cartesian product of two sets, the general definition of cartesian product reduces essentially to the earlier one.

EXAMPLE 2. How does the cartesian product $A \times B \times C$ differ from the cartesian products $A \times (B \times C)$ and $(A \times B) \times C$? Very little. There are obvious bijective correspondences between these sets, indicated as follows:

$$(a, b, c) \longleftrightarrow (a, (b, c)) \longleftrightarrow ((a, b), c).$$

Definition. Given a set X, we define an ***ω-tuple*** of elements of X to be a function

$$\mathbf{x} : \mathbb{Z}_+ \longrightarrow X;$$

we also call such a function a ***sequence***, or an ***infinite sequence***, of elements of X. If $\mathbf{x}$ is an ω-tuple, we often denote the value of $\mathbf{x}$ at i by x_i rather than $\mathbf{x}(i)$, and call it the ith ***coordinate*** of $\mathbf{x}$. We denote $\mathbf{x}$ itself by the symbol

$$(x_1, x_2, \dots) \quad \text{or} \quad (x_n)_{n \in \mathbb{Z}_+}.$$

Now let $\{A_1, A_2, \dots\}$ be a family of sets, indexed with the positive integers; let X be the union of the sets in this family. The ***cartesian product*** of this indexed family of sets, denoted by

$$\prod_{i \in \mathbb{Z}_+} A_i \quad \text{or} \quad A_1 \times A_2 \times \cdots,$$

is defined to be the set of all ω-tuples $(x_1, x_2, \dots)$ of elements of X such that $x_i \in A_i$ for each i.

Nothing in these definitions requires the sets A_i to be different from one another. Indeed, they may all equal the same set X. In that case, the cartesian product $A_1 \times \cdots \times A_m$ is just the set of *all* m-tuples of elements of X, which we denote by X^m. Similarly, the product $A_1 \times A_2 \times \cdots$ is just the set of all ω-tuples of elements of X, which we denote by X^ω.

Later we will define the cartesian product of an *arbitrary* indexed family of sets.

EXAMPLE 3. If $\mathbb{R}$ is the set of real numbers, then $\mathbb{R}^m$ denotes the set of all m-tuples of real numbers; it is often called ***euclidean m-space*** (although Euclid would never recognize it). Analogously, $\mathbb{R}^\omega$ is sometimes called "infinite-dimensional euclidean space"; it is the set of all ω-tuples $(x_1, x_2, \dots)$ of real numbers, that is, the set of all functions $\mathbf{x} : \mathbb{Z}_+ \to \mathbb{R}$.

Exercises

1. Show there is a bijective correspondence of $A \times B$ with $B \times A$.

2. (a) Show that if $n > 1$ there is bijective correspondence of

$$A_1 \times \cdots \times A_n \quad \text{with} \quad (A_1 \times \cdots \times A_{n-1}) \times A_n.$$

(b) Given the indexed family $\{A_1, A_2, \dots\}$, let $B_i = A_{2i-1} \times A_{2i}$ for each positive integer i. Show there is bijective correspondence of $A_1 \times A_2 \times \cdots$ with $B_1 \times B_2 \times \cdots$.

3. Let $A = A_1 \times A_2 \times \cdots$ and $B = B_1 \times B_2 \times \cdots$.

(a) Show that if $B_i \subset A_i$ for all i, then $B \subset A$. (Strictly speaking, if we are given a function mapping the index set $\mathbb{Z}_+$ into the union of the sets B_i, we must change its range before it can be considered as a function mapping $\mathbb{Z}_+$ into the union of the sets A_i. We shall ignore this technicality when dealing with cartesian products).

(b) Show the converse of (a) holds if B is nonempty.
(c) Show that if A is nonempty, each A_i is nonempty. Does the converse hold? (We will return to this question in the exercises of §19.)
(d) What is the relation between the set $A \cup B$ and the cartesian product of the sets $A_i \cup B_i$? What is the relation between the set $A \cap B$ and the cartesian product of the sets $A_i \cap B_i$?

4. Let $m, n \in \mathbb{Z}_+$. Let $X \neq \varnothing$.
(a) If $m \leq n$, find an injective map $f : X^m \to X^n$.
(b) Find a bijective map $g : X^m \times X^n \to X^{m+n}$.
(c) Find an injective map $h : X^n \to X^\omega$.
(d) Find a bijective map $k : X^n \times X^\omega \to X^\omega$.
(e) Find a bijective map $l : X^\omega \times X^\omega \to X^\omega$.
(f) If $A \subset B$, find an injective map $m : (A^\omega)^n \to B^\omega$.

5. Which of the following subsets of $\mathbb{R}^\omega$ can be expressed as the cartesian product of subsets of $\mathbb{R}$?
(a) $\{\mathbf{x} \mid x_i \text{ is an integer for all } i\}$.
(b) $\{\mathbf{x} \mid x_i \geq i \text{ for all } i\}$.
(c) $\{\mathbf{x} \mid x_i \text{ is an integer for all } i \geq 100\}$.
(d) $\{\mathbf{x} \mid x_2 = x_3\}$.

§6 Finite Sets

Finite sets and infinite sets, countable sets and uncountable sets, these are types of sets that you may have encountered before. Nevertheless, we shall discuss them in this section and the next, not only to make sure you understand them thoroughly, but also to elucidate some particular points of logic that will arise later on. First we consider finite sets.

Recall that if n is a positive integer, we use S_n to denote the set of positive integers less than n; it is called a *section* of the positive integers. The sets S_n are the prototypes for what we call the finite sets.

Definition. A set is said to be ***finite*** if there is a bijective correspondence of A with some section of the positive integers. That is, A is finite if it is empty or if there is a bijection

$$f : A \longrightarrow \{1, \ldots, n\}$$

for some positive integer n. In the former case, we say that A has ***cardinality 0***; in the latter case, we say that A has ***cardinality n***.

For instance, the set $\{1, \ldots, n\}$ itself has cardinality n, for it is in bijective correspondence with itself under the identity function.

Now note carefully: *We have not yet shown that the cardinality of a finite set is uniquely determined by the set.* It is of course clear that the empty set must have cardinality zero. But as far as we know, there might exist bijective correspondences of a given nonempty set A with two different sets $\{1, \ldots, n\}$ and $\{1, \ldots, m\}$. The possibility may seem ridiculous, for it is like saying that it is possible for two people to count the marbles in a box and come out with two different answers, *both correct.* Our experience with counting in everyday life suggests that such is impossible, and in fact this is easy to prove when n is a small number such as $1, 2$, or 3. But a direct proof when n is 5 million would be impossibly demanding.

Even empirical demonstration would be difficult for such a large value of n. One might, for instance, construct an experiment by taking a freight car full of marbles and hiring 10 different people to count them independently. If one thinks of the physical problems involved, it seems likely that the counters would not all arrive at the same answer. Of course, the conclusion one could draw is that at least one person made a mistake. But that would mean assuming the correctness of the result one was trying to demonstrate empirically. An alternative explanation could be that there do exist bijective correspondences between the given set of marbles and two different sections of the positive integers.

In real life, we accept the first explanation. We simply take it on faith that our experience in counting comparatively small sets of objects demonstrates a truth that holds for arbitrarily large sets as well.

However, in mathematics (as opposed to real life), one does not have to take this statement on faith. If it is formulated in terms of the existence of bijective correspondences rather than in terms of the physical act of counting, it is capable of mathematical proof. We shall prove shortly that if $n \neq m$, there do not exist bijective functions mapping a given set A onto both the sets $\{1, \ldots, n\}$ and $\{1, \ldots, m\}$.

There are a number of other "intuitively obvious" facts about finite sets that are capable of mathematical proof; we shall prove some of them in this section and leave the rest to the exercises. Here is an easy fact to start with:

Lemma 6.1. *Let n be a positive integer. Let A be a set; let a_0 be an element of A. Then there exists a bijective correspondence f of the set A with the set $\{1, \ldots, n+1\}$ if and only if there exists a bijective correspondence g of the set $A - \{a_0\}$ with the set $\{1, \ldots, n\}$.*

Proof. There are two implications to be proved. Let us first assume that there is a bijective correspondence

$$g : A - \{a_0\} \longrightarrow \{1, \ldots, n\}.$$

We then define a function $f : A \longrightarrow \{1, \ldots, n+1\}$ by setting

$$\begin{aligned} f(x) &= g(x) \qquad \text{for } x \in A - \{a_0\}, \\ f(a_0) &= n + 1. \end{aligned}$$

One checks at once that f is bijective.

To prove the converse, assume there is a bijective correspondence

$$f : A \longrightarrow \{1, \ldots, n+1\}.$$

If f maps a_0 to the number $n+1$, things are especially easy; in that case, the restriction $f|A - \{a_0\}$ is the desired bijective correspondence of $A - \{a_0\}$ with $\{1, \ldots, n\}$. Otherwise, let $f(a_0) = m$, and let a_1 be the point of A such that $f(a_1) = n+1$. Then $a_1 \neq a_0$. Define a new function

$$h : A \longrightarrow \{1, \ldots, n+1\}$$

by setting

$$\begin{aligned} h(a_0) &= n+1, \\ h(a_1) &= m, \\ h(x) &= f(x) \qquad \text{for } x \in A - \{a_0\} - \{a_1\}. \end{aligned}$$

See Figure 6.1. It is easy to check that h is a bijection.

Now we are back in the easy case; the restriction $h|A - \{a_0\}$ is the desired bijection of $A - \{a_0\}$ with $\{1, \ldots, n\}$. ∎

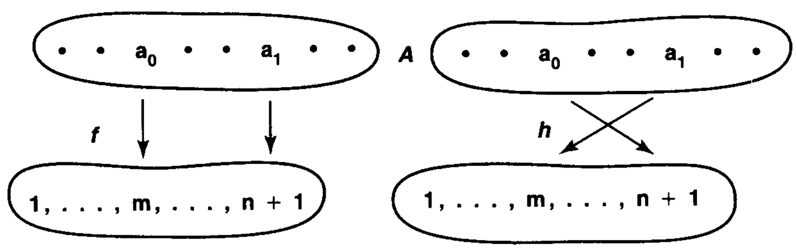


Figure 6.1

From this lemma a number of useful consequences follow:

Theorem 6.2. *Let A be a set; suppose that there exists a bijection $f : A \to \{1, \ldots, n\}$ for some $n \in \mathbb{Z}_+$. Let B be a proper subset of A. Then there exists no bijection $g : B \to \{1, \ldots, n\}$; but (provided $B \neq \varnothing$) there does exist a bijection $h : B \to \{1, \ldots, m\}$ for some $m < n$.*

Proof. The case in which $B = \varnothing$ is trivial, for there cannot exist a bijection of the empty set B with the nonempty set $\{1, \ldots, n\}$.

We prove the theorem "by induction." Let C be the subset of $\mathbb{Z}_+$ consisting of those integers n for which the theorem holds. We shall show that C is inductive. From this we conclude that $C = \mathbb{Z}_+$, so the theorem is true for all positive integers n.

First we show the theorem is true for $n = 1$. In this case A consists of a single element $\{a\}$, and its only proper subset B is the empty set.

Now assume that the theorem is true for n; we prove it true for $n+1$. Suppose that $f : A \to \{1, \dots, n+1\}$ is a bijection, and B is a nonempty proper subset of A. Choose an element a_0 of B and an element a_1 of $A - B$. We apply the preceding lemma to conclude there is a bijection

$$g : A - \{a_0\} \longrightarrow \{1, \dots, n\}.$$

Now $B - \{a_0\}$ is a proper subset of $A - \{a_0\}$, for a_1 belongs to $A - \{a_0\}$ and not to $B - \{a_0\}$. Because the theorem has been assumed to hold for the integer n, we conclude the following:

(1) There exists no bijection $h : B - \{a_0\} \to \{1, \dots, n\}$.

(2) Either $B - \{a_0\} = \varnothing$, or there exists a bijection

$$k : B - \{a_0\} \longrightarrow \{1, \dots, p\} \qquad \text{for some } p < n.$$

The preceding lemma, combined with (1), implies that there is no bijection of B with $\{1, \dots, n+1\}$. This is the first half of what we wanted to proved. To prove the second half, note that if $B - \{a_0\} = \varnothing$, there is a bijection of B with the set $\{1\}$; while if $B - \{a_0\} \neq \varnothing$, we can apply the preceding lemma, along with (2), to conclude that there is a bijection of B with $\{1, \dots, p+1\}$. In either case, there is a bijection of B with $\{1, \dots, m\}$ for some $m < n+1$, as desired. The induction principle now shows that the theorem is true for all $n \in \mathbb{Z}_+$. ∎

Corollary 6.3. *If A is finite, there is no bijection of A with a proper subset of itself.*

Proof. Assume that B is a proper subset of A and that $f : A \to B$ is a bijection. By assumption, there is a bijection $g : A \to \{1, \dots, n\}$ for some n. The composite $g \circ f^{-1}$ is then a bijection of B with $\{1, \dots, n\}$. This contradicts the preceding theorem. ∎

Corollary 6.4. *$\mathbb{Z}_+$ is not finite.*

Proof. The function $f : \mathbb{Z}_+ \to \mathbb{Z}_+ - \{1\}$ defined by $f(n) = n+1$ is a bijection of $\mathbb{Z}_+$ with a proper subset of itself. ∎

Corollary 6.5. *The cardinality of a finite set A is uniquely determined by A.*

Proof. Let $m < n$. Suppose there are bijections

$$f : A \longrightarrow \{1, \dots, n\},$$
$$g : A \longrightarrow \{1, \dots, m\}.$$

Then the composite

$$g \circ f^{-1} : \{1, \dots, n\} \longrightarrow \{1, \dots, m\}$$

is a bijection of the finite set $\{1, \dots, n\}$ with a proper subset of itself, contradicting the corollary just proved. ∎

Corollary 6.6. *If B is a subset of the finite set A, then B is finite. If B is a proper subset of A, then the cardinality of B is less than the cardinality of A.*

Corollary 6.7. *Let B be a nonempty set. Then the following are equivalent:*

(1) *B is finite.*

(2) *There is a surjective function from a section of the positive integers onto B.*

(3) *There is an injective function from B into a section of the positive integers.*

Proof. (1) $\Longrightarrow$ (2). Since B is nonempty, there is, for some n, a bijective function $f : \{1, \ldots, n\} \to B$.

(2) $\Longrightarrow$ (3). If $f : \{1, \ldots, n\} \to B$ is surjective, define $g : B \to \{1, \ldots, n\}$ by the equation

$$g(b) = \text{smallest element of } f^{-1}(\{b\}).$$

Because f is surjective, the set $f^{-1}\{(b)\}$ is nonempty; then the well-ordering property of $\mathbb{Z}_+$ tells us that $g(b)$ is uniquely defined. The map g is injective, for if $b \neq b'$, then the sets $f^{-1}(\{b\})$ and $f^{-1}(\{b'\})$ are disjoint, so their smallest elements must be different.

(3) $\Longrightarrow$ (1). If $g : B \to \{1, \ldots, n\}$ is injective, then changing the range of g gives a bijection of B with a subset of $\{1, \ldots, n\}$. It follows from the preceding corollary that B is finite. ■

Corollary 6.8. *Finite unions and finite cartesian products of finite sets are finite.*

Proof. We first show that if A and B are finite, so is $A \cup B$. The result is trivial if A or B is empty. Otherwise, there are bijections $f : \{1, \ldots, m\} \to A$ and $g : \{1, \ldots, n\} \to B$ for some choice of m and n. Define a function $h : \{1, \ldots, m + n\} \to A \cup B$ by setting $h(i) = f(i)$ for $i = 1, 2, \ldots, m$ and $h(i) = g(i - m)$ for $i = m + 1, \ldots, m + n$. It is easy to check that h is surjective, from which it follows that $A \cup B$ is finite.

Now we show by induction that finiteness of the sets $A_1, \ldots, A_n$ implies finiteness of their union. This result is trivial for $n = 1$. Assuming it true for $n - 1$, we note that $A_1 \cup \cdots \cup A_n$ is the union of the two finite sets $A_1 \cup \cdots \cup A_{n-1}$ and A_n, so the result of the preceding paragraph applies.

Now we show that the cartesian product of two finite sets A and B is finite. Given $a \in A$, the set $\{a\} \times B$ is finite, being in bijective correspondence with B. The set $A \times B$ is the union of these sets; since there are only finitely many of them, $A \times B$ is a finite union of finite sets and thus finite.

To prove that the product $A_1 \times \cdots \times A_n$ is finite if each A_i is finite, one proceeds by induction. ■

Exercises

1. (a) Make a list of all the injective maps

$$f : \{1, 2, 3\} \longrightarrow \{1, 2, 3, 4\}.$$

Show that none is bijective. (This constitutes a *direct* proof that a set A of cardinality three does not have cardinality four.)

(b) How many injective maps

$$f : \{1, \ldots, 8\} \longrightarrow \{1, \ldots, 10\}$$

are there? (You can see why one would not wish to try to prove *directly* that there is no bijective correspondence between these sets.)

2. Show that if B is not finite and $B \subset A$, then A is not finite.
3. Let X be the two-element set $\{0, 1\}$. Find a bijective correspondence between X^ω and a proper subset of itself.
4. Let A be a nonempty finite simply ordered set.
   (a) Show that A has a largest element. [*Hint:* Proceed by induction on the cardinality of A.]
   (b) Show that A has the order type of a section of the positive integers.
5. If $A \times B$ is finite, does it follow that A and B are finite?
6. (a) Let $A = \{1, \ldots, n\}$. Show there is a bijection of $\mathcal{P}(A)$ with the cartesian product X^n, where X is the two-element set $X = \{0, 1\}$.
   (b) Show that if A is finite, then $\mathcal{P}(A)$ is finite.
7. If A and B are finite, show that the set of all functions $f : A \to B$ is finite.

§7 Countable and Uncountable Sets

Just as sections of the positive integers are the prototypes for the finite sets, the set of all the positive integers is the prototype for what we call the *countably infinite* sets. In this section, we shall study such sets; we shall also construct some sets that are neither finite nor countably infinite. This study will lead us into a discussion of what we mean by the process of "inductive definition."

Definition. A set A is said to be ***infinite*** if it is not finite. It is said to be ***countably infinite*** if there is a bijective correspondence

$$f : A \longrightarrow \mathbb{Z}_+.$$

EXAMPLE 1. The set $\mathbb{Z}$ of all integers is countably infinite. One checks easily that the function $f : \mathbb{Z} \to \mathbb{Z}_+$ defined by

$$f(n) = \begin{cases} 2n & \text{if } n > 0, \\ -2n + 1 & \text{if } n \leq 0 \end{cases}$$

is a bijection.

EXAMPLE 2. The product $\mathbb{Z}_+ \times \mathbb{Z}_+$ is countably infinite. If we represent the elements of the product $\mathbb{Z}_+ \times \mathbb{Z}_+$ by the integer points in the first quadrant, then the left-hand portion of Figure 7.1 suggests how to "count" the points, that is, how to put them in bijective correspondence with the positive integers. A picture is not a proof, of course, but this picture suggests a proof. First, we define a bijection $f : \mathbb{Z}_+ \times \mathbb{Z}_+ \to A$, where A is the subset of $\mathbb{Z}_+ \times \mathbb{Z}_+$ consisting of pairs (x, y) for which $y \leq x$, by the equation

$$f(x, y) = (x + y - 1, y).$$

Then we construct a bijection of A with the positive integers, defining $g : A \to \mathbb{Z}_+$ by the formula

$$g(x, y) = \frac{1}{2}(x - 1)x + y.$$

We leave it to you to show that f and g are bijections.

Another proof that $\mathbb{Z}_+ \times \mathbb{Z}_+$ is countably infinite will be given later.

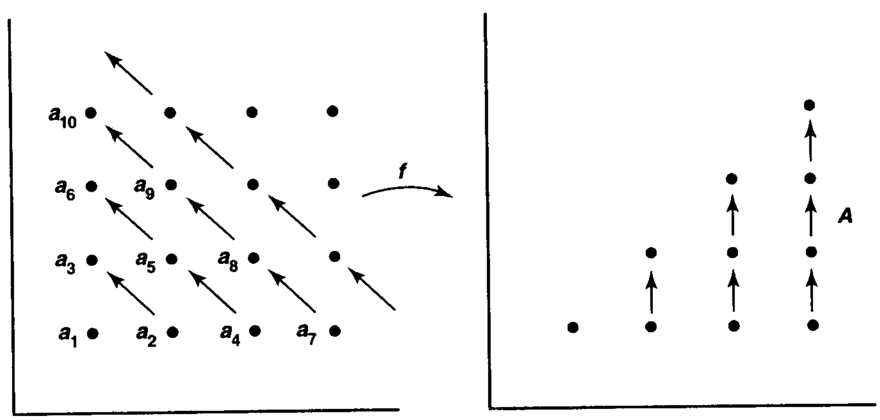


Figure 7.1

Definition. A set is said to be ***countable*** if it is either finite or countably infinite. A set that is not countable is said to be ***uncountable***.

There is a very useful criterion for showing that a set is countable. It is the following:

Theorem 7.1. *Let B be a nonempty set. Then the following are equivalent:*

(1) *B is countable.*

(2) *There is a surjective function $f : \mathbb{Z}_+ \to B$.*

(3) *There is an injective function $g : B \to \mathbb{Z}_+$.*

Proof. (1) $\Longrightarrow$ (2). Suppose that B is countable. If B is countably infinite, there is a bijection $f : \mathbb{Z}_+ \to B$ by definition, and we are through. If B is finite, there is a bijection $h : \{1, \dots, n\} \to B$ for some $n \geq 1$. (Recall that $B \neq \varnothing$.) We can extend h to a surjection $f : \mathbb{Z}_+ \to B$ by defining

$$f(i) = \begin{cases} h(i) & \text{for } 1 \leq i \leq n, \\ h(1) & \text{for } i > n. \end{cases}$$

(2) $\Longrightarrow$ (3). Let $f : \mathbb{Z}_+ \to B$ be a surjection. Define $g : B \to \mathbb{Z}_+$ by the equation

$$g(b) = \text{smallest element of } f^{-1}(\{b\}).$$

Because f is surjective, $f^{-1}(\{b\})$ is nonempty; thus g is well defined. The map g is injective, for if $b \neq b'$, the sets $f^{-1}(\{b\})$ and $f^{-1}(\{b'\})$ are disjoint, so their smallest elements are different.

(3) $\Longrightarrow$ (1). Let $g : B \to \mathbb{Z}_+$ be an injection; we wish to prove B is countable. By changing the range of g, we can obtain a bijection of B with a subset of $\mathbb{Z}_+$. Thus to prove our result, it suffices to show that every subset of $\mathbb{Z}_+$ is countable. So let C be a subset of $\mathbb{Z}_+$.

If C is finite, it is countable by definition. So what we need to prove is that every infinite subset C of $\mathbb{Z}_+$ is countably infinite. This statement is certainly plausible. For the elements of C can easily be arranged in an infinite sequence; one simply takes the set $\mathbb{Z}_+$ in its usual order and "erases" all the elements of $\mathbb{Z}_+$ that are not in C!

The plausibility of this argument may make one overlook its informality. Providing a formal proof requires a certain amount of care. We state this result as a separate lemma, which follows. ■

Lemma 7.2. *If C is an infinite subset of $\mathbb{Z}_+$, then C is countably infinite.*

Proof. We define a bijection $h : \mathbb{Z}_+ \to C$. We proceed by induction. Define $h(1)$ to be the smallest element of C; it exists because every nonempty subset C of $\mathbb{Z}_+$ has a smallest element. Then assuming that $h(1), \dots, h(n-1)$ are defined, define

$$h(n) = \text{ smallest element of } [C - h(\{1, \dots, n-1\})].$$

The set $C - h(\{1, \dots, n-1\})$ is not empty; for if it were empty, then $h : \{1, \dots, n-1\} \to C$ would be surjective, so that C would be finite (by Corollary 6.7). Thus $h(n)$ is well defined. By induction, we have defined $h(n)$ for all $n \in \mathbb{Z}_+$.

To show that h is injective is easy. Given $m < n$, note that $h(m)$ belongs to the set $h(\{1, \dots, n-1\})$, whereas $h(n)$, by definition, does not. Hence $h(n) \neq h(m)$.

To show that h is surjective, let c be any element of C; we show that c lies in the image set of h. First note that $h(\mathbb{Z}_+)$ cannot be contained in the finite set $\{1, \dots, c\}$, because $h(\mathbb{Z}_+)$ is infinite (since h is injective). Therefore, there is an n in $\mathbb{Z}_+$, such that $h(n) > c$. Let m be the *smallest* element of $\mathbb{Z}_+$, such that $h(m) \geq c$. Then for all $i < m$, we must have $h(i) < c$. Thus, c does not belong to the set $h(\{1, \dots, m-1\})$. Since $h(m)$ is defined as the smallest element of the set $C - h(\{1, \dots, m-1\})$, we must have $h(m) \leq c$. Putting the two inequalities together, we have $h(m) = c$, as desired. ■

There is a point in the preceding proof where we stretched the principles of logic a bit. It occurred at the point where we said that "using the induction principle" we had defined the function h for all positive integers n. You may have seen arguments like this used before, with no questions raised concerning their legitimacy. We have already used such an argument ourselves, in the exercises of §4, when we defined a^n.

But there is a problem here. After all, the induction principle states only that if A is an inductive set of positive integers, then $A = \mathbb{Z}_+$. To use the principle to prove a theorem "by induction," one begins the proof with the statement "Let A be the set of all positive integers n for which the theorem is true," and then one goes ahead to prove that A is inductive, so that A must be all of $\mathbb{Z}_+$.

In the preceding theorem, however, we were not really proving a theorem by induction, but defining something by induction. How then should we start the proof? Can we start by saying, "Let A be the set of all integers n for which the function h is defined"? But that's silly; the symbol h has no *meaning* at the outset of the proof. It only takes on meaning in the course of the proof. So something more is needed.

What is needed is another principle, which we call the ***principle of recursive definition***. In the proof of the preceding theorem, we wished to assert the following:

Given the infinite subset C of $\mathbb{Z}_+$, there is a unique function $h : \mathbb{Z}_+ \to C$ satisfying the formula:

$$(*) \qquad \begin{aligned} h(1) &= \text{smallest element of } C, \\ h(i) &= \text{smallest element of } [C - h(\{1, \dots, i-1\})] \quad \text{for all } i > 1. \end{aligned}$$

The formula $(*)$ is called a ***recursion formula*** for h; it defines the function h *in terms of itself.* A definition given by such a formula is called a ***recursive definition***.

Now one can get into logical difficulties when one tries to define something recursively. Not all recursive formulas make sense. The recursive formula

$$h(i) = \text{smallest element of } [C - h(\{1, \dots, i+1\})],$$

for example, is self-contradictory; although $h(i)$ necessarily is an element of the set $h(\{1, \dots, i+1\})$, this formula says that it does not belong to the set. Another example is the classic paradox:

> Let the barber of Seville shave every man of Seville who does not shave himself. Who shall shave the barber?

In this statement, the barber appears twice, once in the phrase "the barber of Seville" and once as an element of the set "men of Seville"; this definition of whom the barber shall shave is a recursive one. It also happens to be self-contradictory.

Some recursive formulas do make sense, however. Specifically, one has the following principle:

Principle of recursive definition. *Let A be a set. Given a formula that defines $h(1)$ as a unique element of A, and for $i > 1$ defines $h(i)$ uniquely as an element of A in terms of the values of h for positive integers less than i, this formula determines a unique function $h : \mathbb{Z}_+ \to A$.*

This principle is the one we actually used in the proof of Lemma 7.2. You can simply accept it on faith if you like. It may however be proved rigorously, using the principle of induction. We shall formulate it more precisely in the next section and indicate how it is proved. Mathematicians seldom refer to this principle specifically. They are much more likely to write a proof like our proof of Lemma 7.2 above, a proof in which they invoke the "induction principle" to define a function when what they are really using is the principle of recursive definition. We shall avoid undue pedantry in this book by following their example.

Corollary 7.3. *A subset of a countable set is countable.*

Proof. Suppose $A \subset B$, where B is countable. There is an injection f of B into $\mathbb{Z}_+$; the restriction of f to A is an injection of A into $\mathbb{Z}_+$. ∎

Corollary 7.4. *The set $\mathbb{Z}_+ \times \mathbb{Z}_+$ is countably infinite.*

Proof. In view of Theorem 7.1, it suffices to construct an injective map $f : \mathbb{Z}_+ \times \mathbb{Z}_+ \to \mathbb{Z}_+$. We define f by the equation

$$f(n, m) = 2^n 3^m.$$

It is easy to check that f is injective. For suppose that $2^n 3^m = 2^p 3^q$. If $n < p$, then $3^m = 2^{p-n} 3^q$, contradicting the fact that 3^m is odd for all m. Therefore, $n = p$. As a result, $3^m = 3^q$, Then if $m < q$, it follows that $1 = 3^{q-m}$, another contradiction. Hence $m = q$. ∎

EXAMPLE 3. The set $\mathbb{Q}_+$ of positive rational numbers is countably infinite. For we can define a surjection $g : \mathbb{Z}_+ \times \mathbb{Z}_+ \to \mathbb{Q}_+$ by the equation

$$g(n, m) = m/n.$$

Because $\mathbb{Z}_+ \times \mathbb{Z}_+$ is countable, there is a surjection $f : \mathbb{Z}_+ \to \mathbb{Z}_+ \times \mathbb{Z}_+$. Then the composite $g \circ f : \mathbb{Z}_+ \to \mathbb{Q}_+$ is a surjection, so that $\mathbb{Q}_+$ is countable. And, of course, $\mathbb{Q}_+$ is infinite because it contains $\mathbb{Z}_+$.

We leave it as an exercise to show the set $\mathbb{Q}$ of *all* rational numbers is countably infinite.

Theorem 7.5. *A countable union of countable sets is countable.*

Proof. Let $\{A_n\}_{n \in J}$ be an indexed family of countable sets, where the index set J is either $\{1, \ldots, N\}$ or $\mathbb{Z}_+$. Assume that each set A_n is nonempty, for convenience; this assumption does not change anything.

Because each A_n is countable, we can choose, for each n, a surjective function $f_n : \mathbb{Z}_+ \to A_n$. Similarly, we can choose a surjective function $g : \mathbb{Z}_+ \to J$. Now define

$$h : \mathbb{Z}_+ \times \mathbb{Z}_+ \to \bigcup_{n \in J} A_n$$

by the equation

$$h(k, m) = f_{g(k)}(m).$$

It is easy to check that h is surjective. Since $\mathbb{Z}_+ \times \mathbb{Z}_+$ is in bijective correspondence with $\mathbb{Z}_+$, the countability of the union follows from Theorem 7.1. ∎

Theorem 7.6. *A finite product of countable sets is countable.*

Proof. First let us show that the product of two countable sets A and B is countable. The result is trivial if A or B is empty. Otherwise, choose surjective functions $f : \mathbb{Z}_+ \to A$ and $g : \mathbb{Z}_+ \to B$. Then the function $h : \mathbb{Z}_+ \times \mathbb{Z}_+ \to A \times B$ defined by the equation $h(n, m) = (f(n), g(m))$ is surjective, so that $A \times B$ is countable.

In general, we proceed by induction. Assuming that $A_1 \times \cdots \times A_{n-1}$ is countable if each A_i is countable, we prove the same thing for the product $A_1 \times \cdots \times A_n$. First, note that there is a bijective correspondence

$$g : A_1 \times \cdots \times A_n \longrightarrow (A_1 \times \cdots \times A_{n-1}) \times A_n$$

defined by the equation

$$g(x_1, \ldots, x_n) = ((x_1, \ldots, x_{n-1}), x_n).$$

Because the set $A_1 \times \cdots \times A_{n-1}$ is countable by the induction assumption and A_n is countable by hypothesis, the product of these two sets is countable, as proved in the preceding paragraph. We conclude that $A_1 \times \cdots \times A_n$ is countable as well. ∎

It is very tempting to assert that countable products of countable sets should be countable; but this assertion is in fact not true:

Theorem 7.7. *Let X denote the two element set $\{0, 1\}$. Then the set X^ω is uncountable.*

Proof. We show that, given any function

$$g : \mathbb{Z}_+ \longrightarrow X^\omega,$$

g is not surjective. For this purpose, let us denote $g(n)$ as follows :

$$g(n) = (x_{n1}, x_{n2}, x_{n3}, \ldots x_{nm}, \ldots),$$

where each x_{ij} is either 0 or 1. Then we define an element $\mathbf{y} = (y_1, y_2, \ldots, y_n, \ldots)$ of X^ω by letting

$$y_n = \begin{cases} 0 & \text{if } x_{nn} = 1, \\ 1 & \text{if } x_{nn} = 0. \end{cases}$$

(If we write the numbers x_{ni} in a rectangular array, the particular elements x_{nn} appear as the diagonal entries in this array; we choose $\mathbf{y}$ so that its nth coordinate *differs* from the diagonal entry x_{nn}.)

Now $\mathbf{y}$ is an element of X^ω, and $\mathbf{y}$ does not lie in the image of g; given n, the point $g(n)$ and the point $\mathbf{y}$ differ in at least one coordinate, namely, the nth. Thus, g is not surjective. ∎

The cartesian product $\{0, 1\}^\omega$ is one example of an uncountable set. Another is the set $\mathcal{P}(\mathbb{Z}_+)$, as the following theorem implies:

Theorem 7.8. *Let A be a set. There is no injective map $f : \mathcal{P}(A) \to A$, and there is no surjective map $g : A \to \mathcal{P}(A)$.*

Proof. In general, if B is a nonempty set, the existence of an injective map $f : B \to C$ implies the existence of a surjective map $g : C \to B$; one defines $g(c) = f^{-1}(c)$ for each c in the image set of f, and defines g arbitrarily on the rest of C.

Therefore, it suffices to prove that given a map $g : A \to \mathcal{P}(A)$, the map g is not surjective. For each $a \in A$, the image $g(a)$ of a is a subset of A, which may or may not contain the point a itself. Let B be the subset of A consisting of all those points a such that $g(a)$ does not contain a;

$$B = \{a \mid a \in A - g(a)\}.$$

Now, B may be empty, or it may be all of A, but that does not matter. We assert that B is a subset of A that does not lie in the image of g. For suppose that $B = g(a_0)$ for some $a_0 \in A$. We ask the question: Does a_0 belong to B or not? By definition of B,

$$a_0 \in B \iff a_0 \in A - g(a_0) \iff a_0 \in A - B.$$

In either case, we have a contradiction. ∎

Now we have proved the existence of uncountable sets. But we have not yet mentioned the most familiar uncountable set of all—the set of real numbers. You have probably seen the uncountability of $\mathbb{R}$ demonstrated already. If one assumes that every real number can be represented uniquely by an infinite decimal (with the proviso that a representation ending in an infinite string of 9's is forbidden), then the uncountability of the reals can be proved by a variant of the diagonal procedure used in the proof of Theorem 7.7. But this proof is in some ways not very satisfying. One reason is that the infinite decimal representation of a real number is not at all an elementary consequence of the axioms but requires a good deal of labor to prove. Another reason is that the uncountability of $\mathbb{R}$ does not, in fact, depend on the infinite decimal expansion of $\mathbb{R}$ or indeed on any of the algebraic properties of $\mathbb{R}$; it depends on only the order properties of $\mathbb{R}$. We shall demonstrate the uncountability of $\mathbb{R}$, using only its order properties, in a later chapter.

Exercises

1. Show that $\mathbb{Q}$ is countably infinite.

2. Show that the maps f and g of Examples 1 and 2 are bijections.

3. Let X be the two-element set $\{0, 1\}$. Show there is a bijective correspondence between the set $\mathcal{P}(\mathbb{Z}_+)$ and the cartesian product X^ω.

4. (a) A real number x is said to be ***algebraic*** (over the rationals) if it satisfies some polynomial equation of positive degree

$$x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = 0$$

with rational coefficients a_i. Assuming that each polynomial equation has only finitely many roots, show that the set of algebraic numbers is countable.

(b) A real number is said to be ***transcendental*** if it is not algebraic. Assuming the reals are uncountable, show that the transcendental numbers are uncountable. (It is a somewhat surprising fact that only two transcendental numbers are familiar to us: e and π. Even proving these two numbers transcendental is highly nontrivial.)

5. Determine, for each of the following sets, whether or not it is countable. Justify your answers.

(a) The set A of all functions $f : \{0, 1\} \to \mathbb{Z}_+$.

(b) The set B_n of all functions $f : \{1, \ldots, n\} \to \mathbb{Z}_+$.

(c) The set $C = \bigcup_{n \in \mathbb{Z}_+} B_n$.

(d) The set D of all functions $f : \mathbb{Z}_+ \to \mathbb{Z}_+$.

(e) The set E of all functions $f : \mathbb{Z}_+ \to \{0, 1\}$.

(f) The set F of all functions $f : \mathbb{Z}_+ \to \{0, 1\}$ that are "eventually zero." [We say that f is ***eventually zero*** if there is a positive integer N such that $f(n) = 0$ for all $n \geq N$.]

(g) The set G of all functions $f : \mathbb{Z}_+ \to \mathbb{Z}_+$ that are eventually 1.

(h) The set H of all functions $f : \mathbb{Z}_+ \to \mathbb{Z}_+$ that are eventually constant.

(i) The set I of all two-element subsets of $\mathbb{Z}_+$.

(j) The set J of all finite subsets of $\mathbb{Z}_+$.

6. We say that two sets A and B ***have the same cardinality*** if there is a bijection of A with B.

(a) Show that if $B \subset A$ and if there is an injection

$$f : A \longrightarrow B,$$

then A and B have the same cardinality. [*Hint:* Define $A_1 = A$, $B_1 = B$, and for $n > 1$, $A_n = f(A_{n-1})$ and $B_n = f(B_{n-1})$. (Recursive definition again!) Note that $A_1 \supset B_1 \supset A_2 \supset B_2 \supset A_3 \supset \cdots$. Define a bijection $h : A \to B$ by the rule

$$h(x) = \begin{cases} f(x) & \text{if } x \in A_n - B_n \text{ for some } n, \\ x & \text{otherwise.]} \end{cases}$$

(b) *Theorem (Schroeder-Bernstein theorem). If there are injections $f : A \to C$ and $g : C \to A$, then A and C have the same cardinality.*

7. Show that the sets D and E of Exercise 5 have the same cardinality.

8. Let X denote the two-element set $\{0, 1\}$; let $\mathcal{B}$ be the set of *countable* subsets of X^ω. Show that X^ω and $\mathcal{B}$ have the same cardinality.

9. (a) The formula

$$(*)\qquad \begin{aligned} h(1) &= 1, \\ h(2) &= 2, \\ h(n) &= [h(n+1)]^2 - [h(n-1)]^2 \qquad \text{for } n \geq 2 \end{aligned}$$

is not one to which the principle of recursive definition applies. Show that nevertheless there does exist a function $h : \mathbb{Z}_+ \to \mathbb{R}$ satisfying this formula. [*Hint:* Reformulate $(*)$ so that the principle will apply and require h to be positive.]

(b) Show that the formula $(*)$ of part (a) does not determine h uniquely. [*Hint:* If h is a positive function satisfying $(*)$, let $f(i) = h(i)$ for $i \neq 3$, and let $f(3) = -h(3)$.]

(c) Show that there is no function $h : \mathbb{Z}_+ \to \mathbb{R}$ satisfying the formula

$$\begin{aligned} h(1) &= 1, \\ h(2) &= 2, \\ h(n) &= [h(n+1)]^2 + [h(n-1)]^2 \qquad \text{for } n \geq 2. \end{aligned}$$

*§8 The Principle of Recursive Definition

Before considering the general form of the principle of recursive definition, let us first prove it in a specific case, that of Lemma 7.2. That should make the underlying idea of the proof much clearer when we consider the general case.

So, given the infinite subset C of $\mathbb{Z}_+$, let us consider the following recursion formula for a function $h : \mathbb{Z}_+ \to C$:

$$(*)\qquad \begin{aligned} h(1) &= \text{smallest element of } C, \\ h(i) &= \text{smallest element of } [C - h(\{1, \ldots, i-1\})] \qquad \text{for } i > 1. \end{aligned}$$

We shall prove that there exists a unique function $h : \mathbb{Z}_+ \to C$ satisfying this recursion formula.

The first step is to prove that there exist functions defined on *sections* $\{1, \ldots, n\}$ of $\mathbb{Z}_+$ that satisfy $(*)$:

Lemma 8.1. *Given $n \in \mathbb{Z}_+$, there exists a function*

$$f : \{1, \ldots, n\} \to C$$

that satisfies $(*)$ for all i in its domain.

Proof. The point of this lemma is that it is a statement that depends on n; therefore, it is capable of being proved by induction. Let A be the set of all n for which the lemma holds. We show that A is inductive. It then follows that $A = \mathbb{Z}_+$.

The lemma is true for $n = 1$, since the function $f : \{1\} \to C$ defined by the equation

$$f(1) = \text{smallest element of } C$$

satisfies $(*)$.

Supposing the lemma to be true for $n - 1$, we prove it true for n. By hypothesis, there is a function $f' : \{1, \ldots, n - 1\} \to C$ satisfying $(*)$ for all i in its domain. Define $f : \{1, \ldots, n\} \to C$ by the equations

$$\begin{aligned} f(i) &= f'(i) \qquad \text{for } i \in \{1, \ldots, n - 1\}, \\ f(n) &= \text{smallest element of } [C - f'(\{1, \ldots, n - 1\})]. \end{aligned}$$

Since C is infinite, f' is not surjective; hence the set $C - f'(\{1, \ldots, n - 1\})$ is not empty, and $f(n)$ is well defined. Note that this definition is an acceptable one; it does not define f in terms of *itself* but in terms of the given function f'.

It is easy to check that f satisfies $(*)$ for all i in its domain. The function f satisfies $(*)$ for $i \leq n - 1$ because it equals f' there. And f satisfies $(*)$ for $i = n$ because, by definition,

$$f(n) = \text{smallest element of } [C - f'(\{1, \ldots, n - 1\})]$$

and $f'(\{1, \ldots, n - 1\}) = f(\{1, \ldots, n - 1\})$. ∎

Lemma 8.2. *Suppose that $f : \{1, \ldots, n\} \to C$ and $g : \{1, \ldots, m\} \to C$ both satisfy $(*)$ for all i in their respective domains. Then $f(i) = g(i)$ for all i in both domains.*

Proof. Suppose not. Let i be the *smallest* integer for which $f(i) \neq g(i)$. The integer i is not 1, because

$$f(1) = \text{smallest element of } C = g(1),$$

by $(*)$. Now for all $j < i$, we have $f(j) = g(j)$. Because f and g satisfy $(*)$,

$$\begin{aligned} f(i) &= \text{smallest element of } [C - f(\{1, \ldots, i - 1\})], \\ g(i) &= \text{smallest element of } [C - g(\{1, \ldots, i - 1\})]. \end{aligned}$$

Since $f(\{1, \ldots, i - 1\}) = g(\{1, \ldots, i - 1\})$, we have $f(i) = g(i)$, contrary to the choice of i. ∎

Theorem 8.3. *There exists a unique function $h : \mathbb{Z}_+ \to C$ satisfying $(*)$ for all $i \in \mathbb{Z}_+$.*

Proof. By Lemma 8.1, there exists for each n a function that maps $\{1, \ldots, n\}$ into C and satisfies $(*)$ for all i in its domain. Given n, Lemma 8.2 shows that this function is unique; two such functions having the same domain must be equal. Let $f_n : \{1, \ldots, n\} \to C$ denote this unique function.

Now comes the crucial step. We define a function $h : \mathbb{Z}_+ \to C$ by defining its rule to be the *union* U of the rules of the functions f_n. The rule for f_n is a subset of $\{1, \ldots, n\} \times C$; therefore, U is a subset of $\mathbb{Z}_+ \times C$. We must show that U is the rule for a function $h : \mathbb{Z}_+ \to C$.

That is, we must show that each element i of $\mathbb{Z}_+$ appears as the first coordinate of exactly one element of U. This is easy. The integer i lies in the domain of f_n if and only if $n > i$. Therefore, the set of elements of U of which i is the first coordinate is precisely the set of all pairs of the form $(i, f_n(i))$, for $n \geq i$. Now Lemma 8.2 tells us that $f_n(i) = f_m(i)$ if $n, m \geq i$. Therefore, all these elements of U are equal; that is, there is only one element of U that has i as its first coordinate.

To show that h satisfies $(*)$ is also easy; it is a consequence of the following facts:

$$h(i) = f_n(i) \qquad \text{for } i \leq n,$$
$$f_n \text{ satisfies } (*) \text{ for all } i \text{ in its domain.}$$

The proof of uniqueness is a copy of the proof of Lemma 8.2. ∎

Now we formulate the general principle of recursive definition. There are no new ideas involved in its proof, so we leave it as an exercise.

Theorem 8.4 (Principle of recursive definition). *Let A be a set; let a_0 be an element of A. Suppose ρ is a function that assigns, to each function f mapping a nonempty section of the positive integers into A, an element of A. Then there exists a unique function*

$$h : \mathbb{Z}_+ \to A$$

such that

$$(*) \qquad \begin{aligned} h(1) &= a_0, \\ h(i) &= \rho(h|\{1, \ldots, i-1\}) \qquad \text{for } i > 1. \end{aligned}$$

The formula $(*)$ is called a ***recursion formula*** for h. It specifies $h(1)$, and it expresses the value of h at $i > 1$ in terms of the values of h for positive integers less than i.

EXAMPLE 1. Let us show that Theorem 8.3 is a special case of this theorem. Given the infinite subset C of $\mathbb{Z}_+$, let a_0 be the smallest element of C, and define ρ by the equation

$$\rho(f) = \text{smallest element of } [C - (\text{image set of } f)].$$

Because C is infinite and f is a function mapping a finite set into C, the image set of f is not all of C; therefore, ρ is well defined. By Theorem 8.4 there exists a function $h : \mathbb{Z}_+ \to C$ such that $h(1) = a_0$, and for $i > 1$,

$$\begin{aligned} h(i) &= \rho(h|\{1, \dots, i-1\}) \\ &= \text{smallest element of } [C - (\text{image set of } h|\{1, \dots, i-1\})] \\ &= \text{smallest element of } [C - h(\{1 \dots, i-1\})], \end{aligned}$$

as desired.

EXAMPLE 2. Given $a \in \mathbb{R}$, we "defined" a^n, in the exercises of §4, by the recursion formula

$$\begin{aligned} a^1 &= a, \\ a^n &= a^{n-1} \cdot a. \end{aligned}$$

We wish to apply Theorem 8.4 to define a function $h : \mathbb{Z}_+ \to \mathbb{R}$ rigorously such that $h(n) = a^n$. To apply this theorem, let a_0 denote the element a of $\mathbb{R}$, and define ρ by the equation $\rho(f) = f(m) \cdot a$, where $f : \{1, \dots, m\} \to \mathbb{R}$. Then there exists a unique function $h : \mathbb{Z}_+ \to \mathbb{R}$ such that

$$\begin{aligned} h(1) &= a_0, \\ h(i) &= \rho(h|\{1, \dots, i-1\}) \qquad \text{for } i > 1. \end{aligned}$$

This means that $h(1) = a$, and $h(i) = h(i-1) \cdot a$ for $i > 1$. If we denote $h(i)$ by a^i, we have

$$\begin{aligned} a^1 &= a, \\ a^i &= a^{i-1} \cdot a, \end{aligned}$$

as desired.

Exercises

1. Let $(b_1, b_2, \dots)$ be an infinite sequence of real numbers. The sum $\sum_{k=1}^n b_k$ is defined by induction as follows :

$$\begin{aligned} \sum_{k=1}^n b_k &= b_1 & \text{for } n = 1, \\ \sum_{k=1}^n b_k &= (\sum_{k=1}^{n-1} b_k) + b_n & \text{for } n > 1. \end{aligned}$$

Let A be the set of real numbers; choose ρ so that Theorem 8.4 applies to define this sum rigorously. We sometimes denote the sum $\sum_{k=1}^n b_k$ by the symbol $b_1 + b_2 + \cdots + b_n$.

2. Let $(b_1, b_2, \dots)$ be an infinite sequence of real numbers. We define the product $\prod_{k=1}^{n} b_k$ by the equations

$$\prod_{k=1}^{1} b_k = b_1,$$
$$\prod_{k=1}^{n} b_k = (\prod_{k=1}^{n-1} b_k) \cdot b_n \qquad \text{for } n > 1.$$

Use Theorem 8.4 to define this product rigorously. We sometimes denote the product $\prod_{k=1}^{n} b_k$ by the symbol $b_1 b_2 \cdots b_n$.

3. Obtain the definitions of a^n and $n!$ for $n \in \mathbb{Z}_+$ as special cases of Exercise 2.

4. The *Fibonacci numbers* of number theory are defined recursively by the formula

$$\lambda_1 = \lambda_2 = 1,$$
$$\lambda_n = \lambda_{n-1} + \lambda_{n-2} \qquad \text{for } n > 2.$$

Define them rigorously by use of Theorem 8.4.

5. Show that there is a unique function $h : \mathbb{Z}_+ \to \mathbb{R}_+$ satisfying the formula

$$h(1) = 3,$$
$$h(i) = [h(i-1) + 1]^{1/2} \qquad \text{for } i > 1.$$

6. (a) Show that there is no function $h : \mathbb{Z}_+ \to \mathbb{R}_+$ satisfying the formula

$$h(1) = 3,$$
$$h(i) = [h(i-1) - 1]^{1/2} \qquad \text{for } i > 1.$$

Explain why this example does not violate the principle of recursive definition.

(b) Consider the recursion formula

$$h(1) = 3,$$
$$h(i) = \begin{Bmatrix} [h(i-1) - 1]^{1/2} & \text{if } h(i-1) > 1 \\ 5 & \text{if } h(i-1) \leq 1 \end{Bmatrix} \qquad \text{for } i > 1.$$

Show that there exists a unique function $h : \mathbb{Z}_+ \to \mathbb{R}_+$ satisfying this formula.

7. Prove Theorem 8.4.

8. Verify the following version of the principle of recursive definition: Let A be a set. Let ρ be a function assigning, to every function f mapping a section S_n of $\mathbb{Z}_+$ into A, an element $\rho(f)$ of A. Then there is a unique function $h : \mathbb{Z}_+ \to A$ such that $h(n) = \rho(h|S_n)$ for each $n \in \mathbb{Z}_+$.

§9 Infinite Sets and the Axiom of Choice

We have already obtained several criteria for a set to be infinite. We know, for instance, that a set A is infinite if it has a countably infinite subset, or if there is a bijection of A with a proper subset of itself. It turns out that either of these properties is sufficient to characterize infinite sets. This we shall now prove. The proof will lead us into a discussion of a point of logic we have not yet mentioned—the axiom of choice.

Theorem 9.1. *Let A be a set. The following statements about A are equivalent:*

(1) *There exists an injective function $f : \mathbb{Z}_+ \to A$.*

(2) *There exists a bijection of A with a proper subset of itself.*

(3) *A is infinite.*

Proof. We prove the implications (1) $\Rightarrow$ (2) $\Rightarrow$ (3) $\Rightarrow$ (1). To prove that (1) $\Rightarrow$ (2), we assume there is an injective function $f : \mathbb{Z}_+ \to A$. Let the image set $f(\mathbb{Z}_+)$ be denoted by B; and let $f(n)$ be denoted by a_n. Because f is injective, $a_n \neq a_m$ if $n \neq m$. Define

$$g : A \longrightarrow A - \{a_1\}$$

by the equations

$$\begin{aligned} g(a_n) &= a_{n+1} && \text{for } a_n \in B, \\ g(x) &= x && \text{for } x \in A - B. \end{aligned}$$

The map g is indicated schematically in Figure 9.1; one checks easily that it is a bijection.

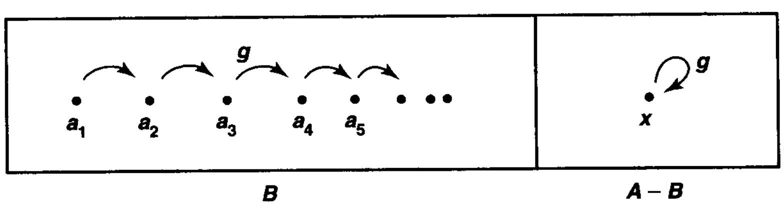


Figure 9.1

The implication (2) $\Rightarrow$ (3) is just the contrapositive of Corollary 6.3, so it has already been proved. To prove that (3) $\Rightarrow$ (1), we assume that A is infinite and construct "by induction" an injective function $f : \mathbb{Z}_+ \to A$.

First, since the set A is not empty, we can choose a point a_1 of A; define $f(1)$ to be the point so chosen.

Then, assuming that we have defined $f(1), \dots, f(n-1)$, we wish to define $f(n)$. The set $A - f(\{1, \dots, n-1\})$ is not empty; for if it were empty, the map $f : \{1, \dots, n-1\} \to A$ would be a surjection and A would be finite. Hence, we can choose an element of the set $A - f(\{1, \dots, n-1\})$ and define $f(n)$ to be this element. "Using the induction principle", we have defined f for all $n \in \mathbb{Z}_+$.

It is easy to see that f is injective. For suppose that $m < n$. Then $f(m)$ belongs to the set $f(\{1, \dots, n-1\})$, whereas $f(n)$, by definition, does not. Therefore, $f(n) \neq f(m)$. ∎

Let us try to reformulate this "induction" proof more carefully, so as to make explicit our use of the principle of recursive definition.

Given the infinite set A, we attempt to define $f : \mathbb{Z}_+ \to A$ recursively by the formula

$$(*) \qquad \begin{aligned} f(1) &= a_1, \\ f(i) &= \text{an arbitrary element of } [A - f(\{1, \dots, i-1\})] \qquad \text{for } i > 1. \end{aligned}$$

But this is not an acceptable recursion formula at all! For it does not define $f(i)$ *uniquely* in terms of $f|\{1, \dots, i-1\}$.

In this respect this formula differs notably from the recursion formula we considered in proving Lemma 7.2. There we had an infinite subset C of $\mathbb{Z}_+$, and we defined h by the formula

$$\begin{aligned} h(1) &= \text{smallest element of } C, \\ h(i) &= \text{smallest element of } [C - h(\{1, \dots, i-1\})] \qquad \text{for } i > 1. \end{aligned}$$

This formula does define $h(i)$ uniquely in terms of $h|\{1, \dots, i-1\}$.

Another way of seeing that $(*)$ is not an acceptable recursion formula is to note that if it were, the principle of recursive definition would imply that there is a *unique* function $f : \mathbb{Z}_+ \to A$ satisfying $(*)$. But by no stretch of the imagination does $(*)$ specify f uniquely. In fact, this "definition" of f involves infinitely many arbitrary choices.

What we are saying is that the proof we have given for Theorem 9.1 is not actually a proof. Indeed, on the basis of the properties of set theory we have discussed up to now, it is not *possible* to prove this theorem. Something more is needed.

Previously, we described certain definite allowable methods for specifying sets:

(1) Defining a set by listing its elements, or by taking a given set A and specifying a subset B of it by giving a property that the elements of B are to satisfy.

(2) Taking unions or intersections of the elements of a given collection of sets, or taking the difference of two sets.

(3) Taking the set of all subsets of a given set.

(4) Taking cartesian products of sets.

Now the rule for the function f is really a set: a subset of $\mathbb{Z}_+ \times A$. Therefore, to prove the existence of the function f, we must construct the appropriate subset of $\mathbb{Z}_+ \times A$, using the allowed methods for forming sets. The methods already given simply are not adequate for this purpose. We need a new way of asserting the existence of a set. So, we add to the list of allowed methods of forming sets the following:

Axiom of choice. *Given a collection $\mathcal{A}$ of disjoint nonempty sets, there exists a set C consisting of exactly one element from each element of $\mathcal{A}$; that is, a set C such that C is contained in the union of the elements of $\mathcal{A}$, and for each $A \in \mathcal{A}$, the set $C \cap A$ contains a single element.*

The set C can be thought of as having been obtained by choosing one element from each of the sets in $\mathcal{A}$.

The axiom of choice certainly seems an innocent-enough assertion. And, in fact, most mathematicians today accept it as part of the set theory on which they base their mathematics. But in years past a good deal of controversy raged around this particular assertion concerning set theory, for there are theorems one can prove with its aid that some mathematicians were reluctant to accept. One such is the well-ordering theorem, which we shall discuss shortly. For the present we shall simply use the choice axiom to clear up the difficulty we mentioned in the preceding proof. First, we prove an easy consequence of the axiom of choice:

Lemma 9.2 (Existence of a choice function). *Given a collection $\mathcal{B}$ of nonempty sets (not necessarily disjoint), there exists a function*

$$c : \mathcal{B} \longrightarrow \bigcup_{B \in \mathcal{B}} B$$

such that $c(B)$ is an element of B, for each $B \in \mathcal{B}$.

The function c is called a ***choice function*** for the collection $\mathcal{B}$.

The difference between this lemma and the axiom of choice is that in this lemma the sets of the collection $\mathcal{B}$ are not required to be disjoint. For example, one can allow $\mathcal{B}$ to be the collection of *all* nonempty subsets of a given set.

Proof of the lemma. Given an element B of $\mathcal{B}$, we define a set B' as follows:

$$B' = \{(B, x) \mid x \in B\}.$$

That is, B' is the collection of all ordered pairs, where the first coordinate of the ordered pair is the set B, and the second coordinate is an element of B. The set B' is a subset of the cartesian product

$$\mathcal{B} \times \bigcup_{B \in \mathcal{B}} B.$$

Because B contains at least one element x, the set B' contains at least the element (B, x), so it is nonempty.

Now we claim that if B_1 and B_2 are two different sets in $\mathcal{B}$, then the corresponding sets B_1' and B_2' are disjoint. For the typical element of B_1' is a pair of the form (B_1, x_1) and the typical element of B_2' is a pair of the form (B_2, x_2). No two such elements can be equal, for their first coordinates are different. Now let us form the collection

$$\mathcal{C} = \{B' \mid B \in \mathcal{B}\};$$

it is a collection of disjoint nonempty subsets of

$$\mathcal{B} \times \bigcup_{B \in \mathcal{B}} B.$$

By the choice axiom, there exists a set c consisting of exactly one element from each element of $\mathcal{C}$. Our claim is that c *is* the rule for the desired choice function.

In the first place, c is a subset of

$$\mathcal{B} \times \bigcup_{B \in \mathcal{B}} B.$$

In the second place, c contains exactly one element from each set B'; therefore, for each $B \in \mathcal{B}$, the set c contains exactly one ordered pair (B, x) whose first coordinate is B. Thus c is indeed the rule for a function from the collection $\mathcal{B}$ to the set $\bigcup_{B \in \mathcal{B}} B$. Finally, if $(B, x) \in c$, then x belongs to B, so that $c(B) \in B$, as desired. ■

A second proof of Theorem 9.1. Using this lemma, one can make the proof of Theorem 9.1 more precise. Given the infinite set A, we wish to construct an injective function $f : \mathbb{Z}_+ \to A$. Let us form the collection $\mathcal{B}$ of all nonempty subsets of A. The lemma just proved asserts the existence of a choice function for $\mathcal{B}$; that is, a function

$$c : \mathcal{B} \longrightarrow \bigcup_{B \in \mathcal{B}} B = A$$

such that $c(B) \in B$ for each $B \in \mathcal{B}$. Let us now define a function $f : \mathbb{Z}_+ \to A$ by the recursion formula

$$(*) \qquad \begin{aligned} f(1) &= c(A), \\ f(i) &= c(A - f(\{1, \dots, i-1\})) \qquad \text{for } i > 1. \end{aligned}$$

Because A is infinite, the set $A - f(\{1, \dots, i-1\})$ is nonempty; therefore, the right side of this equation makes sense. Since this formula defines $f(i)$ uniquely in terms of $f|\{1, \dots, i-1\}$, the principle of recursive definition applies. We conclude that there exists a unique function $f : \mathbb{Z}_+ \to A$ satisfying $(*)$ for all $i \in \mathbb{Z}_+$. Injectivity of f follows as before. ■

Having emphasized that in order to construct a proof of Theorem 9.1 that is logically correct, one must make specific use of a choice function, we now backtrack and admit that in practice most mathematicians do no such thing. They go on with no qualms giving proofs like our first version, proofs that involve an infinite number of arbitrary choices. They know that they are really using the choice axiom; and they know that if it were necessary, they could put their proofs into a logically more satisfactory form by introducing a choice function specifically. But usually they do not bother.

And neither will we. You will find few further specific uses of a choice function in this book; we shall introduce a choice function only when the proof would become confusing without it. But there will be many proofs in which we make an infinite number of arbitrary choices, and in each such case we will actually be using the choice axiom implicitly.

Now we must confess that in an earlier section of this book there is a proof in which we constructed a certain function by making an infinite number of arbitrary choices. And we slipped that proof in without even mentioning the choice axiom. Our apologies for the deception. We leave it to you to ferret out which proof it was!

Let us make one final comment on the choice axiom. There are two forms of this axiom. One can be called the ***finite axiom of choice***; it asserts that given a *finite* collection $\mathcal{A}$ of disjoint nonempty sets, there exists a set C consisting of exactly one element from each element of $\mathcal{A}$. One needs this weak form of the choice axiom all the time; we have used it freely in the preceding sections with no comment. No mathematician has any qualms about the finite choice axiom; it is part of everyone's set theory. Said differently, no one has qualms about a proof that involves only finitely many arbitrary choices.

The stronger form of the axiom of choice, the one that applies to an *arbitrary* collection $\mathcal{A}$ of nonempty sets, is the one that is properly called "the axiom of choice." When a mathematician writes, "This proof depends on the choice axiom," it is invariably this stronger form of the axiom that is meant.

Exercises

1. Define an injective map $f : \mathbb{Z}_+ \to X^\omega$, where X is the two-element set $\{0, 1\}$, without using the choice axiom.

2. Find if possible a choice function for each of the following collections, without using the choice axiom:
   (a) The collection $\mathcal{A}$ of nonempty subsets of $\mathbb{Z}_+$.
   (b) The collection $\mathcal{B}$ of nonempty subsets of $\mathbb{Z}$.
   (c) The collection $\mathcal{C}$ of nonempty subsets of the rational numbers $\mathbb{Q}$.
   (d) The collection $\mathcal{D}$ of nonempty subsets of X^ω, where $X = \{0, 1\}$.

3. Suppose that A is a set and $\{f_n\}_{n\in\mathbb{Z}_+}$ is a given indexed family of injective functions

$$f_n : \{1, \ldots, n\} \longrightarrow A.$$

   Show that A is infinite. Can you define an injective function $f : \mathbb{Z}_+ \to A$ without using the choice axiom?

4. There was a theorem in §7 whose proof involved an infinite number of arbitrary choices. Which one was it? Rewrite the proof so as to make explicit the use of the choice axiom. (Several of the earlier exercises have used the choice axiom also.)

5. (a) Use the choice axiom to show that if $f : A \to B$ is surjective, then f has a right inverse $h : B \to A$.
(b) Show that if $f : A \to B$ is injective and A is not empty, then f has a left inverse. Is the axiom of choice needed?

6. Most of the famous paradoxes of naive set theory are associated in some way or other with the concept of the "set of all sets." None of the rules we have given for forming sets allows us to consider such a set. And for good reason—the concept itself is self-contradictory. For suppose that $\mathcal{A}$ denotes the "set of all sets."
(a) Show that $\mathcal{P}(\mathcal{A}) \subset \mathcal{A}$; derive a contradiction.
(b) (*Russell's paradox.*) Let $\mathcal{B}$ be the subset of $\mathcal{A}$ consisting of all sets that are not elements of themselves;

$$\mathcal{B} = \{A \mid A \in \mathcal{A} \text{ and } A \notin A\}.$$

(Of course, there may be *no* set A such that $A \in A$; if such is the case, then $\mathcal{B} = \mathcal{A}$.) Is $\mathcal{B}$ an element of itself or not?

7. Let A and B be two nonempty sets. If there is an injection of B into A, but no injection of A into B, we say that A has ***greater cardinality*** than B.
(a) Conclude from Theorem 9.1 that every uncountable set has greater cardinality than $\mathbb{Z}_+$.
(b) Show that if A has greater cardinality than B, and B has greater cardinality than C, then A has greater cardinality than C.
(c) Find a sequence $A_1, A_2, \ldots$ of infinite sets, such that for each $n \in \mathbb{Z}_+$, the set A_{n+1} has greater cardinality than A_n.
(d) Find a set that for every n has cardinality greater than A_n.

***8.** Show that $\mathcal{P}(\mathbb{Z}_+)$ and $\mathbb{R}$ have the same cardinality. [*Hint:* You may use the fact that every real number has a decimal expansion, which is unique if expansions that end in an infinite string of 9's are forbidden.]

A famous conjecture of set theory, called the *continuum hypothesis*, asserts that there exists no set having greater cardinality than $\mathbb{Z}_+$ and lesser cardinality than $\mathbb{R}$. The *generalized continuum hypothesis* asserts that, given the infinite set A, there is no set having greater cardinality than A and lesser cardinality than $\mathcal{P}(A)$. Surprisingly enough, both of these assertions have been shown to be independent of the usual axioms for set theory. For a readable expository account, see [Sm].

§10 Well-Ordered Sets

One of the useful properties of the set $\mathbb{Z}_+$ of positive integers is the fact that each of its nonempty subsets has a smallest element. Generalizing this property leads to the concept of a well-ordered set.

Definition. A set A with an order relation $<$ is said to be ***well-ordered*** if every nonempty subset of A has a smallest element.

EXAMPLE 1. Consider the set $\{1, 2\} \times \mathbb{Z}_+$ in the dictionary ordering. Schematically, it can be represented as one infinite sequence followed by another infinite sequence:

$$a_1, a_2, a_3, \ldots; \; b_1, b_2, b_3, \ldots$$

with the understanding that each element is less than every element to the right of it. It is not difficult to see that every nonempty subset C of this ordered set has a smallest element: If C contains any one of the elements a_n, we simply take the smallest element of the intersection of C with the sequence $a_1, a_2, \ldots$; while if C contains no a_n, then it is a subset of the sequence $b_1, b_2, \ldots$ and as such has a smallest element.

EXAMPLE 2. Consider the set $\mathbb{Z}_+ \times \mathbb{Z}_+$ in the dictionary order. Schematically, it can be represented as an infinite sequence of infinite sequences. We show that it is well-ordered. Let X be a nonempty subset of $\mathbb{Z}_+ \times \mathbb{Z}_+$. Let A be the subset of $\mathbb{Z}_+$ consisting of all *first coordinates* of elements of X. Now A has a smallest element; call it a_0. Then the collection

$$\{b \mid a_0 \times b \in X\}$$

is a nonempty subset of $\mathbb{Z}_+$; let b_0 be its smallest element. By definition of the dictionary order, $a_0 \times b_0$ is the smallest element of X. See Figure 10.1.

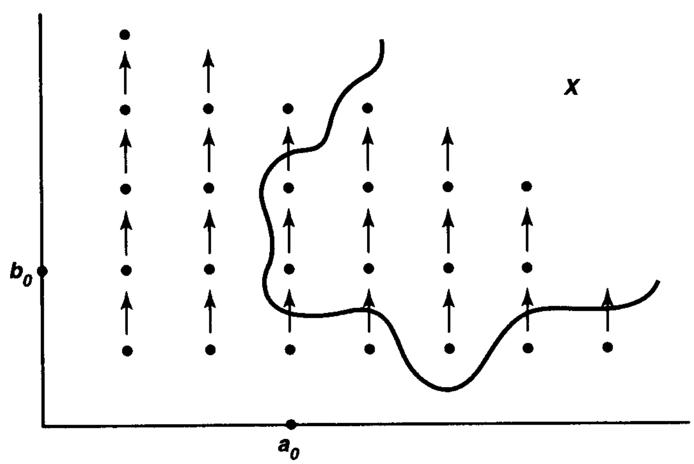


Figure 10.1

EXAMPLE 3. The set of integers is not well-ordered in the usual order; the subset consisting of the negative integers has no smallest element. Nor is the set of real numbers in the interval $0 \leq x \leq 1$ well-ordered; the subset consisting of those x for which $0 < x < 1$ has no smallest element (although it has a greatest lower bound, of course).

There are several ways of constructing well-ordered sets. Two of them are the following:

(1) If A is a well-ordered set, then any subset of A is well-ordered in the restricted order relation.

(2) If A and B are well-ordered sets, then $A \times B$ is well-ordered in the dictionary order.

The proof of (1) is trivial; the proof of (2) follows the pattern given in Example 2.

It follows that the set $\mathbb{Z}_+ \times (\mathbb{Z}_+ \times \mathbb{Z}_+)$ is well-ordered in the dictionary order; it can be represented as an infinite sequence of infinite sequences of infinite sequences. Similarly, $(\mathbb{Z}_+)^4$ is well-ordered in the dictionary order. And so on. But if you try to generalize to an infinite product of $\mathbb{Z}_+$ with itself, you will run into trouble. We shall examine this situation shortly.

Now, given a set A without an order relation, it is natural to ask whether there exists an order relation for A that makes it into a well-ordered set. If A is finite, any bijection

$$f : A \longrightarrow \{1, \dots, n\}$$

can be used to define an order relation on A; under this relation, A has the same order type as the ordered set $\{1, \dots, n\}$. In fact, every order relation on a finite set can be obtained in this way:

Theorem 10.1. *Every nonempty finite ordered set has the order type of a section $\{1, \dots, n\}$ of $\mathbb{Z}_+$, so it is well-ordered.*

Proof. This was given as an exercise in §6; we prove it here. First, we show that every finite ordered set A has a largest element. If A has one element, this is trivial. Supposing it true for sets having $n - 1$ elements, let A have n elements and let $a_0 \in A$. Then $A - \{a_0\}$ has a largest element a_1, and the larger of $\{a_0, a_1\}$ is the largest element of A.

Second, we show there is an order-preserving bijection of A with $\{1, \dots, n\}$ for some n. If A has one element, this fact is trivial. Suppose that it is true for sets having $n - 1$ elements. Let b be the largest element of A. By hypothesis, there is an order-preserving bijection

$$f' : A - \{b\} \longrightarrow \{1, \dots, n - 1\}.$$

Define an order-preserving bijection $f : A \to \{1, \dots, n\}$ by setting

$$\begin{aligned} f(x) &= f'(x) \qquad \text{for } x \neq b, \\ f(b) &= n. \end{aligned}$$

■

Thus, a finite ordered set has only one possible order type. For an infinite set, things are quite different. The well-ordered sets

$$\begin{aligned} &\mathbb{Z}_+, \\ &\{1, \dots, n\} \times \mathbb{Z}_+, \\ &\mathbb{Z}_+ \times \mathbb{Z}_+, \\ &\mathbb{Z}_+ \times (\mathbb{Z}_+ \times \mathbb{Z}_+) \end{aligned}$$

are all countably infinite, but they all have different order types, as you can check.

All the examples we have given of well-ordered sets are orderings of countable sets. It is natural to ask whether one can find a well-ordered uncountable set.

The obvious uncountable set to try is the countably infinite product

$$X = \mathbb{Z}_+ \times \mathbb{Z}_+ \times \cdots = (\mathbb{Z}_+)^\omega$$

of $\mathbb{Z}_+$ with itself. One can generalize the dictionary order to this set in a natural way, by defining

$$(a_1, a_2, \ldots) < (b_1, b_2, \ldots)$$

if for some $n \geq 1$,

$$a_i = b_i, \qquad \text{for } i < n \text{ and } a_n < b_n.$$

This is, in fact, an order relation on the set X; but unfortunately it is not a well-ordering. Consider the set A of all elements $\mathbf{x}$ of X of the form

$$\mathbf{x} = (1, \ldots, 1, 2, 1, 1, \ldots),$$

where exactly one coordinate of $\mathbf{x}$ equals 2, and the others are all equal to 1. The set A clearly has no smallest element.

Thus, the dictionary order at least does not give a well-ordering of the set $(\mathbb{Z}_+)^\omega$. Is there some other order relation on this set that is a well-ordering? No one has ever constructed a specific well-ordering of $(\mathbb{Z}_+)^\omega$. Nevertheless, there is a famous theorem that says such a well-ordering exists:

Theorem (Well-ordering theorem). *If A is a set, there exists an order relation on A that is a well-ordering.*

This theorem was proved by Zermelo in 1904, and it startled the mathematical world. There was considerable debate as to the correctness of the proof; the lack of any constructive procedure for well-ordering an arbitrary uncountable set led many to be skeptical. When the proof was analyzed closely, the only point at which it was found that there might be some question was a construction involving an infinite number of arbitrary choices, that is, a construction involving—the choice axiom.

Some mathematicians rejected the choice axiom as a result, and for many years a legitimate question about a new theorem was: Does its proof involve the choice axiom or not? A theorem was considered to be on somewhat shaky ground if one had to use the choice axiom in its proof. Present-day mathematicians, by and large, do not have such qualms. They accept the axiom of choice as a reasonable assumption about set theory, and they accept the well-ordering theorem along with it.

The proof that the choice axiom implies the well-ordering theorem is rather long (although not exceedingly difficult) and primarily of interest to logicians; we shall omit it. If you are interested, a proof is outlined in the supplementary exercises at the end of the chapter. Instead, we shall simply assume the well-ordering theorem whenever we need it. Consider it to be an additional axiom of set theory if you like!

We shall in fact need the full strength of this assumption only occasionally. Most of the time, all we need is the following weaker result:

Corollary. *There exists an uncountable well-ordered set.*

We now use this result to construct a particular well-ordered set that will prove to be very useful.

Definition. Let X be a well-ordered set. Given $\alpha \in X$, let S_α denote the set

$$S_\alpha = \{x \mid x \in X \text{ and } x < \alpha\}.$$

It is called the ***section*** of X by α.

Lemma 10.2. *There exists a well-ordered set A having a largest element Ω, such that the section S_Ω of A by Ω is uncountable but every other section of A is countable.*

Proof. We begin with an uncountable well-ordered set B. Let C be the well-ordered set $\{1, 2\} \times B$ in the dictionary order; then some section of C is uncountable. (Indeed, the section of C by any element of the form $2 \times b$ is uncountable.) Let Ω be the smallest element of C for which the section of C by Ω is uncountable. Then let A consist of this section along with the element Ω. ∎

Note that S_Ω is an uncountable well-ordered set every section of which is countable. Its order type is in fact uniquely determined by this condition. We shall call it a *minimal uncountable well-ordered set.* Furthermore, we shall denote the well-ordered set $A = S_\Omega \cup \{\Omega\}$ by the symbol $\bar{S}_\Omega$ (for reasons to be seen later).

The most useful property of the set S_Ω for our purposes is expressed in the following theorem:

Theorem 10.3. *If A is a countable subset of S_Ω, then A has an upper bound in S_Ω.*

Proof. Let A be a countable subset of S_Ω. For each $a \in A$, the section S_a is countable. Therefore, the union $B = \bigcup_{a\in A} S_a$ is also countable. Since S_Ω is uncountable, the set B is not all of S_Ω; let x be a point of S_Ω that is not in B. Then x is an upper bound for A. For if $x < a$ for some a in A, then x belongs to S_a and hence to B, contrary to choice. ∎

Exercises

1. Show that every well-ordered set has the least upper bound property.

2. (a) Show that in a well-ordered set, every element except the largest (if one exists) has an immediate successor.
   (b) Find a set in which every element has an immediate successor that is not well-ordered.

3. Both $\{1, 2\} \times \mathbb{Z}_+$ and $\mathbb{Z}_+ \times \{1, 2\}$ are well-ordered in the dictionary order. Do they have the same order type?

4. (a) Let $\mathbb{Z}_-$ denote the set of negative integers in the usual order. Show that a simply ordered set A fails to be well-ordered if and only if it contains a subset having the same order type as $\mathbb{Z}_-$.
   (b) Show that if A is simply ordered and every countable subset of A is well-ordered, then A is well-ordered.

5. Show the well-ordering theorem implies the choice axiom.

6. Let S_Ω be the minimal uncountable well-ordered set.
   (a) Show that S_Ω has no largest element.
   (b) Show that for every $\alpha \in S_\Omega$, the subset $\{x \mid \alpha < x\}$ is uncountable.
   (c) Let X_0 be the subset of S_Ω consisting of all elements x such that x has no immediate predecessor. Show that X_0 is uncountable.

7. Let J be a well-ordered set. A subset J_0 of J is said to be ***inductive*** if for every $\alpha \in J$,

$$(S_\alpha \subset J_0) \Longrightarrow \alpha \in J_0.$$

Theorem (The principle of transfinite induction). If J is a well-ordered set and J_0 is an inductive subset of J, then $J_0 = J$.

8. (a) Let A_1 and A_2 be disjoint sets, well-ordered by $<_1$ and $<_2$, respectively. Define an order relation on $A_1 \cup A_2$ by letting $a < b$ either if $a, b \in A_1$ and $a <_1 b$, or if $a, b \in A_2$ and $a <_2 b$, or if $a \in A_1$ and $b \in A_2$. Show that this is a well-ordering.
   (b) Generalize (a) to an arbitrary family of disjoint well-ordered sets, indexed by a well-ordered set.

9. Consider the subset A of $(\mathbb{Z}_+)^\omega$ consisting of all infinite sequences of positive integers $\mathbf{x} = (x_1, x_2, \dots)$ that end in an infinite string of 1's. Give A the following order: $\mathbf{x} < \mathbf{y}$ if $x_n < y_n$ and $x_i = y_i$ for $i > n$. We call this the "antidictionary order" on A.
   (a) Show that for every n, there is a section of A that has the same order type as $(\mathbb{Z}_+)^n$ in the dictionary order.
   (b) Show A is well-ordered.

10. *Theorem. Let J and C be well-ordered sets; assume that there is no surjective function mapping a section of J onto C. Then there exists a unique function $h : J \to C$ satisfying the equation*

$$(*) \qquad h(x) = \textit{smallest}\,[C - h(S_x)]$$

for each $x \in J$, where S_x is the section of J by x.

Proof.

(a) If h and k map sections of J, or all of J, into C and satisfy $(*)$ for all x in their respective domains, show that $h(x) = k(x)$ for all x in both domains.

(b) If there exists a function $h : S_\alpha \to C$ satisfying $(*)$, show that there exists a function $k : S_\alpha \cup \{\alpha\} \to C$ satisfying $(*)$.

(c) If $K \subset J$ and for all $\alpha \in K$ there exists a function $h_\alpha : S_\alpha \to C$ satisfying $(*)$, show that there exists a function

$$k : \bigcup_{\alpha \in K} S_\alpha \longrightarrow C$$

satisfying $(*)$.

(d) Show by transfinite induction that for every $\beta \in J$, there exists a function $h_\beta : S_\beta \to C$ satisfying $(*)$. [*Hint:* If β has an immediate predecessor α, then $S_\beta = S_\alpha \cup \{\alpha\}$. If not, S_β is the union of all S_α with $\alpha < \beta$.]

(e) Prove the theorem.

11. Let A and B be two sets. Using the well-ordering theorem, prove that either they have the same cardinality, or one has cardinality greater than the other. [*Hint:* If there is no surjection $f : A \to B$, apply the preceding exercise.]

*§11 The Maximum Principle†

We have already indicated that the axiom of choice leads to the deep theorem that every set can be well-ordered. The axiom of choice has other consequences that are even more important in mathematics. Collectively referred to as "maximum principles," they come in many versions. Formulated independently by a number of mathematicians, including F. Hausdorff, K. Kuratowski, S. Bochner, and M. Zorn, during the years 1914–1935, they were typically proved as consequences of the well-ordering theorem. Later, it was realized that they were in fact *equivalent* to the well-ordering theorem. We consider several of them here.

First, we make a definition. Given a set A, a relation $\prec$ on A is called a ***strict partial order*** on A if it has the following two properties:

(1) (Nonreflexivity) The relation $a \prec a$ never holds.

(2) (Transitivity) If $a \prec b$ and $b \prec c$, then $a \prec c$.

These are just the second and third of the properties of a simple order (see §3); the comparability property is the one that is omitted. In other words, a strict partial order behaves just like a simple order except that it need not be true that for every pair of distinct points x and y in the set, either $x \prec y$ or $y \prec x$.

If $\prec$ is a strict partial order on a set A, it can easily happen that some subset B of A is simply ordered by the relation; all that is needed is for every pair of elements of B to be comparable under $\prec$.

†This section will be assumed in Chapters 5 and 14.

Now we can state the following principle, which was first formulated by Hausdorff in 1914.

Theorem (The maximum principle). *Let A be a set; let $\prec$ be a strict partial order on A. Then there exists a maximal simply ordered subset B of A.*

Said differently, there exists a subset B of A such that B is simply ordered by $\prec$ and such that no subset of A that properly contains B is simply ordered by $\prec$.

EXAMPLE 1. If $\mathcal{A}$ is any collection of sets, the relation "is a proper subset of" is a strict partial order on $\mathcal{A}$. Suppose that $\mathcal{A}$ is the collection of all circular regions (interiors of circles) in the plane. One maximal simply ordered subcollection of $\mathcal{A}$ consists of all circular regions with centers at the origin. Another maximal simply ordered subcollection consists of all circular regions bounded by circles tangent from the right to the y-axis at the origin. See Figure 11.1.

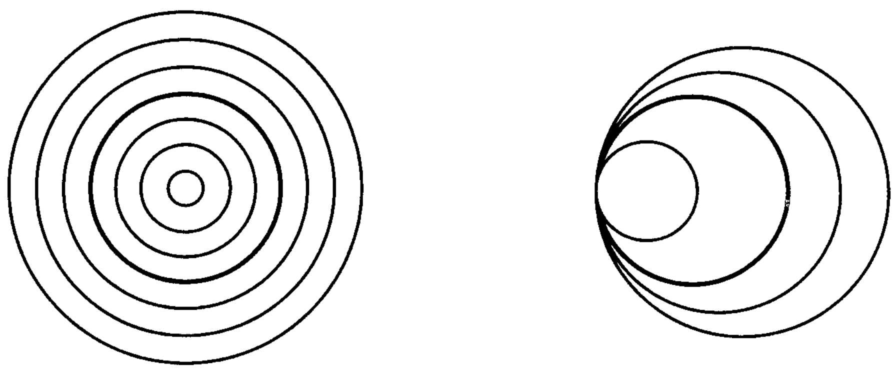


Figure 11.1

EXAMPLE 2. If (x_0, y_0) and (x_1, y_1) are two points of the plane $\mathbb{R}^2$, define

$$(x_0, y_0) \prec (x_1, y_1)$$

if $y_0 = y_1$ and $x_0 < x_1$. This is a partial ordering of $\mathbb{R}^2$ under which two points are comparable only if they lie on the same horizontal line. The maximal simply ordered sets are the horizontal lines in $\mathbb{R}^2$.

One can give an intuitive "proof" of the maximum principle that is rather appealing. It involves a step-by-step procedure, which one can describe in physical terms as follows. Suppose we take a box, and put into it some of the elements of A according to the following plan: First we pick an arbitrary element of A and put it in the box. Then we pick another element of A. If it is comparable with the element in the box, we put it in the box too; otherwise, we throw it away. At the general step, we will have a collection of elements in the box and a collection of elements that have been tossed away. Take one of the remaining elements of A. If it is comparable with everything in the box, toss it in the box, too; otherwise, throw it away. Similarly continue. After you have checked all the elements of A, the elements you have in the box will be comparable with one another, and thus they will form a simply ordered set. Every element not in the box will be noncomparable with at least one element in the box, for that was why it was tossed away. Hence, the simply ordered set in the box is maximal, for no larger subset of A can satisfy the comparability condition.

Now of course the weak point in the preceding "proof" comes when we said, "After you have checked all the elements of A." How do you know you ever "get through" checking all the elements of A? If A should happen to be countable, it is not hard to make this intuitive proof into a real proof. Let us take the countably infinite case; the finite case is even easier. Index the elements of A bijectively with the positive integers, so that $A = \{a_1, a_2 \dots\}$. This indexing gives a way of deciding what order to test the elements of A in, and how to know when one has tested them all.

Now we define a function $h : \mathbb{Z}_+ \to \{0, 1\}$, by letting it assign the value 0 to i if we "put a_i in the box," and the value 1 if we "throw a_i away." This means that $h(1) = 0$, and for $i > 1$, we have $h(i) = 0$ if and only if a_i is comparable with every element of the set

$$\{a_j \mid j < i \text{ and } h(j) = 0\}.$$

By the principle of recursive definition, this formula determines a unique function $h : \mathbb{Z}_+ \to \{0, 1\}$. It is easy to check that the set of those a_j for which $h(j) = 0$ is a maximal simply ordered subset of A.

If A is not countable, a variant of this procedure will work, *if we allow ourselves to use the well-ordering theorem.* Instead of indexing the elements of A with the set $\mathbb{Z}_+$, we index them (in a bijective fashion) with the elements of some well-ordered set J, so that $A = \{a_\alpha \mid \alpha \in J\}$. For this we need the well-ordering theorem, so that we know there is a bijection between A and some well-ordered set J. Then we can proceed as in the previous paragraph, letting α replace i in the argument. Strictly speaking, you need to generalize the principle of recursive definition to well-ordered sets as well, but that is not particularly difficult. (See the Supplementary Exercises.)

Thus, the well-ordering theorem implies the maximum principle.

Although the maximum principle of Hausdorff was the first to be formulated and is probably the simplest to understand, there is another such principle that is nowadays the one most frequently quoted. It is popularly called "Zorn's Lemma," although Kuratowski (1922) and Bochner (1922) preceded Zorn (1935) in enunciating and proving versions of it. For a history and discussion of the tangled history of these ideas, see [C] or [Mo]. To state this principle, we need some terminology.

Definition. Let A be a set and let $\prec$ be a strict partial order on A. If B is a subset of A, an ***upper bound*** on B is an element c of A such that for every b in B, either $b = c$ or $b \prec c$. A ***maximal element*** of A is an element m of A such that for no element a of A does the relation $m \prec a$ hold.

Zorn's Lemma. *Let A be a set that is strictly partially ordered. If every simply ordered subset of A has an upper bound in A, then A has a maximal element.*

Zorn's lemma is an easy consequence of the maximum principle: Given A, the maximum principle implies that A has a maximal simply ordered subset B. The hypothesis of Zorn's lemma tells us that B has an upper bound c in A. The element c is then automatically a maximal element of A. For if $c \prec d$ for some element d of A, then the set $B \cup \{d\}$, which properly contains B, is simply ordered because $b \prec d$ for every $b \in B$. This fact contradicts maximality of B.

It is also true that the maximum principle is an easy consequence of Zorn's lemma. See Exercises 5–7.

One final remark. We have defined what we mean by a strict partial order on a set, but we have not said what a partial order itself is. Let $\prec$ be a strict partial order on a set A. Suppose that we define $a \preceq b$ if either $a \prec b$ or $a = b$. Then the relation $\preceq$ is called a ***partial order*** on A. For example, the inclusion relation $\subset$ on a collection of sets is a partial order, whereas proper inclusion is a strict partial order.

Many authors prefer to deal with partial orderings rather than strict partial orderings; the maximum principle and Zorn's lemma are often expressed in these terms. Which formulation is used is simply a matter of taste and convenience.

Exercises

1. If a and b are real numbers, define $a \prec b$ if $b - a$ is positive and rational. Show this is a strict partial order on $\mathbb{R}$. What are the maximal simply ordered subsets?

2. (a) Let $\prec$ be a strict partial order on the set A. Define a relation on A by letting $a \preceq b$ if either $a \prec b$ or $a = b$. Show that this relation has the following properties, which are called the ***partial order axioms***:

(i) $a \preceq a$ for all $a \in A$.

(ii) $a \preceq b$ and $b \preceq a \Longrightarrow a = b$.

(iii) $a \preceq b$ and $b \preceq c \Longrightarrow a \preceq c$.

(b) Let P be a relation on A that satisfies properties (i)–(iii). Define a relation S on A by letting aSb if aPb and $a \neq b$. Show that S is a strict partial order on A.

3. Let A be a set with a strict partial order $\prec$; let $x \in A$. Suppose that we wish to find a maximal simply ordered subset B of A that contains x. One plausible way of attempting to define B is to let B equal the set of all those elements of A that are *comparable* with x;

$$B = \{y \mid y \in A \text{ and either } x \prec y \text{ or } y \prec x\}.$$

But this will not always work. In which of Examples 1 and 2 will this procedure succeed and in which will it not?

4. Given two points (x_0, y_0) and (x_1, y_1) of $\mathbb{R}^2$, define

$$(x_0, y_0) \prec (x_1, y_1)$$

if $x_0 < x_1$ and $y_0 \leq y_1$. Show that the curves $y = x^3$ and $y = 2$ are maximal simply ordered subsets of $\mathbb{R}^2$, and the curve $y = x^2$ is not. Find all maximal simply ordered subsets.

5. Show that Zorn's lemma implies the following:
Lemma (Kuratowski). Let $\mathcal{A}$ be a collection of sets. Suppose that for every subcollection $\mathcal{B}$ of $\mathcal{A}$ that is simply ordered by proper inclusion, the union of the elements of $\mathcal{B}$ belongs to $\mathcal{A}$. Then $\mathcal{A}$ has an element that is properly contained in no other element of $\mathcal{A}$.

6. A collection $\mathcal{A}$ of subsets of a set X is said to be of *finite type* provided that a subset B of X belongs to $\mathcal{A}$ if and only if every finite subset of B belongs to $\mathcal{A}$. Show that the Kuratowski lemma implies the following:
Lemma (Tukey, 1940). Let $\mathcal{A}$ be a collection of sets. If $\mathcal{A}$ is of finite type, then $\mathcal{A}$ has an element that is properly contained in no other element of $\mathcal{A}$.

7. Show that the Tukey lemma implies the Hausdorff maximum principle. [*Hint:* If $\prec$ is a strict partial order on A, let $\mathcal{A}$ be the collection of all subsets of A that are simply ordered by $\prec$. Show that $\mathcal{A}$ is of finite type.]

8. A typical use of Zorn's lemma in algebra is the proof that every vector space has a basis. Recall that if A is a subset of the vector space V, we say a vector belongs to the *span* of A if it equals a finite linear combination of elements of A. The set A is *independent* if the only finite linear combination of elements of A that equals the zero vector is the trivial one having all coefficients zero. If A is independent and if every vector in V belongs to the span of A, then A is a *basis* for V.
(a) If A is independent and $v \in V$ does not belong to the span of A, show $A \cup \{v\}$ is independent.
(b) Show the collection of all independent sets in V has a maximal element.
(c) Show that V has a basis.

*Supplementary Exercises: Well-Ordering

In the following exercises, we ask you to prove the equivalence of the choice axiom, the well-ordering theorem, and the maximum principle. We comment that of these exercises, only Exercise 7 uses the choice axiom.

1. *Theorem (General principle of recursive definition). Let J be a well-ordered set; let C be a set. Let $\mathcal{F}$ be the set of all functions mapping sections of J into C. Given a function $\rho : \mathcal{F} \to C$, there exists a unique function $h : J \to C$ such that $h(\alpha) = \rho(h|S_\alpha)$ for each $\alpha \in J$.*
[*Hint:* Follow the pattern outlined in Exercise 10 of §10.]

2. (a) Let J and E be well-ordered sets; let $h : J \to E$. Show the following two statements are equivalent:
(i) h is order preserving and its image is E or a section of E.

(ii) $h(\alpha) =$ smallest $[E - h(S_\alpha)]$ for all α.

[*Hint:* Show that each of these conditions implies that $h(S_\alpha)$ is a section of E; conclude that it must be the section by $h(\alpha)$.]

(b) If E is a well-ordered set, show that no section of E has the order type of E, nor do two different sections of E have the same order type. [*Hint:* Given J, there is at most one order-preserving map of J into E whose image is E or a section of E.]

3. Let J and E be well-ordered sets; suppose there is an order-preserving map $k : J \to E$. Using Exercises 1 and 2, show that J has the order type of E or a section of E. [*Hint:* Choose $e_0 \in E$. Define $h : J \to E$ by the recursion formula

$$h(\alpha) = \text{smallest}\,[E - h(S_\alpha)] \quad \text{if} \quad h(S_\alpha) \neq E,$$

and $h(\alpha) = e_0$ otherwise. Show that $h(\alpha) \leq k(\alpha)$ for all α; conclude that $h(S_\alpha) \neq E$ for all α.]

4. Use Exercises 1–3 to prove the following:

(a) If A and B are well-ordered sets, then exactly one of the following three conditions holds: A and B have the same order type, or A has the order type of a section of B, or B has the order type of a section of A. [*Hint:* Form a well-ordered set containing both A and B, as in Exercise 8 of §10; then apply the preceding exercise.]

(b) Suppose that A and B are well-ordered sets that are uncountable, such that every section of A and of B is countable. Show A and B have the same order type.

5. Let X be a set; let $\mathcal{A}$ be the collection of all pairs $(A, <)$, where A is a subset of X and $<$ is a well-ordering of A. Define

$$(A, <) \prec (A', <')$$

if $(A, <)$ *equals* a section of $(A', <')$.

(a) Show that $\prec$ is a strict partial order on $\mathcal{A}$.

(b) Let $\mathcal{B}$ be a subcollection of $\mathcal{A}$ that is simply ordered by $\prec$. Define B' to be the union of the sets B, for all $(B, <) \in \mathcal{B}$; and define $<'$ to be the union of the relations $<$, for all $(B, <) \in \mathcal{B}$. Show that $(B', <')$ is a well-ordered set.

6. Use Exercises 1 and 5 to prove the following:

Theorem. *The maximum principle is equivalent to the well-ordering theorem.*

7. Use Exercises 1–5 to prove the following:

Theorem. *The choice axiom is equivalent to the well-ordering theorem.*

Proof. Let X be a set; let c be a fixed choice function for the nonempty subsets of X. If T is a subset of X and $<$ is a relation on T, we say that $(T, <)$ is a ***tower*** in X if $<$ is a well-ordering of T and if for each $x \in T$,

$$x = c(X - S_x(T)),$$

where $S_x(T)$ is the section of T by x.

(a) Let $(T_1, <_1)$ and $(T_2, <_2)$ be two towers in X. Show that either these two ordered sets are the same, or one equals a section of the other. [*Hint:* Switching indices if necessary, we can assume that $h : T_1 \to T_2$ is order preserving and $h(T_1)$ equals either T_2 or a section of T_2. Use Exercise 2 to show that $h(x) = x$ for all x.]

(b) If $(T, <)$ is a tower in X and $T \neq X$, show there is a tower in X of which $(T, <)$ is a section.

(c) Let $\{(T_k, <_k) | k \in K\}$ be the collection of all towers in X. Let

$$T = \bigcup_{k \in K} T_k \quad \text{and} \quad < = \bigcup_{k \in K} (<_k).$$

Show that $(T, <)$ is a tower in X. Conclude that $T = X$.

8. Using Exercises 1–4, construct an uncountable well-ordered set, as follows. Let $\mathcal{A}$ be the collection of all pairs $(A, <)$, where A is a subset of $\mathbb{Z}_+$ and $<$ is a well-ordering of A. (We allow A to be empty.) Define $(A, <) \sim (A', <')$ if $(A, <)$ and $(A', <')$ have the same order type. It is trivial to show this is an equivalence relation. Let $[(A, <)]$ denote the equivalence class of $(A, <)$; let E denote the collection of these equivalence classes. Define

$$[(A, <)] \ll [(A', <')]$$

if $(A, <)$ has the order type of a *section* of $(A', <')$.

(a) Show that the relation $\ll$ is well defined and is a simple order on E. Note that the equivalence class $[(\varnothing, \varnothing)]$ is the smallest element of E.

(b) Show that if $\alpha = [(A, <)]$ is an element of E, then $(A, <)$ has the same order type as the section $S_\alpha(E)$ of E by α. [*Hint:* Define a map $f : A \to E$ by setting $f(x) = [(S_x(A), \text{restriction of } <)]$ for each $x \in A$.]

(c) Conclude that E is well-ordered by $\ll$.

(d) Show that E is uncountable. [*Hint:* If $h : E \to \mathbb{Z}_+$ is a bijection, then h gives rise to a well-ordering of $\mathbb{Z}_+$.]

This same argument, with $\mathbb{Z}_+$ replaced by an arbitrary well-ordered set X, proves (without use of the choice axiom) the existence of a well-ordered set E whose cardinality is greater than that of X.

This exercise shows that one can construct an uncountable well-ordered set, and hence the minimal uncountable well-ordered set, by an explicit construction that does not use the choice axiom. However, this result is less interesting than it might appear. The crucial property of S_Ω, the one we use repeatedly, is the fact that every countable subset of S_Ω has an upper bound in S_Ω. That fact depends, in turn, on the fact that a countable union of countable sets is countable. And the proof of *that* result (if you examine it carefully) involves an infinite number of arbitrary choices—that is, it depends on the choice axiom.

Said differently, without the choice axiom we may be able to construct the minimal uncountable well-ordered set, but we can't use it for anything!

Chapter 2

Topological Spaces and Continuous Functions

The concept of topological space grew out of the study of the real line and euclidean space and the study of continuous functions on these spaces. In this chapter, we define what a topological space is, and we study a number of ways of constructing a topology on a set so as to make it into a topological space. We also consider some of the elementary concepts associated with topological spaces. Open and closed sets, limit points, and continuous functions are introduced as natural generalizations of the corresponding ideas for the real line and euclidean space.

§12 Topological Spaces

The definition of a topological space that is now standard was a long time in being formulated. Various mathematicians—Fréchet, Hausdorff, and others—proposed different definitions over a period of years during the first decades of the twentieth century, but it took quite a while before mathematicians settled on the one that seemed most suitable. They wanted, of course, a definition that was as broad as possible, so that it would include as special cases all the various examples that were useful in mathematics—euclidean space, infinite-dimensional euclidean space, and function spaces among them—but they also wanted the definition to be narrow enough that the standard theorems about these familiar spaces would hold for topological spaces in general. This is always the problem when one is trying to formulate a new mathematical concept, to decide how general its definition should be. The definition finally settled on may seem a bit abstract, but as you work through the various ways of constructing topological spaces, you will get a better feeling for what the concept means.

Definition. A ***topology*** on a set X is a collection $\mathcal{T}$ of subsets of X having the following properties:

(1) $\varnothing$ and X are in $\mathcal{T}$.

(2) The union of the elements of any subcollection of $\mathcal{T}$ is in $\mathcal{T}$.

(3) The intersection of the elements of any finite subcollection of $\mathcal{T}$ is in $\mathcal{T}$.

A set X for which a topology $\mathcal{T}$ has been specified is called a ***topological space***.

Properly speaking, a topological space is an ordered pair $(X, \mathcal{T})$ consisting of a set X and a topology $\mathcal{T}$ on X, but we often omit specific mention of $\mathcal{T}$ if no confusion will arise.

If X is a topological space with topology $\mathcal{T}$, we say that a subset U of X is an ***open set*** of X if U belongs to the collection $\mathcal{T}$. Using this terminology, one can say that a topological space is a set X together with a collection of subsets of X, called *open sets*, such that $\varnothing$ and X are both open, and such that arbitrary unions and finite intersections of open sets are open.

EXAMPLE 1. Let X be a three-element set, $X = \{a, b, c\}$. There are many possible topologies on X, some of which are indicated schematically in Figure 12.1. The diagram in the upper right-hand corner indicates the topology in which the open sets are X, $\varnothing$, $\{a, b\}$, $\{b\}$, and $\{b, c\}$. The topology in the upper left-hand corner contains only X and $\varnothing$, while the topology in the lower right-hand corner contains every subset of X. You can get other topologies on X by permuting a, b, and c.

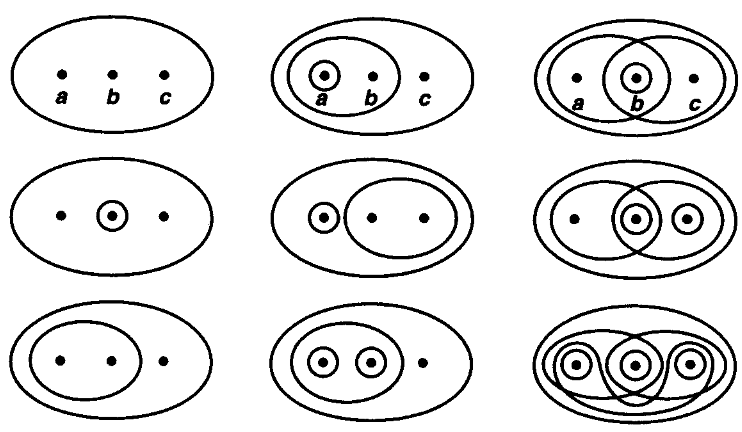


Figure 12.1

From this example, you can see that even a three-element set has many different topologies. But not every collection of subsets of X is a topology on X. Neither of the collections indicated in Figure 12.2 is a topology, for instance.

Figure 12.2

EXAMPLE 2. If X is any set, the collection of *all* subsets of X is a topology on X; it is called the ***discrete topology***. The collection consisting of X and $\varnothing$ only is also a topology on X; we shall call it the ***indiscrete topology***, or the ***trivial topology***.

EXAMPLE 3. Let X be a set; let $\mathcal{T}_f$ be the collection of all subsets U of X such that $X - U$ either is finite or is all of X. Then $\mathcal{T}_f$ is a topology on X, called the ***finite complement topology***. Both X and $\varnothing$ are in $\mathcal{T}_f$, since $X - X$ is finite and $X - \varnothing$ is all of X. If $\{U_\alpha\}$ is an indexed family of nonempty elements of $\mathcal{T}_f$, to show that $\bigcup U_\alpha$ is in $\mathcal{T}_f$, we compute

$$X - \bigcup U_\alpha = \bigcap (X - U_\alpha).$$

The latter set is finite because each set $X - U_\alpha$ is finite. If $U_1, \ldots, U_n$ are nonempty elements of $\mathcal{T}_f$, to show that $\bigcap U_i$ is in $\mathcal{T}_f$, we compute

$$X - \bigcap_{i=1}^{n} U_i = \bigcup_{i=1}^{n} (X - U_i).$$

The latter set is a finite union of finite sets and, therefore, finite.

EXAMPLE 4. Let X be a set; let $\mathcal{T}_c$ be the collection of all subsets U of X such that $X - U$ either is countable or is all of X. Then $\mathcal{T}_c$ is a topology on X, as you can check.

Definition. Suppose that $\mathcal{T}$ and $\mathcal{T}'$ are two topologies on a given set X. If $\mathcal{T}' \supset \mathcal{T}$, we say that $\mathcal{T}'$ is ***finer*** than $\mathcal{T}$; if $\mathcal{T}'$ *properly* contains $\mathcal{T}$, we say that $\mathcal{T}'$ is ***strictly finer*** than $\mathcal{T}$. We also say that $\mathcal{T}$ is ***coarser*** than $\mathcal{T}'$, or ***strictly coarser***, in these two respective situations. We say $\mathcal{T}$ is ***comparable*** with $\mathcal{T}'$ if either $\mathcal{T}' \supset \mathcal{T}$ or $\mathcal{T} \supset \mathcal{T}'$.

This terminology is suggested by thinking of a topological space as being something like a truckload full of gravel—the pebbles and all unions of collections of pebbles being the open sets. If now we smash the pebbles into smaller ones, the collection of open sets has been enlarged, and the topology, like the gravel, is said to have been made finer by the operation.

Two topologies on X need not be comparable, of course. In Figure 12.1 preceding, the topology in the upper right-hand corner is strictly finer than each of the three topologies in the first column and strictly coarser than each of the other topologies in the third column. But it is not comparable with any of the topologies in the second column.

Other terminology is sometimes used for this concept. If $\mathcal{T}' \supset \mathcal{T}$, some mathematicians would say that $\mathcal{T}'$ is ***larger*** than $\mathcal{T}$, and $\mathcal{T}$ is ***smaller*** than $\mathcal{T}'$. This is certainly acceptable terminology, if not as vivid as the words "finer" and "coarser."

Many mathematicians use the words "weaker" and "stronger" in this context. Unfortunately, some of them (particularly analysts) are apt to say that $\mathcal{T}'$ is stronger than $\mathcal{T}$ if $\mathcal{T}' \supset \mathcal{T}$, while others (particularly topologists) are apt to say that $\mathcal{T}'$ is weaker than $\mathcal{T}$ in the same situation! If you run across the terms "strong topology" or "weak topology" in some book, you will have to decide from the context which inclusion is meant. We shall not use these terms in this book.

§13 Basis for a Topology

For each of the examples in the preceding section, we were able to specify the topology by describing the entire collection $\mathcal{T}$ of open sets. Usually this is too difficult. In most cases, one specifies instead a smaller collection of subsets of X and defines the topology in terms of that.

Definition. If X is a set, a ***basis*** for a topology on X is a collection $\mathcal{B}$ of subsets of X (called ***basis elements***) such that

(1) For each $x \in X$, there is at least one basis element B containing x.

(2) If x belongs to the intersection of two basis elements B_1 and B_2, then there is a basis element B_3 containing x such that $B_3 \subset B_1 \cap B_2$.

If $\mathcal{B}$ satisfies these two conditions, then we define the ***topology $\mathcal{T}$ generated by $\mathcal{B}$*** as follows: A subset U of X is said to be open in X (that is, to be an element of $\mathcal{T}$) if for each $x \in U$, there is a basis element $B \in \mathcal{B}$ such that $x \in B$ and $B \subset U$. Note that each basis element is itself an element of $\mathcal{T}$.

We will check shortly that the collection $\mathcal{T}$ is indeed a topology on X. But first let us consider some examples.

EXAMPLE 1. Let $\mathcal{B}$ be the collection of all circular regions (interiors of circles) in the plane. Then $\mathcal{B}$ satisfies both conditions for a basis. The second condition is illustrated in Figure 13.1. In the topology generated by $\mathcal{B}$, a subset U of the plane is open if every x in U lies in some circular region contained in U.

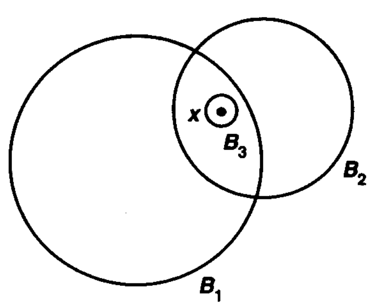


Figure 13.1

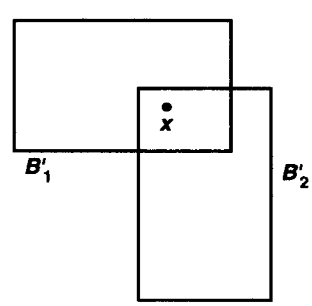


Figure 13.2

EXAMPLE 2. Let $\mathcal{B}'$ be the collection of all rectangular regions (interiors of rectangles) in the plane, where the rectangles have sides parallel to the coordinate axes. Then $\mathcal{B}'$ satisfies both conditions for a basis. The second condition is illustrated in Figure 13.2; in this case, the condition is trivial, because the intersection of any two basis elements is itself a basis element (or empty). As we shall see later, the basis $\mathcal{B}'$ generates the same topology on the plane as the basis $\mathcal{B}$ given in the preceding example.

EXAMPLE 3. If X is any set, the collection of all one-point subsets of X is a basis for the discrete topology on X.

Let us check now that the collection $\mathcal{T}$ generated by the basis $\mathcal{B}$ is, in fact, a topology on X. If U is the empty set, it satisfies the defining condition of openness vacuously. Likewise, X is in $\mathcal{T}$, since for each $x \in X$ there is some basis element B containing x and contained in X. Now let us take an indexed family $\{U_\alpha\}_{\alpha \in J}$, of elements of $\mathcal{T}$ and show that

$$U = \bigcup_{\alpha \in J} U_\alpha$$

belongs to $\mathcal{T}$. Given $x \in U$, there is an index α such that $x \in U_\alpha$. Since U_α is open, there is a basis element B such that $x \in B \subset U_\alpha$. Then $x \in B$ and $B \subset U$, so that U is open, by definition.

Now let us take *two* elements U_1 and U_2 of $\mathcal{T}$ and show that $U_1 \cap U_2$ belongs to $\mathcal{T}$. Given $x \in U_1 \cap U_2$, choose a basis element B_1 containing x such that $B_1 \subset U_1$; choose also a basis element B_2 containing x such that $B_2 \subset U_2$. The second condition for a basis enables us to choose a basis element B_3 containing x such that $B_3 \subset B_1 \cap B_2$. See Figure 13.3. Then $x \in B_3$ and $B_3 \subset U_1 \cap U_2$, so $U_1 \cap U_2$ belongs to $\mathcal{T}$, by definition.

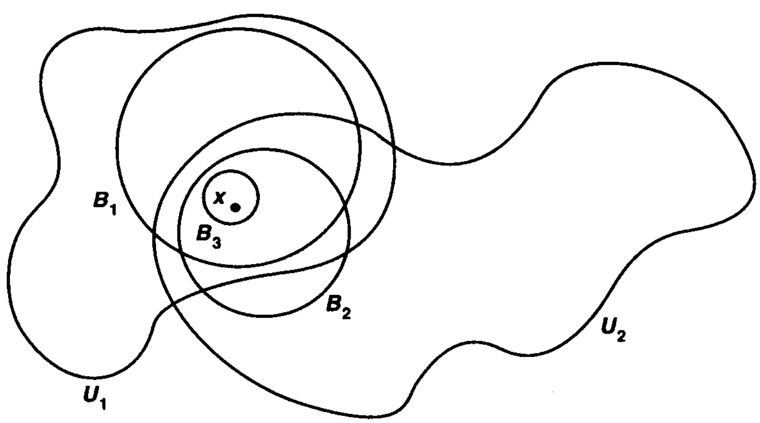


Figure 13.3

Finally, we show by induction that any finite intersection $U_1 \cap \cdots \cap U_n$ of elements of $\mathcal{T}$ is in $\mathcal{T}$. This fact is trivial for $n = 1$; we suppose it true for $n - 1$ and prove it for n. Now

$$(U_1 \cap \cdots \cap U_n) = (U_1 \cap \cdots \cap U_{n-1}) \cap U_n.$$

By hypothesis, $U_1 \cap \cdots \cap U_{n-1}$ belongs to $\mathcal{T}$; by the result just proved, the intersection of $U_1 \cap \cdots \cap U_{n-1}$ and U_n also belongs to $\mathcal{T}$.

Thus we have checked that collection of open sets generated by a basis $\mathcal{B}$ is, in fact, a topology.

Another way of describing the topology generated by a basis is given in the following lemma:

Lemma 13.1. *Let X be a set; let $\mathcal{B}$ be a basis for a topology $\mathcal{T}$ on X. Then $\mathcal{T}$ equals the collection of all unions of elements of $\mathcal{B}$.*

Proof. Given a collection of elements of $\mathcal{B}$, they are also elements of $\mathcal{T}$. Because $\mathcal{T}$ is a topology, their union is in $\mathcal{T}$. Conversely, given $U \in \mathcal{T}$, choose for each $x \in U$ an element B_x of $\mathcal{B}$ such that $x \in B_x \subset U$. Then $U = \bigcup_{x \in U} B_x$, so U equals a union of elements of $\mathcal{B}$. ∎

This lemma states that every open set U in X can be expressed as a union of basis elements. This expression for U is not, however, unique. Thus the use of the term "basis" in topology differs drastically from its use in linear algebra, where the equation expressing a given vector as a linear combination of basis vectors *is* unique.

We have described in two different ways how to go from a basis to the topology it generates. Sometimes we need to go in the reverse direction, from a topology to a basis generating it. Here is one way of obtaining a basis for a given topology; we shall use it frequently.

Lemma 13.2. *Let X be a topological space. Suppose that $\mathcal{C}$ is a collection of open sets of X such that for each open set U of X and each x in U, there is an element C of $\mathcal{C}$ such that $x \in C \subset U$. Then $\mathcal{C}$ is a basis for the topology of X.*

Proof. We must show that $\mathcal{C}$ is a basis. The first condition for a basis is easy: Given $x \in X$, since X is itself an open set, there is by hypothesis an element C of $\mathcal{C}$ such that $x \in C \subset X$. To check the second condition, let x belong to $C_1 \cap C_2$, where C_1 and C_2 are elements of $\mathcal{C}$. Since C_1 and C_2 are open, so is $C_1 \cap C_2$. Therefore, there exists by hypothesis an element C_3 in $\mathcal{C}$ such that $x \in C_3 \subset C_1 \cap C_2$.

Let $\mathcal{T}$ be the collection of open sets of X; we must show that the topology $\mathcal{T}'$ generated by $\mathcal{C}$ equals the topology $\mathcal{T}$. First, note that if U belongs to $\mathcal{T}$ and if $x \in U$, then there is by hypothesis an element C of $\mathcal{C}$ such that $x \in C \subset U$. It follows that U belongs to the topology $\mathcal{T}'$, by definition. Conversely, if W belongs to the topology $\mathcal{T}'$, then W equals a union of elements of $\mathcal{C}$, by the preceding lemma. Since each element of $\mathcal{C}$ belongs to $\mathcal{T}$ and $\mathcal{T}$ is a topology, W also belongs to $\mathcal{T}$. ∎

When topologies are given by bases, it is useful to have a criterion in terms of the bases for determining whether one topology is finer than another. One such criterion is the following:

Lemma 13.3. *Let $\mathcal{B}$ and $\mathcal{B}'$ be bases for the topologies $\mathcal{T}$ and $\mathcal{T}'$, respectively, on X. Then the following are equivalent:*

(1) *$\mathcal{T}'$ is finer than $\mathcal{T}$.*

(2) *For each $x \in X$ and each basis element $B \in \mathcal{B}$ containing x, there is a basis element $B' \in \mathcal{B}'$ such that $x \in B' \subset B$.*

Proof. (2) $\Rightarrow$ (1). Given an element U of $\mathcal{T}$, we wish to show that $U \in \mathcal{T}'$. Let $x \in U$. Since $\mathcal{B}$ generates $\mathcal{T}$, there is an element $B \in \mathcal{B}$ such that $x \in B \subset U$. Condition (2) tells us there exists an element $B' \in \mathcal{B}'$ such that $x \in B' \subset B$. Then $x \in B' \subset U$, so $U \in \mathcal{T}'$, by definition.

(1) $\Rightarrow$ (2). We are given $x \in X$ and $B \in \mathcal{B}$, with $x \in B$. Now B belongs to $\mathcal{T}$ by definition and $\mathcal{T} \subset \mathcal{T}'$ by condition (1); therefore, $B \in \mathcal{T}'$. Since $\mathcal{T}'$ is generated by $\mathcal{B}'$, there is an element $B' \in \mathcal{B}'$ such that $x \in B' \subset B$. ∎

Some students find this condition hard to remember. "Which way does the inclusion *go*?" they ask. It may be easier to remember if you recall the analogy between a topological space and a truckload full of gravel. Think of the pebbles as the basis elements of the topology; after the pebbles are smashed to dust, the dust particles are the basis elements of the new topology. The new topology is finer than the old one, and each dust particle was contained inside a pebble, as the criterion states.

EXAMPLE 4. One can now see that the collection $\mathcal{B}$ of all circular regions in the plane generates the same topology as the collection $\mathcal{B}'$ of all rectangular regions; Figure 13.4 illustrates the proof. We shall treat this example more formally when we study metric spaces.

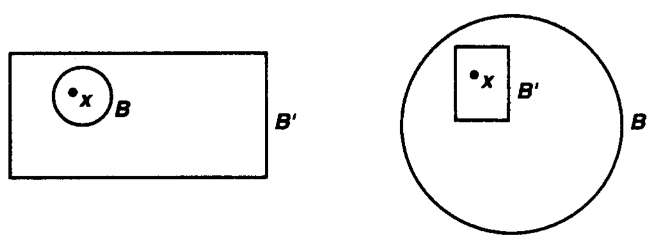


Figure 13.4

We now define three topologies on the real line $\mathbb{R}$, all of which are of interest.

Definition. If $\mathcal{B}$ is the collection of all open intervals in the real line,

$$(a, b) = \{x \mid a < x < b\},$$

the topology generated by $\mathcal{B}$ is called the ***standard topology*** on the real line. Whenever we consider $\mathbb{R}$, we shall suppose it is given this topology unless we specifically state otherwise. If $\mathcal{B}'$ is the collection of all half-open intervals of the form

$$[a, b) = \{x \mid a \leq x < b\},$$

where $a < b$, the topology generated by $\mathcal{B}'$ is called the ***lower limit topology*** on $\mathbb{R}$. When $\mathbb{R}$ is given the lower limit topology, we denote it by $\mathbb{R}_\ell$. Finally let K denote the set of all numbers of the form $1/n$, for $n \in \mathbb{Z}_+$, and let $\mathcal{B}''$ be the collection of all open intervals (a, b), along with all sets of the form $(a, b) - K$. The topology generated by $\mathcal{B}''$ will be called the ***K-topology*** on $\mathbb{R}$. When $\mathbb{R}$ is given this topology, we denote it by $\mathbb{R}_K$.

It is easy to see that all three of these collections are bases; in each case, the intersection of two basis elements is either another basis element or is empty. The relation between these topologies is the following:

Lemma 13.4. *The topologies of $\mathbb{R}_\ell$ and $\mathbb{R}_K$ are strictly finer than the standard topology on $\mathbb{R}$, but are not comparable with one another.*

Proof. Let $\mathcal{T}$, $\mathcal{T}'$, and $\mathcal{T}''$ be the topologies of $\mathbb{R}$, $\mathbb{R}_\ell$, and $\mathbb{R}_K$, respectively. Given a basis element (a, b) for $\mathcal{T}$ and a point x of (a, b), the basis element $[x, b)$ for $\mathcal{T}'$ contains x and lies in (a, b). On the other hand, given the basis element $[x, d)$ for $\mathcal{T}'$, there is no open interval (a, b) that contains x and lies in $[x, d)$. Thus $\mathcal{T}'$ is strictly finer than $\mathcal{T}$.

A similar argument applies to $\mathbb{R}_K$. Given a basis element (a, b) for $\mathcal{T}$ and a point x of (a, b), this same interval is a basis element for $\mathcal{T}''$ that contains x. On the other hand, given the basis element $B = (-1, 1) - K$ for $\mathcal{T}''$ and the point 0 of B, there is no open interval that contains 0 and lies in B.

We leave it to you to show that the topologies of $\mathbb{R}_\ell$ and $\mathbb{R}_K$ are not comparable. ∎

A question may occur to you at this point. Since the topology generated by a basis $\mathcal{B}$ may be described as the collection of arbitrary unions of elements of $\mathcal{B}$, what happens if you start with a given collection of sets and take finite intersections of them as well as arbitrary unions? This question leads to the notion of a subbasis for a topology.

Definition. A ***subbasis*** $\mathcal{S}$ for a topology on X is a collection of subsets of X whose union equals X. The ***topology generated by the subbasis*** $\mathcal{S}$ is defined to be the collection $\mathcal{T}$ of all unions of finite intersections of elements of $\mathcal{S}$.

We must of course check that $\mathcal{T}$ is a topology. For this purpose it will suffice to show that the collection $\mathcal{B}$ of all finite intersections of elements of $\mathcal{S}$ is a basis, for then the collection $\mathcal{T}$ of all unions of elements of $\mathcal{B}$ is a topology, by Lemma 13.1. Given $x \in X$, it belongs to an element of $\mathcal{S}$ and hence to an element of $\mathcal{B}$; this is the first condition for a basis. To check the second condition, let

$$B_1 = S_1 \cap \cdots \cap S_m \quad \text{and} \quad B_2 = S_1' \cap \cdots \cap S_n'$$

be two elements of $\mathcal{B}$. Their intersection

$$B_1 \cap B_2 = (S_1 \cap \cdots \cap S_m) \cap (S_1' \cap \cdots \cap S_n')$$

is also a finite intersection of elements of $\mathcal{S}$, so it belongs to $\mathcal{B}$.

Exercises

1. Let X be a topological space; let A be a subset of X. Suppose that for each $x \in A$ there is an open set U containing x such that $U \subset A$. Show that A is open in X.
2. Consider the nine topologies on the set $X = \{a, b, c\}$ indicated in Example 1 of §12. Compare them; that is, for each pair of topologies, determine whether they are comparable, and if so, which is the finer.
3. Show that the collection $\mathcal{T}_c$ given in Example 4 of §12 is a topology on the set X. Is the collection

$$\mathcal{T}_\infty = \{U \mid X - U \text{ is infinite or empty or all of } X\}$$

   a topology on X?
4. (a) If $\{\mathcal{T}_\alpha\}$ is a family of topologies on X, show that $\bigcap \mathcal{T}_\alpha$ is a topology on X. Is $\bigcup \mathcal{T}_\alpha$ a topology on X?
   (b) Let $\{\mathcal{T}_\alpha\}$ be a family of topologies on X. Show that there is a unique smallest topology on X containing all the collections $\mathcal{T}_\alpha$, and a unique largest topology contained in all $\mathcal{T}_\alpha$.
   (c) If $X = \{a, b, c\}$, let

$$\mathcal{T}_1 = \{\varnothing, X, \{a\}, \{a, b\}\} \quad \text{and} \quad \mathcal{T}_2 = \{\varnothing, X, \{a\}, \{b, c\}\}.$$

   Find the smallest topology containing $\mathcal{T}_1$ and $\mathcal{T}_2$, and the largest topology contained in $\mathcal{T}_1$ and $\mathcal{T}_2$.
5. Show that if $\mathcal{A}$ is a basis for a topology on X, then the topology generated by $\mathcal{A}$ equals the intersection of all topologies on X that contain $\mathcal{A}$. Prove the same if $\mathcal{A}$ is a subbasis.
6. Show that the topologies of $\mathbb{R}_\ell$ and $\mathbb{R}_K$ are not comparable.
7. Consider the following topologies on $\mathbb{R}$:

$$\begin{aligned}
\mathcal{T}_1 &= \text{the standard topology,}\\
\mathcal{T}_2 &= \text{the topology of } \mathbb{R}_K,\\
\mathcal{T}_3 &= \text{the finite complement topology,}\\
\mathcal{T}_4 &= \text{the upper limit topology, having all sets } (a, b] \text{ as basis,}\\
\mathcal{T}_5 &= \text{the topology having all sets } (-\infty, a) = \{x \mid x < a\} \text{ as basis.}
\end{aligned}$$

   Determine, for each of these topologies, which of the others it contains.
8. (a) Apply Lemma 13.2 to show that the countable collection

$$\mathcal{B} = \{(a, b) \mid a < b,\ a \text{ and } b \text{ rational}\}$$

is a basis that generates the standard topology on $\mathbb{R}$.

(b) Show that the collection

$$\mathcal{C} = \{[a, b) \mid a < b, a \text{ and } b \text{ rational}\}$$

is a basis that generates a topology different from the lower limit topology on $\mathbb{R}$.

§14 The Order Topology

If X is a simply ordered set, there is a standard topology for X, defined using the order relation. It is called the *order topology*; in this section, we consider it and study some of its properties.

Suppose that X is a set having a simple order relation $<$. Given elements a and b of X such that $a < b$, there are four subsets of X that are called the ***intervals*** determined by a and b. They are the following :

$$(a, b) = \{x \mid a < x < b\},$$
$$(a, b] = \{x \mid a < x \leq b\},$$
$$[a, b) = \{x \mid a \leq x < b\},$$
$$[a, b] = \{x \mid a \leq x \leq b\}.$$

The notation used here is familiar to you already in the case where X is the real line, but these are intervals in an arbitrary ordered set. A set of the first type is called an ***open interval*** in X, a set of the last type is called a ***closed interval*** in X, and sets of the second and third types are called ***half-open intervals***. The use of the term "open" in this connection suggests that open intervals in X should turn out to be open sets when we put a topology on X. And so they will.

Definition. Let X be a set with a simple order relation; assume X has more than one element. Let $\mathcal{B}$ be the collection of all sets of the following types:

(1) All open intervals (a, b) in X.

(2) All intervals of the form $[a_0, b)$, where a_0 is the smallest element (if any) of X.

(3) All intervals of the form $(a, b_0]$, where b_0 is the largest element (if any) of X.

The collection $\mathcal{B}$ is a basis for a topology on X, which is called the ***order topology***.

If X has no smallest element, there are no sets of type (2), and if X has no largest element, there are no sets of type (3).

One has to check that $\mathcal{B}$ satisfies the requirements for a basis. First, note that every element x of X lies in at least one element of $\mathcal{B}$: The smallest element (if any) lies in all sets of type (2), the largest element (if any) lies in all sets of type (3), and every other element lies in a set of type (1). Second, note that the intersection of any two sets of the preceding types is again a set of one of these types, or is empty. Several cases need to be checked; we leave it to you.

EXAMPLE 1. The standard topology on $\mathbb{R}$, as defined in the preceding section, is just the order topology derived from the usual order on $\mathbb{R}$.

EXAMPLE 2. Consider the set $\mathbb{R} \times \mathbb{R}$ in the dictionary order; we shall denote the general element of $\mathbb{R} \times \mathbb{R}$ by $x \times y$, to avoid difficulty with notation. The set $\mathbb{R} \times \mathbb{R}$ has neither a largest nor a smallest element, so the order topology on $\mathbb{R} \times \mathbb{R}$ has as basis the collection of all open intervals of the form $(a \times b, c \times d)$ for $a < c$, and for $a = c$ and $b < d$. These two types of intervals are indicated in Figure 14.1. The subcollection consisting of only intervals of the second type is also a basis for the order topology on $\mathbb{R} \times \mathbb{R}$, as you can check.

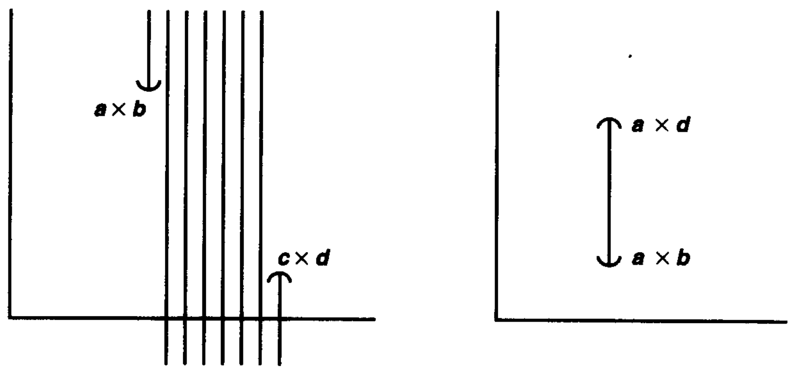


Figure 14.1

EXAMPLE 3. The positive integers $\mathbb{Z}_+$ form an ordered set with a smallest element. The order topology on $\mathbb{Z}_+$ is the discrete topology, for every one-point set is open: If $n > 1$, then the one-point set $\{n\} = (n-1, n+1)$ is a basis element; and if $n = 1$, the one-point set $\{1\} = [1, 2)$ is a basis element.

EXAMPLE 4. The set $X = \{1, 2\} \times \mathbb{Z}_+$ in the dictionary order is another example of an ordered set with a smallest element. Denoting $1 \times n$ by a_n and $2 \times n$ by b_n, we can represent X by

$$a_1, a_2, \ldots; b_1, b_2, \ldots.$$

The order topology on X is *not* the discrete topology. Most one-point sets are open, but there is an exception—the one-point set $\{b_1\}$. Any open set containing b_1 must contain a basis element about b_1 (by definition), and any basis element containing b_1 contains points of the a_i sequence.

Definition. If X is an ordered set, and a is an element of X, there are four subsets of X that are called the ***rays*** determined by a. They are the following:

$$(a, +\infty) = \{x \mid x > a\},$$
$$(-\infty, a) = \{x \mid x < a\},$$
$$[a, +\infty) = \{x \mid x \geq a\},$$
$$(-\infty, a] = \{x \mid x \leq a\}.$$

Sets of the first two types are called ***open rays***, and sets of the last two types are called ***closed rays***.

The use of the term "open" suggests that open rays in X are open sets in the order topology. And so they are. Consider, for example, the ray $(a, +\infty)$. If X has a largest element b_0, then $(a, +\infty)$ equals the basis element $(a, b_0]$. If X has no largest element, then $(a, +\infty)$ equals the union of all basis elements of the form (a, x), for $x > a$. In either case, $(a, +\infty)$ is open. A similar argument applies to the ray $(-\infty, a)$.

The open rays, in fact, form a subbasis for the order topology on X, as we now show. Because the open rays are open in the order topology, the topology they generate is contained in the order topology. On the other hand, every basis element for the order topology equals a finite intersection of open rays; the interval (a, b) equals the intersection of $(-\infty, b)$ and $(a, +\infty)$, while $[a_0, b)$ and $(a, b_0]$, if they exist, are themselves open rays. Hence the topology generated by the open rays contains the order topology.

§15 The Product Topology on $X \times Y$

If X and Y are topological spaces, there is a standard way of defining a topology on the cartesian product $X \times Y$. We consider this topology now and study some of its properties.

Definition. Let X and Y be topological spaces. The ***product topology*** on $X \times Y$ is the topology having as basis the collection $\mathcal{B}$ of all sets of the form $U \times V$, where U is an open subset of X and V is an open subset of Y.

Let us check that $\mathcal{B}$ is a basis. The first condition is trivial, since $X \times Y$ is itself a basis element. The second condition is almost as easy, since the intersection of any two basis elements $U_1 \times V_1$ and $U_2 \times V_2$ is another basis element. For

$$(U_1 \times V_1) \cap (U_2 \times V_2) = (U_1 \cap U_2) \times (V_1 \cap V_2),$$

and the latter set is a basis element because $U_1 \cap U_2$ and $V_1 \cap V_2$ are open in X and Y, respectively. See Figure 15.1.

Note that the collection $\mathcal{B}$ is not a topology on $X \times Y$. The union of the two rectangles pictured in Figure 15.1, for instance, is not a product of two sets, so it cannot belong to $\mathcal{B}$; however, it is open in $X \times Y$.

Each time we introduce a new concept, we shall try to relate it to the concepts that have been previously introduced. In the present case, we ask: What can one say if the topologies on X and Y are given by bases? The answer is as follows:

Theorem 15.1. *If $\mathcal{B}$ is a basis for the topology of X and $\mathcal{C}$ is a basis for the topology of Y, then the collection*

$$\mathcal{D} = \{B \times C \mid B \in \mathcal{B} \text{ and } C \in \mathcal{C}\}$$

is a basis for the topology of $X \times Y$.

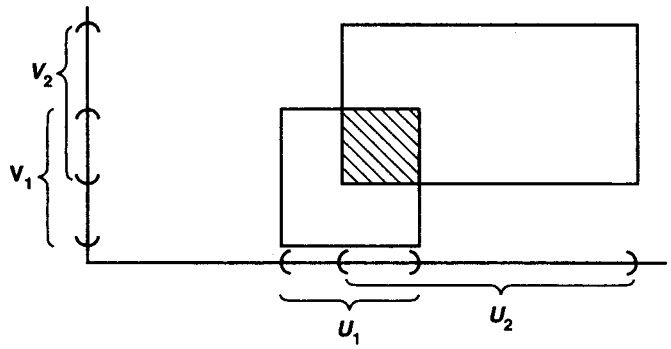


Figure 15.1

Proof. We apply Lemma 13.2. Given an open set W of $X \times Y$ and a point $x \times y$ of W, by definition of the product topology there is a basis element $U \times V$ such that $x \times y \in U \times V \subset W$. Because $\mathcal{B}$ and $\mathcal{C}$ are bases for X and Y, respectively, we can choose an element B of $\mathcal{B}$ such that $x \in B \subset U$, and an element C of $\mathcal{C}$ such that $y \in C \subset V$. Then $x \times y \in B \times C \subset W$. Thus the collection $\mathcal{D}$ meets the criterion of Lemma 13.2, so $\mathcal{D}$ is a basis for $X \times Y$. ∎

EXAMPLE 1. We have a standard topology on $\mathbb{R}$: the order topology. The product of this topology with itself is called the ***standard topology*** on $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$. It has as basis the collection of all products of open sets of $\mathbb{R}$, but the theorem just proved tells us that the much smaller collection of all products $(a, b) \times (c, d)$ of open intervals in $\mathbb{R}$ will also serve as a basis for the topology of $\mathbb{R}^2$. Each such set can be pictured as the interior of a rectangle in $\mathbb{R}^2$. Thus the standard topology on $\mathbb{R}^2$ is just the one we considered in Example 2 of §13.

It is sometimes useful to express the product topology in terms of a subbasis. To do this, we first define certain functions called projections.

Definition. Let $\pi_1 : X \times Y \to X$ be defined by the equation

$$\pi_1(x, y) = x;$$

let $\pi_2 : X \times Y \to Y$ be defined by the equation

$$\pi_2(x, y) = y.$$

The maps π_1 and π_2 are called the ***projections*** of $X \times Y$ onto its first and second factors, respectively.

We use the word "onto" because π_1 and π_2 are surjective (unless one of the spaces X or Y happens to be empty, in which case $X \times Y$ is empty and our whole discussion is empty as well!).

If U is an open subset of X, then the set $\pi_1^{-1}(U)$ is precisely the set $U \times Y$, which is open in $X \times Y$. Similarly, if V is open in Y, then

$$\pi_2^{-1}(V) = X \times V,$$

which is also open in $X \times Y$. The intersection of these two sets is the set $U \times V$, as indicated in Figure 15.2. This fact leads to the following theorem:

Theorem 15.2. *The collection*

$$\mathcal{S} = \{\pi_1^{-1}(U) \mid U \text{ open in } X\} \cup \{\pi_2^{-1}(V) \mid V \text{ open in } Y\}$$

is a subbasis for the product topology on $X \times Y$.

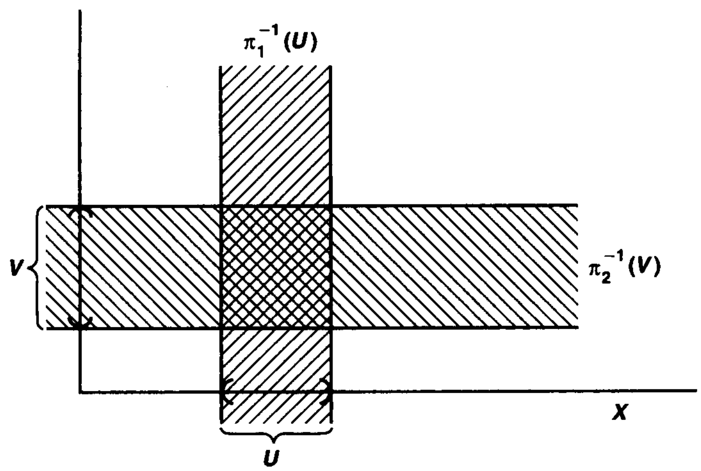


Figure 15.2

Proof. Let $\mathcal{T}$ denote the product topology on $X \times Y$; let $\mathcal{T}'$ be the topology generated by $\mathcal{S}$. Because every element of $\mathcal{S}$ belongs to $\mathcal{T}$, so do arbitrary unions of finite intersections of elements of $\mathcal{S}$. Thus $\mathcal{T}' \subset \mathcal{T}$. On the other hand, every basis element $U \times V$ for the topology $\mathcal{T}$ is a finite intersection of elements of $\mathcal{S}$, since

$$U \times V = \pi_1^{-1}(U) \cap \pi_2^{-1}(V).$$

Therefore, $U \times V$ belongs to $\mathcal{T}'$, so that $\mathcal{T} \subset \mathcal{T}'$ as well. ■

§16 The Subspace Topology

Definition. Let X be a topological space with topology $\mathcal{T}$. If Y is a subset of X, the collection

$$\mathcal{T}_Y = \{Y \cap U \mid U \in \mathcal{T}\}$$

is a topology on Y, called the ***subspace topology***. With this topology, Y is called a ***subspace*** of X; its open sets consist of all intersections of open sets of X with Y.

It is easy to see that $\mathcal{T}_Y$ is a topology. It contains $\varnothing$ and Y because

$$\varnothing = Y \cap \varnothing \quad \text{and} \quad Y = Y \cap X,$$

where $\varnothing$ and X are elements of $\mathcal{T}$. The fact that it is closed under finite intersections and arbitrary unions follows from the equations

$$(U_1 \cap Y) \cap \cdots \cap (U_n \cap Y) = (U_1 \cap \cdots \cap U_n) \cap Y,$$
$$\bigcup_{\alpha \in J} (U_\alpha \cap Y) = (\bigcup_{\alpha \in J} U_\alpha) \cap Y.$$

Lemma 16.1. *If $\mathcal{B}$ is a basis for the topology of X then the collection*

$$\mathcal{B}_Y = \{B \cap Y \mid B \in \mathcal{B}\}$$

is a basis for the subspace topology on Y.

Proof. Given U open in X and given $y \in U \cap Y$, we can choose an element B of $\mathcal{B}$ such that $y \in B \subset U$. Then $y \in B \cap Y \subset U \cap Y$. It follows from Lemma 13.2 that $\mathcal{B}_Y$ is a basis for the subspace topology on Y. ■

When dealing with a space X and a subspace Y, one needs to be careful when one uses the term "open set". Does one mean an element of the topology of Y or an element of the topology of X? We make the following definition : If Y is a subspace of X, we say that a set U is ***open in Y*** (or open *relative to* Y) if it belongs to the topology of Y; this implies in particular that it is a subset of Y. We say that U is ***open in X*** if it belongs to the topology of X.

There is a special situation in which every set open in Y is also open in X:

Lemma 16.2. *Let Y be a subspace of X. If U is open in Y and Y is open in X, then U is open in X.*

Proof. Since U is open in Y, $U = Y \cap V$ for some set V open in X. Since Y and V are both open in X, so is $Y \cap V$. ■

Now let us explore the relation between the subspace topology and the order and product topologies. For product topologies, the result is what one might expect; for order topologies, it is not.

Theorem 16.3. *If A is a subspace of X and B is a subspace of Y, then the product topology on $A \times B$ is the same as the topology $A \times B$ inherits as a subspace of $X \times Y$.*

Proof. The set $U \times V$ is the general basis element for $X \times Y$, where U is open in X and V is open in Y. Therefore, $(U \times V) \cap (A \times B)$ is the general basis element for the subspace topology on $A \times B$. Now

$$(U \times V) \cap (A \times B) = (U \cap A) \times (V \cap B).$$

Since $U \cap A$ and $V \cap B$ are the general open sets for the subspace topologies on A and B, respectively, the set $(U \cap A) \times (V \cap B)$ is the general basis element for the product topology on $A \times B$.

The conclusion we draw is that the bases for the subspace topology on $A \times B$ and for the product topology on $A \times B$ are the same. Hence the topologies are the same. ∎

Now let X be an ordered set in the order topology, and let Y be a subset of X. The order relation on X, when restricted to Y, makes Y into an ordered set. However, *the resulting order topology on Y need not be the same as the topology that Y inherits as a subspace of X.* We give one example where the subspace and order topologies on Y agree, and two examples where they do not.

EXAMPLE 1. Consider the subset $Y = [0, 1]$ of the real line $\mathbb{R}$, in the *subspace* topology. The subspace topology has as basis all sets of the form $(a, b) \cap Y$, where (a, b) is an open interval in $\mathbb{R}$. Such a set is of one of the following types:

$$(a, b) \cap Y = \begin{cases} (a, b) & \text{if } a \text{ and } b \text{ are in } Y, \\ [0, b) & \text{if only } b \text{ is in } Y, \\ (a, 1] & \text{if only } a \text{ is in } Y, \\ Y \text{ or } \varnothing & \text{if neither } a \text{ nor } b \text{ is in } Y. \end{cases}$$

By definition, each of these sets is open in Y. But sets of the second and third types are not open in the larger space $\mathbb{R}$.

Note that these sets form a basis for the *order* topology on Y. Thus, we see that in the case of the set $Y = [0, 1]$, its subspace topology (as a subspace of $\mathbb{R}$) and its order topology are the same.

EXAMPLE 2. Let Y be the subset $[0, 1) \cup \{2\}$ of $\mathbb{R}$. In the subspace topology on Y the one-point set $\{2\}$ is open, because it is the intersection of the open set $(\frac{3}{2}, \frac{5}{2})$ with Y. But in the order topology on Y, the set $\{2\}$ is not open. Any basis element for the order topology on Y that contains 2 is of the form

$$\{x \mid x \in Y \text{ and } a < x \leq 2\}$$

for some $a \in Y$; such a set necessarily contains points of Y less than 2.

EXAMPLE 3. Let $I = [0, 1]$. The dictionary order on $I \times I$ is just the restriction to $I \times I$ of the dictionary order on the plane $\mathbb{R} \times \mathbb{R}$. However, the dictionary order topology on $I \times I$ is not the same as the subspace topology on $I \times I$ obtained from the dictionary order topology on $\mathbb{R} \times \mathbb{R}$! For example, the set $\{1/2\} \times (1/2, 1]$ is open in $I \times I$ in the subspace topology, but not in the order topology, as you can check. See Figure 16.1.

The set $I \times I$ in the dictionary order topology will be called the ***ordered square***, and denoted by I_o^2.

The anomaly illustrated in Examples 2 and 3 does not occur for intervals or rays in an ordered set X. This we now prove.

Given an ordered set X, let us say that a subset Y of X is ***convex*** in X if for each pair of points $a < b$ of Y, the entire interval (a, b) of points of X lies in Y. Note that intervals and rays in X are convex in X.

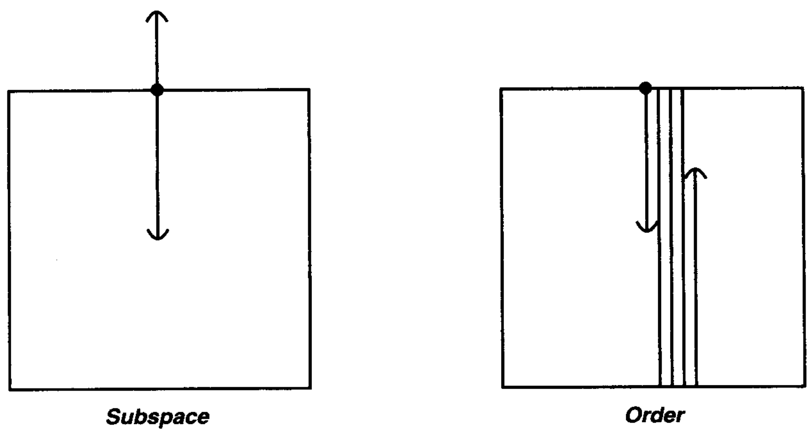


Figure 16.1

Theorem 16.4. *Let X be an ordered set in the order topology; let Y be a subset of X that is convex in X. Then the order topology on Y is the same as the topology Y inherits as a subspace of X.*

Proof. Consider the ray $(a, +\infty)$ in X. What is its intersection with Y? If $a \in Y$, then

$$(a, +\infty) \cap Y = \{x \mid x \in Y \text{ and } x > a\};$$

this is an open ray of the ordered set Y. If $a \notin Y$, then a is either a lower bound on Y or an upper bound on Y, since Y is convex. In the former case, the set $(a, +\infty) \cap Y$ equals all of Y; in the latter case, it is empty.

A similar remark shows that the intersection of the ray $(-\infty, a)$ with Y is either an open ray of Y, or Y itself, or empty. Since the sets $(a, +\infty) \cap Y$ and $(-\infty, a) \cap Y$ form a subbasis for the subspace topology on Y, and since each is open in the order topology, the order topology contains the subspace topology.

To prove the reverse, note that any open ray of Y equals the intersection of an open ray of X with Y, so it is open in the subspace topology on Y. Since the open rays of Y are a subbasis for the order topology on Y, this topology is contained in the subspace topology. ∎

To avoid ambiguity, let us agree that whenever X is an ordered set in the order topology and Y is a subset of X, *we shall assume that Y is given the subspace topology* unless we specifically state otherwise. If Y is convex in X, this is the same as the order topology on Y; otherwise, it may not be.

Exercises

1. Show that if Y is a subspace of X, and A is a subset of Y, then the topology A

inherits as a subspace of Y is the same as the topology it inherits as a subspace of X.

2. If $\mathcal{T}$ and $\mathcal{T}'$ are topologies on X and $\mathcal{T}'$ is strictly finer than $\mathcal{T}$, what can you say about the corresponding subspace topologies on the subset Y of X?
3. Consider the set $Y = [-1, 1]$ as a subspace of $\mathbb{R}$. Which of the following sets are open in Y? Which are open in $\mathbb{R}$?
$$A = \{x \mid \tfrac{1}{2} < |x| < 1\},$$
$$B = \{x \mid \tfrac{1}{2} < |x| \leq 1\},$$
$$C = \{x \mid \tfrac{1}{2} \leq |x| < 1\},$$
$$D = \{x \mid \tfrac{1}{2} \leq |x| \leq 1\},$$
$$E = \{x \mid 0 < |x| < 1 \text{ and } 1/x \notin \mathbb{Z}_+\}.$$
4. A map $f : X \to Y$ is said to be an ***open map*** if for every open set U of X, the set $f(U)$ is open in Y. Show that $\pi_1 : X \times Y \to X$ and $\pi_2 : X \times Y \to Y$ are open maps.
5. Let X and X' denote a single set in the topologies $\mathcal{T}$ and $\mathcal{T}'$, respectively; let Y and Y' denote a single set in the topologies $\mathcal{U}$ and $\mathcal{U}'$, respectively. Assume these sets are nonempty.
   (a) Show that if $\mathcal{T}' \supset \mathcal{T}$ and $\mathcal{U}' \supset \mathcal{U}$, then the product topology on $X' \times Y'$ is finer than the product topology on $X \times Y$.
   (b) Does the converse of (a) hold? Justify your answer.
6. Show that the countable collection
$$\{(a, b) \times (c, d) \mid a < b \text{ and } c < d, \text{ and } a, b, c, d \text{ are rational}\}$$
is a basis for $\mathbb{R}^2$.
7. Let X be an ordered set. If Y is a proper subset of X that is convex in X, does it follow that Y is an interval or a ray in X?
8. If L is a straight line in the plane, describe the topology L inherits as a subspace of $\mathbb{R}_\ell \times \mathbb{R}$ and as a subspace of $\mathbb{R}_\ell \times \mathbb{R}_\ell$. In each case it is a familiar topology.
9. Show that the dictionary order topology on the set $\mathbb{R} \times \mathbb{R}$ is the same as the product topology $\mathbb{R}_d \times \mathbb{R}$, where $\mathbb{R}_d$ denotes $\mathbb{R}$ in the discrete topology. Compare this topology with the standard topology on $\mathbb{R}^2$.
10. Let $I = [0, 1]$. Compare the product topology on $I \times I$, the dictionary order topology on $I \times I$, and the topology $I \times I$ inherits as a subspace of $\mathbb{R} \times \mathbb{R}$ in the dictionary order topology.

§17 Closed Sets and Limit Points

Now that we have a few examples at hand, we can introduce some of the basic concepts associated with topological spaces. In this section, we treat the notions of *closed set,* *closure* of a set, and *limit point*. These lead naturally to consideration of a certain axiom for topological spaces called the *Hausdorff axiom*.

Closed Sets

A subset A of a topological space X is said to be ***closed*** if the set $X - A$ is open.

EXAMPLE 1. The subset $[a, b]$ of $\mathbb{R}$ is closed because its complement

$$\mathbb{R} - [a, b] = (-\infty, a) \cup (b, +\infty),$$

is open. Similarly, $[a, +\infty)$ is closed, because its complement $(-\infty, a)$ is open. These facts justify our use of the terms "closed interval" and "closed ray." The subset $[a, b)$ of $\mathbb{R}$ is neither open nor closed.

EXAMPLE 2. In the plane $\mathbb{R}^2$, the set

$$\{x \times y \mid x \geq 0 \text{ and } y \geq 0\}$$

is closed, because its complement is the union of the two sets

$$(-\infty, 0) \times \mathbb{R} \quad \text{and} \quad \mathbb{R} \times (-\infty, 0),$$

each of which is a product of open sets of $\mathbb{R}$ and is, therefore, open in $\mathbb{R}^2$.

EXAMPLE 3. In the finite complement topology on a set X, the closed sets consist of X itself and all finite subsets of X.

EXAMPLE 4. In the discrete topology on the set X, every set is open; it follows that every set is closed as well.

EXAMPLE 5. Consider the following subset of the real line:

$$Y = [0, 1] \cup (2, 3),$$

in the subspace topology. In this space, the set $[0, 1]$ is open, since it is the intersection of the open set $(-\frac{1}{2}, \frac{3}{2})$ of $\mathbb{R}$ with Y. Similarly, $(2, 3)$ is open as a subset of Y; it is even open as a subset of $\mathbb{R}$. Since $[0, 1]$ and $(2, 3)$ are complements in Y of each other, we conclude that both $[0, 1]$ and $(2, 3)$ are closed as subsets of Y.

These examples suggest that an answer to the mathematician's riddle: "How is a set different from a door?" should be: "A door must be either open or closed, and cannot be both, while a set can be open, or closed, or both, or neither!"

The collection of closed subsets of a space X has properties similar to those satisfied by the collection of open subsets of X:

Theorem 17.1. *Let X be a topological space. Then the following conditions hold:*

(1) *$\varnothing$ and X are closed.*

(2) *Arbitrary intersections of closed sets are closed.*

(3) *Finite unions of closed sets are closed.*

Proof. (1) $\varnothing$ and X are closed because they are the complements of the open sets X and $\varnothing$, respectively.

(2) Given a collection of closed sets $\{A_\alpha\}_{\alpha \in J}$, we apply DeMorgan's law,

$$X - \bigcap_{\alpha \in J} A_\alpha = \bigcup_{\alpha \in J} (X - A_\alpha).$$

Since the sets $X - A_\alpha$ are open by definition, the right side of this equation represents an arbitrary union of open sets, and is thus open. Therefore, $\bigcap A_\alpha$ is closed.

(3) Similarly, if A_i is closed for $i = 1, \dots, n$, consider the equation

$$X - \bigcup_{i=1}^{n} A_i = \bigcap_{i=1}^{n} (X - A_i).$$

The set on the right side of this equation is a finite intersection of open sets and is therefore open. Hence $\bigcup A_i$ is closed. ∎

Instead of using open sets, one could just as well specify a topology on a space by giving a collection of sets (to be called "closed sets") satisfying the three properties of this theorem. One could then define open sets as the complements of closed sets and proceed just as before. This procedure has no particular advantage over the one we have adopted, and most mathematicians prefer to use open sets to define topologies.

Now when dealing with subspaces, one needs to be careful in using the term "closed set." If Y is a subspace of X, we say that a set A is ***closed in Y*** if A is a subset of Y and if A is closed in the subspace topology of Y (that is, if $Y - A$ is open in Y). We have the following theorem:

Theorem 17.2. *Let Y be a subspace of X. Then a set A is closed in Y if and only if it equals the intersection of a closed set of X with Y.*

Proof. Assume that $A = C \cap Y$, where C is closed in X. (See Figure 17.1.) Then $X - C$ is open in X, so that $(X - C) \cap Y$ is open in Y, by definition of the subspace topology. But $(X - C) \cap Y = Y - A$. Hence $Y - A$ is open in Y, so that A is closed in Y. Conversely, assume that A is closed in Y. (See Figure 17.2.) Then $Y - A$ is open in Y, so that by definition it equals the intersection of an open set U of X with Y. The set $X - U$ is closed in X, and $A = Y \cap (X - U)$, so that A equals the intersection of a closed set of X with Y, as desired. ∎

A set A that is closed in the subspace Y may or may not be closed in the larger space X. As was the case with open sets, there is a criterion for A to be closed in X; we leave the proof to you:

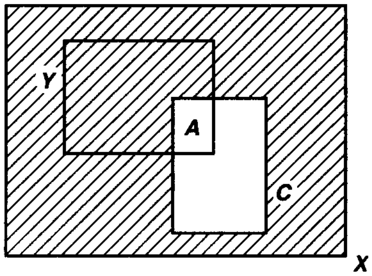


Figure 17.1

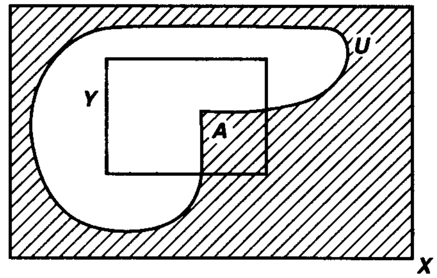


Figure 17.2

Theorem 17.3. *Let Y be a subspace of X. If A is closed in Y and Y is closed in X, then A is closed in X.*

Closure and Interior of a Set

Given a subset A of a topological space X, the ***interior*** of A is defined as the union of all open sets contained in A, and the ***closure*** of A is defined as the intersection of all closed sets containing A.

The interior of A is denoted by $\text{Int}\, A$ and the closure of A is denoted by $\text{Cl}\, A$ or by $\bar{A}$. Obviously $\text{Int}\, A$ is an open set and $\bar{A}$ is a closed set; furthermore,

$$\text{Int}\, A \subset A \subset \bar{A}.$$

If A is open, $A = \text{Int}\, A$; while if A is closed, $A = \bar{A}$.

We shall not make much use of the interior of a set, but the closure of a set will be quite important.

When dealing with a topological space X and a subspace Y, one needs to exercise care in taking closures of sets. If A is a subset of Y, the closure of A in Y and the closure of A in X will in general be different. *In such a situation, we reserve the notation $\bar{A}$ to stand for the closure of A in X.* The closure of A in Y can be expressed in terms of $\bar{A}$, as the following theorem shows:

Theorem 17.4. *Let Y be a subspace of X; let A be a subset of Y; let $\bar{A}$ denote the closure of A in X. Then the closure of A in Y equals $\bar{A} \cap Y$.*

Proof. Let B denote the closure of A in Y. The set $\bar{A}$ is closed in X, so $\bar{A} \cap Y$ is closed in Y by Theorem 17.2. Since $\bar{A} \cap Y$ contains A, and since by definition B equals the intersection of *all* closed subsets of Y containing A, we must have $B \subset (\bar{A} \cap Y)$.

On the other hand, we know that B is closed in Y. Hence by Theorem 17.2, $B = C \cap Y$ for some set C closed in X. Then C is a closed set of X containing A; because $\bar{A}$ is the intersection of *all* such closed sets, we conclude that $\bar{A} \subset C$. Then $(\bar{A} \cap Y) \subset (C \cap Y) = B$. ∎

The definition of the closure of a set does not give us a convenient way for actually finding the closures of specific sets, since the collection of all closed sets in X, like the collection of all open sets, is usually much too big to work with. Another way of describing the closure of a set, useful because it involves only a basis for the topology of X, is given in the following theorem.

First let us introduce some convenient terminology. We shall say that a set A ***intersects*** a set B if the intersection $A \cap B$ is not empty.

Theorem 17.5. *Let A be a subset of the topological space X.*

(a) Then $x \in \bar{A}$ if and only if every open set U containing x intersects A.

(b) Supposing the topology of X is given by a basis, then $x \in \bar{A}$ if and only if every basis element B containing x intersects A.

Proof. Consider the statement in (a). It is a statement of the form $P \Leftrightarrow Q$. Let us transform each implication to its contrapositive, thereby obtaining the logically equivalent statement (not P) $\Leftrightarrow$ (not Q). Written out, it is the following:

$x \notin \bar{A} \Longleftrightarrow$ there exists an open set U containing x that does not intersect A.

In this form, our theorem is easy to prove. If x is not in $\bar{A}$, the set $U = X - \bar{A}$ is an open set containing x that does not intersect A, as desired. Conversely, if there exists an open set U containing x which does not intersect A, then $X - U$ is a closed set containing A. By definition of the closure $\bar{A}$, the set $X - U$ must contain $\bar{A}$; therefore, x cannot be in $\bar{A}$.

Statement (b) follows readily. If every open set containing x intersects A, so does every basis element B containing x, because B is an open set. Conversely, if every basis element containing x intersects A, so does every open set U containing x, because U contains a basis element that contains x. ∎

Mathematicians often use some special terminology here. They shorten the statement "U is an open set containing x" to the phrase

"U is a ***neighborhood*** of x."

Using this terminology, one can write the first half of the preceding theorem as follows:

> If A is a subset of the topological space X, then $x \in \bar{A}$ if and only if every neighborhood of x intersects A.

EXAMPLE 6. Let X be the real line $\mathbb{R}$. If $A = (0, 1]$, then $\bar{A} = [0, 1]$, for every neighborhood of 0 intersects A, while every point outside $[0, 1]$ has a neighborhood disjoint from A. Similar arguments apply to the following subsets of X:

If $B = \{1/n \mid n \in \mathbb{Z}_+\}$, then $\bar{B} = \{0\} \cup B$. If $C = \{0\} \cup (1, 2)$, then $\bar{C} = \{0\} \cup [1, 2]$. If $\mathbb{Q}$ is the set of rational numbers, then $\bar{\mathbb{Q}} = \mathbb{R}$. If $\mathbb{Z}_+$ is the set of positive integers, then $\bar{\mathbb{Z}}_+ = \mathbb{Z}_+$. If $\mathbb{R}_+$ is the set of positive reals, then the closure of $\mathbb{R}_+$ is the set $\mathbb{R}_+ \cup \{0\}$. (This is the reason we introduced the notation $\bar{\mathbb{R}}_+$ for the set $\mathbb{R}_+ \cup \{0\}$, back in §2.)

EXAMPLE 7. Consider the subspace $Y = (0, 1]$ of the real line $\mathbb{R}$. The set $A = (0, \frac{1}{2})$ is a subset of Y; its closure in $\mathbb{R}$ is the set $[0, \frac{1}{2}]$, and its closure in Y is the set $[0, \frac{1}{2}] \cap Y = (0, \frac{1}{2}]$.

Some mathematicians use the term "neighborhood" differently. They say that A is a neighborhood of x if A merely *contains* an open set containing x. We shall not follow this practice.

Limit Points

There is yet another way of describing the closure of a set, a way that involves the important concept of limit point, which we consider now.

If A is a subset of the topological space X and if x is a point of X, we say that x is a ***limit point*** (or "cluster point," or "point of accumulation") of A if every neighborhood of x intersects A in some point *other than x itself.* Said differently, x is a limit point of A if it belongs to the closure of $A - \{x\}$. The point x may lie in A or not; for this definition it does not matter.

EXAMPLE 8. Consider the real line $\mathbb{R}$. If $A = (0, 1]$, then the point 0 is a limit point of A and so is the point $\frac{1}{2}$. In fact, every point of the interval $[0, 1]$ is a limit point of A, but no other point of $\mathbb{R}$ is a limit point of A.

If $B = \{1/n \mid n \in \mathbb{Z}_+\}$, then 0 is the only limit point of B. Every other point x of $\mathbb{R}$ has a neighborhood that either does not intersect B at all, or it intersects B only in the point x itself. If $C = \{0\} \cup (1, 2)$, then the limit points of C are the points of the interval $[1, 2]$. If $\mathbb{Q}$ is the set of rational numbers, every point of $\mathbb{R}$ is a limit point of $\mathbb{Q}$. If $\mathbb{Z}_+$ is the set of positive integers, no point of $\mathbb{R}$ is a limit point of $\mathbb{Z}_+$. If $\mathbb{R}_+$ is the set of positive reals, then every point of $\{0\} \cup \mathbb{R}_+$ is a limit point of $\mathbb{R}_+$.

Comparison of Examples 6 and 8 suggests a relationship between the closure of a set and the limit points of a set. That relationship is given in the following theorem:

Theorem 17.6. *Let A be a subset of the topological space X; let A' be the set of all limit points of A. Then*

$$\bar{A} = A \cup A'.$$

Proof. If x is in A', every neighborhood of x intersects A (in a point different from x). Therefore, by Theorem 17.5, x belongs to $\bar{A}$. Hence $A' \subset \bar{A}$. Since by definition $A \subset \bar{A}$, it follows that $A \cup A' \subset \bar{A}$.

To demonstrate the reverse inclusion, we let x be a point of $\bar{A}$ and show that $x \in A \cup A'$. If x happens to lie in A, it is trivial that $x \in A \cup A'$; suppose that x does not lie in A. Since $x \in \bar{A}$, we know that every neighborhood U of x intersects A; because $x \notin A$, the set U must intersect A in a point different from x. Then $x \in A'$, so that $x \in A \cup A'$, as desired. ∎

Corollary 17.7. *A subset of a topological space is closed if and only if it contains all its limit points.*

Proof. The set A is closed if and only if $A = \bar{A}$, and the latter holds if and only if $A' \subset A$. ∎

Hausdorff Spaces

One's experience with open and closed sets and limit points in the real line and the plane can be misleading when one considers more general topological spaces. For example, in the spaces $\mathbb{R}$ and $\mathbb{R}^2$, each one-point set $\{x_0\}$ is closed. This fact is easily proved; every point different from x_0 has a neighborhood not intersecting $\{x_0\}$, so that $\{x_0\}$ is its own closure. But this fact is not true for arbitrary topological spaces. Consider the topology on the three-point set $\{a, b, c\}$ indicated in Figure 17.3. In this space, the one-point set $\{b\}$ is not closed, for its complement is *not* open.

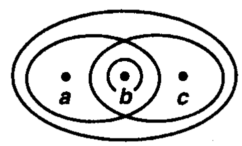


Figure 17.3

Similarly, one's experience with the properties of convergent sequences in $\mathbb{R}$ and $\mathbb{R}^2$ can be misleading when one deals with more general topological spaces. In an arbitrary topological space, one says that a sequence $x_1, x_2, \dots$ of points of the space X ***converges*** to the point x of X provided that, corresponding to each neighborhood U of x, there is a positive integer N such that $x_n \in U$ for all $n \geq N$. In $\mathbb{R}$ and $\mathbb{R}^2$, a sequence cannot converge to more than one point, but in an arbitrary space, it can. In the space indicated in Figure 17.3, for example, the sequence defined by setting $x_n = b$ for all n converges not only to the point b, but also to the point a and to the point c!

Topologies in which one-point sets are not closed, or in which sequences can converge to more than one point, are considered by many mathematicians to be somewhat strange. They are not really very interesting, for they seldom occur in other branches of mathematics. And the theorems that one can prove about topological spaces are rather limited if such examples are allowed. Therefore, one often imposes an additional condition that will rule out examples like this one, bringing the class of spaces under consideration closer to those to which one's geometric intuition applies. The condition was suggested by the mathematician Felix Hausdorff, so mathematicians have come to call it by his name.

Definition. A topological space X is called a ***Hausdorff space*** if for each pair x_1, x_2 of distinct points of X, there exist neighborhoods U_1, and U_2 of x_1 and x_2, respectively, that are disjoint.

Theorem 17.8. *Every finite point set in a Hausdorff space X is closed.*

Proof. It suffices to show that every one-point set $\{x_0\}$ is closed. If x is a point of X different from x_0, then x and x_0 have disjoint neighborhoods U and V, respectively. Since U does not intersect $\{x_0\}$, the point x cannot belong to the closure of the set $\{x_0\}$. As a result, the closure of the set $\{x_0\}$ is $\{x_0\}$ itself, so that it is closed. ∎

The condition that finite point sets be closed is in fact weaker than the Hausdorff condition. For example, the real line $\mathbb{R}$ in the finite complement topology is not a Hausdorff space, but it is a space in which finite point sets are closed. The condition that finite point sets be closed has been given a name of its own: it is called the ***T_1 axiom***. (We shall explain the reason for this strange terminology in Chapter 4.) The T_1 axiom will appear in this book in a few exercises, and in just one theorem, which is the following:

Theorem 17.9. *Let X be a space satisfying the T_1 axiom; let A be a subset of X. Then the point x is a limit point of A if and only if every neighborhood of x contains infinitely many points of A.*

Proof. If every neighborhood of x intersects A in infinitely many points, it certainly intersects A in some point other than x itself, so that x is a limit point of A.

Conversely, suppose that x is a limit point of A, and suppose some neighborhood U of x intersects A in only finitely many points. Then U also intersects $A - \{x\}$ in finitely many points; let $\{x_1, \dots, x_m\}$ be the points of $U \cap (A - \{x\})$. The set $X - \{x_1, \dots, x_m\}$ is an open set of X, since the finite point set $\{x_1, \dots, x_m\}$ is closed; then

$$U \cap (X - \{x_1, \dots, x_m\})$$

is a neighborhood of x that intersects the set $A - \{x\}$ not at all. This contradicts the assumption that x is a limit point of A. ∎

One reason for our lack of interest in the T_1 axiom is the fact that many of the interesting theorems of topology require not just that axiom, but the full strength of the Hausdorff axiom. Furthermore, most of the spaces that are important to mathematicians are Hausdorff spaces. The following two theorems give some substance to these remarks.

Theorem 17.10. *If X is a Hausdorff space, then a sequence of points of X converges to at most one point of X.*

Proof. Suppose that x_n is a sequence of points of X that converges to x. If $y \neq x$, let U and V be disjoint neighborhoods of x and y, respectively. Since U contains x_n for all but finitely many values of n, the set V cannot. Therefore, x_n cannot converge to y. ∎

If the sequence x_n of points of the Hausdorff space X converges to the point x of X, we often write $x_n \to x$, and we say that x is the ***limit*** of the sequence x_n.

The proof of the following result is left to the exercises.

Theorem 17.11. *Every simply ordered set is a Hausdorff space in the order topology. The product of two Hausdorff spaces is a Hausdorff space. A subspace of a Hausdorff space is a Hausdorff space.*

The Hausdorff condition is generally considered to be a very mild extra condition to impose on a topological space. Indeed, in a first course in topology some mathematicians go so far as to impose this condition at the outset, refusing to consider spaces that are not Hausdorff spaces. We shall not go this far, but we shall certainly assume the Hausdorff condition whenever it is needed in a proof without having any qualms about limiting seriously the range of applications of the results.

The Hausdorff condition is one of a number of extra conditions one can impose on a topological space. Each time one imposes such a condition, one can prove stronger theorems, but one limits the class of spaces to which the theorems apply. Much of the research that has been done in topology since its beginnings has centered on the problem of finding conditions that will be strong enough to enable one to prove interesting theorems about spaces satisfying those conditions, and yet not so strong that they limit severely the range of applications of the results.

We shall study a number of such conditions in the next two chapters. The Hausdorff condition and the T_1 axiom are but two of a collection of conditions similar to one another that are called collectively the *separation axioms*. Other conditions include the *countability axioms*, and various *compactness* and *connectedness* conditions. Some of these are quite stringent requirements, as you will see.

Exercises

1. Let $\mathcal{C}$ be a collection of subsets of the set X. Suppose that $\varnothing$ and X are in $\mathcal{C}$, and that finite unions and arbitrary intersections of elements of $\mathcal{C}$ are in $\mathcal{C}$. Show that the collection

$$\mathcal{T} = \{X - C \mid C \in \mathcal{C}\}$$

   is a topology on X.

2. Show that if A is closed in Y and Y is closed in X, then A is closed in X.
3. Show that if A is closed in X and B is closed in Y, then $A \times B$ is closed in $X \times Y$.
4. Show that if U is open in X and A is closed in X, then $U - A$ is open in X, and $A - U$ is closed in X.
5. Let X be an ordered set in the order topology. Show that $\overline{(a, b)} \subset [a, b]$. Under what conditions does equality hold?

6. Let A, B, and A_α denote subsets of a space X. Prove the following:
(a) If $A \subset B$, then $\bar{A} \subset \bar{B}$.
(b) $\overline{A \cup B} = \bar{A} \cup \bar{B}$.
(c) $\overline{\bigcup A_\alpha} \supset \bigcup \bar{A}_\alpha$; give an example where equality fails.

7. Criticize the following "proof" that $\overline{\bigcup A_\alpha} \subset \bigcup \bar{A}_\alpha$: if $\{A_\alpha\}$ is a collection of sets in X and if $x \in \overline{\bigcup A_\alpha}$, then every neighborhood U of x intersects $\bigcup A_\alpha$. Thus U must intersect some A_α, so that x must belong to the closure of some A_α. Therefore, $x \in \bigcup \bar{A}_\alpha$.

8. Let A, B, and A_α denote subsets of a space X. Determine whether the following equations hold; if an equality fails, determine whether one of the inclusions $\supset$ or $\subset$ holds.
(a) $\overline{A \cap B} = \bar{A} \cap \bar{B}$.
(b) $\overline{\bigcap A_\alpha} = \bigcap \bar{A}_\alpha$.
(c) $\overline{A - B} = \bar{A} - \bar{B}$.

9. Let $A \subset X$ and $B \subset Y$. Show that in the space $X \times Y$,
$$\overline{A \times B} = \bar{A} \times \bar{B}.$$

10. Show that every order topology is Hausdorff.

11. Show that the product of two Hausdorff spaces is Hausdorff.

12. Show that a subspace of a Hausdorff space is Hausdorff.

13. Show that X is Hausdorff if and only if the ***diagonal*** $\Delta = \{x \times x \mid x \in X\}$ is closed in $X \times X$.

14. In the finite complement topology on $\mathbb{R}$, to what point or points does the sequence $x_n = 1/n$ converge?

15. Show the T_1 axiom is equivalent to the condition that for each pair of points of X, each has a neighborhood not containing the other.

16. Consider the five topologies on $\mathbb{R}$ given in Exercise 7 of §13.
(a) Determine the closure of the set $K = \{1/n \mid n \in \mathbb{Z}_+\}$ under each of these topologies.
(b) Which of these topologies satisfy the Hausdorff axiom? the T_1 axiom?

17. Consider the lower limit topology on $\mathbb{R}$ and the topology given by the basis $\mathcal{C}$ of Exercise 8 of §13. Determine the closures of the intervals $A = (0, \sqrt{2})$ and $B = (\sqrt{2}, 3)$ in these two topologies.

18. Determine the closures of the following subsets of the ordered square:
$$A = \{(1/n) \times 0 \mid n \in \mathbb{Z}_+\},$$
$$B = \{(1 - 1/n) \times \tfrac{1}{2} \mid n \in \mathbb{Z}_+\},$$
$$C = \{x \times 0 \mid 0 < x < 1\},$$
$$D = \{x \times \tfrac{1}{2} \mid 0 < x < 1\},$$
$$E = \{\tfrac{1}{2} \times y \mid 0 < y < 1\}.$$

19. If $A \subset X$, we define the ***boundary*** of A by the equation

$$\operatorname{Bd} A = \bar{A} \cap \overline{(X - A)}.$$

(a) Show that $\operatorname{Int} A$ and $\operatorname{Bd} A$ are disjoint, and $\bar{A} = \operatorname{Int} A \cup \operatorname{Bd} A$.
(b) Show that $\operatorname{Bd} A = \varnothing \Leftrightarrow A$ is both open and closed.
(c) Show that U is open $\Leftrightarrow \operatorname{Bd} U = \bar{U} - U$.
(d) If U is open, is it true that $U = \operatorname{Int}(\bar{U})$? Justify your answer.

20. Find the boundary and the interior of each of the following subsets of $\mathbb{R}^2$:
(a) $A = \{x \times y \mid y = 0\}$
(b) $B = \{x \times y \mid x > 0 \text{ and } y \neq 0\}$
(c) $C = A \cup B$
(d) $D = \{x \times y \mid x \text{ is rational}\}$
(e) $E = \{x \times y \mid 0 < x^2 - y^2 \leq 1\}$
(f) $F = \{x \times y \mid x \neq 0 \text{ and } y \leq 1/x\}$

***21.** (Kuratowski) Consider the collection of all subsets A of the topological space X. The operations of closure $A \to \bar{A}$ and complementation $A \to X - A$ are functions from this collection to itself.
(a) Show that starting with a given set A, one can form no more than 14 distinct sets by applying these two operations successively.
(b) Find a subset A of $\mathbb{R}$ (in its usual topology) for which the maximum of 14 is obtained.

§18 Continuous Functions

The concept of continuous function is basic to much of mathematics. Continuous functions on the real line appear in the first pages of any calculus book, and continuous functions in the plane and in space follow not far behind. More general kinds of continuous functions arise as one goes further in mathematics. In this section, we shall formulate a definition of continuity that will include all these as special cases, and we shall study various properties of continuous functions. Many of these properties are direct generalizations of things you learned about continuous functions in calculus and analysis.

Continuity of a Function

Let X and Y be topological spaces. A function $f : X \to Y$ is said to be ***continuous*** if for each open subset V of Y, the set $f^{-1}(V)$ is an open subset of X.

Recall that $f^{-1}(V)$ is the set of all points x of X for which $f(x) \in V$; it is empty if V does not intersect the image set $f(X)$ of f.

Continuity of a function depends not only upon the function f itself, but also on the topologies specified for its domain and range. If we wish to emphasize this fact, we can say that f is continuous *relative* to specific topologies on X and Y.

Let us note that if the topology of the range space Y is given by a basis $\mathcal{B}$, then to prove continuity of f it suffices to show that the inverse image of every *basis element* is open: The arbitrary open set V of Y can be written as a union of basis elements

$$V = \bigcup_{\alpha \in J} B_\alpha.$$

Then

$$f^{-1}(V) = \bigcup_{\alpha \in J} f^{-1}(B_\alpha),$$

so that $f^{-1}(V)$ is open if each set $f^{-1}(B_\alpha)$ is open.

If the topology on Y is given by a subbasis $\mathcal{S}$, to prove continuity of f it will even suffice to show that the inverse image of each *subbasis* element is open: The arbitrary basis element B for Y can be written as a finite intersection $S_1 \cap \cdots \cap S_n$ of subbasis elements; it follows from the equation

$$f^{-1}(B) = f^{-1}(S_1) \cap \cdots \cap f^{-1}(S_n)$$

that the inverse image of every basis element is open.

EXAMPLE 1. Let us consider a function like those studied in analysis, a "real-valued function of a real variable,"

$$f : \mathbb{R} \longrightarrow \mathbb{R}.$$

In analysis, one defines continuity of f via the "ϵ-δ definition," a bugaboo over the years for every student of mathematics. As one would expect, the ϵ-δ definition and ours are equivalent. To prove that our definition implies the ϵ-δ definition, for instance, we proceed as follows:

Given x_0 in $\mathbb{R}$, and given $\epsilon > 0$, the interval $V = (f(x_0) - \epsilon, f(x_0) + \epsilon)$ is an open set of the range space $\mathbb{R}$. Therefore, $f^{-1}(V)$ is an open set in the domain space $\mathbb{R}$. Because $f^{-1}(V)$ contains the point x_0, it contains some basis element (a, b) about x_0. We choose δ to be the smaller of the two numbers $x_0 - a$ and $b - x_0$. Then if $|x - x_0| < \delta$, the point x must be in (a, b), so that $f(x) \in V$, and $|f(x) - f(x_0)| < \epsilon$, as desired.

Proving that the ϵ-δ definition implies our definition is no harder; we leave it to you. We shall return to this example when we study metric spaces.

EXAMPLE 2. In calculus one considers the property of continuity for many kinds of functions. For example, one studies functions of the following types:

$f : \mathbb{R} \longrightarrow \mathbb{R}^2$ (curves in the plane)

$f : \mathbb{R} \longrightarrow \mathbb{R}^3$ (curves in space)

$f : \mathbb{R}^2 \longrightarrow \mathbb{R}$ (functions $f(x, y)$ of two real variables)

$f : \mathbb{R}^3 \longrightarrow \mathbb{R}$ (functions $f(x, y, z)$ of three real variables)

$f : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ (vector fields $\mathbf{v}(x, y)$ in the plane).

Each of them has a notion of continuity defined for it. Our general definition of continuity includes all these as special cases; this fact will be a consequence of general theorems we shall prove concerning continuous functions on product spaces and on metric spaces.

EXAMPLE 3. Let $\mathbb{R}$ denote the set of real numbers in its usual topology, and let $\mathbb{R}_\ell$ denote the same set in the lower limit topology. Let

$$f : \mathbb{R} \longrightarrow \mathbb{R}_\ell$$

be the identity function; $f(x) = x$ for every real number x. Then f is not a continuous function; the inverse image of the open set $[a, b)$ of $\mathbb{R}_\ell$ equals itself, which is not open in $\mathbb{R}$. On the other hand, the identity function

$$g : \mathbb{R}_\ell \longrightarrow \mathbb{R}$$

is continuous, because the inverse image of (a, b) is itself, which is open in $\mathbb{R}_\ell$.

In analysis, one studies several different but equivalent ways of formulating the definition of continuity. Some of these generalize to arbitrary spaces, and they are considered in the theorems that follow. The familiar "ϵ-δ" definition and the "convergent sequence definition" do not generalize to arbitrary spaces; they will be treated when we study metric spaces.

Theorem 18.1. *Let X and Y be topological spaces; let $f : X \to Y$. Then the following are equivalent:*

(1) *f is continuous.*

(2) *For every subset A of X, one has $f(\bar{A}) \subset \overline{f(A)}$.*

(3) *For every closed set B of Y, the set $f^{-1}(B)$ is closed in X.*

(4) *For each $x \in X$ and each neighborhood V of $f(x)$, there is a neighborhood U of x such that $f(U) \subset V$.*

If the condition in (4) holds for the point x of X, we say that f is ***continuous at the point x***.

Proof. We show that (1) $\Rightarrow$ (2) $\Rightarrow$ (3) $\Rightarrow$ (1) and that (1) $\Rightarrow$ (4) $\Rightarrow$ (1).

(1) $\Rightarrow$ (2). Assume that f is continuous. Let A be a subset of X. We show that if $x \in \bar{A}$, then $f(x) \in \overline{f(A)}$. Let V be a neighborhood of $f(x)$. Then $f^{-1}(V)$ is an open set of X containing x; it must intersect A in some point y. Then V intersects $f(A)$ in the point $f(y)$, so that $f(x) \in \overline{f(A)}$, as desired.

(2) $\Rightarrow$ (3). Let B be closed in Y and let $A = f^{-1}(B)$. We wish to prove that A is closed in X; we show that $\bar{A} = A$. By elementary set theory, we have $f(A) = f(f^{-1}(B)) \subset B$. Therefore, if $x \in \bar{A}$,

$$f(x) \in f(\bar{A}) \subset \overline{f(A)} \subset \bar{B} = B,$$

so that $x \in f^{-1}(B) = A$. Thus $\bar{A} \subset A$, so that $\bar{A} = A$, as desired.

(3) $\Rightarrow$ (1). Let V be an open set of Y. Set $B = Y - V$. Then

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

Now B is a closed set of Y. Then $f^{-1}(B)$ is closed in X by hypothesis, so that $f^{-1}(V)$ is open in X, as desired.

(1) $\Rightarrow$ (4). Let $x \in X$ and let V be a neighborhood of $f(x)$. Then the set $U = f^{-1}(V)$ is a neighborhood of x such that $f(U) \subset V$.

(4) $\Rightarrow$ (1). Let V be an open set of Y; let x be a point of $f^{-1}(V)$. Then $f(x) \in V$, so that by hypothesis there is a neighborhood U_x of x such that $f(U_x) \subset V$. Then $U_x \subset f^{-1}(V)$. It follows that $f^{-1}(V)$ can be written as the union of the open sets U_x, so that it is open. ∎

Homeomorphisms

Let X and Y be topological spaces; let $f : X \to Y$ be a bijection. If both the function f and the inverse function

$$f^{-1} : Y \to X$$

are continuous, then f is called a ***homeomorphism***.

The condition that f^{-1} be continuous says that for each open set U of X, the inverse image of U under the map $f^{-1} : Y \to X$ is open in Y. But the *inverse image* of U under the map f^{-1} is the same as the *image* of U under the map f. See Figure 18.1. So another way to define a homeomorphism is to say that it is a bijective correspondence $f : X \to Y$ such that $f(U)$ is open if and only if U is open.

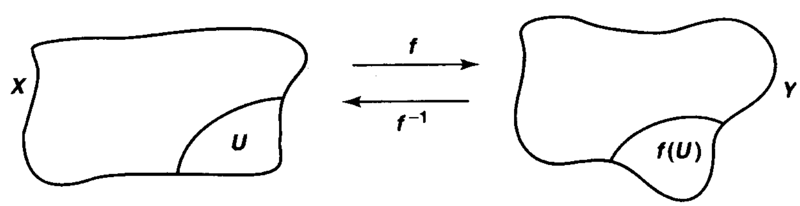


Figure 18.1

This remark shows that a homeomorphism $f : X \to Y$ gives us a bijective correspondence not only between X and Y but between the collections of open sets of X and of Y. As a result, any property of X that is entirely expressed in terms of the topology of X (that is, in terms of the open sets of X) yields, via the correspondence f, the corresponding property for the space Y. Such a property of X is called a ***topological property*** of X.

You may have studied in modern algebra the notion of an *isomorphism* between algebraic objects such as groups or rings. An isomorphism is a bijective correspondence that preserves the algebraic structure involved. The analogous concept in topology is that of *homeomorphism*; it is a bijective correspondence that preserves the topological structure involved.

Now suppose that $f : X \to Y$ is an injective continuous map, where X and Y are topological spaces. Let Z be the image set $f(X)$, considered as a subspace of Y; then the function $f' : X \to Z$ obtained by restricting the range of f is bijective. If f' happens to be a homeomorphism of X with Z, we say that the map $f : X \to Y$ is a ***topological imbedding***, or simply an ***imbedding***, of X in Y.

EXAMPLE 4. The function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = 3x + 1$ is a homeomorphism. See Figure 18.2. If we define $g : \mathbb{R} \to \mathbb{R}$ by the equation

$$g(y) = \frac{1}{3}(y - 1)$$

then one can check easily that $f(g(y)) = y$ and $g(f(x)) = x$ for all real numbers x and y. It follows that f is bijective and that $g = f^{-1}$; the continuity of f and g is a familiar result from calculus.

EXAMPLE 5. The function $F : (-1, 1) \to \mathbb{R}$ defined by

$$F(x) = \frac{x}{1 - x^2}$$

is a homeomorphism. See Figure 18.3. We have already noted in Example 9 of §3 that F is a bijective order-preserving correspondence; its inverse is the function G defined by

$$G(y) = \frac{2y}{1 + (1 + 4y^2)^{1/2}}.$$

The fact that F is a homeomorphism can be proved in two ways. One way is to note that because F is order preserving and bijective, F carries a basis element for the order topology in $(-1, 1)$ onto a basis element for the order topology in $\mathbb{R}$ and vice versa. As a result, F is automatically a homeomorphism of $(-1, 1)$ with $\mathbb{R}$ (both in the order topology). Since the order topology on $(-1, 1)$ and the usual (subspace) topology agree, F is a homeomorphism of $(-1, 1)$ with $\mathbb{R}$.

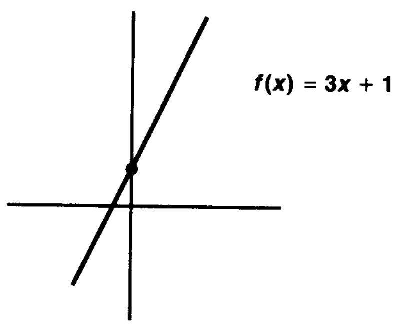


Figure 18.2

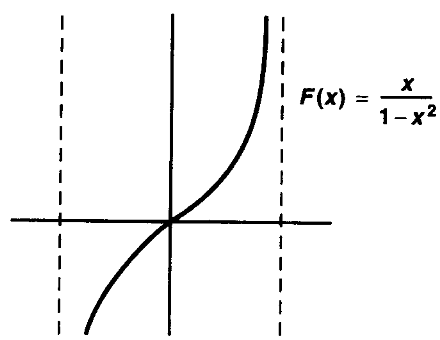


Figure 18.3

A second way to show F a homeomorphism is to use the continuity of the algebraic functions and the square-root function to show that both F and G are continuous. These are familiar facts from calculus.

EXAMPLE 6. A bijective function $f : X \to Y$ can be continuous without being a homeomorphism. One such function is the identity map $g : \mathbb{R}_\ell \to \mathbb{R}$ considered in Example 3. Another is the following: Let S^1 denote the ***unit circle***,

$$S^1 = \{x \times y \mid x^2 + y^2 = 1\},$$

considered as a subspace of the plane $\mathbb{R}^2$, and let

$$F : [0, 1) \longrightarrow S^1$$

be the map defined by $f(t) = (\cos 2\pi t, \sin 2\pi t)$. The fact that f is bijective and continuous follows from familiar properties of the trigonometric functions. But the function f^{-1} is not continuous. The image under f of the open set $U = [0, \frac{1}{4})$ of the domain, for instance, is not open in S^1, for the point $p = f(0)$ lies in no open set V of $\mathbb{R}^2$ such that $V \cap S^1 \subset f(U)$. See Figure 18.4.

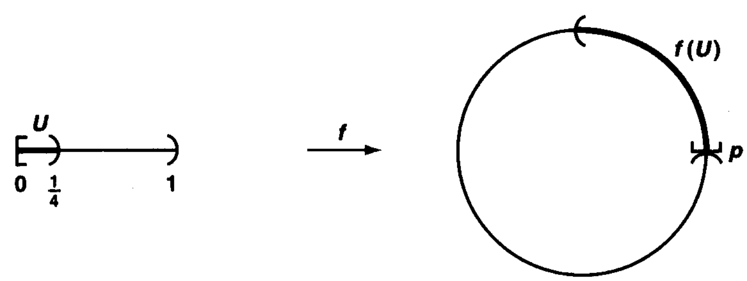


Figure 18.4

EXAMPLE 7. Consider the function

$$g : [0, 1) \longrightarrow \mathbb{R}^2$$

obtained from the function f of the preceding example by expanding the range. The map g is an example of a continuous injective map that is not an imbedding.

Constructing Continuous Functions

How does one go about constructing continuous functions from one topological space to another? There are a number of methods used in analysis, of which some generalize to arbitrary topological spaces and others do not. We study first some constructions that do hold for general topological spaces, deferring consideration of the others until later.

Theorem 18.2 (Rules for constructing continuous functions). *Let X, Y, and Z be topological spaces.*

(a) *(Constant function) If $f : X \to Y$ maps all of X into the single point y_0 of Y, then f is continuous.*

(b) *(Inclusion) If A is a subspace of X, the inclusion function $j : A \to X$ is continuous.*

(c) *(Composites) If $f : X \to Y$ and $g : Y \to Z$ are continuous, then the map $g \circ f : X \to Z$ is continuous.*

(d) *(Restricting the domain) If $f : X \to Y$ is continuous, and if A is a subspace of X, then the restricted function $f|A : A \to Y$ is continuous.*

(e) *(Restricting or expanding the range) Let $f : X \to Y$ be continuous. If Z is a subspace of Y containing the image set $f(X)$, then the function $g : X \to Z$ obtained by restricting the range of f is continuous. If Z is a space having Y as a subspace, then the function $h : X \to Z$ obtained by expanding the range of f is continuous.*

(f) *(Local formulation of continuity) The map $f : X \to Y$ is continuous if X can be written as the union of open sets U_α such that $f|U_\alpha$ is continuous for each α.*

Proof. (a) Let $f(x) = y_0$ for every x in X. Let V be open in Y. The set $f^{-1}(V)$ equals X or $\varnothing$, depending on whether V contains y_0 or not. In either case, it is open.

(b) If U is open in X, then $j^{-1}(U) = U \cap A$, which is open in A by definition of the subspace topology.

(c) If U is open in Z, then $g^{-1}(U)$ is open in Y and $f^{-1}(g^{-1}(U))$ is open in X. But

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

by elementary set theory.

(d) The function $f|A$ equals the composite of the inclusion map $j : A \to X$ and the map $f : X \to Y$, both of which are continuous.

(e) Let $f : X \to Y$ be continuous. If $f(X) \subset Z \subset Y$, we show that the function $g : X \to Z$ obtained from f is continuous. Let B be open in Z. Then $B = Z \cap U$ for some open set U of Y. Because Z contains the entire image set $f(X)$,

$$f^{-1}(U) = g^{-1}(B),$$

by elementary set theory. Since $f^{-1}(U)$ is open, so is $g^{-1}(B)$.

To show $h : X \to Z$ is continuous if Z has Y as a subspace, note that h is the composite of the map $f : X \to Y$ and the inclusion map $j : Y \to Z$.

(f) By hypothesis, we can write X as a union of open sets U_α, such that $f|U_\alpha$ is continuous for each α. Let V be an open set in Y. Then

$$f^{-1}(V) \cap U_\alpha = (f|U_\alpha)^{-1}(V),$$

because both expressions represent the set of those points x lying in U_α for which $f(x) \in V$. Since $f|U_\alpha$ is continuous, this set is open in U_α, and hence open in X. But

$$f^{-1}(V) = \bigcup_\alpha (f^{-1}(V) \cap U_\alpha),$$

so that $f^{-1}(V)$ is also open in X. ∎

Theorem 18.3 (The pasting lemma). *Let $X = A \cup B$, where A and B are closed in X. Let $f : A \to Y$ and $g : B \to Y$ be continuous. If $f(x) = g(x)$ for every $x \in A \cap B$, then f and g combine to give a continuous function $h : X \to Y$, defined by setting $h(x) = f(x)$ if $x \in A$, and $h(x) = g(x)$ if $x \in B$.*

Proof. Let C be a closed subset of Y. Now

$$h^{-1}(C) = f^{-1}(C) \cup g^{-1}(C),$$

by elementary set theory. Since f is continuous, $f^{-1}(C)$ is closed in A and, therefore, closed in X. Similarly, $g^{-1}(C)$ is closed in B and therefore closed in X. Their union $h^{-1}(C)$ is thus closed in X. ∎

This theorem also holds if A and B are open sets in X; this is just a special case of the "local formulation of continuity" rule given in preceding theorem.

EXAMPLE 8. Let us define a function $h : \mathbb{R} \to \mathbb{R}$ by setting

$$h(x) = \begin{cases} x & \text{for } x \leq 0, \\ x/2 & \text{for } x \geq 0. \end{cases}$$

Each of the "pieces" of this definition is a continuous function, and they agree on the overlapping part of their domains, which is the one-point set $\{0\}$. Since their domains are closed in $\mathbb{R}$, the function h is continuous. One needs the "pieces" of the function to agree on the overlapping part of their domains in order to have a function at all. The equations

$$k(x) = \begin{cases} x - 2 & \text{for } x \leq 0, \\ x + 2 & \text{for } x \geq 0, \end{cases}$$

for instance, do not define a function. On the other hand, one needs some limitations on the sets A and B to guarantee continuity. The equations

$$l(x) = \begin{cases} x - 2 & \text{for } x < 0, \\ x + 2 & \text{for } x \geq 0, \end{cases}$$

for instance, do define a function l mapping $\mathbb{R}$ into $\mathbb{R}$, and both of the pieces are continuous. But l is not continuous; the inverse image of the open set $(1, 3)$, for instance, is the nonopen set $[0, 1)$. See Figure 18.5.

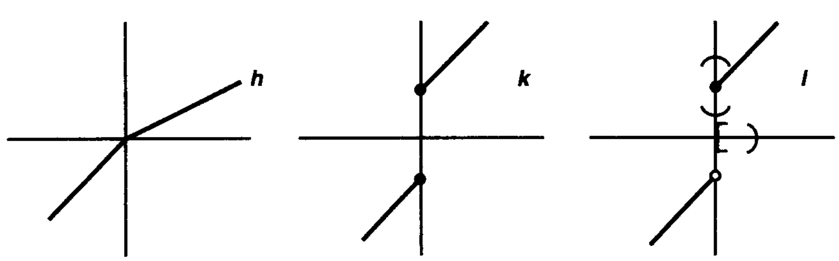


Figure 18.5

Theorem 18.4 (Maps into products). *Let $f : A \to X \times Y$ be given by the equation*

$$f(a) = (f_1(a), f_2(a)).$$

Then f is continuous if and only if the functions

$$f_1 : A \longrightarrow X \quad \text{and} \quad f_2 : A \longrightarrow Y$$

are continuous.

The maps f_1 and f_2 are called the ***coordinate functions*** of f.

Proof. Let $\pi_1 : X \times Y \to X$ and $\pi_2 : X \times Y \to Y$ be projections onto the first and second factors, respectively. These maps are continuous. For $\pi_1^{-1}(U) = U \times Y$ and $\pi_2^{-1}(V) = X \times V$, and these sets are open if U and V are open. Note that for each $a \in A$,

$$f_1(a) = \pi_1(f(a)) \quad \text{and} \quad f_2(a) = \pi_2(f(a)).$$

If the function f is continuous, then f_1 and f_2 are composites of continuous functions and therefore continuous. Conversely, suppose that f_1 and f_2 are continuous. We show that for each basis element $U \times V$ for the topology of $X \times Y$, its inverse image $f^{-1}(U \times V)$ is open. A point a is in $f^{-1}(U \times V)$ if and only if $f(a) \in U \times V$, that is, if and only if $f_1(a) \in U$ and $f_2(a) \in V$. Therefore,

$$f^{-1}(U \times V) = f_1^{-1}(U) \cap f_2^{-1}(V).$$

Since both of the sets $f_1^{-1}(U)$ and $f_2^{-1}(V)$ are open, so is their intersection. ∎

There is no useful criterion for the continuity of a map $f : A \times B \to X$ whose *domain* is a product space. One might conjecture that f is continuous if it is continuous "in each variable separately," but this conjecture is not true. (See Exercise 12.)

EXAMPLE 9. In calculus, a *parametrized curve* in the plane is defined to be a continuous map $f : [a, b] \to \mathbb{R}^2$. It is often expressed in the form $f(t) = (x(t), y(t))$; and one frequently uses the fact that f is a continuous function of t if both x and y are. Similarly, a *vector field* in the plane

$$\begin{aligned}\mathbf{v}(x, y) &= P(x, y)\mathbf{i} + Q(x, y)\mathbf{j} \\ &= (P(x, y), Q(x, y))\end{aligned}$$

is said to be continuous if both P and Q are continuous functions, or equivalently, if $\mathbf{v}$ is continuous as a map of $\mathbb{R}^2$ into $\mathbb{R}^2$. Both of these statements are simply special cases of the preceding theorem.

One way of forming continuous functions that is used a great deal in analysis is to take sums, differences, products, or quotients of continuous real-valued functions. It is a standard theorem that if $f, g : X \to \mathbb{R}$ are continuous, then $f + g$, $f - g$, and $f \cdot g$ are continuous, and f/g is continuous if $g(x) \neq 0$ for all x. We shall consider this theorem in §21.

Yet another method for constructing continuous functions that is familiar from analysis is to take the limit of an infinite sequence of functions. There is a theorem to the effect that if a sequence of continuous real-valued functions of a real variable converges uniformly to a limit function, then the limit function is necessarily continuous. This theorem is called the *Uniform Limit Theorem*. It is used, for instance, to demonstrate the continuity of the trigonometric functions, when one defines these functions rigorously using the infinite series definitions of the sine and cosine. This theorem generalizes to a theorem about maps of an arbitrary topological space X into a metric space Y. We shall prove it in §21.

Exercises

1. Prove that for functions $f : \mathbb{R} \to \mathbb{R}$, the ϵ-δ definition of continuity implies the open set definition.

2. Suppose that $f : X \to Y$ is continuous. If x is a limit point of the subset A of X, is it necessarily true that $f(x)$ is a limit point of $f(A)$?

3. Let X and X' denote a single set in the two topologies $\mathcal{T}$ and $\mathcal{T}'$, respectively. Let $i : X' \to X$ be the identity function.
(a) Show that i is continuous $\Leftrightarrow$ $\mathcal{T}'$ is finer than $\mathcal{T}$.
(b) Show that i is a homeomorphism $\Leftrightarrow$ $\mathcal{T}' = \mathcal{T}$.

4. Given $x_0 \in X$ and $y_0 \in Y$, show that the maps $f : X \to X \times Y$ and $g : Y \to X \times Y$ defined by

$$f(x) = x \times y_0 \qquad \text{and} \qquad g(y) = x_0 \times y$$

are imbeddings.

5. Show that the subspace (a, b) of $\mathbb{R}$ is homeomorphic with $(0, 1)$ and the subspace $[a, b]$ of $\mathbb{R}$ is homeomorphic with $[0, 1]$.

6. Find a function $f : \mathbb{R} \to \mathbb{R}$ that is continuous at precisely one point.

7. (a) Suppose that $f : \mathbb{R} \to \mathbb{R}$ is "continuous from the right," that is,

$$\lim_{x \to a^+} f(x) = f(a),$$

for each $a \in \mathbb{R}$. Show that f is continuous when considered as a function from $\mathbb{R}_\ell$ to $\mathbb{R}$.
(b) Can you conjecture what functions $f : \mathbb{R} \to \mathbb{R}$ are continuous when considered as maps from $\mathbb{R}$ to $\mathbb{R}_\ell$? As maps from $\mathbb{R}_\ell$ to $\mathbb{R}_\ell$? We shall return to this question in Chapter 3.

8. Let Y be an ordered set in the order topology. Let $f, g : X \to Y$ be continuous.
(a) Show that the set $\{x \mid f(x) \leq g(x)\}$ is closed in X.

(b) Let $h : X \to Y$ be the function

$$h(x) = \min\{f(x), g(x)\}.$$

Show that h is continuous. [*Hint:* Use the pasting lemma.]

9. Let $\{A_\alpha\}$ be a collection of subsets of X; let $X = \bigcup_\alpha A_\alpha$. Let $f : X \to Y$; suppose that $f|A_\alpha$, is continuous for each α.

(a) Show that if the collection $\{A_\alpha\}$ is finite and each set A_α is closed, then f is continuous.

(b) Find an example where the collection $\{A_\alpha\}$ is countable and each A_α is closed, but f is not continuous.

(c) An indexed family of sets $\{A_\alpha\}$ is said to be ***locally finite*** if each point x of X has a neighborhood that intersects A_α for only finitely many values of α. Show that if the family $\{A_\alpha\}$ is locally finite and each A_α is closed, then f is continuous.

10. Let $f : A \to B$ and $g : C \to D$ be continuous functions. Let us define a map $f \times g : A \times C \to B \times D$ by the equation

$$(f \times g)(a \times c) = f(a) \times g(c).$$

Show that $f \times g$ is continuous.

11. Let $F : X \times Y \to Z$. We say that F is ***continuous in each variable separately*** if for each y_0 in Y, the map $h : X \to Z$ defined by $h(x) = F(x \times y_0)$ is continuous, and for each x_0 in X, the map $k : Y \to Z$ defined by $k(y) = F(x_0 \times y)$ is continuous. Show that if F is continuous, then F is continuous in each variable separately.

12. Let $F : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be defined by the equation

$$F(x \times y) = \begin{cases} xy/(x^2 + y^2) & \text{if } x \times y \neq 0 \times 0. \\ 0 & \text{if } x \times y = 0 \times 0. \end{cases}$$

(a) Show that F is continuous in each variable separately.

(b) Compute the function $g : \mathbb{R} \to \mathbb{R}$ defined by $g(x) = F(x \times x)$.

(c) Show that F is not continuous.

13. Let $A \subset X$; let $f : A \to Y$ be continuous; let Y be Hausdorff. Show that if f may be extended to a continuous function $g : \bar{A} \to Y$, then g is uniquely determined by f.

§19 The Product Topology

We now return, for the remainder of the chapter, to the consideration of various methods for imposing topologies on sets.

Previously, we defined a topology on the product $X \times Y$ of two topological spaces. In the present section, we generalize this definition to more general cartesian products.

So let us consider the cartesian products

$$X_1 \times \cdots \times X_n \quad \text{and} \quad X_1 \times X_2 \times \cdots,$$

where each X_i is a topological space. There are two possible ways to proceed. One way is to take as basis all sets of the form $U_1 \times \cdots \times U_n$ in the first case, and of the form $U_1 \times U_2 \times \cdots$ in the second case, where U_i is an open set of X_i for each i. This procedure does indeed define a topology on the cartesian product; we shall call it the *box topology*.

Another way to proceed is to generalize the subbasis formulation of the definition, given in §15. In this case, we take as a subbasis all sets of the form $\pi_i^{-1}(U_i)$, where i is any index and U_i is an open set of X_i. We shall call this topology the *product topology*.

How do these topologies differ? Consider the typical basis element B for the second topology. It is a finite intersection of subbasis elements $\pi_i^{-1}(U_i)$, say for $i = i_1, \dots, i_k$. Then a point $\mathbf{x}$ belongs to B if and only if $\pi_i(\mathbf{x})$ belongs to U_i for $i = i_1, \dots, i_k$; there is no restriction on $\pi_i(x)$ for other values of i.

It follows that these two topologies agree for the finite cartesian product and differ for the infinite product. What is not clear is why we seem to prefer the second topology. This is the question we shall explore in this section.

Before proceeding, however, we shall introduce a more general notion of cartesian product. So far, we have defined the cartesian product of an indexed family of sets only in the cases where the index set was the set $\{1, \dots, n\}$ or the set $\mathbb{Z}_+$. Now we consider the case where the index set is completely arbitrary.

Definition. Let J be an index set. Given a set X, we define a ***J-tuple*** of elements of X to be a function $\mathbf{x} : J \to X$. If α is an element of J, we often denote the value of $\mathbf{x}$ at α by x_α rather than $\mathbf{x}(\alpha)$; we call it the αth ***coordinate*** of $\mathbf{x}$. And we often denote the function $\mathbf{x}$ itself by the symbol

$$(x_\alpha)_{\alpha \in J},$$

which is as close as we can come to a "tuple notation" for an arbitrary index set J. We denote the set of all J-tuples of elements of X by X^J.

Definition. Let $\{A_\alpha\}_{\alpha \in J}$ be an indexed family of sets; let $X = \bigcup_{\alpha \in J} A_\alpha$. The ***cartesian product*** of this indexed family, denoted by

$$\prod_{\alpha \in J} A_\alpha,$$

is defined to be the set of all J-tuples $(x_\alpha)_{\alpha \in J}$ of elements of X such that $x_\alpha \in A_\alpha$ for each $\alpha \in J$. That is, it is the set of all functions

$$\mathbf{x} : J \to \bigcup_{\alpha \in J} A_\alpha$$

such that $\mathbf{x}(\alpha) \in A_\alpha$ for each $\alpha \in J$.

Occasionally we denote the product simply by $\prod A_\alpha$, and its general element by (x_α), if the index set is understood.

If all the sets A_α are equal to one set X, then the cartesian product $\prod_{\alpha\in J} A_\alpha$ is just the set X^J of *all* J-tuples of elements of X. We sometimes use "tuple notation" for the elements of X^J, and sometimes we use functional notation, depending on which is more convenient.

Definition. Let $\{X_\alpha\}_{\alpha\in J}$ be an indexed family of topological spaces. Let us take as a basis for a topology on the product space

$$\prod_{\alpha\in J} X_\alpha$$

the collection of all sets of the form

$$\prod_{\alpha\in J} U_\alpha,$$

where U_α is open in X_α, for each $\alpha \in J$. The topology generated by this basis is called the ***box topology***.

This collection satisfies the first condition for a basis because $\prod X_\alpha$ is itself a basis element; and it satisfies the second condition because the intersection of any two basis elements is another basis element:

$$\Big(\prod_{\alpha\in J} U_\alpha\Big) \cap \Big(\prod_{\alpha\in J} V_\alpha\Big) = \prod_{\alpha\in J}(U_\alpha \cap V_\alpha).$$

Now we generalize the subbasis formulation of the definition. Let

$$\pi_\beta : \prod_{\alpha\in J} X_\alpha \to X_\beta$$

be the function assigning to each element of the product space its βth coordinate,

$$\pi_\beta((x_\alpha)_{\alpha\in J}) = x_\beta;$$

it is called the ***projection mapping*** associated with the index β.

Definition. Let $\mathcal{S}_\beta$ denote the collection

$$\mathcal{S}_\beta = \{\pi_\beta^{-1}(U_\beta) \mid U_\beta \text{ open in } X_\beta\},$$

and let $\mathcal{S}$ denote the union of these collections,

$$\mathcal{S} = \bigcup_{\beta\in J} \mathcal{S}_\beta.$$

The topology generated by the subbasis $\mathcal{S}$ is called the ***product topology***. In this topology $\prod_{\alpha\in J} X_\alpha$ is called a ***product space***.

To compare these topologies, we consider the basis $\mathcal{B}$ that $\mathcal{S}$ generates. The collection $\mathcal{B}$ consists of all finite intersections of elements of $\mathcal{S}$. If we intersect elements belonging to the same one of the sets $\mathcal{S}_\beta$, we do not get anything new, because

$$\pi_\beta^{-1}(U_\beta) \cap \pi_\beta^{-1}(V_\beta) = \pi_\beta^{-1}(U_\beta \cap V_\beta);$$

the intersection of two elements of $\mathcal{S}_\beta$, or of finitely many such elements, is again an element of $\mathcal{S}_\beta$. We get something new only when we intersect elements from different sets $\mathcal{S}_\beta$. The typical element of the basis $\mathcal{B}$ can thus be described as follows: Let β_1, $\dots$, β_n be a finite set of distinct indices from the index set J, and let U_{β_i} be an open set in X_{β_i} for $i = 1, \dots, n$. Then

$$B = \pi_{\beta_1}^{-1}(U_{\beta_1}) \cap \pi_{\beta_2}^{-1}(U_{\beta_2}) \cap \cdots \cap \pi_{\beta_n}^{-1}(U_{\beta_n})$$

is the typical element of $\mathcal{B}$.

Now a point $\mathbf{x} = (x_\alpha)$ is in B if and only if its β_1th coordinate is in U_{β_1}, its β_2th coordinate is in U_{β_2}, and so on. There is no restriction whatever on the αth coordinate of $\mathbf{x}$ if α is not one of the indices $\beta_1, \dots, \beta_n$. As a result, we can write B as the product

$$B = \prod_{\alpha \in J} U_\alpha,$$

where U_α denotes the entire space X_α if $\alpha \neq \beta_1, \dots, \beta_n$.

All this is summarized in the following theorem:

Theorem 19.1 (Comparison of the box and product topologies). *The box topology on $\prod X_\alpha$ has as basis all sets of the form $\prod U_\alpha$, where U_α is open in X_α for each α. The product topology on $\prod X_\alpha$ has as basis all sets of the form $\prod U_\alpha$, where U_α is open in X_α for each α and U_α equals X_α except for finitely many values of α.*

Two things are immediately clear. First, for finite products $\prod_{\alpha=1}^{n} X_\alpha$ the two topologies are precisely the same. Second, the box topology is in general finer than the product topology.

What is not so clear is why we prefer the product topology to the box topology. The answer will appear as we continue our study of topology. We shall find that a number of important theorems about finite products will also hold for arbitrary products if we use the product topology, but not if we use the box topology. As a result, the product topology is extremely important in mathematics. The box topology is not so important; we shall use it primarily for constructing counterexamples. Therefore, we make the following convention:

> *Whenever we consider the product $\prod X_\alpha$, we shall assume it is given the product topology unless we specifically state otherwise.*

Some of the theorems we proved for the product $X \times Y$ hold for the product $\prod X_\alpha$ no matter which topology we use. We list them here; most of the proofs are left to the exercises.

Theorem 19.2. *Suppose the topology on each space X_α is given by a basis $\mathcal{B}_\alpha$. The collection of all sets of the form*

$$\prod_{\alpha \in J} B_\alpha,$$

where $B_\alpha \in \mathcal{B}_\alpha$ for each α, will serve as a basis for the box topology on $\prod_{\alpha \in J} X_\alpha$.

The collection of all sets of the same form, where $B_\alpha \in \mathcal{B}_\alpha$ for finitely many indices α and $B_\alpha = X_\alpha$ for all the remaining indices, will serve as a basis for the product topology $\prod_{\alpha \in J} X_\alpha$.

EXAMPLE 1. Consider euclidean n-space $\mathbb{R}^n$. A basis for $\mathbb{R}$ consists of all open intervals in $\mathbb{R}$; hence a basis for the topology of $\mathbb{R}^n$ consists of all products of the form

$$(a_1, b_1) \times (a_2, b_2) \times \cdots \times (a_n, b_n).$$

Since $\mathbb{R}^n$ is a finite product, the box and product topologies agree. Whenever we consider $\mathbb{R}^n$, we will assume that it is given this topology, unless we specifically state otherwise.

Theorem 19.3. *Let A_α be a subspace of X_α, for each $\alpha \in J$. Then $\prod A_\alpha$ is a subspace of $\prod X_\alpha$ if both products are given the box topology, or if both products are given the product topology.*

Theorem 19.4. *If each space X_α is a Hausdorff space, then $\prod X_\alpha$ is a Hausdorff space in both the box and product topologies.*

Theorem 19.5. *Let $\{X_\alpha\}$ be an indexed family of spaces; let $A_\alpha \subset X_\alpha$ for each α. If $\prod X_\alpha$ is given either the product or the box topology, then*

$$\prod \bar{A}_\alpha = \overline{\prod A_\alpha}.$$

Proof. Let $\mathbf{x} = (x_\alpha)$ be a point of $\prod \bar{A}_\alpha$; we show that $\mathbf{x} \in \overline{\prod A_\alpha}$. Let $U = \prod U_\alpha$ be a basis element for either the box or product topology that contains $\mathbf{x}$. Since $x_\alpha \in \bar{A}_\alpha$, we can choose a point $y_\alpha \in U_\alpha \cap A_\alpha$ for each α. Then $\mathbf{y} = (y_\alpha)$ belongs to both U and $\prod A_\alpha$. Since U is arbitrary, it follows that $\mathbf{x}$ belongs to the closure of $\prod A_\alpha$.

Conversely, suppose $\mathbf{x} = (x_\alpha)$ lies in the closure of $\prod A_\alpha$, in either topology. We show that for any given index β, we have $x_\beta \in \bar{A}_\beta$. Let V_β be an arbitrary open set of X_β containing x_β. Since $\pi_\beta^{-1}(V_\beta)$ is open in $\prod X_\alpha$ in either topology, it contains a point $\mathbf{y} = (y_\alpha)$ of $\prod A_\alpha$. Then y_β belongs to $V_\beta \cap A_\beta$. It follows that $x_\beta \in \bar{A}_\beta$. ∎

So far, no reason has appeared for preferring the product to the box topology. It is when we try to generalize our previous theorem about continuity of maps into product spaces that a difference first arises. Here is a theorem that does not hold if $\prod X_\alpha$ is given the box topology:

Theorem 19.6. *Let $f : A \to \prod_{\alpha \in J} X_\alpha$ be given by the equation*

$$f(a) = (f_\alpha(a))_{\alpha \in J},$$

where $f_\alpha : A \to X_\alpha$ for each α. Let $\prod X_\alpha$ have the product topology. Then the function f is continuous if and only if each function f_α is continuous.

Proof. Let π_β be the projection of the product onto its βth factor. The function π_β is continuous, for if U_β is open in X_β, the set $\pi_\beta^{-1}(U_\beta)$ is a subbasis element for the product topology on X_α. Now suppose that $f : A \to \prod X_\alpha$ is continuous. The function f_β equals the composite $\pi_\beta \circ f$; being the composite of two continuous functions, it is continuous.

Conversely, suppose that each coordinate function f_α is continuous. To prove that f is continuous, it suffices to prove that the inverse image under f of each subbasis element is open in A; we remarked on this fact when we defined continuous functions. A typical subbasis element for the product topology on $\prod X_\alpha$ is a set of the form $\pi_\beta^{-1}(U_\beta)$, where β is some index and U_β is open in X_β. Now

$$f^{-1}(\pi_\beta^{-1}(U_\beta)) = f_\beta^{-1}(U_\beta),$$

because $f_\beta = \pi_\beta \circ f$. Since f_β is continuous, this set is open in A, as desired. ∎

Why does this theorem fail if we use the box topology? Probably the most convincing thing to do is to look at an example.

EXAMPLE 2. Consider $\mathbb{R}^\omega$, the countably infinite product of $\mathbb{R}$ with itself. Recall that

$$\mathbb{R}^\omega = \prod_{n \in \mathbb{Z}_+} X_n,$$

where $X_n = \mathbb{R}$ for each n. Let us define a function $f : \mathbb{R} \to \mathbb{R}^\omega$ by the equation

$$f(t) = (t, t, t, \ldots);$$

the nth coordinate function of f is the function $f_n(t) = t$. Each of the coordinate functions $f_n : \mathbb{R} \to \mathbb{R}$ is continuous; therefore, the function f is continuous if $\mathbb{R}^\omega$ is given the product topology. But f is not continuous if $\mathbb{R}^\omega$ is given the box topology. Consider, for example, the basis element

$$B = (-1, 1) \times (-\frac{1}{2}, \frac{1}{2}) \times (-\frac{1}{3}, \frac{1}{3}) \times \cdots$$

for the box topology. We assert that $f^{-1}(B)$ is not open in $\mathbb{R}$. If $f^{-1}(B)$ were open in $\mathbb{R}$, it would contain some interval $(-\delta, \delta)$ about the point 0. This would mean that $f((-\delta, \delta)) \subset B$, so that, applying π_n to both sides of the inclusion,

$$f_n((-\delta, \delta)) = (-\delta, \delta) \subset (-1/n, 1/n)$$

for all n, a contradiction.

Exercises

1. Prove Theorem 19.2.
2. Prove Theorem 19.3.
3. Prove Theorem 19.4.
4. Show that $(X_1 \times \cdots \times X_{n-1}) \times X_n$ is homeomorphic with $X_1 \times \cdots \times X_n$.
5. One of the implications stated in Theorem 19.6 holds for the box topology. Which one?
6. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots$ be a sequence of the points of the product space $\prod X_\alpha$. Show that this sequence converges to the point $\mathbf{x}$ if and only if the sequence $\pi_\alpha(\mathbf{x}_1), \pi_\alpha(\mathbf{x}_2), \ldots$ converges to $\pi_\alpha(\mathbf{x})$ for each α. Is this fact true if one uses the box topology instead of the product topology?
7. Let $\mathbb{R}^\infty$ be the subset of $\mathbb{R}^\omega$ consisting of all sequences that are "eventually zero," that is, all sequences $(x_1, x_2, \ldots)$ such that $x_i \neq 0$ for only finitely many values of i. What is the closure of $\mathbb{R}^\infty$ in $\mathbb{R}^\omega$ in the box and product topologies? Justify your answer.
8. Given sequences $(a_1, a_2, \ldots)$ and $(b_1, b_2, \ldots)$ of real numbers with $a_i > 0$ for all i, define $h : \mathbb{R}^\omega \to \mathbb{R}^\omega$ by the equation
$$h((x_1, x_2, \ldots)) = (a_1 x_1 + b_1, a_2 x_2 + b_2, \ldots).$$
Show that if $\mathbb{R}^\omega$ is given the product topology, h is a homeomorphism of $\mathbb{R}^\omega$ with itself. What happens if $\mathbb{R}^\omega$ is given the box topology?
9. Show that the choice axiom is equivalent to the statement that for any indexed family $\{A_\alpha\}_{\alpha \in J}$ of nonempty sets, with $J \neq 0$, the cartesian product
$$\prod_{\alpha \in J} A_\alpha$$
is not empty.
10. Let A be a set; let $\{X_\alpha\}_{\alpha \in J}$ be an indexed family of spaces; and let $\{f_\alpha\}_{\alpha \in J}$ be an indexed family of functions $f_\alpha : A \to X_\alpha$.
    (a) Show there is a unique coarsest topology $\mathcal{T}$ on A relative to which each of the functions f_α is continuous.
    (b) Let
    $$\mathcal{S}_\beta = \{f_\beta^{-1}(U_\beta) \mid U_\beta \text{ is open in } X_\beta\},$$
    and let $\mathcal{S} = \bigcup \mathcal{S}_\beta$. Show that $\mathcal{S}$ is a subbasis for $\mathcal{T}$.
    (c) Show that a map $g : Y \to A$ is continuous relative to $\mathcal{T}$ if and only if each map $f_\alpha \circ g$ is continuous.
    (d) Let $f : A \to \prod X_\alpha$ be defined by the equation
    $$f(a) = (f_\alpha(a))_{\alpha \in J};$$
    let Z denote the subspace $f(A)$ of the product space $\prod X_\alpha$. Show that the image under f of each element of $\mathcal{T}$ is an open set of Z.

§20 The Metric Topology

One of the most important and frequently used ways of imposing a topology on a set is to define the topology in terms of a metric on the set. Topologies given in this way lie at the heart of modern analysis, for example. In this section, we shall define the metric topology and shall give a number of examples. In the next section, we shall consider some of the properties that metric topologies satisfy.

Definition. A ***metric*** on a set X is a function

$$d : X \times X \longrightarrow R$$

having the following properties:

(1) $d(x, y) \geq 0$ for all $x, y \in X$; equality holds if and only if $x = y$.

(2) $d(x, y) = d(y, x)$ for all $x, y \in X$.

(3) (Triangle inequality) $d(x, y) + d(y, z) \geq d(x, z)$, for all $x, y, z \in X$.

Given a metric d on X, the number $d(x, y)$ is often called the ***distance*** between x and y in the metric d. Given $\epsilon > 0$, consider the set

$$B_d(x, \epsilon) = \{y \mid d(x, y) < \epsilon\}$$

of all points y whose distance from x is less than ϵ. It is called the ***ϵ-ball centered at x***. Sometimes we omit the metric d from the notation and write this ball simply as $B(x, \epsilon)$, when no confusion will arise.

Definition. If d is a metric on the set X, then the collection of all ϵ-balls $B_d(x, \epsilon)$, for $x \in X$ and $\epsilon > 0$, is a basis for a topology on X, called the ***metric topology*** induced by d.

The first condition for a basis is trivial, since $x \in B(x, \epsilon)$ for any $\epsilon > 0$. Before checking the second condition for a basis, we show that if y is a point of the basis element $B(x, \epsilon)$, then there is a basis element $B(y, \delta)$ *centered* at y that is contained in $B(x, \epsilon)$. Define δ to be the positive number $\epsilon - d(x, y)$. Then $B(y, \delta) \subset B(x, \epsilon)$, for if $z \in B(y, \delta)$, then $d(y, z) < \epsilon - d(x, y)$, from which we conclude that

$$d(x, z) \leq d(x, y) + d(y, z) < \epsilon.$$

See Figure 20.1.

Now to check the second condition for a basis, let B_1 and B_2 be two basis elements and let $y \in B_1 \cap B_2$. We have just shown that we can choose positive numbers δ_1 and δ_2 so that $B(y, \delta_1) \subset B_1$ and $B(y, \delta_2) \subset B_2$. Letting δ be the smaller of δ_1 and δ_2, we conclude that $B(y, \delta) \subset B_1 \cap B_2$.

Using what we have just proved, we can rephrase the definition of the metric topology as follows:

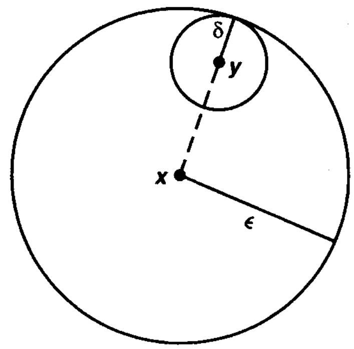


Figure 20.1

A set U is open in the metric topology induced by d if and only if for each $y \in U$, there is a $\delta > 0$ such that $B_d(y, \delta) \subset U$.

Clearly this condition implies that U is open. Conversely, if U is open, it contains a basis element $B = B_d(x, \epsilon)$ containing y, and B in turn contains a basis element $B_d(y, \delta)$ centered at y.

EXAMPLE 1. Given a set X, define

$$d(x, y) = 1 \quad \text{if } x \neq y,$$
$$d(x, y) = 0 \quad \text{if } x = y.$$

It is trivial to check that d is a metric. The topology it induces is the discrete topology; the basis element $B(x, 1)$, for example, consists of the point x alone.

EXAMPLE 2. The standard metric on the real numbers $\mathbb{R}$ is defined by the equation

$$d(x, y) = |x - y|.$$

It is easy to check that d is a metric. The topology it induces is the same as the order topology: Each basis element (a, b) for the order topology is a basis element for the metric topology; indeed,

$$(a, b) = B(x, \epsilon),$$

where $x = (a + b)/2$ and $\epsilon = (b - a)/2$. And conversely, each ϵ-ball $B(x, \epsilon)$ equals an open interval: the interval $(x - \epsilon, x + \epsilon)$.

Definition. If X is a topological space, X is said to be ***metrizable*** if there exists a metric d on the set X that induces the topology of X. A ***metric space*** is a metrizable space X together with a specific metric d that gives the topology of X.

Many of the spaces important for mathematics are metrizable, but some are not. Metrizability is always a highly desirable attribute for a space to possess, for the existence of a metric gives one a valuable tool for proving theorems about the space.

It is, therefore, a problem of fundamental importance in topology to find conditions on a topological space that will guarantee it is metrizable. One of our goals in Chapter 4 will be to find such conditions; they are expressed there in the famous theorem called *Urysohn's metrization theorem.* Further metrization theorems appear in Chapter 6. In the present section we shall content ourselves with proving merely that $\mathbb{R}^n$ and $\mathbb{R}^\omega$ are metrizable.

Although the metrizability problem is an important problem in topology, the study of metric spaces as such does not properly belong to topology as much as it does to analysis. Metrizability of a space depends only on the topology of the space in question, but properties that involve a specific metric for X in general do not. For instance, one can make the following definition in a metric space:

Definition. Let X be a metric space with metric d. A subset A of X is said to be ***bounded*** if there is some number M such that

$$d(a_1, a_2) \leq M$$

for every pair a_1, a_2 of points of A. If A is bounded and nonempty, the ***diameter*** of A is defined to be the number

$$\operatorname{diam} A = \sup\{d(a_1, a_2) \mid a_1, a_2 \in A\}.$$

Boundedness of a set is not a topological property, for it depends on the particular metric d that is used for X. For instance, if X is a metric space with metric d, then there exists a metric $\bar{d}$ that gives the topology of X, relative to which *every* subset of X is bounded. It is defined as follows:

Theorem 20.1. *Let X be a metric space with metric d. Define $\bar{d} : X \times X \to \mathbb{R}$ by the equation*

$$\bar{d}(x, y) = \min\{d(x, y), 1\}.$$

Then $\bar{d}$ is a metric that induces the same topology as d.

The metric $\bar{d}$ is called the ***standard bounded metric*** corresponding to d.

Proof. Checking the first two conditions for a metric is trivial. Let us check the triangle inequality:

$$\bar{d}(x, z) \leq \bar{d}(x, y) + \bar{d}(y, z).$$

Now if either $d(x, y) \geq 1$ or $d(y, z) \geq 1$, then the right side of this inequality is at least 1; since the left side is (by definition) at most 1, the inequality holds. It remains to consider the case in which $d(x, y) < 1$ and $d(y, z) < 1$. In this case, we have

$$d(x, z) \leq d(x, y) + d(y, z) = \bar{d}(x, y) + \bar{d}(y, z).$$

Since $\bar{d}(x, z) \leq d(x, z)$ by definition, the triangle inequality holds for $\bar{d}$.

Now we note that in any metric space, the collection of ϵ-balls with $\epsilon < 1$ forms a basis for the metric topology, for every basis element containing x contains such an ϵ-ball centered at x. It follows that d and $\bar{d}$ induce the same topology on X, because the collections of ϵ-balls with $\epsilon < 1$ under these two metrics are the same collection. ∎

Now we consider some familiar spaces and show they are metrizable.

Definition. Given $\mathbf{x} = (x_1, \dots, x_n)$ in $\mathbb{R}^n$, we define the ***norm*** of $\mathbf{x}$ by the equation

$$\|x\| = (x_1^2 + \cdots + x_n^2)^{1/2};$$

and we define the ***euclidean metric*** d on $\mathbb{R}^n$ by the equation

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| = [(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2]^{1/2}.$$

We define the ***square metric*** ρ by the equation

$$\rho(\mathbf{x}, \mathbf{y}) = \max\{|x_1 - y_1|, \dots, |x_n - y_n|\}.$$

The proof that d is a metric requires some work; it is probably already familiar to you. If not, a proof is outlined in the exercises. We shall seldom have occasion to use this metric on $\mathbb{R}^n$.

To show that ρ is a metric is easier. Only the triangle inequality is nontrivial. From the triangle inequality for $\mathbb{R}$ it follows that for each positive integer i,

$$|x_i - z_i| \leq |x_i - y_i| + |y_i - z_i|.$$

Then by definition of ρ,

$$|x_i - z_i| \leq \rho(\mathbf{x}, \mathbf{y}) + \rho(\mathbf{y}, \mathbf{z}).$$

As a result

$$\rho(\mathbf{x}, \mathbf{z}) = \max\{|x_i - z_i|\} \leq \rho(\mathbf{x}, \mathbf{y}) + \rho(\mathbf{y}, \mathbf{z}),$$

as desired.

On the real line $\mathbb{R} = \mathbb{R}^1$, these two metrics coincide with the standard metric for $\mathbb{R}$. In the plane $\mathbb{R}^2$, the basis elements under d can be pictured as circular regions, while the basis elements under ρ can be pictured as square regions.

We now show that each of these metrics induces the usual topology on $\mathbb{R}^n$. We need the following lemma:

Lemma 20.2. *Let d and d' be two metrics on the set X; let $\mathcal{T}$ and $\mathcal{T}'$ be the topologies they induce, respectively. Then $\mathcal{T}'$ is finer than $\mathcal{T}$ if and only if for each x in X and each $\epsilon > 0$, there exists a $\delta > 0$ such that*

$$B_{d'}(x, \delta) \subset B_d(x, \epsilon).$$

Proof. Suppose that $\mathcal{T}'$ is finer than $\mathcal{T}$. Given the basis element $B_d(x, \epsilon)$ for $\mathcal{T}$, there is by Lemma 13.3 a basis element B' for the topology $\mathcal{T}'$ such that $x \in B' \subset B_d(x, \epsilon)$. Within B' we can find a ball $B_{d'}(x, \delta)$ centered at x.

Conversely, suppose the δ-ϵ condition holds. Given a basis element B for $\mathcal{T}$ containing x, we can find within B a ball $B_d(x, \epsilon)$ centered at x. By the given condition, there is a δ such that $B_{d'}(x, \delta) \subset B_d(x, \epsilon)$. Then Lemma 13.3 applies to show $\mathcal{T}'$ is finer than $\mathcal{T}$. ■

Theorem 20.3. *The topologies on $\mathbb{R}^n$ induced by the euclidean metric d and the square metric ρ are the same as the product topology on $\mathbb{R}^n$.*

Proof. Let $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{y} = (y_1, \dots, y_n)$ be two points of $\mathbb{R}^n$. It is simple algebra to check that

$$\rho(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{y}) \leq \sqrt{n}\rho(\mathbf{x}, \mathbf{y}).$$

The first inequality shows that

$$B_d(\mathbf{x}, \epsilon) \subset B_\rho(\mathbf{x}, \epsilon)$$

for all $\mathbf{x}$ and ϵ, since if $d(\mathbf{x}, \mathbf{y}) < \epsilon$, then $\rho(\mathbf{x}, \mathbf{y}) < \epsilon$ also. Similarly, the second inequality shows that

$$B_\rho(\mathbf{x}, \epsilon/\sqrt{n}) \subset B_d(\mathbf{x}, \epsilon)$$

for all $\mathbf{x}$ and ϵ. It follows from the preceding lemma that the two metric topologies are the same.

Now we show that the product topology is the same as that given by the metric ρ. First, let

$$B = (a_1, b_1) \times \cdots \times (a_n, b_n)$$

be a basis element for the product topology, and let $\mathbf{x} = (x_1, \dots, x_n)$ be an element of B. For each i, there is an ϵ_i such that

$$(x_i - \epsilon_i, x_i + \epsilon_i) \subset (a_i, b_i);$$

choose $\epsilon = \min\{\epsilon_1, \dots, \epsilon_n\}$. Then $B_\rho(\mathbf{x}, \epsilon) \subset B$, as you can readily check. As a result, the ρ-topology is finer than the product topology.

Conversely, let $B_\rho(\mathbf{x}, \epsilon)$ be a basis element for the ρ-topology. Given the element $\mathbf{y} \in B_\rho(\mathbf{x}, \epsilon)$, we need to find a basis element B for the product topology such that

$$\mathbf{y} \in B \subset B_\rho(\mathbf{x}, \epsilon).$$

But this is trivial, for

$$B_\rho(\mathbf{x}, \epsilon) = (x_1 - \epsilon, x_1 + \epsilon) \times \cdots \times (x_n - \epsilon, x_n + \epsilon)$$

is itself a basis element for the product topology. ■

Now we consider the infinite cartesian product $\mathbb{R}^\omega$. It is natural to try to generalize the metrics d and ρ to this space. For instance, one can attempt to define a metric d on $\mathbb{R}^\omega$ by the equation

$$d(\mathbf{x}, \mathbf{y}) = \left[\sum_{i=1}^{\infty}(x_i - y_i)^2\right]^{1/2}.$$

But this equation does not always make sense, for the series in question need not converge. (This equation does define a metric on a certain important subset of $\mathbb{R}^\omega$, however; see the exercises.)

Similarly, one can attempt to generalize the square metric ρ to $\mathbb{R}^\omega$ by defining

$$\rho(\mathbf{x}, \mathbf{y}) = \sup\{|x_n - y_n|\}.$$

Again, this formula does not always make sense. If however we replace the usual metric $d(x, y) = |x - y|$ on $\mathbb{R}$ by its bounded counterpart $\bar{d}(x, y) = \min\{|x - y|, 1\}$, then this definition *does* make sense; it gives a metric on $\mathbb{R}^\omega$ called the *uniform metric*.

The uniform metric can be defined more generally on the cartesian product $\mathbb{R}^J$ for arbitrary J, as follows:

Definition. Given an index set J, and given points $\mathbf{x} = (x_\alpha)_{\alpha \in J}$ and $\mathbf{y} = (y_\alpha)_{\alpha \in J}$ of $\mathbb{R}^J$, let us define a metric $\bar{\rho}$ on $\mathbb{R}^J$ by the equation

$$\bar{\rho}(\mathbf{x}, \mathbf{y}) = \sup\{\bar{d}(x_\alpha, y_\alpha) \mid \alpha \in J\},$$

where $\bar{d}$ is the standard bounded metric on $\mathbb{R}$. It is easy to check that $\bar{\rho}$ is indeed a metric; it is called the ***uniform metric*** on $\mathbb{R}^J$, and the topology it induces is called the ***uniform topology***.

The relation between this topology and the product and box topologies is the following:

Theorem 20.4. *The uniform topology on $\mathbb{R}^J$ is finer than the product topology and coarser than the box topology; these three topologies are all different if J is infinite.*

Proof. Suppose that we are given a point $\mathbf{x} = (x_\alpha)_{\alpha \in J}$ and a product topology basis element $\prod U_\alpha$ about $\mathbf{x}$. Let $\alpha_1, \dots, \alpha_n$ be the indices for which $U_\alpha \neq \mathbb{R}$. Then for each i, choose $\epsilon_i > 0$ so that the ϵ_i-ball centered at x_{α_i} in the $\bar{d}$ metric is contained in U_{α_i}; this we can do because U_{α_i} is open in $\mathbb{R}$. Let $\epsilon = \min\{\epsilon_1, \dots, \epsilon_n\}$; then the ϵ-ball centered at $\mathbf{x}$ in the $\bar{\rho}$ metric is contained in $\prod U_\alpha$. For if $\mathbf{z}$ is a point of $\mathbb{R}^J$ such that $\bar{\rho}(\mathbf{x}, \mathbf{z}) < \epsilon$, then $\bar{d}(x_\alpha, z_\alpha) < \epsilon$ for all α, so that $\mathbf{z} \in \prod U_\alpha$. It follows that the uniform topology is finer than the product topology.

On the other hand, let B be the ϵ-ball centered at $\mathbf{x}$ in the $\bar{\rho}$ metric. Then the box neighborhood

$$U = \prod(x_\alpha - \tfrac{1}{2}\epsilon, x_\alpha + \tfrac{1}{2}\epsilon)$$

of $\mathbf{x}$ is contained in B. For if $\mathbf{y} \in U$, then $\bar{d}(x_\alpha, y_\alpha) < \frac{1}{2}\epsilon$ for all α, so that $\bar{\rho}(\mathbf{x}, \mathbf{y}) \leq \frac{1}{2}\epsilon$.

Showing these three topologies are different if J is infinite is a task we leave to the exercises. ∎

In the case where J is infinite, we still have not determined whether $\mathbb{R}^J$ is metrizable in either the box or the product topology. It turns out that the only one of these cases where $\mathbb{R}^J$ *is* metrizable is the case where J is countable and $\mathbb{R}^J$ has the product topology. As we shall see.

Theorem 20.5. *Let $\bar{d}(a, b) = \min\{|a - b|, 1\}$ be the standard bounded metric on $\mathbb{R}$. If $\mathbf{x}$ and $\mathbf{y}$ are two points of $\mathbb{R}^\omega$, define*

$$D(\mathbf{x}, \mathbf{y}) = \sup\left\{\frac{\bar{d}(x_i, y_i)}{i}\right\}.$$

Then D is a metric that induces the product topology on $\mathbb{R}^\omega$.

Proof. The properties of a metric are satisfied trivially except for the triangle inequality, which is proved by noting that for all i,

$$\frac{\bar{d}(x_i, z_i)}{i} \leq \frac{\bar{d}(x_i, y_i)}{i} + \frac{\bar{d}(y_i, z_i)}{i} \leq D(\mathbf{x}, \mathbf{y}) + D(\mathbf{y}, \mathbf{z}),$$

so that

$$\sup\left\{\frac{\bar{d}(x_i, z_i)}{i}\right\} \leq D(\mathbf{x}, \mathbf{y}) + D(\mathbf{y}, \mathbf{z}).$$

The fact that D gives the product topology requires a little more work. First, let U be open in the metric topology and let $\mathbf{x} \in U$; we find an open set V in the product topology such that $\mathbf{x} \in V \subset U$. Choose an ϵ-ball $B_D(\mathbf{x}, \epsilon)$ lying in U. Then choose N large enough that $1/N < \epsilon$. Finally, let V be the basis element for the product topology

$$V = (x_1 - \epsilon, x_1 + \epsilon) \times \cdots \times (x_N - \epsilon, x_N + \epsilon) \times \mathbb{R} \times \mathbb{R} \times \cdots.$$

We assert that $V \subset B_D(\mathbf{x}, \epsilon)$: Given any $\mathbf{y}$ in $\mathbb{R}^\omega$,

$$\frac{\bar{d}(x_i, y_i)}{i} \leq \frac{1}{N} \quad \text{for } i \geq N.$$

Therefore,

$$D(\mathbf{x}, \mathbf{y}) \leq \max\left\{\frac{\bar{d}(x_1, y_1)}{1}, \ldots, \frac{\bar{d}(x_N, y_N)}{N}, \frac{1}{N}\right\}.$$

If $\mathbf{y}$ is in V, this expression is less than ϵ, so that $V \subset B_D(\mathbf{x}, \epsilon)$, as desired.

Conversely, consider a basis element

$$U = \prod_{i \in \mathbb{Z}_+} U_i$$

for the product topology, where U_i is open in $\mathbb{R}$ for $i = \alpha_1, \dots, \alpha_n$ and $U_i = \mathbb{R}$ for all other indices i. Given $\mathbf{x} \in U$, we find an open set V of the metric topology such that $\mathbf{x} \in V \subset U$. Choose an interval $(x_i - \epsilon_i, x_i + \epsilon_i)$ in $\mathbb{R}$ centered about x_i and lying in U_i for $i = \alpha_1, \dots, \alpha_n$; choose each $\epsilon_i \leq 1$. Then define

$$\epsilon = \min\{\epsilon_i / i \mid i = \alpha_1, \dots, \alpha_n\}.$$

We assert that

$$\mathbf{x} \in B_D(\mathbf{x}, \epsilon) \subset U.$$

Let $\mathbf{y}$ be a point of $B_D(\mathbf{x}, \epsilon)$. Then for all i,

$$\frac{\bar{d}(x_i, y_i)}{i} \leq D(\mathbf{x}, \mathbf{y}) < \epsilon.$$

Now if $i = \alpha_1, \dots, \alpha_n$, then $\epsilon \leq \epsilon_i / i$, so that $\bar{d}(x_i, y_i) < \epsilon_i \leq 1$; it follows that $|x_i - y_i| < \epsilon_i$. Therefore, $\mathbf{y} \in \prod U_i$, as desired. ■

Exercises

1. (a) In $\mathbb{R}^n$, define

$$d'(\mathbf{x}, \mathbf{y}) = |x_1 - y_1| + \cdots + |x_n - y_n|.$$

Show that d' is a metric that induces the usual topology of $\mathbb{R}^n$. Sketch the basis elements under d' when $n = 2$.

(b) More generally, given $p \geq 1$, define

$$d'(\mathbf{x}, \mathbf{y}) = \left[\sum_{i=1}^{n} |x_i - y_i|^p \right]^{1/p}$$

for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Assume that d' is a metric. Show that it induces the usual topology on $\mathbb{R}^n$.

2. Show that $\mathbb{R} \times \mathbb{R}$ in the dictionary order topology is metrizable.

3. Let X be a metric space with metric d.

(a) Show that $d : X \times X \to \mathbb{R}$ is continuous.

(b) Let X' denote a space having the same underlying set as X. Show that if $d : X' \times X' \to \mathbb{R}$ is continuous, then the topology of X' is finer than the topology of X.

One can summarize the result of this exercise as follows: If X has a metric d, then the topology induced by d is the coarsest topology relative to which the function d is continuous.

4. Consider the product, uniform, and box topologies on $\mathbb{R}^\omega$.

(a) In which topologies are the following functions from $\mathbb{R}$ to $\mathbb{R}^\omega$ continuous?

$$f(t) = (t, 2t, 3t, \dots),$$
$$g(t) = (t, t, t, \dots),$$
$$h(t) = (t, \tfrac{1}{2}t, \tfrac{1}{3}t, \dots).$$

(b) In which topologies do the following sequences converge?

$$\begin{aligned}
\mathbf{w}_1 &= (1, 1, 1, 1, \dots), & \mathbf{x}_1 &= (1, 1, 1, 1, \dots),\\
\mathbf{w}_2 &= (0, 2, 2, 2, \dots), & \mathbf{x}_2 &= (0, \tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}, \dots),\\
\mathbf{w}_3 &= (0, 0, 3, 3, \dots), & \mathbf{x}_3 &= (0, 0, \tfrac{1}{3}, \tfrac{1}{3} \dots),\\
&\dots & &\dots\\
\mathbf{y}_1 &= (1, 0, 0, 0, \dots), & \mathbf{z}_1 &= (1, 1, 0, 0, \dots),\\
\mathbf{y}_2 &= (\tfrac{1}{2}, \tfrac{1}{2}, 0, 0, \dots), & \mathbf{z}_2 &= (\tfrac{1}{2}, \tfrac{1}{2}, 0, 0, \dots),\\
\mathbf{y}_3 &= (\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}, 0, \dots), & \mathbf{z}_3 &= (\tfrac{1}{3}, \tfrac{1}{3}, 0, 0, \dots),\\
&\dots & &\dots
\end{aligned}$$

5. Let $\mathbb{R}^\infty$ be the subset of $\mathbb{R}^\omega$ consisting of all sequences that are eventually zero. What is the closure of $\mathbb{R}^\infty$ in $\mathbb{R}^\omega$ in the uniform topology? Justify your answer.

6. Let $\bar{\rho}$ be the uniform metric on $\mathbb{R}^\omega$. Given $\mathbf{x} = (x_1, x_2, \dots) \in \mathbb{R}^\omega$ and given $0 < \epsilon < 1$, let

$$U(\mathbf{x}, \epsilon) = (x_1 - \epsilon, x_1 + \epsilon) \times \cdots \times (x_n - \epsilon, x_n + \epsilon) \times \cdots.$$

(a) Show that $U(\mathbf{x}, \epsilon)$ is not equal to the ϵ-ball $B_{\bar{\rho}}(\mathbf{x}, \epsilon)$.

(b) Show that $U(\mathbf{x}, \epsilon)$ is not even open in the uniform topology.

(c) Show that

$$B_{\bar{\rho}}(\mathbf{x}, \epsilon) = \bigcup_{\delta < \epsilon} U(\mathbf{x}, \delta).$$

7. Consider the map $h : \mathbb{R}^\omega \to \mathbb{R}^\omega$ defined in Exercise 8 of §19; give $\mathbb{R}^\omega$ the uniform topology. Under what conditions on the numbers a_i and b_i is h continuous? a homeomorphism?

8. Let X be the subset of $\mathbb{R}^\omega$ consisting of all sequences $\mathbf{x}$ such that $\sum x_i^2$ converges. Then the formula

$$d(\mathbf{x}, \mathbf{y}) = \left[\sum_{i=1}^{\infty} (x_i - y_i)^2\right]^{1/2}$$

defines a metric on X. (See Exercise 10.) On X we have the three topologies it inherits from the box, uniform, and product topologies on $\mathbb{R}^\omega$. We have also the topology given by the metric d, which we call the *ℓ^2-topology*. (Read "little ell two.")

(a) Show that on X, we have the inclusions

$$\text{box topology} \supset \ell^2\text{-topology} \supset \text{uniform topology}.$$

(b) The set $\mathbb{R}^\infty$ of all sequences that are eventually zero is contained in X. Show that the four topologies that $\mathbb{R}^\infty$ inherits as a subspace of X are all distinct.

(c) The set

$$H = \prod_{n \in \mathbb{Z}_+} [0, 1/n]$$

is contained in X; it is called the ***Hilbert cube***. Compare the four topologies that H inherits as a subspace of X.

9. Show that the euclidean metric d on $\mathbb{R}^n$ is a metric, as follows: If $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $c \in \mathbb{R}$, define

$$\begin{aligned} \mathbf{x} + \mathbf{y} &= (x_1 + y_1, \dots, x_n + y_n), \\ c\mathbf{x} &= (cx_1, \dots, cx_n), \\ \mathbf{x} \cdot \mathbf{y} &= x_1 y_1 + \cdots + x_n y_n. \end{aligned}$$

(a) Show that $\mathbf{x} \cdot (\mathbf{y} + \mathbf{z}) = (\mathbf{x} \cdot \mathbf{y}) + (\mathbf{x} \cdot \mathbf{z})$.

(b) Show that $|\mathbf{x} \cdot \mathbf{y}| \leq \|\mathbf{x}\| \|\mathbf{y}\|$. [*Hint:* If $\mathbf{x}, \mathbf{y} \neq 0$, let $a = 1/\|\mathbf{x}\|$ and $b = 1/\|\mathbf{y}\|$, and use the fact that $\|a\mathbf{x} \pm b\mathbf{y}\| \geq 0$.]

(c) Show that $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$. [*Hint:* Compute $(\mathbf{x} + \mathbf{y}) \cdot (\mathbf{x} + \mathbf{y})$ and apply (b).]

(d) Verify that d is a metric.

10. Let X denote the subset of $\mathbb{R}^\omega$ consisting of all sequences $(x_1, x_2, \dots)$ such that $\sum x_i^2$ converges. (You may assume the standard facts about infinite series. In case they are not familiar to you, we shall give them in Exercise 11 of the next section.)

(a) Show that if $\mathbf{x}, \mathbf{y} \in X$, then $\sum |x_i y_i|$ converges. [*Hint:* Use (b) of Exercise 9 to show that the partial sums are bounded.]

(b) Let $c \in \mathbb{R}$. Show that if $\mathbf{x}, \mathbf{y} \in X$, then so are $\mathbf{x} + \mathbf{y}$ and $c\mathbf{x}$.

(c) Show that

$$d(\mathbf{x}, \mathbf{y}) = \left[\sum_{i=1}^{\infty} (x_i - y_i)^2 \right]^{1/2}$$

is a well-defined metric on X.

***11.** Show that if d is a metric for X, then

$$d'(x, y) = d(x, y)/(1 + d(x, y))$$

is a bounded metric that gives the topology of X. [*Hint:* If $f(x) = x/(1+x)$ for $x > 0$, use the mean-value theorem to show that $f(a + b) - f(b) \leq f(a)$.]

§21 The Metric Topology (continued)

In this section, we discuss the relation of the metric topology to the concepts we have previously introduced.

Subspaces of metric spaces behave the way one would wish them to; if A is a subspace of the topological space X and d is a metric for X, then the restriction of d to $A \times A$ is a metric for the topology of A. This we leave to you to check.

About *order topologies* there is nothing to be said; some are metrizable (for instance, $\mathbb{Z}_+$ and $\mathbb{R}$), and others are not, as we shall see.

The *Hausdorff axiom* is satisfied by every metric topology. If x and y are distinct points of the metric space (X, d), we let $\epsilon = \frac{1}{2}d(x, y)$; then the triangle inequality implies that $B_d(x, \epsilon)$ and $B_d(y, \epsilon)$ are disjoint.

The *product topology* we have already considered in special cases; we have proved that the products $\mathbb{R}^n$ and $\mathbb{R}^\omega$ are metrizable. It is true in general that countable products of metrizable spaces are metrizable; the proof follows a pattern similar to the proof for $\mathbb{R}^\omega$, so we leave it to the exercises.

About *continuous functions* there is a good deal to be said. Consideration of this topic will occupy the remainder of the section.

When we study continuous functions on metric spaces, we are about as close to the study of calculus and analysis as we shall come in this book. There are two things we want to do at this point.

First, we want to show that the familiar "ϵ-δ definition" of continuity carries over to general metric spaces, and so does the "convergent sequence definition" of continuity.

Second, we want to consider two additional methods for constructing continuous functions, besides those discussed in §18. One is the process of taking sums, differences, products, and quotients of continuous real-valued functions. The other is the process of taking limits of uniformly convergent sequences of continuous functions.

Theorem 21.1. *Let $f : X \to Y$; let X and Y be metrizable with metrics d_X and d_Y, respectively. Then continuity of f is equivalent to the requirement that given $x \in X$ and given $\epsilon > 0$, there exists $\delta > 0$ such that*

$$d_X(x, y) < \delta \Longrightarrow d_Y(f(x), f(y)) < \epsilon.$$

Proof. Suppose that f is continuous. Given x and ϵ, consider the set

$$f^{-1}(B(f(x), \epsilon)),$$

which is open in X and contains the point x. It contains some δ-ball $B(x, \delta)$ centered at x. If y is in this δ-ball, then $f(y)$ is in the ϵ-ball centered at $f(x)$, as desired.

Conversely, suppose that the ϵ-δ condition is satisfied. Let V be open in Y; we show that $f^{-1}(V)$ is open in X. Let x be a point of the set $f^{-1}(V)$. Since $f(x) \in V$, there is an ϵ-ball $B(f(x), \epsilon)$ centered at $f(x)$ and contained in V. By the ϵ-δ condition, there is a δ-ball $B(x, \delta)$ centered at x such that $f(B(x, \delta)) \subset B(f(x), \epsilon)$. Then $B(x, \delta)$ is a neighborhood of x contained in $f^{-1}(V)$, so that $f^{-1}(V)$ is open, as desired. ∎

Now we turn to the convergent sequence definition of continuity. We begin by considering the relation between convergent sequences and closures of sets. It is certainly believable, from one's experience in analysis, that if x lies in the closure of a subset A of the space X, then there should exist a sequence of points of A converging to x. This is not true in general, but it is true for metrizable spaces.

Lemma 21.2 (The sequence lemma). *Let X be a topological space; let $A \subset X$. If there is a sequence of points of A converging to x, then $x \in \bar{A}$; the converse holds if X is metrizable.*

Proof. Suppose that $x_n \to x$, where $x_n \in A$. Then every neighborhood U of x contains a point of A, so $x \in \bar{A}$ by Theorem 17.5. Conversely, suppose that X is metrizable and $x \in \bar{A}$. Let d be a metric for the topology of X. For each positive integer n, take the neighborhood $B_d(x, 1/n)$ of radius $1/n$ of x, and choose x_n to be a point of its intersection with A. We assert that the sequence x_n converges to x: Any open set U containing x contains an ϵ-ball $B_d(x, \epsilon)$ centered at x; if we choose N so that $1/N < \epsilon$, then U contains x_i for all $i \geq N$. ∎

Theorem 21.3. *Let $f : X \to Y$. If the function f is continuous, then for every convergent sequence $x_n \to x$ in X, the sequence $f(x_n)$ converges to $f(x)$. The converse holds if X is metrizable.*

Proof. Assume that f is continuous. Given $x_n \to x$, we wish to show that $f(x_n) \to f(x)$. Let V be a neighborhood of $f(x)$. Then $f^{-1}(V)$ is a neighborhood of x, and so there is an N such that $x_n \in f^{-1}(V)$ for $n \geq N$. Then $f(x_n) \in V$ for $n \geq N$.

To prove the converse, assume that the convergent sequence condition is satisfied. Let A be a subset of X; we show that $f(\bar{A}) \subset \overline{f(A)}$. If $x \in \bar{A}$, then there is a sequence x_n of points of A converging to x (by the preceding lemma). By assumption, the sequence $f(x_n)$ converges to $f(x)$. Since $f(x_n) \in f(A)$, the preceding lemma implies that $f(x) \in \overline{f(A)}$. (Note that metrizability of Y is not needed.) Hence $f(\bar{A}) \subset \overline{f(A)}$, as desired. ∎

Incidentally, in proving Lemma 21.2 and Theorem 21.3 we did not use the full strength of the hypothesis that the space X is metrizable. All we really needed was the countable collection $B_d(x, 1/n)$ of balls about x. This fact leads us to make a new definition.

A space X is said to have a ***countable basis at the point*** x if there is a countable collection $\{U_n\}_{n \in \mathbb{Z}_+}$ of neighborhoods of x such that any neighborhood U of x contains *at* *least one* of the sets U_n. A space X that has a countable basis at each of its points is said to satisfy the ***first countability axiom***.

If X has a countable basis $\{U_n\}$ at x, then the proof of Lemma 21.2 goes through; one simply replaces the ball $B_d(x, 1/n)$ throughout by the set

$$B_n = U_1 \cap U_2 \cap \cdots \cap U_n.$$

The proof of Theorem 21.3 goes through unchanged.

A metrizable space always satisfies the first countability axiom, but the converse is not true, as we shall see. Like the Hausdorff axiom, the first countability axiom is a requirement that we sometimes impose on a topological space in order to prove stronger theorems about the space. We shall study it in more detail in Chapter 4.

Now we consider additional methods for constructing continuous functions. We need the following lemma:

Lemma 21.4. *The addition, subtraction, and multiplication operations are continuous functions from $\mathbb{R} \times \mathbb{R}$ into $\mathbb{R}$; and the quotient operation is a continuous function from $\mathbb{R} \times (\mathbb{R} - \{0\})$ into $\mathbb{R}$.*

You have probably seen this lemma proved before; it is a standard "ϵ-δ argument." If not, a proof is outlined in Exercise 12 below; you should have no trouble filling in the details.

Theorem 21.5. *If X is a topological space, and if $f, g : X \to \mathbb{R}$ are continuous functions, then $f + g$, $f - g$, and $f \cdot g$ are continuous. If $g(x) \neq 0$ for all x, then f/g is continuous.*

Proof. The map $h : X \to \mathbb{R} \times \mathbb{R}$ defined by

$$h(x) = f(x) \times g(x)$$

is continuous, by Theorem 18.4. The function $f + g$ equals the composite of h and the addition operation

$$+ : \mathbb{R} \times \mathbb{R} \to \mathbb{R};$$

therefore $f + g$ is continuous. Similar arguments apply to $f - g$, $f \cdot g$, and f/g. ∎

Finally, we come to the notion of uniform convergence.

Definition. Let $f_n : X \to Y$ be a sequence of functions from the set X to the metric space Y. Let d be the metric for Y. We say that the sequence (f_n) ***converges uniformly*** to the function $f : X \to Y$ if given $\epsilon > 0$, there exists an integer N such that

$$d(f_n(x), f(x)) < \epsilon$$

for all $n > N$ and all x in X.

Uniformity of convergence depends not only on the topology of Y but also on its metric. We have the following theorem about uniformly convergent sequences:

Theorem 21.6 (Uniform limit theorem). *Let $f_n : X \to Y$ be a sequence of continuous functions from the topological space X to the metric space Y. If (f_n) converges uniformly to f, then f is continuous.*

Proof. Let V be open in Y; let x_0 be a point of $f^{-1}(V)$. We wish to find a neighborhood U of x_0 such that $f(U) \subset V$.

Let $y_0 = f(x_0)$. First choose ϵ so that the ϵ-ball $B(y_0, \epsilon)$ is contained in V. Then, using uniform convergence, choose N so that for all $n \geq N$ and all $x \in X$,

$$d(f_n(x), f(x)) < \epsilon/3.$$

Finally, using continuity of f_N, choose a neighborhood U of x_0 such that f_N carries U into the $\epsilon/3$ ball in Y centered at $f_N(x_0)$.

We claim that f carries U into $B(y_0, \epsilon)$ and hence into V, as desired. For this purpose, note that if $x \in U$, then

$$\begin{aligned} d(f(x), f_N(x)) &< \epsilon/3 && \text{(by choice of } N\text{)},\\ d(f_N(x), f_N(x_0)) &< \epsilon/3 && \text{(by choice of } U\text{)},\\ d(f_N(x_0), f(x_0)) &< \epsilon/3 && \text{(by choice of } N\text{)}. \end{aligned}$$

Adding and using the triangle inequality, we see that $d(f(x), f(x_0)) < \epsilon$, as desired. ∎

Let us remark that the notion of uniform convergence is related to the definition of the uniform metric, which we gave in the preceding section. Consider, for example, the space $\mathbb{R}^X$ of all functions $f : X \to \mathbb{R}$, in the uniform metric $\bar{\rho}$. It is not difficult to see that a sequence of functions $f_n : X \to \mathbb{R}$ converges uniformly to f if and only if the sequence (f_n) converges to f when they are considered as elements of the metric space $(\mathbb{R}^X, \bar{\rho})$. We leave the proof to the exercises.

We conclude the section with some examples of spaces that are not metrizable.

EXAMPLE 1. *$\mathbb{R}^\omega$ in the box topology is not metrizable.*

We shall show that the sequence lemma does not hold for $\mathbb{R}^\omega$. Let A be the subset of $\mathbb{R}^\omega$ consisting of those points all of whose coordinates are positive:

$$A = \{(x_1, x_2, \dots) \mid x_i > 0 \text{ for all } i \in \mathbb{Z}_+\}.$$

Let $\mathbf{0}$ be the "origin" in $\mathbb{R}^\omega$, that is, the point $(0, 0, \dots)$ each of whose coordinates is zero. In the box topology, $\mathbf{0}$ belongs to $\bar{A}$; for if

$$B = (a_1, b_1) \times (a_2, b_2) \times \cdots$$

is any basis element containing $\mathbf{0}$, then B intersects A. For instance, the point

$$(\tfrac{1}{2}b_1, \tfrac{1}{2}b_2 \dots)$$

belongs to $B \cap A$.

But we assert that there is no sequence of points of A converging to $\mathbf{0}$. For let $(\mathbf{a}_n)$ be a sequence of points of A, where

$$\mathbf{a}_n = (x_{1n}, x_{2n}, \dots, x_{in}, \dots).$$

Every coordinate x_{in} is positive, so we can construct a basis element B' for the box topology on $\mathbb{R}$ by setting

$$B' = (-x_{11}, x_{11}) \times (-x_{22}, x_{22}) \times \cdots.$$

Then B' contains the origin $\mathbf{0}$, but it contains no member of the sequence $(\mathbf{a}_n)$; the point $\mathbf{a}_n$ cannot belong to B' because its nth coordinate x_{nn} does not belong to the interval $(-x_{nn}, x_{nn})$. Hence the sequence $(\mathbf{a}_n)$ cannot converge to $\mathbf{0}$ in the box topology.

EXAMPLE 2. *An uncountable product of $\mathbb{R}$ with itself is not metrizable.*

Let J be an uncountable index set; we show that $\mathbb{R}^J$ does not satisfy the sequence lemma (in the product topology).

Let A be the subset of $\mathbb{R}^J$ consisting of all points (x_α) such that $x_\alpha = 1$ for all but finitely many values of α. Let $\mathbf{0}$ be the "origin" in $\mathbb{R}^J$, the point each of whose coordinates is 0.

We assert that $\mathbf{0}$ belongs to the closure of A. Let $\prod U_\alpha$ be a basis element containing $\mathbf{0}$. Then $U_\alpha \neq \mathbb{R}$ for only finitely many values of α, say for $\alpha = \alpha_1, \dots, \alpha_n$. Let (x_α) be the point of A defined by letting $x_\alpha = 0$ for $\alpha = \alpha_1, \dots, \alpha_n$ and $x_\alpha = 1$ for all other values of α; then $(x_\alpha) \in A \cap \prod U_\alpha$, as desired.

But there is no sequence of points of A converging to $\mathbf{0}$. For let $\mathbf{a}_n$ be a sequence of points of A. Given n, let J_n denote the subset of J consisting of those indices α for which the αth coordinate of $\mathbf{a}_n$ is different from 1. The union of all the sets J_n is a countable union of finite sets and therefore countable. Because J itself is uncountable, there is an index in J, say β, that does not lie in any of the sets J_n. This means that for *each* of the points $\mathbf{a}_n$, its βth coordinate equals 1.

Now let U_β be the open interval $(-1, 1)$ in $\mathbb{R}$, and let U be the open set $\pi_\beta^{-1}(U_\beta)$ in $\mathbb{R}^J$. The set U is a neighborhood of $\mathbf{0}$ that contains none of the points $\mathbf{a}_n$; therefore, the sequence $\mathbf{a}_n$ cannot converge to $\mathbf{0}$.

Exercises

1. Let $A \subset X$. If d is a metric for the topology of X, show that $d|A \times A$ is a metric for the subspace topology on A.
2. Let X and Y be metric spaces with metrics d_X and d_Y, respectively. Let $f : X \to Y$ have the property that for every pair of points x_1, x_2 of X,

$$d_Y(f(x_1), f(x_2)) = d_X(x_1, x_2).$$

Show that f is an imbedding. It is called an ***isometric imbedding*** of X in Y.

3. Let X_n be a metric space with metric d_n, for $n \in \mathbb{Z}_+$.
   (a) Show that

$$\rho(x, y) = \max\{d_1(x_1, y_1), \dots, d_n(x_n, y_n)\}$$

is a metric for the product space $X_1 \times \cdots \times X_n$.

(b) Let $\bar{d}_i = \min\{d_i, 1\}$. Show that

$$D(x, y) = \sup\{\bar{d}_i(x_i, y_i)/i\}$$

is a metric for the product space $\prod X_i$.

4. Show that $\mathbb{R}_\ell$ and the ordered square satisfy the first countability axiom. (This result does not, of course, imply that they are metrizable.)

5. *Theorem.* *Let $x_n \to x$ and $y_n \to y$ in the space $\mathbb{R}$. Then*

$$x_n + y_n \to x + y,$$
$$x_n - y_n \to x - y,$$
$$x_n y_n \to xy,$$

and provided that each $y_n \neq 0$ and $y \neq 0$,

$$x_n / y_n \to x/y.$$

[*Hint:* Apply Lemma 21.4; recall from the exercises of §19 that if $x_n \to x$ and $y_n \to y$, then $x_n \times y_n \to x \times y$.]

6. Define $f_n : [0, 1] \to \mathbb{R}$ by the equation $f_n(x) = x^n$. Show that the sequence $(f_n(x))$ converges for each $x \in [0, 1]$, but that the sequence (f_n) does not converge uniformly.

7. Let X be a set, and let $f_n : X \to \mathbb{R}$ be a sequence of functions. Let $\bar{\rho}$ be the uniform metric on the space $\mathbb{R}^X$. Show that the sequence (f_n) converges uniformly to the function $f : X \to \mathbb{R}$ if and only if the sequence (f_n) converges to f as elements of the metric space $(\mathbb{R}^X, \bar{\rho})$.

8. Let X be a topological space and let Y be a metric space. Let $f_n : X \to Y$ be a sequence of continuous functions. Let x_n be a sequence of points of X converging to x. Show that if the sequence (f_n) converges uniformly to f, then $(f_n(x_n))$ converges to $f(x)$.

9. Let $f_n : \mathbb{R} \to \mathbb{R}$ be the function

$$f_n(x) = \frac{1}{n^3[x - (1/n)]^2 + 1}.$$

See Figure 21.1. Let $f : \mathbb{R} \to \mathbb{R}$ be the zero function.

(a) Show that $f_n(x) \to f(x)$ for each $x \in \mathbb{R}$.

(b) Show that f_n does not converge uniformly to f. (This shows that the converse of Theorem 21.6 does not hold; the limit function f may be continuous even though the convergence is not uniform.)

10. Using the closed set formulation of continuity (Theorem 18.1), show that the following are closed subsets of $\mathbb{R}^2$:

$$A = \{x \times y \mid xy = 1\},$$
$$S^1 = \{x \times y \mid x^2 + y^2 = 1\},$$
$$B^2 = \{x \times y \mid x^2 + y^2 \leq 1\}.$$

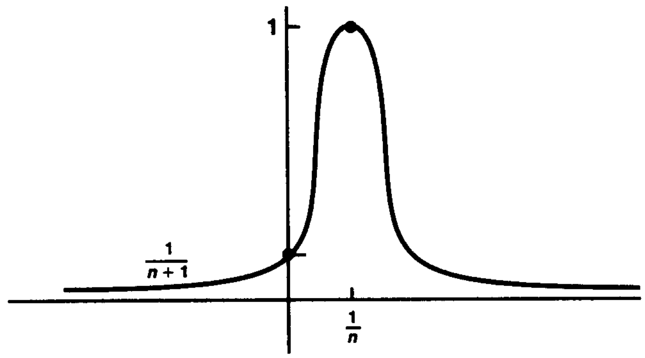


Figure 21.1

The set B^2 is called the (closed) ***unit ball*** in $\mathbb{R}^2$.

11. Prove the following standard facts about infinite series:

(a) Show that if (s_n) is a bounded sequence of real numbers and $s_n \leq s_{n+1}$ for each n, then (s_n) converges.

(b) Let (a_n) be a sequence of real numbers; define

$$s_n = \sum_{i=1}^{n} a_i.$$

If $s_n \to s$, we say that the ***infinite series***

$$\sum_{i=1}^{\infty} a_i$$

converges to s also. Show that if $\sum a_i$ converges to s and $\sum b_i$ converges to t, then $\sum(ca_i + b_i)$ converges to $cs + t$.

(c) Prove the ***comparison test*** for infinite series: If $|a_i| \leq b_i$ for each i, and if the series $\sum b_i$ converges, then the series $\sum a_i$ converges. [*Hint:* Show that the series $\sum |a_i|$ and $\sum c_i$ converge, where $c_i = |a_i| + a_i$.]

(d) Given a sequence of functions $f_n : X \to \mathbb{R}$, let

$$s_n(x) = \sum_{i=1}^{n} f_i(x).$$

Prove the ***Weierstrass M-test*** for uniform convergence: *If $|f_i(x)| \leq M_i$ for all $x \in X$ and all i, and if the series $\sum M_i$ converges, then the sequence (s_n) converges uniformly to a function s.* [*Hint:* Let $r_n = \sum_{i=n+1}^{\infty} M_i$. Show that if $k > n$, then $|s_k(x) - s_n(x)| \leq r_n$; conclude that $|s(x) - s_n(x)| \leq r_n$.]

12. Prove continuity of the algebraic operations on $\mathbb{R}$, as follows: Use the metric $d(a, b) = |a - b|$ on $\mathbb{R}$ and the metric on $\mathbb{R}^2$ given by the equation

$$\rho((x, y), (x_0, y_0)) = \max\{|x - x_0|, |y - y_0|\}.$$

(a) Show that addition is continuous. [*Hint:* Given ϵ, let $\delta = \epsilon/2$ and note that

$$d(x + y, x_0 + y_0) \leq |x - x_0| + |y - y_0|.]$$

(b) Show that multiplication is continuous. [*Hint:* Given (x_0, y_0) and $0 < \epsilon < 1$, let

$$3\delta = \epsilon/(|x_0| + |y_0| + 1)$$

and note that

$$d(xy, x_0y_0) \leq |x_0||y - y_0| + |y_0||x - x_0| + |x - x_0||y - y_0|.]$$

(c) Show that the operation of taking reciprocals is a continuous map from $\mathbb{R} - \{0\}$ to $\mathbb{R}$. [*Hint:* Show the inverse image of the interval (a, b) is open. Consider five cases, according as a and b are positive, negative, or zero.]

(d) Show that the subtraction and quotient operations are continuous.

*§22 The Quotient Topology†

Unlike the topologies we have already considered in this chapter, the quotient topology is not a natural generalization of something you have already studied in analysis. Nevertheless, it is easy enough to motivate. One motivation comes from geometry, where one often has occasion to use "cut-and-paste" techniques to construct such geometric objects as surfaces. The *torus* (surface of a doughnut), for example, can be constructed by taking a rectangle and "pasting" its edges together appropriately, as in Figure 22.1. And the *sphere* (surface of a ball) can be constructed by taking a disc and collapsing its entire boundary to a single point; see Figure 22.2. Formalizing these constructions involves the concept of quotient topology.

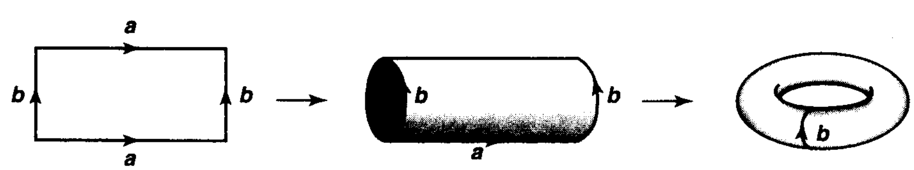


Figure 22.1

†This section will be used throughout Part II of the book. It also is referred to in a number of exercises of Part I.

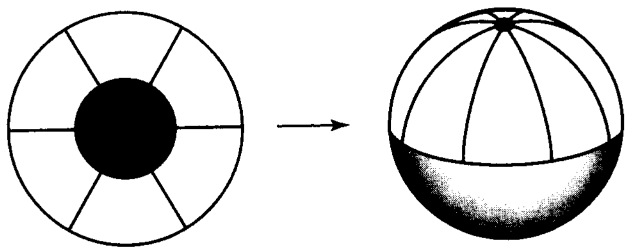


Figure 22.2

Definition. Let X and Y be topological spaces; let $p : X \to Y$ be a surjective map. The map p is said to be a ***quotient map*** provided a subset U of Y is open in Y if and only if $p^{-1}(U)$ is open in X.

This condition is stronger than continuity; some mathematicians call it "strong continuity." An equivalent condition is to require that a subset A of Y be closed in Y if and only if $p^{-1}(A)$ is closed in X. Equivalence of the two conditions follows from equation

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

Another way of describing a quotient map is as follows: We say that a subset C of X is ***saturated*** (with respect to the surjective map $p : X \to Y$) if C contains every set $p^{-1}(\{y\})$ that it intersects. Thus C is saturated if it equals the complete inverse image of a subset of Y. To say that p is a quotient map is equivalent to saying that p is continuous and p maps *saturated* open sets of X to open sets of Y (or saturated closed sets of X to closed sets of Y).

Two special kinds of quotient maps are the *open maps* and the *closed maps*. Recall that a map $f : X \to Y$ is said to be an ***open map*** if for each open set U of X, the set $f(U)$ is open in Y. It is said to be a ***closed map*** if for each closed set A of X, the set $f(A)$ is closed in Y. It follows immediately from the definition that if $p : X \to Y$ is a surjective continuous map that is either open or closed, then p is a quotient map. There are quotient maps that are neither open nor closed. (See Exercise 3.)

EXAMPLE 1. Let X be the subspace $[0, 1] \cup [2, 3]$ of $\mathbb{R}$, and let Y be the subspace $[0, 2]$ of $\mathbb{R}$. The map $p : X \to Y$ defined by

$$p(x) = \begin{cases} x & \text{for } x \in [0, 1], \\ x - 1 & \text{for } x \in [2, 3] \end{cases}$$

is readily seen to be surjective, continuous, and closed. Therefore it is a quotient map. It is not, however, an open map; the image of the open set $[0, 1]$ of X is not open in Y.

Note that if A is the subspace $[0, 1) \cup [2, 3]$ of X, then the map $q : A \to Y$ obtained by restricting p is continuous and surjective, but it is not a quotient map. For the set $[2, 3]$ is open in A and is saturated with respect to q, but its image is not open in Y.

EXAMPLE 2. Let $\pi_1 : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be projection onto the first coordinate; then π_1 is continuous and surjective. Furthermore, π_1 is an open map. For if $U \times V$ is a nonempty basis element for $\mathbb{R} \times \mathbb{R}$, then $\pi_1(U \times V) = U$ is open in $\mathbb{R}$; it follows that π_1 carries open sets of $\mathbb{R} \times \mathbb{R}$ to open sets of $\mathbb{R}$. However, π_1 is not a closed map. The subset

$$C = \{x \times y \mid xy = 1\}$$

of $\mathbb{R} \times \mathbb{R}$ is closed, but $\pi_1(C) = \mathbb{R} - \{0\}$, which is not closed in $\mathbb{R}$.

Note that if A is the subspace of $\mathbb{R} \times \mathbb{R}$ that is the union of C and the origin $\{\mathbf{0}\}$, then the map $q : A \to \mathbb{R}$ obtained by restricting π_1 is continuous and surjective, but it is not a quotient map. For the one-point set $\{\mathbf{0}\}$ is open in A and is saturated with respect to q, but its image is not open in $\mathbb{R}$.

Now we show how the notion of quotient map can be used to construct a topology on a set.

Definition. If X is a space and A is a *set* and if $p : X \to A$ is a surjective map, then there exists exactly one topology $\mathcal{T}$ on A relative to which p is a quotient map; it is called the ***quotient topology*** induced by p.

The topology $\mathcal{T}$ is of course defined by letting it consist of those subsets U of A such that $p^{-1}(U)$ is open in X. It is easy to check that $\mathcal{T}$ is a topology. The sets $\varnothing$ and A are open because $p^{-1}(\varnothing) = \varnothing$ and $p^{-1}(A) = X$. The other two conditions follow from the equations

$$p^{-1}\Big(\bigcup_{\alpha\in J} U_\alpha\Big) = \bigcup_{\alpha\in J} p^{-1}(U_\alpha),$$

$$p^{-1}\Big(\bigcap_{i=1}^{n} U_i\Big) = \bigcap_{i=1}^{n} p^{-1}(U_i).$$

EXAMPLE 3. Let p be the map of the real line $\mathbb{R}$ onto the three-point set $A = \{a, b, c\}$ defined by

$$p(x) = \begin{cases} a & \text{if } x > 0, \\ b & \text{if } x < 0, \\ c & \text{if } x = 0. \end{cases}$$

You can check that the quotient topology on A induced by p is the one indicated in Figure 22.3.

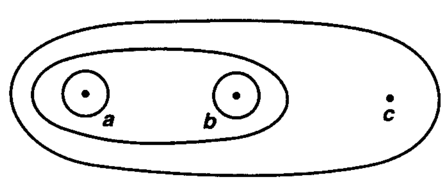


Figure 22.3

There is a special situation in which the quotient topology occurs particularly frequently. It is the following:

Definition. Let X be a topological space, and let X^* be a partition of X into disjoint subsets whose union is X. Let $p : X \to X^*$ be the surjective map that carries each point of X to the element of X^* containing it. In the quotient topology induced by p, the space X^* is called a ***quotient space*** of X.

Given X^*, there is an equivalence relation on X of which the elements of X^* are the equivalence classes. One can think of X^* as having been obtained by "identifying" each pair of equivalent points. For this reason, the quotient space X^* is often called an *identification space*, or a *decomposition space*, of the space X.

We can describe the topology of X^* in another way. A subset U of X^* is a collection of equivalence classes, and the set $p^{-1}(U)$ is just the union of the equivalence classes belonging to U. Thus the typical open set of X^* is a collection of equivalence classes whose *union* is an open set of X.

EXAMPLE 4. Let X be the closed unit ball

$$\{x \times y \mid x^2 + y^2 \leq 1\}$$

in $\mathbb{R}^2$, and let X^* be the partition of X consisting of all the one-point sets $\{x \times y\}$ for which $x^2 + y^2 < 1$, along with the set $S^1 = \{x \times y\} \mid x^2 + y^2 = 1\}$. Typical saturated open sets in X are pictured by the shaded regions in Figure 22.4. One can show that X^* is homeomorphic with the subspace of $\mathbb{R}^3$ called the ***unit 2-sphere***, defined by

$$S^2 = \{(x, y, z) \mid x^2 + y^2 + z^2 = 1\}.$$

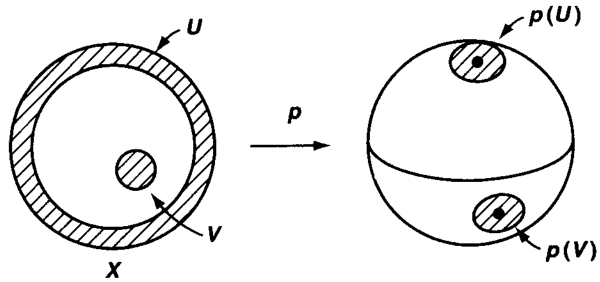


Figure 22.4

EXAMPLE 5. Let X be the rectangle $[0, 1] \times [0, 1]$. Define a partition X^* of X as follows: It consists of all the one-point sets $\{x \times y\}$ where $0 < x < 1$ and $0 < y < 1$, the following types of two-point sets:

$$\{x \times 0, x \times 1\} \quad \text{where } 0 < x < 1,$$
$$\{0 \times y, 1 \times y\} \quad \text{where } 0 < y < 1,$$

and the four-point set

$$\{0 \times 0, 0 \times 1, 1 \times 0, 1 \times 1\}.$$

Typical saturated open sets in X are pictured by the shaded regions in Figure 22.5; each is an open set of X that equals a union of elements of X^*.

The image of each of these sets under p is an open set of X^*, as indicated in Figure 22.6. This description of X^* is just the mathematical way of saying what we expressed in pictures when we pasted the edges of a rectangle together to form a torus.

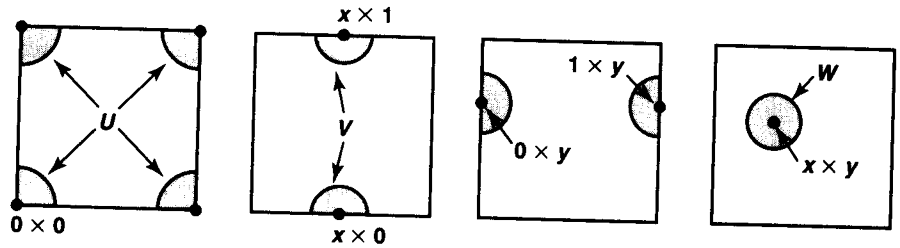


Figure 22.5

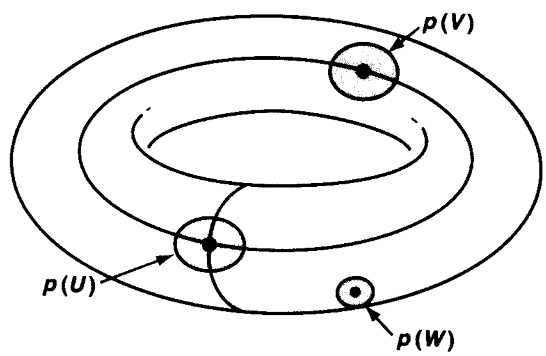


Figure 22.6

Now we explore the relationship between the notions of quotient map and quotient space and the concepts introduced previously. It is interesting to note that this relationship is not as simple as one might wish.

We have already noted that *subspaces* do not behave well; if $p : X \to Y$ is a quotient map and A is a subspace of X, then the map $q : A \to p(A)$ obtained by restricting p need not be a quotient map. One has, however, the following theorem:

Theorem 22.1. *Let $p : X \to Y$ be a quotient map; let A be a subspace of X that is saturated with respect to p; let $q : A \to p(A)$ be the map obtained by restricting p.*

(1) *If A is either open or closed in X, then q is a quotient map.*

*(2) *If p is either an open map or a closed map, then q is a quotient map.*

Proof. *Step 1.* We verify first the following two equations:

$$q^{-1}(V) = p^{-1}(V) \qquad \text{if } V \subset p(A);$$
$$p(U \cap A) = p(U) \cap p(A) \qquad \text{if } U \subset X.$$

To check the first equation, we note that since $V \subset p(A)$ and A is saturated, $p^{-1}(V)$ is contained in A. It follows that both $p^{-1}(V)$ and $q^{-1}(V)$ equal all points of A that are mapped by p into V. To check the second equation, we note that for any two subsets U and A of X, we have the inclusion

$$p(U \cap A) \subset p(U) \cap p(A).$$

To prove the reverse inclusion, suppose $y = p(u) = p(a)$, for $u \in U$ and $a \in A$. Since A is saturated, A contains the set $p^{-1}(p(a))$, so that in particular A contains u. Then $y = p(u)$, where $u \in U \cap A$.

Step 2. Now suppose A is open or p is open. Given the subset V of $p(A)$, we assume that $q^{-1}(V)$ is open in A and show that V is open in $p(A)$.

Suppose first that A is open. Since $q^{-1}(V)$ is open in A and A is open in X, the set $q^{-1}(V)$ is open in X. Since $q^{-1}(V) = p^{-1}(V)$, the latter set is open in X, so that V is open in Y because p is a quotient map. In particular, V is open in $p(A)$.

Now suppose p is open. Since $q^{-1}(V) = p^{-1}(V)$ and $q^{-1}(V)$ is open in A, we have $p^{-1}(V) = U \cap A$ for some set U open in X. Now $p(p^{-1}(V)) = V$ because p is surjective; then

$$V = p(p^{-1}(V)) = p(U \cap A) = p(U) \cap p(A).$$

The set $p(U)$ is open in Y because p is an open map; hence V is open in $p(A)$.

Step 3. The proof when A or p is closed is obtained by replacing the word "open" by the word "closed" throughout Step 2. ∎

Now we consider other concepts introduced previously. *Composites of maps* behave nicely; it is easy to check that the composite of two quotient maps is a quotient map; this fact follows from the equation

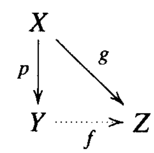
$$p^{-1}(q^{-1}(U)) = (q \circ p)^{-1}(U).$$

On the other hand, *products of maps* do not behave well; the cartesian product of two quotient maps need not be a quotient map. See Example 7 following. One needs further conditions on either the maps or the spaces in order for this statement to be true. One such, a condition on the spaces, is called *local compactness*; we shall study it later. Another, a condition on the maps, is the condition that both the maps p and q be open maps. In that case, it is easy to see that $p \times q$ is also an open map, so it is a quotient map.

Finally, the *Hausdorff condition* does not behave well; even if X is Hausdorff, there is no reason that the quotient space X^* needs to be Hausdorff. There is a simple condition for X^* to satisfy the T_1 axiom; one simply requires that each element of the partition X^* be a closed subset of X. Conditions that will ensure X^* is Hausdorff are harder to find. This is one of the more delicate questions concerning quotient spaces; we shall return to it several times later in the book.

Perhaps the most important result in the study of quotient spaces has to do with the problem of constructing *continuous functions* on a quotient space. We consider that problem now. When we studied product spaces, we had a criterion for determining whether a map $f : Z \to \prod X_\alpha$ *into* a product space was continuous. Its counterpart in the theory of quotient spaces is a criterion for determining when a map $f : X^* \to Z$ *out of* a quotient space is continuous. One has the following theorem:

Theorem 22.2. *Let $p : X \to Y$ be a quotient map. Let Z be a space and let $g : X \to Z$ be a map that is constant on each set $p^{-1}(\{y\})$, for $y \in Y$. Then g induces a map $f : Y \to Z$ such that $f \circ p = g$. The induced map f is continuous if and only if g is continuous; f is a quotient map if and only if g is a quotient map.*



Proof. For each $y \in Y$, the set $g(p^{-1}(\{y\}))$ is a one-point set in Z (since g is constant on $p^{-1}(\{y\})$). If we let $f(y)$ denote this point, then we have defined a map $f : Y \to Z$ such that for each $x \in X$, $f(p(x)) = g(x)$. If f is continuous, then $g = f \circ p$ is continuous. Conversely, suppose g is continuous. Given an open set V of Z, $g^{-1}(V)$ is open in X. But $g^{-1}(V) = p^{-1}(f^{-1}(V))$; because p is a quotient map, it follows that $f^{-1}(V)$ is open in Y. Hence f is continuous.

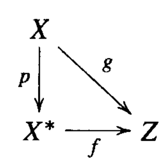
If f is a quotient map, then g is the composite of two quotient maps and is thus a quotient map. Conversely, suppose that g is a quotient map. Since g is surjective, so is f. Let V be a subset of Z; we show that V is open in Z if $f^{-1}(V)$ is open in Y. Now the set $p^{-1}(f^{-1}(V))$ is open in X because p is continuous. Since this set equals $g^{-1}(V)$, the latter is open in X. Then because g is a quotient map, V is open in Z. ∎

Corollary 22.3. *Let $g : X \to Z$ be a surjective continuous map. Let X^* be the following collection of subsets of X:*

$$X^* = \{g^{-1}(\{z\}) \mid z \in Z\}.$$

Give X^* the quotient topology.

(a) *The map g induces a bijective continuous map $f : X^* \to Z$, which is a homeomorphism if and only if g is a quotient map.*



(b) *If Z is Hausdorff, so is X^*.*

Proof. By the preceding theorem, g induces a continuous map $f : X^* \to Z$; it is clear that f is bijective. Suppose that f is a homeomorphism. Then both f and the projection map $p : X \to X^*$ are quotient maps, so that their composite q is a quotient map. Conversely, suppose that g is a quotient map. Then it follows from the preceding theorem that f is a quotient map. Being bijective, f is thus a homeomorphism.

Suppose Z is Hausdorff. Given distinct points of X^*, their images under f are distinct and thus possess disjoint neighborhoods U and V. Then $f^{-1}(U)$ and $f^{-1}(V)$ are disjoint neighborhoods of the two given points of X^*. ■

EXAMPLE 6. Let X be the subspace of $\mathbb{R}^2$ that is the union of the line segments $[0, 1] \times \{n\}$, for $n \in \mathbb{Z}_+$, and let Z be the subspace of $\mathbb{R}^2$ consisting of all points of the form $x \times (x/n)$ for $x \in [0, 1]$ and $n \in \mathbb{Z}_+$. Then X is the union of countably many disjoint line segments, and Z is the union of countably many line segments having an end point in common. See Figure 22.7.

Define a map $g : X \to Z$ by the equation $g(x \times n) = x \times (x/n)$; then g is surjective and continuous. The quotient space X^* whose elements are the sets $g^{-1}(\{z\})$ is simply the space obtained from X by identifying the subset $\{0\} \times \mathbb{Z}_+$ to a point. The map g induces a bijective continuous map $f : X^* \to Z$. But f is *not* a homeomorphism.

To verify this fact, it suffices to show that g is not a quotient map. Consider the sequence of points $x_n = (1/n) \times n$ of X. The set $A = \{x_n\}$ is a closed subset of X because it has no limit points. Also, it is saturated with respect to g. On the other hand, the set $g(A)$ is not closed in Z, for it consists of the points $z_n = (1/n) \times (1/n^2)$; this set has the origin as a limit point.

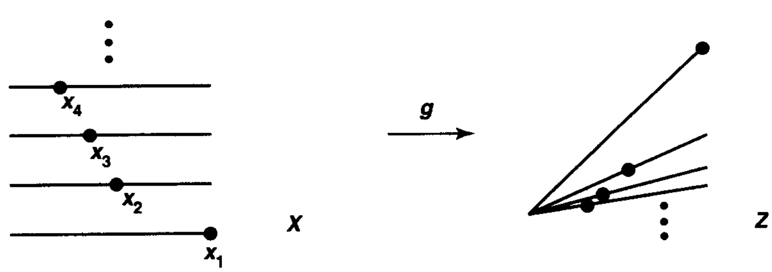


Figure 22.7

EXAMPLE 7. *The product of two quotient maps need not be a quotient map.*

We give an example that involves non-Hausdorff spaces in the exercises. Here is another involving spaces that are nicer.

Let $X = \mathbb{R}$ and let X^* be the quotient space obtained from X by identifying the subset $\mathbb{Z}_+$ to a point b; let $p : X \to X^*$ be the quotient map. Let $\mathbb{Q}$ be the subspace of $\mathbb{R}$ consisting of the rational numbers; let $i : \mathbb{Q} \to \mathbb{Q}$ be the identity map. We show that

$$p \times i : X \times \mathbb{Q} \to X^* \times \mathbb{Q}$$

is not a quotient map.

For each n, let $c_n = \sqrt{2}/n$, and consider the straight lines in $\mathbb{R}^2$ with slopes 1 and -1, respectively, through the point $n \times c_n$. Let U_n consist of all points of $X \times \mathbb{Q}$ that lie above both of these lines or beneath both of them, and also between the vertical lines $x = n - 1/4$ and $x = n + 1/4$. Then U_n is open in $X \times \mathbb{Q}$; it contains the set $\{n\} \times \mathbb{Q}$ because c_n is not rational. See Figure 22.8.

Let U be the union of the sets U_n; then U is open in $X \times \mathbb{Q}$. It is saturated with respect to $p \times i$ because it contains the entire set $\mathbb{Z}_+ \times \{q\}$ for each $q \in \mathbb{Q}$. We assume that $U' = (p \times i)(U)$ is open in $X^* \times \mathbb{Q}$ and derive a contradiction.

Because U contains, in particular, the set $\mathbb{Z}_+ \times 0$, the set U' contains the point $b \times 0$. Hence U' contains an open set of the form $W \times I_\delta$, where W is a neighborhood of b in X^* and I_δ consists of all rational numbers y with $|y| < \delta$. Then

$$p^{-1}(W) \times I_\delta \subset U.$$

Choose n large enough that $c_n < \delta$. Then since $p^{-1}(W)$ is open in X and contains $\mathbb{Z}_+$, we can choose $\epsilon < 1/4$ so that the interval $(n - \epsilon, n + \epsilon)$ is contained in $p^{-1}(W)$. Then U contains the subset $V = (n - \epsilon, n + \epsilon) \times I_\delta$ of $X \times \mathbb{Q}$. But the figure makes clear that there are many points $x \times y$ of V that do not lie in U! (One such is the point $x \times y$, where $x = n + \frac{1}{2}\epsilon$ and y is a rational number with $|y - c_n| < \frac{1}{2}\epsilon$.)

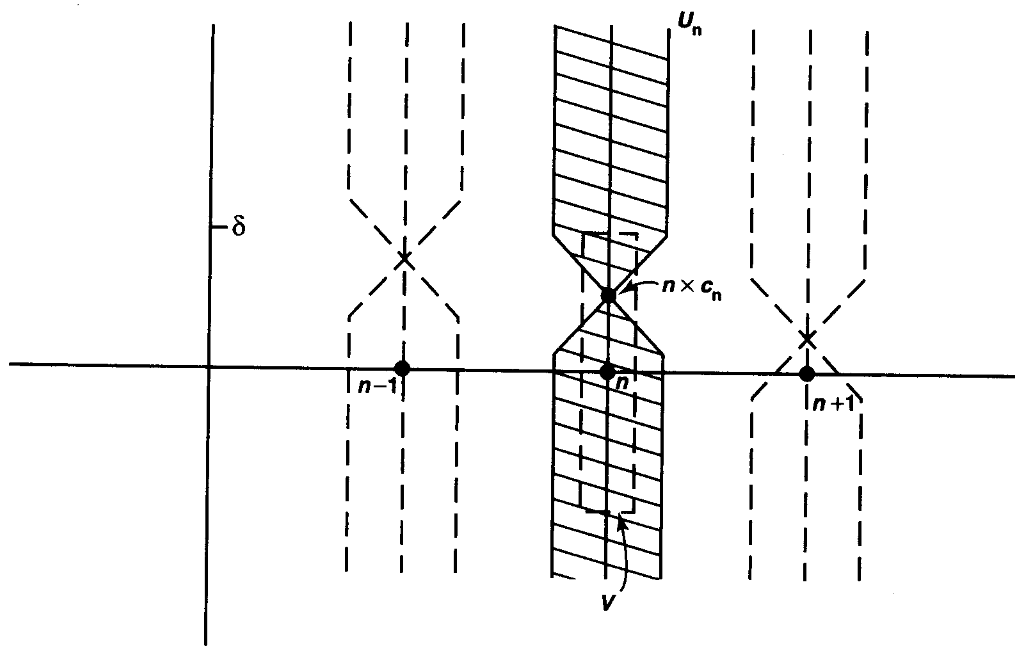


Figure 22.8

Exercises

1. Check the details of Example 3.
2. (a) Let $p : X \to Y$ be a continuous map. Show that if there is a continuous map $f : Y \to X$ such that $p \circ f$ equals the identity map of Y, then p is a quotient map.

   (b) If $A \subset X$, a ***retraction*** of X onto A is a continuous map $r : X \to A$ such that $r(a) = a$ for each $a \in A$. Show that a retraction is a quotient map.

3. Let $\pi_1 : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be projection on the first coordinate. Let A be the subspace of $\mathbb{R} \times \mathbb{R}$ consisting of all points $x \times y$ for which either $x \geq 0$ or $y = 0$ (or both); let $q : A \to \mathbb{R}$ be obtained by restricting π_1. Show that q is a quotient map that is neither open nor closed.

4. (a) Define an equivalence relation on the plane $X = \mathbb{R}^2$ as follows:

$$x_0 \times y_0 \sim x_1 \times y_1 \quad \text{if } x_0 + y_0^2 = x_1 + y_1^2.$$

Let X^* be the corresponding quotient space. It is homeomorphic to a familiar space; what is it? [*Hint:* Set $g(x \times y) = x + y^2$.]

(b) Repeat (a) for the equivalence relation

$$x_0 \times y_0 \sim x_1 \times y_1 \quad \text{if } x_0^2 + y_0^2 = x_1^2 + y_1^2.$$

5. Let $p : X \to Y$ be an open map. Show that if A is open in X, then the map $q : A \to p(A)$ obtained by restricting p is an open map.

6. Recall that $\mathbb{R}_K$ denotes the real line in the K-topology. (See §13.) Let Y be the quotient space obtained from $\mathbb{R}_K$ by collapsing the set K to a point; let $p : \mathbb{R}_K \to Y$ be the quotient map.

(a) Show that Y satisfies the T_1 axiom, but is not Hausdorff.

(b) Show that $p \times p : \mathbb{R}_K \times \mathbb{R}_K \to Y \times Y$ is not a quotient map. [*Hint:* The diagonal is not closed in $Y \times Y$, but its inverse image is closed in $\mathbb{R}_K \times \mathbb{R}_K$.]

*Supplementary Exercises: Topological Groups

In these exercises we consider topological groups and some of their properties. The quotient topology gets its name from the special case that arises when one forms the quotient of a topological group by a subgroup.

A ***topological group*** G is a group that is also a topological space satisfying the T_1 axiom, such that the map of $G \times G$ into G sending $x \times y$ into $x \cdot y$, and the map of G into G sending x into x^{-1}, are continuous maps. Throughout the following exercises, let G denote a topological group.

1. Let H denote a group that is also a topological space satisfying the T_1 axiom. Show that H is a topological group if and only if the map of $H \times H$ into H sending $x \times y$ into $x \cdot y^{-1}$ is continuous.

2. Show that the following are topological groups:

(a) $(\mathbb{Z}, +)$

(b) $(\mathbb{R}, +)$

(c) $(\mathbb{R}_+, \cdot)$

(d) $(S^1, \cdot)$, where we take S^1 to be the space of all complex numbers z for which $|z| = 1$.

(e) The *general linear group* GL(n), under the operation of matrix multiplication. (GL(n) is the set of all nonsingular n by n matrices, topologized by considering it as a subset of euclidean space of dimension n^2 in the obvious way.)

3. Let H be a subspace of G. Show that if H is also a subgroup of G, then both H and $\bar{H}$ are topological groups.

4. Let α be an element of G. Show that the maps $f_\alpha, g_\alpha : G \to G$ defined by

$$f_\alpha(x) = \alpha \cdot x \quad \text{and} \quad g_\alpha(x) = x \cdot \alpha$$

are homeomorphisms of G. Conclude that G is a *homogeneous space*. (This means that for every pair x, y of points of G, there exists a homeomorphism of G onto itself that carries x to y.)

5. Let H be a subgroup of G. If $x \in G$, define $xH = \{x \cdot h \mid h \in H\}$; this set is called a ***left coset*** of H in G. Let G/H denote the collection of left cosets of H in G; it is a partition of G. Give G/H the quotient topology.

(a) Show that if $\alpha \in G$, the map f_α of the preceding exercise induces a homeomorphism of G/H carrying xH to $(\alpha \cdot x)H$. Conclude that G/H is a homogeneous space.

(b) Show that if H is a closed set in the topology of G, then one-point sets are closed in G/H.

(c) Show that the quotient map $p : G \to G/H$ is open.

(d) Show that if H is closed in the topology of G and is a normal subgroup of G, then G/H is a topological group.

6. The integers $\mathbb{Z}$ are a normal subgroup of $(\mathbb{R}, +)$. The quotient $\mathbb{R}/\mathbb{Z}$ is a familiar topological group; what is it?

7. If A and B are subsets of G, let $A \cdot B$ denote the set of all points $a \cdot b$ for $a \in A$ and $b \in B$. Let A^{-1} denote the set of all points a^{-1}, for $a \in A$.

(a) A neighborhood V of the identity element e is said to be ***symmetric*** if $V = V^{-1}$. If U is a neighborhood of e, show there is a symmetric neighborhood V of e such that $V \cdot V \subset U$. [*Hint:* If W is a neighborhood of e, then $W \cdot W^{-1}$ is symmetric.]

(b) Show that G is Hausdorff. In fact, show that if $x \neq y$, there is a neighborhood V of e such that $V \cdot x$ and $V \cdot y$ are disjoint.

(c) Show that G satisfies the following separation axiom, which is called the ***regularity axiom***: Given a closed set A and a point x not in A, there exist disjoint open sets containing A and x, respectively. [*Hint:* There is a neighborhood V of e such that $V \cdot x$ and $V \cdot A$ are disjoint.]

(d) Let H be a subgroup of G that is closed in the topology of G; let $p : G \to G/H$ be the quotient map. Show that G/H satisfies the regularity axiom. [*Hint:* Examine the proof of (c) when A is saturated.]

Chapter 3

Connectedness and Compactness

In the study of calculus, there are three basic theorems about continuous functions, and on these theorems the rest of calculus depends. They are the following:

Intermediate value theorem. If $f : [a, b] \to \mathbb{R}$ is continuous and if r is a real number between $f(a)$ and $f(b)$, then there exists an element $c \in [a, b]$ such that $f(c) = r$.

Maximum value theorem. If $f : [a, b] \to R$ is continuous, then there exists an element $c \in [a, b]$ such that $f(x) \leq f(c)$ for every $x \in [a, b]$.

Uniform continuity theorem. If $f : [a, b] \to \mathbb{R}$ is continuous, then given $\epsilon > 0$, there exists $\delta > 0$ such that $|f(x_1) - f(x_2)| < \epsilon$ for every pair of numbers x_1, x_2 of $[a, b]$ for which $|x_1 - x_2| < \delta$.

These theorems are used in a number of places. The intermediate value theorem is used for instance in constructing inverse functions, such as $\sqrt[3]{x}$ and $\arcsin x$; and the maximum value theorem is used for proving the mean value theorem for derivatives, upon which the two *fundamental theorems of calculus* depend. The uniform continuity theorem is used, among other things, for proving that every continuous function is integrable.

We have spoken of these three theorems as theorems about continuous functions. But they can also be considered as theorems about the closed interval $[a, b]$ of real numbers. The theorems depend not only on the continuity of f but also on properties of the topological space $[a, b]$.

The property of the space $[a, b]$ on which the intermediate value theorem depends is the property called *connectedness*, and the property on which the other two depend is the property called *compactness*. In this chapter, we shall define these properties for arbitrary topological spaces, and shall prove the appropriate generalized versions of these theorems.

As the three quoted theorems are fundamental for the theory of calculus, so are the notions of connectedness and compactness fundamental in higher analysis, geometry, and topology—indeed, in almost any subject for which the notion of topological space itself is relevant.

§23 Connected Spaces

The definition of connectedness for a topological space is a quite natural one. One says that a space can be "separated" if it can be broken up into two "globs"—disjoint open sets. Otherwise, one says that it is connected. From this simple idea much follows.

Definition. Let X be a topological space. A ***separation*** of X is a pair U, V of disjoint nonempty open subsets of X whose union is X. The space X is said to be ***connected*** if there does not exist a separation of X.

Connectedness is obviously a topological property, since it is formulated entirely in terms of the collection of open sets of X. Said differently, if X is connected, so is any space homeomorphic to X.

Another way of formulating the definition of connectedness is the following:

A space X is connected if and only if the only subsets of X that are both open and closed in X are the empty set and X itself.

For if A is a nonempty proper subset of X that is both open and closed in X, then the sets $U = A$ and $V = X - A$ constitute a separation of X, for they are open, disjoint, and nonempty, and their union is X. Conversely, if U and V form a separation of X, then U is nonempty and different from X, and it is both open and closed in X.

For a subspace Y of a topological space X, there is another useful way of formulating the definition of connectedness:

Lemma 23.1. *If Y is a subspace of X, a separation of Y is a pair of disjoint nonempty sets A and B whose union is Y, neither of which contains a limit point of the other. The space Y is connected if there exists no separation of Y.*

Proof. Suppose first that A and B form a separation of Y. Then A is both open and closed in Y. The closure of A in Y is the set $\bar{A} \cap Y$ (where $\bar{A}$ as usual denotes the closure of A in X). Since A is closed in Y, $A = \bar{A} \cap Y$; or to say the same thing, $\bar{A} \cap B = \varnothing$. Since $\bar{A}$ is the union of A and its limit points, B contains no limit points of A. A similar argument shows that A contains no limit points of B.

Conversely, suppose that A and B are disjoint nonempty sets whose union is Y, neither of which contains a limit point of the other. Then $\bar{A} \cap B = \varnothing$ and $A \cap \bar{B} = \varnothing$;

therefore, we conclude that $\bar{A} \cap Y = A$ and $\bar{B} \cap Y = B$. Thus both A and B are closed in Y, and since $A = Y - B$ and $B = Y - A$, they are open in Y as well. ∎

EXAMPLE 1. Let X denote a two-point space in the indiscrete topology. Obviously there is no separation of X, so X is connected.

EXAMPLE 2. Let Y denote the subspace $[-1, 0) \cup (0, 1]$ of the real line $\mathbb{R}$. Each of the sets $[-1, 0)$ and $(0, 1]$ is nonempty and open in Y (although not in $\mathbb{R}$); therefore, they form a separation of Y. Alternatively, note that neither of these sets contains a limit point of the other. (They do have a limit point 0 in common, but that does not matter.)

EXAMPLE 3. Let X be the subspace $[-1, 1]$ of the real line. The sets $[-1, 0]$ and $(0, 1]$ are disjoint and nonempty, but they do not form a separation of X, because the first set is not open in X. Alternatively, note that the first set contains a limit point, 0, of the second. Indeed, there exists *no* separation of the space $[-1, 1]$. We shall prove this fact shortly.

EXAMPLE 4. The rationals $\mathbb{Q}$ are not connected. Indeed, the only connected subspaces of $\mathbb{Q}$ are the one-point sets: If Y is a subspace of $\mathbb{Q}$ containing two points p and q, one can choose an irrational number a lying between p and q, and write Y as the union of the open sets

$$Y \cap (-\infty, a) \quad \text{and} \quad Y \cap (a, +\infty).$$

EXAMPLE 5. Consider the following subset of the plane $\mathbb{R}^2$:

$$X = \{x \times y \mid y = 0\} \cup \{x \times y \mid x > 0 \text{ and } y = 1/x\}.$$

Then X is not connected; indeed, the two indicated sets form a separation of X because neither contains a limit point of the other. See Figure 23.1.

Figure 23.1

We have given several examples of spaces that are not connected. How can one construct spaces that *are* connected? We shall now prove several theorems that tell how to form new connected spaces from given ones. In the next section we shall apply these theorems to show that some specific spaces, such as intervals in $\mathbb{R}$, and balls and cubes in $\mathbb{R}^n$, are connected. First, a lemma:

Lemma 23.2. *If the sets C and D form a separation of X, and if Y is a connected subspace of X, then Y lies entirely within either C or D.*

Proof. Since C and D are both open in X, the sets $C \cap Y$ and $D \cap Y$ are open in Y. These two sets are disjoint and their union is Y; if they were both nonempty, they would constitute a separation of Y. Therefore, one of them is empty. Hence Y must lie entirely in C or in D. ∎

Theorem 23.3. *The union of a collection of connected subspaces of X that have a point in common is connected.*

Proof. Let $\{A_\alpha\}$ be a collection of connected subspaces of a space X; let p be a point of $\bigcap A_\alpha$. We prove that the space $Y = \bigcup A_\alpha$ is connected. Suppose that $Y = C \cup D$ is a separation of Y. The point p is in one of the sets C or D; suppose $p \in C$. Since A_α is connected, it must lie entirely in either C or D, and it cannot lie in D because it contains the point p of C. Hence $A_\alpha \subset C$ for every α, so that $\bigcup A_\alpha \subset C$, contradicting the fact that D is nonempty. ∎

Theorem 23.4. *Let A be a connected subspace of X. If $A \subset B \subset \bar{A}$, then B is also connected.*

Said differently: If B is formed by adjoining to the connected subspace A some or all of its limit points, then B is connected.

Proof. Let A be connected and let $A \subset B \subset \bar{A}$. Suppose that $B = C \cup D$ is a separation of B. By Lemma 23.2, the set A must lie entirely in C or in D; suppose that $A \subset C$. Then $\bar{A} \subset \bar{C}$; since $\bar{C}$ and D are disjoint, B cannot intersect D. This contradicts the fact that D is a nonempty subset of B. ∎

Theorem 23.5. *The image of a connected space under a continuous map is connected.*

Proof. Let $f : X \to Y$ be a continuous map; let X be connected. We wish to prove the image space $Z = f(X)$ is connected. Since the map obtained from f by restricting its range to the space Z is also continuous, it suffices to consider the case of a continuous surjective map

$$g : X \to Z.$$

Suppose that $Z = A \cup B$ is a separation of Z into two disjoint nonempty sets open in Z. Then $g^{-1}(A)$ and $g^{-1}(B)$ are disjoint sets whose union is X; they are open in X because g is continuous, and nonempty because g is surjective. Therefore, they form a separation of X, contradicting the assumption that X is connected. ∎

Theorem 23.6. *A finite cartesian product of connected spaces is connected.*

Proof. We prove the theorem first for the product of two connected spaces X and Y. This proof is easy to visualize. Choose a "base point" $a \times b$ in the product $X \times Y$. Note that the "horizontal slice" $X \times b$ is connected, being homeomorphic with X, and each "vertical slice" $x \times Y$ is connected, being homeomorphic with Y. As a result, each "T-shaped" space

$$T_x = (X \times b) \cup (x \times Y)$$

is connected, being the union of two connected spaces that have the point $x \times b$ in common. See Figure 23.2. Now form the union $\bigcup_{x \in X} T_x$ of all these T-shaped spaces.

This union is connected because it is the union of a collection of connected spaces that have the point $a \times b$ in common. Since this union equals $X \times Y$, the space $X \times Y$ is connected.

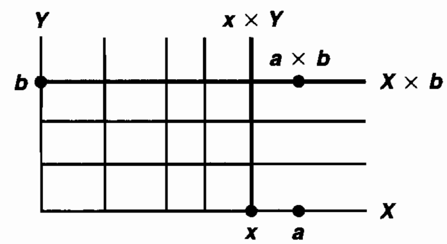


Figure 23.2

The proof for any finite product of connected spaces follows by induction, using the fact (easily proved) that $X_1 \times \cdots \times X_n$ is homeomorphic with $(X_1 \times \cdots \times X_{n-1}) \times X_n$. ∎

It is natural to ask whether this theorem extends to arbitrary products of connected spaces. The answer depends on which topology is used for the product, as the following examples show.

EXAMPLE 6. Consider the cartesian product $\mathbb{R}^\omega$ in the box topology. We can write $\mathbb{R}^\omega$ as the union of the set A consisting of all bounded sequences of real numbers, and the set B of all unbounded sequences. These sets are disjoint, and each is open in the box topology. For if $\mathbf{a}$ is a point of $\mathbb{R}^\omega$, the open set

$$U = (a_1 - 1, a_1 + 1) \times (a_2 - 1, a_2 + 1) \times \cdots$$

consists entirely of bounded sequences if $\mathbf{a}$ is bounded, and of unbounded sequences if $\mathbf{a}$ if unbounded. Thus, even though $\mathbb{R}$ is connected (as we shall prove in the next section), $\mathbb{R}^\omega$ is not connected in the box topology.

EXAMPLE 7. Now consider $\mathbb{R}^\omega$ in the product topology. Assuming that $\mathbb{R}$ is connected, we show that $\mathbb{R}^\omega$ is connected. Let $\tilde{\mathbb{R}}^n$ denote the subspace of $\mathbb{R}^\omega$ consisting of all sequences $\mathbf{x} = (x_1, x_2, \dots)$ such that $x_i = 0$ for $i > n$. The space $\tilde{\mathbb{R}}^n$ is clearly homeomorphic to $\mathbb{R}^n$, so that it is connected, by the preceding theorem. It follows that the space $\mathbb{R}^\infty$ that is the union of the spaces $\tilde{\mathbb{R}}^n$ is connected, for these spaces have the point $\mathbf{0} = (0, 0, \dots)$ in common. We show that the closure of $\mathbb{R}^\infty$ equals all of $\mathbb{R}^\omega$, from which it follows that $\mathbb{R}^\omega$ is connected as well.

Let $\mathbf{a} = (a_1, a_2, \dots)$ be a point of $\mathbb{R}^\omega$. Let $U = \prod U_i$ be a basis element for the product topology that contains $\mathbf{a}$. We show that U intersects $\mathbb{R}^\infty$. There is an integer N such that $U_i = \mathbb{R}$ for $i > N$. Then the point

$$\mathbf{x} = (a_1, \dots, a_n, 0, 0, \dots)$$

of $\mathbb{R}^\infty$ belongs to U, since $a_i \in U_i$ for all i, and $0 \in U_i$ for $i > N$.

The argument just given generalizes to show that an arbitrary product of connected spaces is connected in the product topology. Since we shall not need this result, we leave the proof to the exercises.

Exercises

1. Let $\mathcal{T}$ and $\mathcal{T}'$ be two topologies on X. If $\mathcal{T}' \supset \mathcal{T}$, what does connectedness of X in one topology imply about connectedness in the other?

2. Let $\{A_n\}$ be a sequence of connected subspaces of X, such that $A_n \cap A_{n+1} \neq \varnothing$ for all n. Show that $\bigcup A_n$ is connected.

3. Let $\{A_\alpha\}$ be a collection of connected subspaces of X; let A be a connected subspace of X. Show that if $A \cap A_\alpha \neq \varnothing$ for all α, then $A \cup (\bigcup A_\alpha)$ is connected.

4. Show that if X is an infinite set, it is connected in the finite complement topology.

5. A space is ***totally disconnected*** if its only connected subspaces are one-point sets. Show that if X has the discrete topology, then X is totally disconnected. Does the converse hold?

6. Let $A \subset X$. Show that if C is a connected subspace of X that intersects both A and $X - A$, then C intersects $\operatorname{Bd} A$.

7. Is the space $\mathbb{R}_\ell$ connected? Justify your answer.

8. Determine whether or not $\mathbb{R}^\omega$ is connected in the uniform topology.

9. Let A be a proper subset of X, and let B be a proper subset of Y. If X and Y are connected, show that

$$(X \times Y) - (A \times B)$$

is connected.

10. Let $\{X_\alpha\}_{\alpha \in J}$ be an indexed family of connected spaces; let X be the product space

$$X = \prod_{\alpha \in J} X_\alpha.$$

Let $\mathbf{a} = (a_\alpha)$ be a fixed point of X.

(a) Given any finite subset K of J, let X_K denote the subspace of X consisting of all points $\mathbf{x} = (x_\alpha)$ such that $x_\alpha = a_\alpha$ for $\alpha \notin K$. Show that X_K is connected.

(b) Show that the union Y of the spaces X_K is connected.

(c) Show that X equals the closure of Y; conclude that X is connected.

11. Let $p : X \to Y$ be a quotient map. Show that if each set $p^{-1}(\{y\})$ is connected, and if Y is connected, then X is connected.

12. Let $Y \subset X$; let X and Y be connected. Show that if A and B form a separation of $X - Y$, then $Y \cup A$ and $Y \cup B$ are connected.

§24 Connected Subspaces of the Real Line

The theorems of the preceding section show us how to construct new connected spaces out of given ones. But where can we find some connected spaces to start with? The best place to begin is the real line. We shall prove that $\mathbb{R}$ is connected, and so are the intervals and rays in $\mathbb{R}$.

One application is the intermediate value theorem of calculus, suitably generalized. Another is the result that such familiar spaces as balls and spheres in euclidean space are connected; the proof involves a new notion, called *path connectedness*, which we also discuss.

The fact that intervals and rays in $\mathbb{R}$ are connected may be familiar to you from analysis. We prove it again here, in generalized form. It turns out that this fact does not depend on the algebraic properties of $\mathbb{R}$, but only on its order properties. To make this clear, we shall prove the theorem for an arbitrary ordered set that has the order properties of $\mathbb{R}$. Such a set is called a *linear continuum.*

Definition. A simply ordered set L having more than one element is called a ***linear continuum*** if the following hold:

(1) L has the least upper bound property.

(2) If $x < y$, there exists z such that $x < z < y$.

Theorem 24.1. *If L is a linear continuum in the order topology, then L is connected, and so are intervals and rays in L.*

Proof. Recall that a subspace Y of L is said to be *convex* if for every pair of points a, b of Y with $a < b$, the entire interval $[a, b]$ of points of L lies in Y. We prove that if Y is a convex subspace of L, then Y is connected.

So suppose that Y is the union of the disjoint nonempty sets A and B, each of which is open in Y. Choose $a \in A$ and $b \in B$; suppose for convenience that $a < b$. The interval $[a, b]$ of points of L is contained in Y. Hence $[a, b]$ is the union of the disjoint sets

$$A_0 = A \cap [a, b] \quad \text{and} \quad B_0 = B \cap [a, b],$$

each of which is open in $[a, b]$ in the subspace topology, which is the same as the order topology. The sets A_0 and B_0 are nonempty because $a \in A_0$ and $b \in B_0$. Thus, A_0 and B_0 constitute a separation of $[a, b]$.

Let $c = \sup A_0$. We show that c belongs neither to A_0 nor to B_0, which contradicts the fact that $[a, b]$ is the union of A_0 and B_0.

Case 1. Suppose that $c \in B_0$. Then $c \neq a$, so either $c = b$ or $a < c < b$. In either case, it follows from the fact that B_0 is open in $[a, b]$ that there is some interval of the form $(d, c]$ contained in B_0. If $c = b$, we have a contradiction at once, for d is a smaller upper bound on A_0 than c. If $c < b$, we note that $(c, b]$ does not intersect A_0

(because c is an upper bound on A_0). Then

$$(d, b] = (d, c] \cup (c, b]$$

does not intersect A_0. Again, d is a smaller upper bound on A_0 than c, contrary to construction. See Figure 24.1.

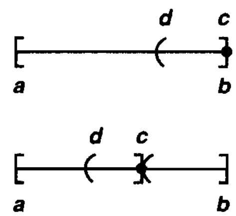


Figure 24.1

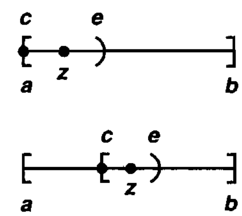


Figure 24.2

Case 2. Suppose that $c \in A_0$. Then $c \neq b$, so either $c = a$ or $a < c < b$. Because A_0 is open in $[a, b]$, there must be some interval of the form $[c, e)$ contained in A_0. See Figure 24.2. Because of order property (2) of the linear continuum L, we can choose a point z of L such that $c < z < e$. Then $z \in A_0$, contrary to the fact that c is an upper bound for A_0. ∎

Corollary 24.2. *The real line $\mathbb{R}$ is connected and so are intervals and rays in $\mathbb{R}$.*

As an application, we prove the intermediate value theorem of calculus, suitably generalized.

Theorem 24.3 (Intermediate value theorem). *Let $f : X \to Y$ be a continuous map, where X is a connected space and Y is an ordered set in the order topology. If a and b are two points of X and if r is a point of Y lying between $f(a)$ and $f(b)$, then there exists a point c of X such that $f(c) = r$.*

The intermediate value theorem of calculus is the special case of this theorem that occurs when we take X to be a closed interval in $\mathbb{R}$ and Y to be $\mathbb{R}$.

Proof. Assume the hypotheses of the theorem. The sets

$$A = f(X) \cap (-\infty, r) \quad \text{and} \quad B = f(X) \cap (r, +\infty)$$

are disjoint, and they are nonempty because one contains $f(a)$ and the other contains $f(b)$. Each is open in $f(X)$, being the intersection of an open ray in Y with $f(X)$. If there were no point c of X such that $f(c) = r$, then $f(X)$ would be the union of the sets A and B. Then A and B would constitute a separation of $f(X)$, contradicting the fact that the image of a connected space under a continuous map is connected. ∎

EXAMPLE 1. One example of a linear continuum different from $\mathbb{R}$ is the ordered square. We check the least upper bound property. (The second property of a linear continuum is trivial to check.) Let A be a subset of $I \times I$; let $\pi_1 : I \times I \to I$ be projection on the first coordinate; let $b = \sup \pi_1(A)$. If $b \in \pi_1(A)$, then A intersects the subset $b \times I$ of $I \times I$. Because $b \times I$ has the order type of I, the set $A \cap (b \times I)$ will have a least upper bound $b \times c$, which will be the least upper bound of A. See Figure 24.3. If $b \notin \pi_1(A)$, then $b \times 0$ is the least upper bound of A; no element of the form $b' \times c$ with $b' < b$ can be an upper bound for A, for then b' would be an upper bound for $\pi_1(A)$.

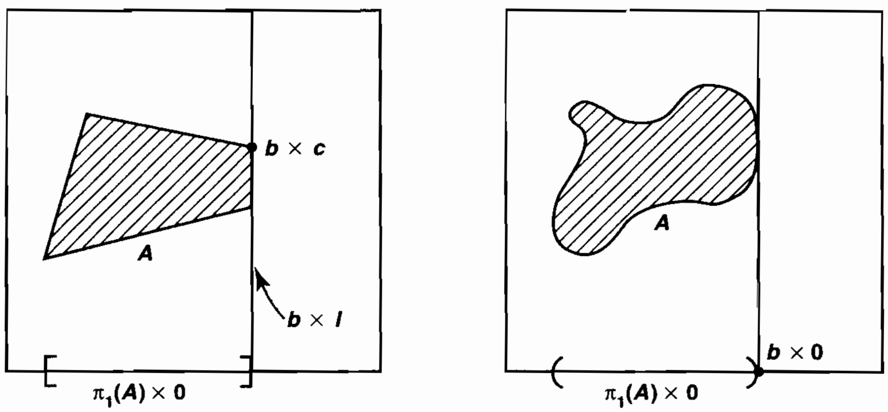


Figure 24.3

EXAMPLE 2. If X is a well-ordered set, then $X \times [0, 1)$ is a linear continuum in the dictionary order; this we leave to you to check. This set can be thought of as having been constructed by "fitting in" a set of the order type of $(0, 1)$ immediately following each element of X.

Connectedness of intervals in $\mathbb{R}$ gives rise to an especially useful criterion for showing that a space X is connected; namely, the condition that every pair of points of X can be joined by a *path* in X:

Definition. Given points x and y of the space X, a ***path*** in X from x to y is a continuous map $f : [a, b] \to X$ of some closed interval in the real line into X, such that $f(a) = x$ and $f(b) = y$. A space X is said to be ***path connected*** if every pair of points of X can be joined by a path in X.

It is easy to see that a path-connected space X is connected. Suppose $X = A \cup B$ is a separation of X. Let $f : [a, b] \to X$ be any path in X. Being the continuous image of a connected set, the set $f([a, b])$ is connected, so that it lies entirely in either A or B. Therefore, there is no path in X joining a point of A to a point of B, contrary to the assumption that X is path connected.

The converse does not hold; a connected space need not be path connected. See Examples 6 and 7 following.

EXAMPLE 3. Define the ***unit ball*** B^n in $\mathbb{R}^n$ by the equation

$$B^n = \{\mathbf{x} \mid \|\mathbf{x}\| \le 1\},$$

where

$$\|\mathbf{x}\| = \|(x_1, \dots, x_n)\| = (x_1^2 + \cdots + x_n^2)^{1/2}.$$

The unit ball is path connected; given any two points $\mathbf{x}$ and $\mathbf{y}$ of B^n, the straight-line path $f : [0, 1] \to \mathbb{R}^n$ defined by

$$f(t) = (1 - t)\mathbf{x} + t\mathbf{y}$$

lies in B^n. For if $\mathbf{x}$ and $\mathbf{y}$ are in B^n and t is in $[0, 1]$,

$$\|f(t)\| \le (1 - t)\|\mathbf{x}\| + t\|\mathbf{y}\| \le 1.$$

A similar argument shows that every open ball $B_d(\mathbf{x}, \epsilon)$ and every closed ball $\bar{B}_d(\mathbf{x}, \epsilon)$ in $\mathbb{R}^n$ is path connected.

EXAMPLE 4. Define ***punctured euclidean space*** to be the space $\mathbb{R}^n - \{\mathbf{0}\}$, where $\mathbf{0}$ is the origin in $\mathbb{R}^n$. If $n > 1$, this space is path connected: Given $\mathbf{x}$ and $\mathbf{y}$ different from $\mathbf{0}$, we can join $\mathbf{x}$ and $\mathbf{y}$ by the straight-line path between them if that path does not go through the origin. Otherwise, we can choose a point $\mathbf{z}$ not on the line joining $\mathbf{x}$ and $\mathbf{y}$, and take the broken-line path from $\mathbf{x}$ to $\mathbf{z}$, and then from $\mathbf{z}$ to $\mathbf{y}$.

EXAMPLE 5. Define the ***unit sphere*** S^{n-1} in $\mathbb{R}^n$ by the equation

$$S^{n-1} = \{\mathbf{x} \mid \|\mathbf{x}\| = 1\}.$$

If $n > 1$, it is path connected. For the map $g : \mathbb{R}^n - \{\mathbf{0}\} \to S^{n-1}$ defined by $g(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|$ is continuous and surjective; and it is easy to show that the continuous image of a path-connected space is path connected.

EXAMPLE 6. *The ordered square I_o^2 is connected but not path connected.*

Being a linear continuum, the ordered square is connected. Let $p = 0 \times 0$ and $q = 1 \times 1$. We suppose there is a path $f : [a, b] \to I_o^2$ joining p and q and derive a contradiction. The image set $f([a, b])$ must contain every point $x \times y$ of I_o^2, by the intermediate value theorem. Therefore, for each $x \in I$, the set

$$U_x = f^{-1}(x \times (0, 1))$$

is a nonempty subset of $[a, b]$; by continuity, it is open in $[a, b]$. See Figure 24.4. Choose, for each $x \in I$, a rational number q_x belonging to U_x. Since the sets U_x are disjoint, the map $x \to q_x$ is an injective mapping of I into $\mathbb{Q}$. This contradicts the fact that the interval I is uncountable (which we shall prove later).

EXAMPLE 7. Let S denote the following subset of the plane.

$$S = \{x \times \sin(1/x) \mid 0 < x \le 1\}.$$

Because S is the image of the connected set $(0, 1]$ under a continuous map, S is connected. Therefore, its closure $\bar{S}$ in $\mathbb{R}^2$ is also connected. The set $\bar{S}$ is a classical example in topology

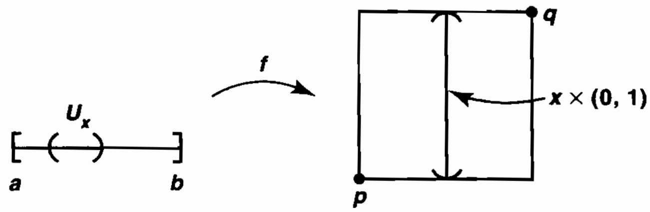


Figure 24.4

called the ***topologist's sine curve***. It is illustrated in Figure 24.5; it equals the union of S and the vertical interval $0 \times [-1, 1]$. We show that $\bar{S}$ is not path connected.

Suppose there is a path $f : [a, c] \to \bar{S}$ beginning at the origin and ending at a point of S. The set of those t for which $f(t) \in 0 \times [-1, 1]$ is closed, so it has a largest element b. Then $f : [b, c] \to \bar{S}$ is a path that maps b into the vertical interval $0 \times [-1, 1]$ and maps the other points of $[b, c]$ to points of S.

Replace $[b, c]$ by $[0, 1]$ for convenience; let $f(t) = (x(t), y(t))$. Then $x(0) = 0$, while $x(t) > 0$ and $y(t) = \sin(1/x(t))$ for $t > 0$. We show there is a sequence of points $t_n \to 0$ such that $y(t_n) = (-1)^n$. Then the sequence $y(t_n)$ does not converge, contradicting continuity of f.

To find t_n, we proceed as follows: Given n, choose u with $0 < u < x(1/n)$ such that $\sin(1/u) = (-1)^n$. Then use the intermediate value theorem to find t_n with $0 < t_n < 1/n$ such that $x(t_n) = u$.

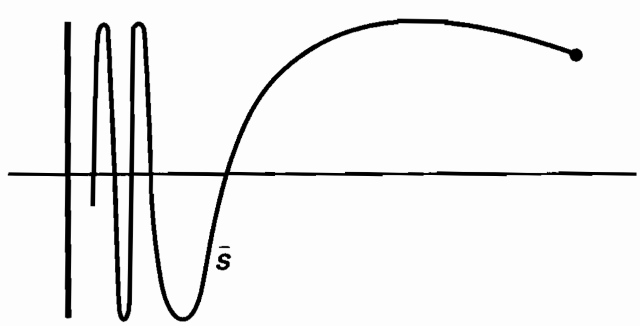


Figure 24.5

Exercises

1. (a) Show that no two of the spaces $(0, 1)$, $(0, 1]$, and $[0, 1]$ are homeomorphic. [*Hint:* What happens if you remove a point from each of these spaces?)]

(b) Suppose that there exist imbeddings $f : X \to Y$ and $g : Y \to X$. Show by means of an example that X and Y need not be homeomorphic.

(c) Show $\mathbb{R}^n$ and $\mathbb{R}$ are not homeomorphic if $n > 1$.

2. Let $f : S^1 \to \mathbb{R}$ be a continuous map. Show there exists a point x of S^1 such that $f(x) = f(-x)$.

3. Let $f : X \to X$ be continuous. Show that if $X = [0, 1]$, there is a point x such that $f(x) = x$. The point x is called a ***fixed point*** of f. What happens if X equals $[0, 1)$ or $(0, 1)$?

4. Let X be an ordered set in the order topology. Show that if X is connected, then X is a linear continuum.

5. Consider the following sets in the dictionary order. Which are linear continua?
(a) $\mathbb{Z}_+ \times [0, 1)$
(b) $[0, 1) \times \mathbb{Z}_+$
(c) $[0, 1) \times [0, 1]$
(d) $[0, 1] \times [0, 1)$

6. Show that if X is a well-ordered set, then $X \times [0, 1)$ in the dictionary order is a linear continuum.

7. (a) Let X and Y be ordered sets in the order topology. Show that if $f : X \to Y$ is order preserving and surjective, then f is a homeomorphism.
(b) Let $X = Y = \bar{\mathbb{R}}_+$. Given a positive integer n, show that the function $f(x) = x^n$ is order preserving and surjective. Conclude that its inverse, the *nth root function*, is continuous.
(c) Let X be the subspace $(-\infty, -1) \cup [0, \infty)$ of $\mathbb{R}$. Show that the function $f : X \to \mathbb{R}$ defined by setting $f(x) = x + 1$ if $x < -1$, and $f(x) = x$ if $x \geq 0$, is order preserving and surjective. Is f a homeomorphism? Compare with (a).

8. (a) Is a product of path-connected spaces necessarily path connected?
(b) If $A \subset X$ and A is path connected, is $\bar{A}$ necessarily path connected?
(c) If $f : X \to Y$ is continuous and X is path connected, is $f(X)$ necessarily path connected?
(d) If $\{A_\alpha\}$ is a collection of path-connected subspaces of X and if $\bigcap A_\alpha \neq \varnothing$, is $\bigcup A_\alpha$ necessarily path connected?

9. Assume that $\mathbb{R}$ is uncountable. Show that if A is a countable subset of $\mathbb{R}^2$, then $\mathbb{R}^2 - A$ is path connected. [*Hint:* How many lines are there passing through a given point of $\mathbb{R}^2$?]

10. Show that if U is an *open* connected subspace of $\mathbb{R}^2$, then U is path connected. [*Hint:* Show that given $x_0 \in U$, the set of points that can be joined to x_0 by a path in U is both open and closed in U.]

11. If A is a connected subspace of X, does it follow that $\text{Int}\, A$ and $\text{Bd}\, A$ are connected? Does the converse hold? Justify your answers.

12.** Recall that S_Ω denotes the minimal uncountable well-ordered set. Let L denote the ordered set $S_\Omega \times [0, 1)$ in the dictionary order, with its smallest element deleted. The set L is a classical example in topology called the ***long line.

Theorem. The long line is path connected and locally homeomorphic to $\mathbb{R}$*, but it cannot be imbedded in* $\mathbb{R}$.

(a) Let X be an ordered set; let $a < b < c$ be points of X. Show that $[a, c)$ has the order type of $[0, 1)$ if and only if both $[a, b)$ and $[b, c)$ have the order type of $[0, 1)$.

(b) Let X be an ordered set. Let $x_0 < x_1 < \cdots$ be an increasing sequence of points of X; suppose $b = \sup\{x_i\}$. Show that $[x_0, b)$ has the order type of $[0, 1)$ if and only if each interval $[x_i, x_{i+1})$ has the order type of $[0, 1)$.

(c) Let a_0 denote the smallest element of S_Ω. For each element a of S_Ω different from a_0, show that the interval $[a_0 \times 0, a \times 0)$ of $S_\Omega \times [0, 1)$ has the order type of $[0, 1)$. [*Hint:* Proceed by transfinite induction. Either a has an immediate predecessor in S_Ω, or there is an increasing sequence a_i in S_Ω with $a = \sup\{a_i\}$.]

(d) Show that L is path connected.

(e) Show that every point of L has a neighborhood homeomorphic with an open interval in $\mathbb{R}$.

(f) Show that L cannot be imbedded in $\mathbb{R}$, or indeed in $\mathbb{R}^n$ for any n. [*Hint:* Any subspace of $\mathbb{R}^n$ has a countable basis for its topology.]

*§25 Components and Local Connectedness†

Given an arbitrary space X, there is a natural way to break it up into pieces that are connected (or path connected). We consider that process now.

Definition. Given X, define an equivalence relation on X by setting $x \sim y$ if there is a connected subspace of X containing both x and y. The equivalence classes are called the ***components*** (or the "connected components") of X.

Symmetry and reflexivity of the relation are obvious. Transitivity follows by noting that if A is a connected subspace containing x and y, and if B is a connected subspace containing y and z, then $A \cup B$ is a subspace containing x and z that is connected because A and B have the point y in common.

The components of X can also be described as follows:

Theorem 25.1. *The components of* X *are connected disjoint subspaces of* X *whose union is* X*, such that each nonempty connected subspace of* X *intersects only one of them.*

Proof. Being equivalence classes, the components of X are disjoint and their union is X. Each connected subspace A of X intersects only one of them. For if A intersects the components C_1 and C_2 of X, say in points x_1 and x_2, respectively, then $x_1 \sim x_2$ by definition; this cannot happen unless $C_1 = C_2$.

†This section will be assumed in Part II of the book.

To show the component C is connected, choose a point x_0 of C. For each point x of C, we know that $x_0 \sim x$, so there is a connected subspace A_x containing x_0 and x. By the result just proved, $A_x \subset C$. Therefore,

$$C = \bigcup_{x \in C} A_x.$$

Since the subspaces A_x are connected and have the point x_0 in common, their union is connected. ∎

Definition. We define another equivalence relation on the space X by defining $x \sim y$ if there is a path in X from x to y. The equivalence classes are called the ***path components*** of X.

Let us show this is an equivalence relation. First we note that if there exists a path $f : [a, b] \to X$ from x to y whose domain is the interval $[a, b]$, then there is also a path g from x to y having the closed interval $[c, d]$ as its domain. (This follows from the fact that any two closed intervals in $\mathbb{R}$ are homeomorphic.) Now the fact that $x \sim x$ for each x in X follows from the existence of the constant path $f : [a, b] \to X$ defined by the equation $f(t) = x$ for all t. Symmetry follows from the fact that if $f : [0, 1] \to X$ is a path from x to y, then the "reverse path" $g : [0, 1] \to X$ defined by $g(t) = f(1 - t)$ is a path from y to x. Finally, transitivity is proved as follows: Let $f : [0, 1] \to X$ be a path from x to y, and let $g : [1, 2] \to X$ be a path from y to z. We can "paste f and g together" to get a path $h : [0, 2] \to X$ from x to z; the path h will be continuous by the "pasting lemma," Theorem 18.3.

One has the following theorem, whose proof is similar to that of the theorem preceding:

Theorem 25.2. *The path components of X are path-connected disjoint subspaces of X whose union is X, such that each nonempty path-connected subspace of X intersects only one of them.*

Note that each component of a space X is closed in X, since the closure of a connected subspace of X is connected. If X has only finitely many components, then each component is also open in X, since its complement is a finite union of closed sets. But in general the components of X need not be open in X.

One can say even less about the path components of X, for they need be neither open nor closed in X. Consider the following examples:

EXAMPLE 1. If $\mathbb{Q}$ is the subspace of $\mathbb{R}$ consisting of the rational numbers, then each component of $\mathbb{Q}$ consists of a single point. None of the components of $\mathbb{Q}$ are open in $\mathbb{Q}$.

EXAMPLE 2. The "topologist's sine curve" $\bar{S}$ of the preceding section is a space that has a single component (since it is connected) and two path components. One path component is the curve S and the other is the vertical interval $V = 0 \times [-1, 1]$. Note that S is open in $\bar{S}$ but not closed, while V is closed but not open.

If one forms a space from $\bar{S}$ by deleting all points of V having rational second coordinate, one obtains a space that has only one component but uncountably many path components.

Connectedness is a useful property for a space to possess. But for some purposes, it is more important that the space satisfy a connectedness condition *locally*. Roughly speaking, local connectedness means that each point has "arbitrarily small" neighborhoods that are connected. More precisely, one has the following definition:

Definition. A space X is said to be ***locally connected at x*** if for every neighborhood U of x, there is a connected neighborhood V of x contained in U. If X is locally connected at each of its points, it is said simply to be ***locally connected***. Similarly, a space X is said to be ***locally path connected at x*** if for every neighborhood U of x, there is a path-connected neighborhood V of x contained in U. If X is locally path connected at each of its points, then it is said to be ***locally path connected***.

EXAMPLE 3. Each interval and each ray in the real line is both connected and locally connected. The subspace $[-1, 0) \cup (0, 1]$ of $\mathbb{R}$ is not connected, but it is locally connected. The topologist's sine curve is connected but not locally connected. The rationals $\mathbb{Q}$ are neither connected nor locally connected.

Theorem 25.3. *A space X is locally connected if and only if for every open set U of X, each component of U is open in X.*

Proof. Suppose that X is locally connected; let U be an open set in X; let C be a component of U. If x is a point of C, we can choose a connected neighborhood V of x such that $V \subset U$. Since V is connected, it must lie entirely in the component C of U. Therefore, C is open in X.

Conversely, suppose that components of open sets in X are open. Given a point x of X and a neighborhood U of x, let C be the component of U containing x. Now C is connected; since it is open in X by hypothesis, X is locally connected at x. ∎

A similar proof holds for the following theorem:

Theorem 25.4. *A space X is locally path connected if and only if for every open set U of X, each path component of U is open in X.*

The relation between path components and components is given in the following theorem:

Theorem 25.5. *If X is a topological space, each path component of X lies in a component of X. If X is locally path connected, then the components and the path components of X are the same.*

Proof. Let C be a component of X; let x be a point of C; let P be the path component of X containing x. Since P is connected, $P \subset C$. We wish to show that if X is locally path connected, $P = C$. Suppose that $P \subsetneq C$. Let Q denote the union of all the path components of X that are different from P and intersect C; each of them necessarily lies in C, so that

$$C = P \cup Q.$$

Because X is locally path connected, each path component of X is open in X. Therefore, P (which is a path component) and Q (which is a union of path components) are open in X, so they constitute a separation of C. This contradicts the fact that C is connected. ∎

Exercises

1. What are the components and path components of $\mathbb{R}_\ell$? What are the continuous maps $f : \mathbb{R} \to \mathbb{R}_\ell$?

2. (a) What are the components and path components of $\mathbb{R}^\omega$ (in the product topology)?

(b) Consider $\mathbb{R}^\omega$ in the uniform topology. Show that $\mathbf{x}$ and $\mathbf{y}$ lie in the same component of $\mathbb{R}^\omega$ if and only if the sequence

$$\mathbf{x} - \mathbf{y} = (x_1 - y_1, x_2 - y_2, \dots)$$

is bounded. [*Hint:* It suffices to consider the case where $\mathbf{y} = \mathbf{0}$.]

(c) Give $\mathbb{R}^\omega$ the box topology. Show that $\mathbf{x}$ and $\mathbf{y}$ lie in the same component of $\mathbb{R}^\omega$ if and only if the sequence $\mathbf{x} - \mathbf{y}$ is "eventually zero." [*Hint:* If $\mathbf{x} - \mathbf{y}$ is not eventually zero, show there is homeomorphism h of $\mathbb{R}^\omega$ with itself such that $h(\mathbf{x})$ is bounded and $h(\mathbf{y})$ is unbounded.]

3. Show that the ordered square is locally connected but not locally path connected. What are the path components of this space?

4. Let X be locally path connected. Show that every connected open set in X is path connected.

5. Let X denote the rational points of the interval $[0, 1] \times 0$ of $\mathbb{R}^2$. Let T denote the union of all line segments joining the point $p = 0 \times 1$ to points of X.

(a) Show that T is path connected, but is locally connected only at the point p.

(b) Find a subset of $\mathbb{R}^2$ that is path connected but is locally connected at none of its points.

6. A space X is said to be ***weakly locally connected at x*** if for every neighborhood U of x, there is a connected subspace of X contained in U that contains a neighborhood of x. Show that if X is weakly locally connected at each of its points, then X is locally connected. [*Hint:* Show that components of open sets are open.]

7. Consider the "infinite broom" X pictured in Figure 25.1. Show that X is not locally connected at p, but is weakly locally connected at p. [*Hint:* Any connected neighborhood of p must contain all the points a_i.]

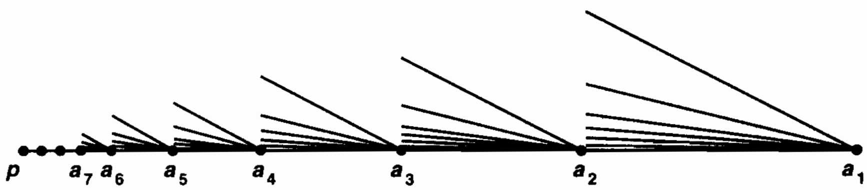


Figure 25.1

8. Let $p : X \to Y$ be a quotient map. Show that if X is locally connected, then Y is locally connected. [*Hint:* If C is a component of the open set U of Y, show that $p^{-1}(C)$ is a union of components of $p^{-1}(U)$.]

9. Let G be a topological group; let C be the component of G containing the identity element e. Show that C is a normal subgroup of G. [*Hint:* If $x \in G$, then xC is the component of G containing x.]

10. Let X be a space. Let us define $x \sim y$ if there is no separation $X = A \cup B$ of X into disjoint open sets such that $x \in A$ and $y \in B$.

(a) Show this relation is an equivalence relation. The equivalence classes are called the ***quasicomponents*** of X.

(b) Show that each component of X lies in a quasicomponent of X, and that the components and quasicomponents of X are the same if X is locally connected.

(c) Let K denote the set $\{1/n \mid n \in \mathbb{Z}_+\}$ and let $-K$ denote the set $\{-1/n \mid n \in \mathbb{Z}_+\}$. Determine the components, path components, and quasicomponents of the following subspaces of $\mathbb{R}^2$:

$$A = (K \times [0, 1]) \cup \{0 \times 0\} \cup \{0 \times 1\}.$$
$$B = A \cup ([0, 1] \times \{0\}).$$
$$C = (K \times [0, 1]) \cup (-K \times [-1, 0]) \cup ([0, 1] \times -K) \cup ([-1, 0] \times K).$$

§26 Compact Spaces

The notion of compactness is not nearly so natural as that of connectedness. From the beginnings of topology, it was clear that the closed interval $[a, b]$ of the real line had a certain property that was crucial for proving such theorems as the maximum value theorem and the uniform continuity theorem. But for a long time, it was not clear how this property should be formulated for an arbitrary topological space. It used to be thought that the crucial property of $[a, b]$ was the fact that every infinite subset of $[a, b]$ has a limit point, and this property was the one dignified with the name of compactness. Later, mathematicians realized that this formulation does not lie at the heart of the matter, but rather that a stronger formulation, in terms of open coverings of the space, is more central. The latter formulation is what we now call compactness.

It is not as natural or intuitive as the former; some familiarity with it is needed before its usefulness becomes apparent.

Definition. A collection $\mathcal{A}$ of subsets of a space X is said to ***cover*** X, or to be a ***covering*** of X, if the union of the elements of $\mathcal{A}$ is equal to X. It is called an ***open covering*** of X if its elements are open subsets of X.

Definition. A space X is said to be ***compact*** if every open covering $\mathcal{A}$ of X contains a finite subcollection that also covers X.

EXAMPLE 1. The real line $\mathbb{R}$ is not compact, for the covering of $\mathbb{R}$ by open intervals

$$\mathcal{A} = \{(n, n+2) \mid n \in \mathbb{Z}\}$$

contains no finite subcollection that covers $\mathbb{R}$.

EXAMPLE 2. The following subspace of $\mathbb{R}$ is compact:

$$X = \{0\} \cup \{1/n \mid n \in \mathbb{Z}_+\}.$$

Given an open covering $\mathcal{A}$ of X, there is an element U of $\mathcal{A}$ containing 0. The set U contains all but finitely many of the points $1/n$; choose, for each point of X not in U, an element of $\mathcal{A}$ containing it. The collection consisting of these elements of $\mathcal{A}$, along with the element U, is a finite subcollection of $\mathcal{A}$ that covers X.

EXAMPLE 3. Any space X containing only finitely many points is necessarily compact, because in this case every open covering of X is finite.

EXAMPLE 4. The interval $(0, 1]$ is not compact; the open covering

$$\mathcal{A} = \{(1/n, 1] \mid n \in \mathbb{Z}_+\}$$

contains no finite subcollection covering $(0, 1]$. Nor is the interval $(0, 1)$ compact; the same argument applies. On the other hand, the interval $[0, 1]$ *is* compact; you are probably already familiar with this fact from analysis. In any case, we shall prove it shortly.

In general, it takes some effort to decide whether a given space is compact or not. First we shall prove some general theorems that show us how to construct new compact spaces out of existing ones. Then in the next section we shall show certain specific spaces are compact. These spaces include all closed intervals in the real line, and all closed and bounded subsets of $\mathbb{R}^n$.

Let us first prove some facts about subspaces. If Y is a subspace of X, a collection $\mathcal{A}$ of subsets of X is said to ***cover*** Y if the union of its elements *contains* Y.

Lemma 26.1. *Let Y be a subspace of X. Then Y is compact if and only if every covering of Y by sets open in X contains a finite subcollection covering Y.*

Proof. Suppose that Y is compact and $\mathcal{A} = \{A_\alpha\}_{\alpha \in J}$ is a covering of Y by sets open in X. Then the collection

$$\{A_\alpha \cap Y \mid \alpha \in J\}$$

is a covering of Y by sets open in Y; hence a finite subcollection

$$\{A_{\alpha_1} \cap Y, \dots, A_{\alpha_n} \cap Y\}$$

covers Y. Then $\{A_{\alpha_1}, \dots, A_{\alpha_n}\}$ is a subcollection of $\mathcal{A}$ that covers Y.

Conversely, suppose the given condition holds; we wish to prove Y compact. Let $\mathcal{A}' = \{A'_\alpha\}$ be a covering of Y by sets open in Y. For each α, choose a set A_α open in X such that

$$A'_\alpha = A_\alpha \cap Y.$$

The collection $\mathcal{A} = \{A_\alpha\}$ is a covering of Y by sets open in X. By hypothesis, some finite subcollection $\{A_{\alpha_1}, \dots, A_{\alpha_n}\}$ covers Y. Then $\{A'_{\alpha_1}, \dots, A'_{\alpha_n}\}$ is a subcollection of $\mathcal{A}'$ that covers Y. ∎

Theorem 26.2. *Every closed subspace of a compact space is compact.*

Proof. Let Y be a closed subspace of the compact space X. Given a covering $\mathcal{A}$ of Y by sets open in X, let us form an open covering $\mathcal{B}$ of X by adjoining to $\mathcal{A}$ the single open set $X - Y$, that is,

$$\mathcal{B} = \mathcal{A} \cup \{X - Y\}.$$

Some finite subcollection of $\mathcal{B}$ covers X. If this subcollection contains the set $X - Y$, discard $X - Y$; otherwise, leave the subcollection alone. The resulting collection is a finite subcollection of $\mathcal{A}$ that covers Y. ∎

Theorem 26.3. *Every compact subspace of a Hausdorff space is closed.*

Proof. Let Y be a compact subspace of the Hausdorff space X. We shall prove that $X - Y$ is open, so that Y is closed.

Let x_0 be a point of $X - Y$. We show there is a neighborhood of x_0 that is disjoint from Y. For each point y of Y, let us choose disjoint neighborhoods U_y and V_y of the points x_0 and y, respectively (using the Hausdorff condition). The collection $\{V_y \mid y \in Y\}$ is a covering of Y by sets open in X; therefore, finitely many of them $V_{y_1}, \dots, V_{y_n}$ cover Y. The open set

$$V = V_{y_1} \cup \dots \cup V_{y_n}$$

contains Y, and it is disjoint from the open set

$$U = U_{y_1} \cap \dots \cap U_{y_n}$$

formed by taking the intersection of the corresponding neighborhoods of x_0. For if z is a point of V, then $z \in V_{y_i}$ for some i, hence $z \notin U_{y_i}$ and so $z \notin U$. See Figure 26.1.

Then U is a neighborhood of x_0 disjoint from Y, as desired. ∎

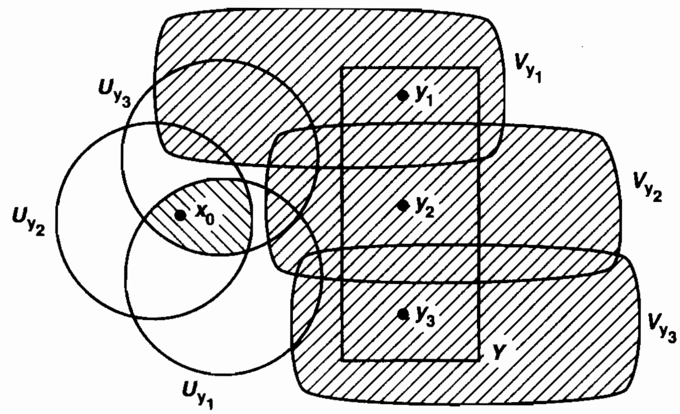


Figure 26.1

The statement we proved in the course of the preceding proof will be useful to us later, so we repeat it here for reference purposes:

Lemma 26.4. *If Y is a compact subspace of the Hausdorff space X and x_0 is not in Y, then there exist disjoint open sets U and V of X containing x_0 and Y, respectively.*

EXAMPLE 5. Once we prove that the interval $[a, b]$ in $\mathbb{R}$ is compact, it follows from Theorem 26.2 that any closed subspace of $[a, b]$ is compact. On the other hand, it follows from Theorem 26.3 that the intervals $(a, b]$ and (a, b) in $\mathbb{R}$ cannot be compact (which we knew already) because they are not closed in the Hausdorff space $\mathbb{R}$.

EXAMPLE 6. One needs the Hausdorff condition in the hypothesis of Theorem 26.3. Consider, for example, the finite complement topology on the real line. The only proper subsets of $\mathbb{R}$ that are closed in this topology are the finite sets. But *every* subset of $\mathbb{R}$ is compact in this topology, as you can check.

Theorem 26.5. *The image of a compact space under a continuous map is compact.*

Proof. Let $f : X \to Y$ be continuous; let X be compact. Let $\mathcal{A}$ be a covering of the set $f(X)$ by sets open in Y. The collection

$$\{f^{-1}(A) \mid A \in \mathcal{A}\}$$

is a collection of sets covering X; these sets are open in X because f is continuous. Hence finitely many of them, say

$$f^{-1}(A_1), \dots, f^{-1}(A_n),$$

cover X. Then the sets $A_1, \dots, A_n$ cover $f(X)$. ∎

One important use of the preceding theorem is as a tool for verifying that a map is a homeomorphism:

Theorem 26.6. *Let $f : X \to Y$ be a bijective continuous function. If X is compact and Y is Hausdorff, then f is a homeomorphism.*

Proof. We shall prove that images of closed sets of X under f are closed in Y; this will prove continuity of the map f^{-1}. If A is closed in X, then A is compact, by Theorem 26.2. Therefore, by the theorem just proved, $f(A)$ is compact. Since Y is Hausdorff, $f(A)$ is closed in Y, by Theorem 26.3. ■

Theorem 26.7. *The product of finitely many compact spaces is compact.*

Proof. We shall prove that the product of two compact spaces is compact; the theorem follows by induction for any finite product.

Step 1. Suppose that we are given spaces X and Y, with Y compact. Suppose that x_0 is a point of X, and N is an open set of $X \times Y$ containing the "slice" $x_0 \times Y$ of $X \times Y$. We prove the following:

There is a neighborhood W of x_0 in X such that N contains the entire set $W \times Y$.

The set $W \times Y$ is often called a ***tube*** about $x_0 \times Y$.

First let us cover $x_0 \times Y$ by basis elements $U \times V$ (for the topology of $X \times Y$) lying in N. The space $x_0 \times Y$ is compact, being homeomorphic to Y. Therefore, we can cover $x_0 \times Y$ by finitely many such basis elements

$$U_1 \times V_1, \dots, U_n \times V_n.$$

(We assume that each of the basis elements $U_i \times V_i$ actually intersects $x_0 \times Y$, since otherwise that basis element would be superfluous; we could discard it from the finite collection and still have a covering of $x_0 \times Y$.) Define

$$W = U_1 \cap \cdots \cap U_n.$$

The set W is open, and it contains x_0 because each set $U_i \times V_i$ intersects $x_0 \times Y$.

We assert that the sets $U_i \times V_i$, which were chosen to cover the slice $x_0 \times Y$, actually cover the tube $W \times Y$. Let $x \times y$ be a point of $W \times Y$. Consider the point $x_0 \times y$ of the slice $x_0 \times Y$ having the same y-coordinate as this point. Now $x_0 \times y$ belongs to $U_i \times V_i$ for some i, so that $y \in V_i$. But $x \in U_j$ for *every* j (because $x \in W$). Therefore, we have $x \times y \in U_i \times V_i$, as desired.

Since all the sets $U_i \times V_i$ lie in N, and since they cover $W \times Y$, the tube $W \times Y$ lies in N also. See Figure 26.2.

Step 2. Now we prove the theorem. Let X and Y be compact spaces. Let $\mathcal{A}$ be an open covering of $X \times Y$. Given $x_0 \in X$, the slice $x_0 \times Y$ is compact and may therefore be covered by finitely many elements $A_1, \dots, A_m$ of $\mathcal{A}$. Their union $N = A_1 \cup \cdots \cup A_m$ is an open set containing $x_0 \times Y$; by Step 1, the open set N contains

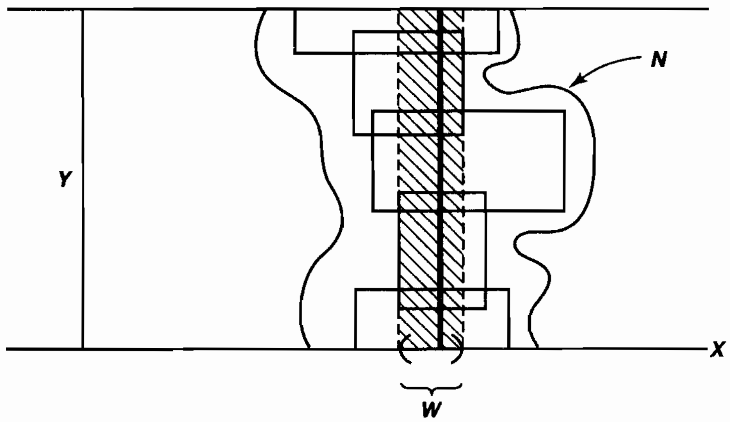


Figure 26.2

a tube $W \times Y$ about $x_0 \times Y$, where W is open in X. Then $W \times Y$ is covered by finitely many elements $A_1, \dots, A_m$ of $\mathcal{A}$.

Thus, for each x in X, we can choose a neighborhood W_x of x such that the tube $W_x \times Y$ can be covered by finitely many elements of $\mathcal{A}$. The collection of all the neighborhoods W_x is an open covering of X; therefore by compactness of X, there exists a finite subcollection

$$\{W_1, \dots, W_k\}$$

covering X. The union of the tubes

$$W_1 \times Y, \dots, W_k \times Y$$

is all of $X \times Y$; since each may be covered by finitely many elements of $\mathcal{A}$, so may $X \times Y$ be covered. ∎

The statement proved in Step 1 of the preceding proof will be useful to us later, so we repeat it here as a lemma, for reference purposes:

Lemma 26.8 (The tube lemma). *Consider the product space $X \times Y$, where Y is compact. If N is an open set of $X \times Y$ containing the slice $x_0 \times Y$ of $X \times Y$, then N contains some tube $W \times Y$ about $x_0 \times Y$, where W is a neighborhood of x_0 in X.*

EXAMPLE 7. The tube lemma is certainly not true if Y is not compact. For example, let Y be the y-axis in $\mathbb{R}^2$, and let

$$N = \{x \times y; |x| < 1/(y^2 + 1)\}.$$

Then N is an open set containing the set $0 \times \mathbb{R}$, but it contains no tube about $0 \times \mathbb{R}$. It is illustrated in Figure 26.3.

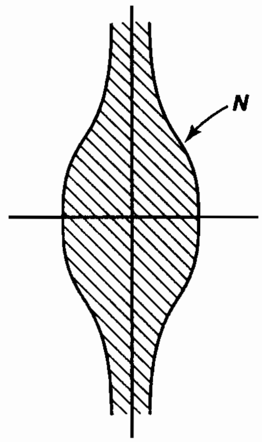


Figure 26.3

There is an obvious question to ask at this point. *Is the product of infinitely many compact spaces compact*? One would hope that the answer is "yes," and in fact it is. The result is important (and difficult) enough to be called by the name of the man who proved it; it is called the *Tychonoff theorem*.

In proving the fact that a cartesian product of connected spaces is connected, one proves it first for finite products and derives the general case from that. In proving that cartesian products of compact spaces are compact, however, there is no way to go directly from finite products to infinite ones. The infinite case demands a new approach, and the proof is a difficult one. Because of its difficulty, and also to avoid losing the main thread of our discussion in this chapter, we have decided to postpone it until later. However, you can study it now if you wish; the section in which it is proved (§37) can be studied immediately after this section without causing any disruption in continuity.

There is one final criterion for a space to be compact, a criterion that is formulated in terms of closed sets rather than open sets. It does not look very natural nor very useful at first glance, but it in fact proves to be useful on a number of occasions. First we make a definition.

Definition. A collection $\mathcal{C}$ of subsets of X is said to have the ***finite intersection property*** if for every finite subcollection

$$\{C_1, \ldots, C_n\}$$

of $\mathcal{C}$, the intersection $C_1 \cap \cdots \cap C_n$ is nonempty.

Theorem 26.9. *Let X be a topological space. Then X is compact if and only if for every collection $\mathcal{C}$ of closed sets in X having the finite intersection property, the intersection $\bigcap_{C \in \mathcal{C}} C$ of all the elements of $\mathcal{C}$ is nonempty.*

Proof. Given a collection $\mathcal{A}$ of subsets of X, let

$$\mathcal{C} = \{X - A \mid A \in \mathcal{A}\}$$

be the collection of their complements. Then the following statements hold:

(1) $\mathcal{A}$ is a collection of open sets if and only if $\mathcal{C}$ is a collection of closed sets.

(2) The collection $\mathcal{A}$ covers X if and only if the intersection $\bigcap_{C \in \mathcal{C}} C$ of all the elements of $\mathcal{C}$ is empty.

(3) The finite subcollection $\{A_1, \dots, A_n\}$ of $\mathcal{A}$ covers X if and only if the intersection of the corresponding elements $C_i = X - A_i$ of $\mathcal{C}$ is empty.

The first statement is trivial, while the second and third follow from DeMorgan's law:

$$X - (\bigcup_{\alpha \in J} A_\alpha) = \bigcap_{\alpha \in J} (X - A_\alpha).$$

The proof of the theorem now proceeds in two easy steps: taking the *contrapositive* (of the theorem), and then the *complement* (of the sets)!

The statement that X is compact is equivalent to saying: "Given any collection $\mathcal{A}$ of open subsets of X, if $\mathcal{A}$ covers X, then some finite subcollection of $\mathcal{A}$ covers X." This statement is equivalent to its contrapositive, which is the following: "Given any collection $\mathcal{A}$ of open sets, if no finite subcollection of $\mathcal{A}$ covers X, then $\mathcal{A}$ does not cover X." Letting $\mathcal{C}$ be, as earlier, the collection $\{X - A \mid A \in \mathcal{A}\}$ and applying (1)–(3), we see that this statement is in turn equivalent to the following: "Given any collection $\mathcal{C}$ of closed sets, if every finite intersection of elements of $\mathcal{C}$ is nonempty, then the intersection of all the elements of $\mathcal{C}$ is nonempty." This is just the condition of our theorem. ■

A special case of this theorem occurs when we have a ***nested sequence*** $C_1 \supset C_2 \supset \cdots \supset C_n \supset C_{n+1} \supset \dots$ of closed sets in a compact space X. If each of the sets C_n is nonempty, then the collection $\mathcal{C} = \{C_n\}_{n \in \mathbb{Z}_+}$ automatically has the finite intersection property. Then the intersection

$$\bigcap_{n \in \mathbb{Z}_+} C_n$$

is nonempty.

We shall use the closed set criterion for compactness in the next section to prove the uncountability of the set of real numbers, in Chapter 5 when we prove the Tychonoff theorem, and again in Chapter 8 when we prove the Baire category theorem.

Exercises

1. (a) Let $\mathcal{T}$ and $\mathcal{T}'$ be two topologies on the set X; suppose that $\mathcal{T}' \supset \mathcal{T}$. What does compactness of X under one of these topologies imply about compactness under the other?

(b) Show that if X is compact Hausdorff under both $\mathcal{T}$ and $\mathcal{T}'$, then either $\mathcal{T}$ and $\mathcal{T}'$ are equal or they are not comparable.

2. (a) Show that in the finite complement topology on $\mathbb{R}$, every subspace is compact.
(b) If $\mathbb{R}$ has the topology consisting of all sets A such that $\mathbb{R} - A$ is either countable or all of $\mathbb{R}$, is $[0, 1]$ a compact subspace?

3. Show that a finite union of compact subspaces of X is compact.

4. Show that every compact subspace of a metric space is bounded in that metric and is closed. Find a metric space in which not every closed bounded subspace is compact.

5. Let A and B be disjoint compact subspaces of the Hausdorff space X. Show that there exist disjoint open sets U and V containing A and B, respectively.

6. Show that if $f : X \to Y$ is continuous, where X is compact and Y is Hausdorff, then f is a closed map (that is, f carries closed sets to closed sets).

7. Show that if Y is compact, then the projection $\pi_1 : X \times Y \to X$ is a closed map.

8. *Theorem.* *Let $f : X \to Y$; let Y be compact Hausdorff. Then f is continuous if and only if the **graph** of f,*

$$G_f = \{x \times f(x) \mid x \in X\},$$

is closed in $X \times Y$. [*Hint:* If G_f is closed and V is a neighborhood of $f(x_0)$, then the intersection of G_f and $X \times (Y - V)$ is closed. Apply Exercise 7.]

9. Generalize the tube lemma as follows:
Theorem. *Let A and B be subspaces of X and Y, respectively; let N be an open set in $X \times Y$ containing $A \times B$. If A and B are compact, then there exist open sets U and V in X and Y, respectively, such that*

$$A \times B \subset U \times V \subset N.$$

10. (a) Prove the following partial converse to the uniform limit theorem:
Theorem. *Let $f_n : X \to \mathbb{R}$ be a sequence of continuous functions, with $f_n(x) \to f(x)$ for each $x \in X$. If f is continuous, and if the sequence f_n is monotone increasing, and if X is compact, then the convergence is uniform.*
[We say that f_n is *monotone increasing* if $f_n(x) \leq f_{n+1}(x)$ for all n and x.]
(b) Give examples to show that this theorem fails if you delete the requirement that X be compact, or if you delete the requirement that the sequence be monotone. [*Hint:* See the exercises of §21.]

11. *Theorem.* *Let X be a compact Hausdorff space. Let $\mathcal{A}$ be a collection of closed connected subsets of X that is simply ordered by proper inclusion. Then*

$$Y = \bigcap_{A \in \mathcal{A}} A$$

is connected. [*Hint:* If $C \cup D$ is a separation of Y, choose disjoint open sets U and V of X containing C and D, respectively, and show that

$$\bigcap_{A \in \mathcal{A}} (A - (U \cup V))$$

is not empty.]

12. Let $p : X \to Y$ be a closed continuous surjective map such that $p^{-1}(\{y\})$ is compact, for each $y \in Y$. (Such a map is called a ***perfect map***.) Show that if Y is compact, then X is compact. [*Hint:* If U is an open set containing $p^{-1}(\{y\})$, there is a neighborhood W of y such that $p^{-1}(W)$ is contained in U.]

13. Let G be a topological group.

(a) Let A and B be subspaces of G. If A is closed and B is compact, show $A \cdot B$ is closed. [*Hint:* If c is not in $A \cdot B$, find a neighborhood W of c such that $W \cdot B^{-1}$ is disjoint from A.]

(b) Let H be a subgroup of G; let $p : G \to G/H$ be the quotient map. If H is compact, show that p is a closed map.

(c) Let H be a compact subgroup of G. Show that if G/H is compact, then G is compact.

§27 Compact Subspaces of the Real Line

The theorems of the preceding section enable us to construct new compact spaces from existing ones, but in order to get very far we have to find some compact spaces to start with. The natural place to begin is the real line; we shall prove that every closed interval in $\mathbb{R}$ is compact. Applications include the extreme value theorem and the uniform continuity theorem of calculus, suitably generalized. We also give a characterization of all compact subspaces of $\mathbb{R}^n$, and a proof of the uncountability of the set of real numbers.

It turns out that in order to prove every closed interval in $\mathbb{R}$ is compact, we need only *one* of the order properties of the real line—the least upper bound property. We shall prove the theorem using only this hypothesis; then it will apply not only to the real line, but to well-ordered sets and other ordered sets as well.

Theorem 27.1. *Let X be a simply ordered set having the least upper bound property. In the order topology, each closed interval in X is compact.*

Proof. *Step 1.* Given $a < b$, let $\mathcal{A}$ be a covering of $[a, b]$ by sets open in $[a, b]$ in the subspace topology (which is the same as the order topology). We wish to prove the existence of a finite subcollection of $\mathcal{A}$ covering $[a, b]$. First we prove the following: If x is a point of $[a, b]$ different from b, then there is a point $y > x$ of $[a, b]$ such that the interval $[x, y]$ can be covered by at most two elements of $\mathcal{A}$.

If x has an immediate successor in X, let y be this immediate successor. Then $[x, y]$ consists of the two points x and y, so that it can be covered by at most two elements of $\mathcal{A}$. If x has no immediate successor in X, choose an element A of $\mathcal{A}$ containing x. Because $x \neq b$ and A is open, A contains an interval of the form $[x, c)$, for some c in $[a, b]$. Choose a point y in (x, c); then the interval $[x, y]$ is covered by the single element A of $\mathcal{A}$.

Step 2. Let C be the set of all points $y > a$ of $[a, b]$ such that the interval $[a, y]$ can be covered by finitely many elements of $\mathcal{A}$. Applying Step 1 to the case $x = a$, we see that there exists at least one such y, so C is not empty. Let c be the least upper bound of the set C; then $a < c \leq b$.

Step 3. We show that c belongs to C; that is, we show that the interval $[a, c]$ can be covered by finitely many elements of $\mathcal{A}$. Choose an element A of $\mathcal{A}$ containing c; since A is open, it contains an interval of the form $(d, c]$ for some d in $[a, b]$. If c is not in C, there must be a point z of C lying in the interval (d, c), because otherwise d would be a smaller upper bound on C than c. See Figure 27.1. Since z is in C, the interval $[a, z]$ can be covered by finitely many, say n, elements of $\mathcal{A}$. Now $[z, c]$ lies in the single element A of $\mathcal{A}$, hence $[a, c] = [a, z] \cup [z, c]$ can be covered by $n + 1$ elements of $\mathcal{A}$. Thus c is in C, contrary to assumption.

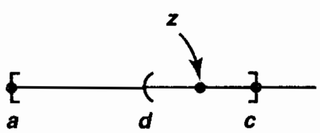


Figure 27.1

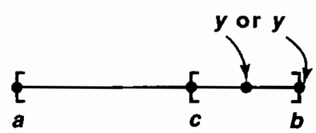


Figure 27.2

Step 4. Finally, we show that $c = b$, and our theorem is proved. Suppose that $c < b$. Applying Step 1 to the case $x = c$, we conclude that there exists a point $y > c$ of $[a, b]$ such that the interval $[c, y]$ can be covered by finitely many elements of $\mathcal{A}$. See Figure 27.2. We proved in Step 3 that c is in C, so $[a, c]$ can be covered by finitely many elements of $\mathcal{A}$. Therefore, the interval

$$[a, y] = [a, c] \cup [c, y]$$

can also be covered by finitely many elements of $\mathcal{A}$. This means that y is in C, contradicting the fact that c is an upper bound on C. ■

Corollary 27.2. *Every closed interval in* $\mathbb{R}$ *is compact.*

Now we characterize the compact subspaces of $\mathbb{R}^n$:

Theorem 27.3. *A subspace A of $\mathbb{R}^n$ is compact if and only if it is closed and is bounded in the euclidean metric d or the square metric ρ.*

Proof. It will suffice to consider only the metric ρ; the inequalities

$$\rho(x, y) \leq d(x, y) \leq \sqrt{n}\rho(x, y)$$

imply that A is bounded under d if and only if it is bounded under ρ.

Suppose that A is compact. Then, by Theorem 26.3, it is closed. Consider the collection of open sets

$$\{B_\rho(\mathbf{0}, m) \mid m \in \mathbb{Z}_+\},$$

whose union is all of $\mathbb{R}^n$. Some finite subcollection covers A. It follows that $A \subset B_\rho(\mathbf{0}, M)$ for some M. Therefore, for any two points x and y of A, we have $\rho(x, y) \leq 2M$. Thus A is bounded under ρ.

Conversely, suppose that A is closed and bounded under ρ; suppose that $\rho(x, y) \leq N$ for every pair x, y of points of A. Choose a point x_0 of A, and let $\rho(x_0, \mathbf{0}) = b$. The triangle inequality implies that $\rho(x, \mathbf{0}) \leq N + b$ for every x in A. If $P = N + b$, then A is a subset of the cube $[-P, P]^n$, which is compact. Being closed, A is also compact. ■

Students often remember this theorem as stating that the collection of compact sets in a *metric space* equals the collection of closed and bounded sets. This statement is clearly ridiculous as it stands, because the question as to which sets are bounded depends for its answer on the metric, whereas which sets are compact depends only on the topology of the space.

EXAMPLE 1. The unit sphere S^{n-1} and the closed unit ball B^n in $\mathbb{R}^n$ are compact because they are closed and bounded. The set

$$A = \{x \times (1/x) \mid 0 < x \leq 1\}$$

is closed in $\mathbb{R}^2$, but it is not compact because it is not bounded. The set

$$S = \{x \times (\sin(1/x)) \mid 0 < x \leq 1\}$$

is bounded in $\mathbb{R}^2$, but it is not compact because it is not closed.

Now we prove the extreme value theorem of calculus, in suitably generalized form.

Theorem 27.4 (Extreme value theorem). *Let $f : X \to Y$ be continuous, where Y is an ordered set in the order topology. If X is compact, then there exist points c and d in X such that $f(c) \leq f(x) \leq f(d)$ for every $x \in X$.*

The extreme value theorem of calculus is the special case of this theorem that occurs when we take X to be a closed interval in $\mathbb{R}$ and Y to be $\mathbb{R}$.

Proof. Since f is continuous and X is compact, the set $A = f(X)$ is compact. We show that A has a largest element M and a smallest element m. Then since m and M belong to A, we must have $m = f(c)$ and $M = f(d)$ for some points c and d of X.

If A has no largest element, then the collection

$$\{(-\infty, a) \mid a \in A\}$$

forms an open covering of A. Since A is compact, some finite subcollection

$$\{(-\infty, a_1), \dots, (-\infty, a_n)\}$$

covers A. If a_i is the largest of the elements $a_1, \dots a_n$, then a_i belongs to none of these sets, contrary to the fact that they cover A.

A similar argument shows that A has a smallest element. ■

Now we prove the uniform continuity theorem of calculus. In the process, we are led to introduce a new notion that will prove to be surprisingly useful, that of a *Lebesgue number* for an open covering of a metric space. First, a preliminary notion:

Definition. Let (X, d) be a metric space; let A be a nonempty subset of X. For each $x \in X$, we define the ***distance from x to A*** by the equation

$$d(x, A) = \inf\{d(x, a) \mid a \in A\}.$$

It is easy to show that for fixed A, the function $d(x, A)$ is a continuous function of x: Given $x, y \in X$, one has the inequalities

$$d(x, A) \leq d(x, a) \leq d(x, y) + d(y, a),$$

for each $a \in A$. It follows that

$$d(x, A) - d(x, y) \leq \inf d(y, a) = d(y, A),$$

so that

$$d(x, A) - d(y, A) \leq d(x, y).$$

The same inequality holds with x and y interchanged; continuity of the function $d(x, A)$ follows.

Now we introduce the notion of Lebesgue number. Recall that the *diameter* of a bounded subset A of a metric space (X, d) is the number

$$\sup\{d(a_1, a_2) \mid a_1, a_2 \in A\}.$$

Lemma 27.5 (The Lebesgue number lemma). *Let $\mathcal{A}$ be an open covering of the metric space (X, d). If X is compact, there is a $\delta > 0$ such that for each subset of X having diameter less than δ, there exists an element of $\mathcal{A}$ containing it.*

The number δ is called a ***Lebesgue number*** for the covering $\mathcal{A}$.

Proof. Let $\mathcal{A}$ be an open covering of X. If X itself is an element of $\mathcal{A}$, then any positive number is a Lebesgue number for $\mathcal{A}$. So assume X is not an element of $\mathcal{A}$.

Choose a finite subcollection $\{A_1, \ldots, A_n\}$ of $\mathcal{A}$ that covers X. For each i, set $C_i = X - A_i$, and define $f : X \to \mathbb{R}$ by letting $f(x)$ be the average of the numbers $d(x, C_i)$. That is,

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} d(x, C_i).$$

We show that $f(x) > 0$ for all x. Given $x \in X$, choose i so that $x \in A_i$. Then choose ϵ so the ϵ-neighborhood of x lies in A_i. Then $d(x, C_i) \geq \epsilon$, so that $f(x) \geq \epsilon/n$.

Since f is continuous, it has a minimum value δ; we show that δ is our required Lebesgue number. Let B be a subset of X of diameter less than δ. Choose a point x_0 of B; then B lies in the δ-neighborhood of x_0. Now

$$\delta \leq f(x_0) \leq d(x_0, C_m),$$

where $d(x_0, C_m)$ is the largest of the numbers $d(x_0, C_i)$. Then the δ-neighborhood of x_0 is contained in the element $A_m = X - C_m$ of the covering $\mathcal{A}$. ■

Definition. A function f from the metric space (X, d_X) to the metric space (Y, d_Y) is said to be ***uniformly continuous*** if given $\epsilon > 0$, there is a $\delta > 0$ such that for every pair of points x_0, x_1 of X,

$$d_X(x_0, x_1) < \delta \Longrightarrow d_Y(f(x_0), f(x_1)) < \epsilon.$$

Theorem 27.6 (Uniform continuity theorem). *Let $f : X \to Y$ be a continuous map of the compact metric space (X, d_X) to the metric space (Y, d_Y). Then f is uniformly continuous.*

Proof. Given $\epsilon > 0$, take the open covering of Y by balls $B(y, \epsilon/2)$ of radius $\epsilon/2$. Let $\mathcal{A}$ be the open covering of X by the inverse images of these balls under f. Choose δ to be a Lebesgue number for the covering $\mathcal{A}$. Then if x_1 and x_2 are two points of X such that $d_X(x_1, x_2) < \delta$, the two-point set $\{x_1, x_2\}$ has diameter less than δ, so that its image $\{f(x_1), f(x_2)\}$ lies in some ball $B(y, \epsilon/2)$. Then $d_Y(f(x_1), f(x_2)) < \epsilon$, as desired. ■

Finally, we prove that the real numbers are uncountable. The interesting thing about this proof is that it involves no algebra at all—no decimal or binary expansions of real numbers or the like—just the order properties of $\mathbb{R}$.

Definition. If X is a space, a point x of X is said to be an ***isolated point*** of X if the one-point set $\{x\}$ is open in X.

Theorem 27.7. *Let X be a nonempty compact Hausdorff space. If X has no isolated points, then X is uncountable.*

Proof. *Step 1.* We show first that given any nonempty open set U of X and any point x of X, there exists a nonempty open set V contained in U such that $x \notin \bar{V}$.

Choose a point y of U different from x; this is possible if x is in U because x is not an isolated point of X and it is possible if x is not in U simply because U is nonempty. Now choose disjoint open sets W_1 and W_2 about x and y, respectively. Then the set $V = W_2 \cap U$ is the desired open set; it is contained in U, it is nonempty because it contains y, and its closure does not contain x. See Figure 27.3.

Step 2. We show that given $f : \mathbb{Z}_+ \to X$, the function f is not surjective. It follows that X is uncountable.

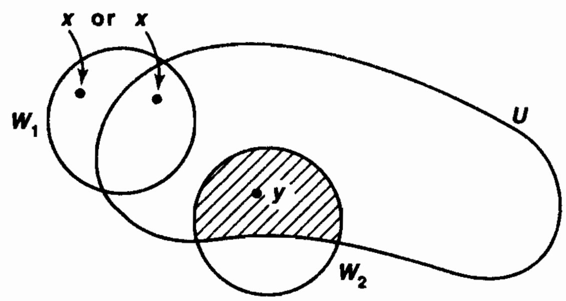


Figure 27.3

Let $x_n = f(n)$. Apply Step 1 to the nonempty open set $U = X$ to choose a nonempty open set $V_1 \subset X$ such that $\bar{V}_1$ does not contain x_1. In general, given V_{n-1} open and nonempty, choose V_n to be a nonempty open set such that $V_n \subset V_{n-1}$ and $\bar{V}_n$ does not contain x_n. Consider the nested sequence

$$\bar{V}_1 \supset \bar{V}_2 \supset \cdots$$

of nonempty closed sets of X. Because X is compact, there is a point $x \in \bigcap \bar{V}_n$, by Theorem 26.9. Now x cannot equal x_n for any n, since x belongs to $\bar{V}_n$ and x_n does not. ∎

Corollary 27.8. *Every closed interval in $\mathbb{R}$ is uncountable.*

Exercises

1. Prove that if X is an ordered set in which every closed interval is compact, then X has the least upper bound property.
2. Let X be a metric space with metric d; let $A \subset X$ be nonempty.
   (a) Show that $d(x, A) = 0$ if and only if $x \in \bar{A}$.
   (b) Show that if A is compact, $d(x, A) = d(x, a)$ for some $a \in A$.
   (c) Define the *ϵ-neighborhood* of A in X to be the set

   $$U(A, \epsilon) = \{x \mid d(x, A) < \epsilon\}.$$

   Show that $U(A, \epsilon)$ equals the union of the open balls $B_d(a, \epsilon)$ for $a \in A$.
   (d) Assume that A is compact; let U be an open set containing A. Show that some ϵ-neighborhood of A is contained in U.
   (e) Show the result in (d) need not hold if A is closed but not compact.
3. Recall that $\mathbb{R}_K$ denotes $\mathbb{R}$ in the K-topology.
   (a) Show that $[0, 1]$ is not compact as a subspace of $\mathbb{R}_K$.

(b) Show that $\mathbb{R}_K$ is connected. [*Hint:* $(-\infty, 0)$ and $(0, \infty)$ inherit their usual topologies as subspaces of $\mathbb{R}_K$.]
(c) Show that $\mathbb{R}_K$ is not path connected.

4. Show that a connected metric space having more than one point is uncountable.

5. Let X be a compact Hausdorff space; let $\{A_n\}$ be a countable collection of closed sets of X. Show that if each set A_n has empty interior in X, then the union $\bigcup A_n$ has empty interior in X. [*Hint:* Imitate the proof of Theorem 27.7.]

This is a special case of the *Baire category theorem*, which we shall study in Chapter 8.

6. Let A_0 be the closed interval $[0, 1]$ in $\mathbb{R}$. Let A_1 be the set obtained from A_0 by deleting its "middle third" $(\frac{1}{3}, \frac{2}{3})$. Let A_2 be the set obtained from A_1 by deleting its "middle thirds" $(\frac{1}{9}, \frac{2}{9})$ and $(\frac{7}{9}, \frac{8}{9})$. In general, define A_n by the equation

$$A_n = A_{n-1} - \bigcup_{k=0}^{\infty} \left(\frac{1+3k}{3^n}, \frac{2+3k}{3^n} \right).$$

The intersection

$$C = \bigcap_{n \in \mathbb{Z}_+} A_n$$

is called the ***Cantor set***; it is a subspace of $[0, 1]$.
(a) Show that C is totally disconnected.
(b) Show that C is compact.
(c) Show that each set A_n is a union of finitely many disjoint closed intervals of length $1/3^n$; and show that the end points of these intervals lie in C.
(d) Show that C has no isolated points.
(e) Conclude that C is uncountable.

§28 Limit Point Compactness

As indicated when we first mentioned compact sets, there are other formulations of the notion of compactness that are frequently useful. In this section we introduce one of them. Weaker in general than compactness, it coincides with compactness for metrizable spaces.

Definition. A space X is said to be ***limit point compact*** if every infinite subset of X has a limit point.

In some ways this property is more natural and intuitive than that of compactness. In the early days of topology, it was given the name "compactness," while the open covering formulation was called "bicompactness." Later, the word "compact" was shifted to apply to the open covering definition, leaving this one to search for a new name. It still has not found a name on which everyone agrees. On historical grounds, some call it "Fréchet compactness"; others call it the "Bolzano-Weierstrass property." We have invented the term "limit point compactness." It seems as good a term as any; at least it describes what the property is about.

Theorem 28.1. *Compactness implies limit point compactness, but not conversely.*

Proof. Let X be a compact space. Given a subset A of X, we wish to prove that if A is infinite, then A has a limit point. We prove the contrapositive—if A has no limit point, then A must be finite.

So suppose A has no limit point. Then A contains all its limit points, so that A is closed. Furthermore, for each $a \in A$ we can choose a neighborhood U_a of a such that U_a intersects A in the point a alone. The space X is covered by the open set $X - A$ and the open sets U_a; being compact, it can be covered by finitely many of these sets. Since $X - A$ does not intersect A, and each set U_a contains only one point of A, the set A must be finite. ∎

EXAMPLE 1. Let Y consist of two points; give Y the topology consisting of Y and the empty set. Then the space $X = \mathbb{Z}_+ \times Y$ is limit point compact, for *every* nonempty subset of X has a limit point. It is not compact, for the covering of X by the open sets $U_n = \{n\} \times Y$ has no finite subcollection covering X.

EXAMPLE 2. Here is a less trivial example. Consider the minimal uncountable well-ordered set S_Ω, in the order topology. The space S_Ω is not compact, since it has no largest element. However, it is limit point compact: Let A be an infinite subset of S_Ω. Choose a subset B of A that is countably infinite. Being countable, the set B has an upper bound b in S_Ω; then B is a subset of the interval $[a_0, b]$ of S_Ω, where a_0 is the smallest element of S_Ω. Since S_Ω has the least upper bound property, the interval $[a_0, b]$ is compact. By the preceding theorem, B has a limit point x in $[a_0, b]$. The point x is also a limit point of A. Thus S_Ω is limit point compact.

We now show these two versions of compactness coincide for metrizable spaces; for this purpose, we introduce yet another version of compactness called *sequential compactness*. This result will be used in Chapter 7.

Definition. Let X be a topological space. If (x_n) is a sequence of points of X, and if

$$n_1 < n_2 < \cdots < n_i < \cdots$$

is an increasing sequence of positive integers, then the sequence (y_i) defined by setting $y_i = x_{n_i}$ is called a ***subsequence*** of the sequence (x_n). The space X is said to be ***sequentially compact*** if every sequence of points of X has a convergent subsequence.

***Theorem 28.2.** *Let X be a metrizable space. Then the following are equivalent:*

(1) *X is compact.*

(2) *X is limit point compact.*

(3) *X is sequentially compact.*

Proof. We have already proved that (1) $\Rightarrow$ (2). To show that (2) $\Rightarrow$ (3), assume that X is limit point compact. Given a sequence (x_n) of points of X, consider the set $A = \{x_n \mid n \in \mathbb{Z}_+\}$. If the set A is finite, then there is a point x such that $x = x_n$ for infinitely many values of n. In this case, the sequence (x_n) has a subsequence that is constant, and therefore converges trivially. On the other hand, if A is infinite, then A has a limit point x. We define a subsequence of (x_n) converging to x as follows: First choose n_1 so that

$$x_{n_1} \in B(x, 1).$$

Then suppose that the positive integer n_{i-1} is given. Because the ball $B(x, 1/i)$ intersects A in infinitely many points, we can choose an index $n_i > n_{i-1}$ such that

$$x_{n_i} \in B(x, 1/i).$$

Then the subsequence $x_{n_1}, x_{n_2}, \dots$ converges to x.

Finally, we show that (3) $\Rightarrow$ (1). This is the hardest part of the proof.

First, we show that if X is sequentially compact, then the Lebesgue number lemma holds for X. (This would follow from compactness, but compactness is what we are trying to prove!) Let $\mathcal{A}$ be an open covering of X. We assume that there is no $\delta > 0$ such that each set of diameter less than δ has an element of $\mathcal{A}$ containing it, and derive a contradiction.

Our assumption implies in particular that for each positive integer n, there exists a set of diameter less than $1/n$ that is not contained in any element of $\mathcal{A}$; let C_n be such a set. Choose a point $x_n \in C_n$, for each n. By hypothesis, some subsequence (x_{n_i}) of the sequence (x_n) converges, say to the point a. Now a belongs to some element A of the collection $\mathcal{A}$; because A is open, we may choose an $\epsilon > 0$ such that $B(a, \epsilon) \subset A$. If i is large enough that $1/n_i < \epsilon/2$, then the set C_{n_i} lies in the $\epsilon/2$-neighborhood of x_{n_i}; if i is also chosen large enough that $d(x_{n_i}, a) < \epsilon/2$, then C_{n_i} lies in the ϵ-neighborhood of a. But this means that $C_{n_i} \subset A$, contrary to hypothesis.

Second, we show that if X is sequentially compact, then given $\epsilon > 0$, there exists a finite covering of X by open ϵ-balls. Once again, we proceed by contradiction. Assume that there exists an $\epsilon > 0$ such that X cannot be covered by finitely many ϵ-balls. Construct a sequence of points x_n of X as follows: First, choose x_1 to be any point of X. Noting that the ball $B(x_1, \epsilon)$ is not all of X (otherwise X could be covered by a single ϵ-ball), choose x_2 to be a point of X not in $B(x_1, \epsilon)$. In general, given $x_1, \dots, x_n$, choose x_{n+1} to be a point not in the union

$$B(x_1, \epsilon) \cup \cdots \cup B(x_n, \epsilon),$$

using the fact that these balls do not cover X. Note that by construction $d(x_{n+1}, x_i) \geq \epsilon$ for $i = 1, \dots, n$. Therefore, the sequence (x_n) can have no convergent subsequence; in fact, any ball of radius $\epsilon/2$ can contain x_n for at most *one* value of n.

Finally, we show that if X is sequentially compact, then X is compact. Let $\mathcal{A}$ be an open covering of X. Because X is sequentially compact, the open covering $\mathcal{A}$ has a Lebesgue number δ. Let $\epsilon = \delta/3$; use sequential compactness of X to find a finite covering of X by open ϵ-balls. Each of these balls has diameter at most $2\delta/3$, so it lies in an element of $\mathcal{A}$. Choosing one such element of $\mathcal{A}$ for each of these ϵ-balls, we obtain a finite subcollection of $\mathcal{A}$ that covers X. ∎

EXAMPLE 3. Recall that $\bar{S}_\Omega$ denotes the minimal uncountable well-ordered set S_Ω with the point Ω adjoined. (In the order topology, Ω is a limit point of S_Ω, which is why we introduced the notation $\bar{S}_\Omega$ for $S_\Omega \cup \{\Omega\}$, back in §10.) It is easy to see that the space $\bar{S}_\Omega$ is not metrizable, for it does not satisfy the sequence lemma: The point Ω is a limit point of S_Ω; but it is not the limit of a sequence of points of S_Ω, for any sequence of points of S_Ω has an upper bound in S_Ω. The space S_Ω, on the other hand, does satisfy the sequence lemma, as you can readily check. Nevertheless, S_Ω is not metrizable, for it is limit point compact but not compact.

Exercises

1. Give $[0, 1]^\omega$ the uniform topology. Find an infinite subset of this space that has no limit point.

2. Show that $[0, 1]$ is not limit point compact as a subspace of $\mathbb{R}_\ell$.

3. Let X be limit point compact.
 (a) If $f : X \to Y$ is continuous, does it follow that $f(X)$ is limit point compact?
 (b) If A is a closed subset of X, does it follow that A is limit point compact?
 (c) If X is a subspace of the Hausdorff space Z, does it follow that X is closed in Z?

 We comment that it is not in general true that the product of two limit point compact spaces is limit point compact, even if the Hausdorff condition is assumed. But the examples are fairly sophisticated. See [S-S], Example 112.

4. A space X is said to be ***countably compact*** if every countable open covering of X contains a finite subcollection that covers X. Show that for a T_1 space X, countable compactness is equivalent to limit point compactness. [*Hint:* If no finite subcollection of U_n covers X, choose $x_n \notin U_1 \cup \cdots \cup U_n$, for each n.]

5. Show that X is countably compact if and only if every nested sequence $C_1 \supset C_2 \supset \cdots$ of closed nonempty sets of X has a nonempty intersection.

6. Let (X, d) be a metric space. If $f : X \to X$ satisfies the condition

$$d(f(x), f(y)) = d(x, y)$$

for all $x, y \in X$, then f is called an ***isometry*** of X. Show that if f is an isometry and X is compact, then f is bijective and hence a homeomorphism. [*Hint:* If $a \notin f(X)$, choose ϵ so that the ϵ-neighborhood of a is disjoint from $f(X)$. Set $x_1 = a$, and $x_{n+1} = f(x_n)$ in general. Show that $d(x_n, x_m) \geq \epsilon$ for $n \neq m$.]

7. Let (X, d) be a metric space. If f satisfies the condition

$$d(f(x), f(y)) < d(x, y)$$

for all $x, y \in X$ with $x \neq y$, then f is called a ***shrinking map***. If there is a number $\alpha < 1$ such that

$$d(f(x), f(y)) \leq \alpha d(x, y)$$

for all $x, y \in X$, then f is called a ***contraction***. A ***fixed point*** of f is a point x such that $f(x) = x$.

(a) If f is a contraction and X is compact, show f has a unique fixed point. [*Hint:* Define $f^1 = f$ and $f^{n+1} = f \circ f^n$. Consider the intersection A of the sets $A_n = f^n(X)$.]

(b) Show more generally that if f is a shrinking map and X is compact, then f has a unique fixed point. [*Hint:* Let A be as before. Given $x \in A$, choose x_n so that $x = f^{n+1}(x_n)$. If a is the limit of some subsequence of the sequence $y_n = f^n(x_n)$, show that $a \in A$ and $f(a) = x$. Conclude that $A = f(A)$, so that $\operatorname{diam} A = 0$.]

(c) Let $X = [0, 1]$. Show that $f(x) = x - x^2/2$ maps X into X and is a shrinking map that is not a contraction. [*Hint:* Use the mean-value theorem of calculus.]

(d) The result in (a) holds if X is a complete metric space, such as $\mathbb{R}$; see the exercises of §43. The result in (b) does not: Show that the map $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = [x + (x^2 + 1)^{1/2}]/2$ is a shrinking map that is not a contraction and has no fixed point.

§29 Local Compactness

In this section we study the notion of local compactness, and we prove the basic theorem that any locally compact Hausdorff space can be imbedded in a certain compact Hausdorff space that is called its *one-point compactification*.

Definition. A space X is said to be ***locally compact at x*** if there is some compact subspace C of X that contains a neighborhood of x. If X is locally compact at each of its points, X is said simply to be ***locally compact***.

Note that a compact space is automatically locally compact.

EXAMPLE 1. The real line $\mathbb{R}$ is locally compact. The point x lies in some interval (a, b), which in turn is contained in the compact subspace $[a, b]$. The subspace $\mathbb{Q}$ of rational numbers is not locally compact, as you can check.

EXAMPLE 2. The space $\mathbb{R}^n$ is locally compact; the point x lies in some basis element $(a_1, b_1) \times \cdots \times (a_n, b_n)$, which in turn lies in the compact subspace $[a_1, b_1] \times \cdots \times [a_n, b_n]$. The space $\mathbb{R}^\omega$ is not locally compact; *none* of its basis elements are contained in compact subspaces. For if

$$B = (a_1, b_1) \times \cdots \times (a_n, b_n) \times \mathbb{R} \times \cdots \times \mathbb{R} \times \cdots$$

were contained in a compact subspace, then its closure

$$\bar{B} = [a_1, b_1] \times \cdots \times [a_n, b_n] \times \mathbb{R} \times \cdots$$

would be compact, which it is not.

EXAMPLE 3. Every simply ordered set X having the least upper bound property is locally compact: Given a basis element for X, it is contained in a closed interval in X, which is compact.

Two of the most well-behaved classes of spaces to deal with in mathematics are the metrizable spaces and the compact Hausdorff spaces. Such spaces have many useful properties, which one can use in proving theorems and making constructions and the like. If a given space is not of one of these types, the next best thing one can hope for is that it is a subspace of one of these spaces. Of course, a subspace of a metrizable space is itself metrizable, so one does not get any new spaces in this way. But a subspace of a compact Hausdorff space need not be compact. Thus arises the question: Under what conditions is a space homeomorphic with a subspace of a compact Hausdorff space? We give one answer here. We shall return to this question in Chapter 5 when we study compactifications in general.

Theorem 29.1. *Let X be a space. Then X is locally compact Hausdorff if and only if there exists a space Y satisfying the following conditions:*

(1) *X is a subspace of Y.*

(2) *The set $Y - X$ consists of a single point.*

(3) *Y is a compact Hausdorff space.*

If Y and Y' are two spaces satisfying these conditions, then there is a homeomorphism of Y with Y' that equals the identity map on X.

Proof. *Step 1.* We first verify uniqueness. Let Y and Y' be two spaces satisfying these conditions. Define $h : Y \to Y'$ by letting h map the single point p of $Y - X$ to the point q of $Y' - X$, and letting h equal the identity on X. We show that if U is open in Y, then $h(U)$ is open in Y'. Symmetry then implies that h is a homeomorphism.

First, consider the case where U does not contain p. Then $h(U) = U$. Since U is open in Y and is contained in X, it is open in X. Because X is open in Y', the set U is also open in Y', as desired.

Second, suppose that U contains p. Since $C = Y - U$ is closed in Y, it is compact as a subspace of Y. Because C is contained in X, it is a compact subspace of X. Then because X is a subspace of Y', the space C is also a compact subspace of Y'. Because Y' is Hausdorff, C is closed in Y', so that $h(U) = Y' - C$ is open in Y', as desired.

Step 2. Now we suppose X is locally compact Hausdorff and construct the space Y. Step 1 gives us an idea how to proceed. Let us take some object that is not a point of X, denote it by the symbol ∞ for convenience, and adjoin it to X, forming the set $Y = X \cup \{\infty\}$. Topologize Y by defining the collection of open sets of Y to consist of (1) all sets U that are open in X, and (2) all sets of the form $Y - C$, where C is a compact subspace of X.

We need to check that this collection is, in fact, a topology on Y. The empty set is a set of type (1), and the space Y is a set of type (2). Checking that the intersection of two open sets is open involves three cases:

$$\begin{aligned} U_1 \cap U_2 & \qquad \text{is of type (1).} \\ (Y - C_1) \cap (Y - C_2) = Y - (C_1 \cup C_2) & \qquad \text{is of type (2).} \\ U_1 \cap (Y - C_1) = U_1 \cap (X - C_1) & \qquad \text{is of type (1),} \end{aligned}$$

because C_1 is closed in X. Similarly, one checks that the union of any collection of open sets is open:

$$\begin{aligned} \bigcup U_\alpha = U & \qquad \text{is of type (1).} \\ \bigcup (Y - C_\beta) = Y - (\bigcap C_\beta) = Y - C & \qquad \text{is of type (2).} \\ (\bigcup U_\alpha) \cup (\bigcup (Y - C_\beta)) = U \cup (Y - C) = Y - (C - U), & \end{aligned}$$

which is of type (2) because $C - U$ is a closed subspace of C and therefore compact.

Now we show that X is a subspace of Y. Given any open set of Y, we show its intersection with X is open in X. If U is of type (1), then $U \cap X = U$; if $Y - C$ is of type (2), then $(Y - C) \cap X = X - C$; both of these sets are open in X. Conversely, any set open in X is a set of type (1) and therefore open in Y by definition.

To show that Y is compact, let $\mathcal{A}$ be an open covering of Y. The collection $\mathcal{A}$ must contain an open set of type (2), say $Y - C$, since none of the open sets of type (1) contain the point ∞. Take all the members of $\mathcal{A}$ different from $Y - C$ and intersect them with X; they form a collection of open sets of X covering C. Because C is compact, finitely many of them cover C; the corresponding finite collection of elements of $\mathcal{A}$ will, along with the element $Y - C$, cover all of Y.

To show that Y is Hausdorff, let x and y be two points of Y. If both of them lie in X, there are disjoint sets U and V open in X containing them, respectively. On the other hand, if $x \in X$ and $y = \infty$, we can choose a compact set C in X containing a neighborhood U of x. Then U and $Y - C$ are disjoint neighborhoods of x and ∞, respectively, in Y.

Step 3. Finally, we prove the converse. Suppose a space Y satisfying conditions (1)–(3) exists. Then X is Hausdorff because it is a subspace of the Hausdorff space Y. Given $x \in X$, we show X is locally compact at x. Choose disjoint open sets U and V of Y containing x and the single point of $Y - X$, respectively. Then the set $C = Y - V$ is closed in Y, so it is a compact subspace of Y. Since C lies in X, it is also compact as a subspace of X; it contains the neighborhood U of x. ∎

If X itself should happen to be compact, then the space Y of the preceding theorem is not very interesting, for it is obtained from X by adjoining a single isolated point. However, if X is not compact, then the point of $Y - X$ is a limit point of X, so that $\bar{X} = Y$.

Definition. If Y is a compact Hausdorff space and X is a proper subspace of Y whose closure equals Y, then Y is said to be a ***compactification*** of X. If $Y - X$ equals a single point, then Y is called the ***one-point compactification*** of X.

We have shown that X has a one-point compactification Y if and only if X is a locally compact Hausdorff space that is not itself compact. We speak of Y as "the" one-point compactification because Y is uniquely determined up to a homeomorphism.

EXAMPLE 4. The one-point compactification of the real line $\mathbb{R}$ is homeomorphic with the circle, as you may readily check. Similarly, the one-point compactification of $\mathbb{R}^2$ is homeomorphic to the sphere S^2. If $\mathbb{R}^2$ is looked at as the space $\mathbb{C}$ of complex numbers, then $\mathbb{C} \cup \{\infty\}$ is called the *Riemann sphere*, or the *extended complex plane*.

In some ways our definition of local compactness is not very satisfying. Usually one says that a space X satisfies a given property "locally" if every $x \in X$ has "arbitrarily small" neighborhoods having the given property. Our definition of local compactness has nothing to do with "arbitrarily small" neighborhoods, so there is some question whether we should call it local compactness at all.

Here is another formulation of local compactness, one more truly "local" in nature; it is equivalent to our definition when X is Hausdorff.

Theorem 29.2. *Let X be a Hausdorff space. Then X is locally compact if and only if given x in X, and given a neighborhood U of x, there is a neighborhood V of x such that $\bar{V}$ is compact and $\bar{V} \subset U$.*

Proof. Clearly this new formulation implies local compactness; the set $C = \bar{V}$ is the desired compact set containing a neighborhood of x. To prove the converse, suppose X is locally compact; let x be a point of X and let U be a neighborhood of x. Take the one-point compactification Y of X, and let C be the set $Y - U$. Then C is closed in Y, so that C is a compact subspace of Y. Apply Lemma 26.4 to choose disjoint open sets V and W containing x and C, respectively. Then the closure $\bar{V}$ of V in Y is compact; furthermore, $\bar{V}$ is disjoint from C, so that $\bar{V} \subset U$, as desired. ■

Corollary 29.3. *Let X be locally compact Hausdorff; let A be a subspace of X. If A is closed in X or open in X, then A is locally compact.*

Proof. Suppose that A is closed in X. Given $x \in A$, let C be a compact subspace of X containing the neighborhood U of x in X. Then $C \cap A$ is closed in C and thus compact, and it contains the neighborhood $U \cap A$ of x in A. (We have not used the Hausdorff condition here.)

Suppose now that A is open in X. Given $x \in A$, we apply the preceding theorem to choose a neighborhood V of x in X such that $\bar{V}$ is compact and $\bar{V} \subset A$. Then $C = \bar{V}$ is a compact subspace of A containing the neighborhood V of x in A. ■

Corollary 29.4. *A space X is homeomorphic to an open subspace of a compact Hausdorff space if and only if X is locally compact Hausdorff.*

Proof. This follows from Theorem 29.1 and Corollary 29.3. ■

Exercises

1. Show that the rationals $\mathbb{Q}$ are not locally compact.
2. Let $\{X_\alpha\}$ be an indexed family of nonempty spaces.
   (a) Show that if $\prod X_\alpha$ is locally compact, then each X_α is locally compact and X_α is compact for all but finitely many values of α.
   (b) Prove the converse, assuming the Tychonoff theorem.
3. Let X be a locally compact space. If $f : X \to Y$ is continuous, does it follow that $f(X)$ is locally compact? What if f is both continuous and open? Justify your answer.
4. Show that $[0, 1]^\omega$ is not locally compact in the uniform topology.
5. If $f : X_1 \to X_2$ is a homeomorphism of locally compact Hausdorff spaces, show f extends to a homeomorphism of their one-point compactifications.
6. Show that the one-point compactification of $\mathbb{R}$ is homeomorphic with the circle S^1.
7. Show that the one-point compactification of S_Ω is homeomorphic with $\bar{S}_\Omega$.
8. Show that the one-point compactification of $\mathbb{Z}_+$ is homeomorphic with the subspace $\{0\} \cup \{1/n \mid n \in \mathbb{Z}_+\}$ of $\mathbb{R}$.
9. Show that if G is a locally compact topological group and H is a subgroup, then G/H is locally compact.
10. Show that if X is a Hausdorff space that is locally compact at the point x, then for each neighborhood U of x, there is a neighborhood V of x such that $\bar{V}$ is compact and $\bar{V} \subset U$.

*11. Prove the following:
   (a) *Lemma.* *If $p : X \to Y$ is a quotient map and if Z is a locally compact Hausdorff space, then the map*

$$\pi = p \times i_Z : X \times Z \longrightarrow Y \times Z$$

   is a quotient map.
   [*Hint:* If $\pi^{-1}(A)$ is open and contains $x \times y$, choose open sets U_1 and V with $\bar{V}$ compact, such that $x \times y \in U_1 \times V$ and $U_1 \times \bar{V} \subset \pi^{-1}(A)$. Given $U_i \times \bar{V} \subset \pi^{-1}(A)$, use the tube lemma to choose an open set U_{i+1} containing $p^{-1}(p(U_i))$ such that $U_{i+1} \times \bar{V} \subset \pi^{-1}(A)$. Let $U = \bigcup U_i$; show that $U \times V$ is a saturated neighborhood of $x \times y$ that is contained in $\pi^{-1}(A)$.]
   An entirely different proof of this result will be outlined in the exercises of §46.
   (b) *Theorem.* *Let $p : A \to B$ and $q : C \to D$ be quotient maps. If B and C are locally compact Hausdorff spaces, then $p \times q : A \times C \to B \times D$ is a quotient map.*

*Supplementary Exercises: Nets

We have already seen that sequences are "adequate" to detect limit points, continuous functions, and compact sets in metrizable spaces. There is a generalization of the notion of sequence, called a *net*, that will do the same thing for an arbitrary topological space. We give the relevant definitions here, and leave the proofs as exercises. Recall that a relation $\preceq$ on a set A is called a *partial order* relation if the following conditions hold:

(1) $\alpha \preceq \alpha$ for all α.

(2) If $\alpha \preceq \beta$ and $\beta \preceq \alpha$, then $\alpha = \beta$.

(3) If $\alpha \preceq \beta$ and $\beta \preceq \gamma$, then $\alpha \preceq \gamma$.

Now we make the following definition:

A ***directed set*** J is a set with a partial order $\preceq$ such that for each pair α, β of elements of J, there exists an element γ of J having the property that $\alpha \preceq \gamma$ and $\beta \preceq \gamma$.

1. Show that the following are directed sets:

(a) Any simply ordered set, under the relation $\leq$.

(b) The collection of all subsets of a set S, partially ordered by inclusion (that is, $A \preceq B$ if $A \subset B$).

(c) A collection $\mathcal{A}$ of subsets of S that is closed under finite intersections, partially ordered by reverse inclusion (that is $A \preceq B$ if $A \supset B$).

(d) The collection of all closed subsets of a space X, partially ordered by inclusion.

2. A subset K of J is said to be ***cofinal*** in J if for each $\alpha \in J$, there exists $\beta \in K$ such that $\alpha \preceq \beta$. Show that if J is a directed set and K is cofinal in J, then K is a directed set.

3. Let X be a topological space. A ***net*** in X is a function f from a directed set J into X. If $\alpha \in J$, we usually denote $f(\alpha)$ by x_α. We denote the net f itself by the symbol $(x_\alpha)_{\alpha\in J}$, or merely by (x_α) if the index set is understood.

The net (x_α) is said to ***converge*** to the point x of X (written $x_\alpha \to x$) if for each neighborhood U of x, there exists $\alpha \in J$ such that

$$\alpha \preceq \beta \Longrightarrow x_\beta \in U.$$

Show that these definitions reduce to familiar ones when $J = \mathbb{Z}_+$.

4. Suppose that

$$(x_\alpha)_{\alpha\in J} \longrightarrow x \text{ in } X \quad \text{and} \quad (y_\alpha)_{\alpha\in J} \longrightarrow y \text{ in } Y.$$

Show that $(x_\alpha \times y_\alpha) \longrightarrow x \times y$ in $X \times Y$.

5. Show that if X is Hausdorff, a net in X converges to at most one point.

6. *Theorem. Let $A \in X$. Then $x \in \bar{A}$ if and only if there is a net of points of A converging to x.*

[*Hint:* To prove the implication $\Rightarrow$, take as index set the collection of all neighborhoods of x, partially ordered by reverse inclusion.]

7. *Theorem. Let $f : X \to Y$. Then f is continuous if and only if for every convergent net (x_α) in X, converging to x, say, the net $(f(x_\alpha))$ converges to $f(x)$.*

8. Let $f : J \to X$ be a net in X; let $f(\alpha) = x_\alpha$. If K is a directed set and $g : K \to J$ is a function such that

(i) $i \preceq j \Rightarrow g(i) \preceq g(j)$,

(ii) $g(K)$ is cofinal in J,

then the composite function $f \circ g : K \to X$ is called a ***subnet*** of (x_α). Show that if the net (x_α) converges to x, so does any subnet.

9. Let $(x_\alpha)_{\alpha \in J}$ be a net in X. We say that x is an ***accumulation point*** of the net (x_α) if for each neighborhood U of x, the set of those α for which $x_\alpha \in U$ is cofinal in J.

Lemma. The net (x_α) has the point x as an accumulation point if and only if some subnet of (x_α) converges to x.

[*Hint:* To prove the implication $\Rightarrow$, let K be the set of all pairs (α, U) where $\alpha \in J$ and U is a neighborhood of x containing x_α. Define $(\alpha, U) \preceq (\beta, V)$ if $\alpha \preceq \beta$ and $V \subset U$. Show that K is a directed set and use it to define the subnet.]

10. *Theorem. X is compact if and only if every net in X has a convergent subnet.*

[*Hint:* To prove the implication $\Rightarrow$, let $B_\alpha = \{x_\beta \mid \alpha \preceq \beta\}$ and show that $\{B_\alpha\}$ has the finite intersection property. To prove $\Leftarrow$, let $\mathcal{A}$ be a collection of closed sets having the finite intersection property, and let $\mathcal{B}$ be the collection of all finite intersections of elements of $\mathcal{A}$, partially ordered by reverse inclusion.]

11. *Corollary. Let G be a topological group; let A and B be subsets of G. If A is closed in G and B is compact, then $A \cdot B$ is closed in G.*

[*Hint:* First give a proof using sequences, assuming that G is metrizable.]

12. Check that the preceding exercises remain correct if condition (2) is omitted from the definition of *directed set*. Many mathematicians use the term "directed set" in this more general sense.